REVIEW ARTICLE

A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues

Toktam Maleki Shahmahmood[1,2], Shohreh Jalaie[3], Zahra Soleymani[1], Fatemeh Haresabadi[2], Parvin Nemati[4]

[1]Department of Speech Therapy, School of Rehabilitation, Tehran University of Medical Sciences, Tehran, [2]Department of Speech Therapy, Faculty of Paramedical Sciences, Mashhad University of Medical Sciences, Mashhad, [3]Department of Physiotherapy, School of Rehabilitation, Tehran University of Medical Sciences, Tehran, Iran, [4]Department of Psychology, Tuebingen University, Tuebingen, Germany

Background: Identification of children with specific language impairment (SLI) has been viewed as both necessity and challenge. Investigators and clinicians use different tests and measures for this purpose. Some of these tests/measures have good psychometric properties, but it is not sufficient for diagnostic purposes. A diagnostic procedure can be used for identification a specific population with confidence only when its sensitivity and specificity are acceptable. In this study, we searched for tests/measures with predefined sensitivity and specificity for identification of preschool children with SLI from their typically developing peers. **Materials and Methods:** A computerized search in bibliographic databases from 2000 to August 2015 was performed with the following keywords: "specific language impairment" or SLI" and "primary language impairment" or 'PLI' with at least one of the followings: "diagnosis," "identification," "accuracy," "sensitivity," and "specificity." In addition, the related citations and reference lists of the selected articles were considered. **Results:** The results of reviewing 23 included studies show that the index measures used in studies vary in accuracy with the sensitivity ranging from 16% to 100% and the specificity ranging from 14% to 100%. **Conclusion:** These varieties in sensitivity and specificity of different tests/measures confirm the necessity of attention to the diagnostic power of tests/measures before their use as diagnostic tool. Further, the results indicate there are some promising tests/measures that the available evidence supports their performances in the diagnosis of SLI in preschool-aged children, yet the place of a reference standard for the diagnosis of SLI is vacant among investigations.

Key words: Accuracy, diagnosis, preschool age, sensitivity, specific language impairment, specificity

How to cite this article: Maleki Shahmahmood T, Jalaie Sh, Soleymani Z, Haresabadi F, Nemati P. A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. J Res Med Sci 2016;21:67.

INTRODUCTION

Specific language impairment (SLI) is a developmental language disorder in the absence of obvious accompanying conditions such as mental retardation, neurological damage, and hearing or emotional impairment.[1] Epidemiological evidence suggests that SLI represents the largest segment of language impairments, estimated at roughly 7% of the general population.[2,3] For most children with SLI, the central section of impairment is grammar;[4,5] nevertheless, the symptom of this condition is so heterogeneous among affected children.[6] Moreover, individuals with SLI often show similar and overlapping sets of symptom with other disorders such as dyslexia or autism.[7] Because of these heterogeneous and overlapping symptoms, the differential diagnosis of young children with SLI from normal developing children and children having other language disorders is a challenge for both clinicians and researchers, but it is either a necessity.[8,9] Applying accurate diagnostic tests/measures is the first step in treatment planning and carrying out epidemiologic research.[10]

Diagnosis of SLI depends on both exclusionary and inclusionary criteria. Exclusionary criteria help in

Access this article online	
Quick Response Code:	Website: www.jmsjournal.net
	DOI: ****

Address for correspondence: Dr. Shohreh Jalaie, Department of Physiotherapy, School of Rehabilitation, Tehran University of Medical Sciences, Pich-e-Shemiran, Enghelab Ave., Tehran, 1148965141, Iran. E-mail: jalaeish@sina.tums.ac.ir

Received: 10-01-2016; **Revised:** 04-03-2016; **Accepted:** 25-05-2016

ruling out other meddlesome conditions, and inclusionary criteria confirm the presence of language disorder in children. In spite of general agreement about exclusionary criteria among clinicians and researchers, there is no consensus about inclusionary criteria.[1] A brief review on literature indicates that the criteria for selecting SLI subjects vary among studies from scores of standardized tests to assessing child's language in naturalistic contexts. The diagnostic performance of these tests/measures is a critical issue. Although some of these test/measures have good psychometric properties (including validity and reliability) and even are capable in showing group differences between language impaired and typically developing (TD) children, these properties of a test/measure are not enough to conclude that it can be introduced as a diagnostic tool. Diagnostic measures of tests must be further explored at individual level rather than group level, which include finding sensitivity and specificity in predefined cutoff point/s.[11,12]

Sensitivity of a test means the degree to which children who previously are classified as SLI (using a reference test), will be identified truly as affected by the test and specificity, means the degree to which children who are independently classified as having normal development will be identified as unaffected by the test. According to Plante and Vance, sensitivity and specificity values of ≥90% are considered good, 80–89% considered adequate, and below 80% considered unacceptable.[13]

Sensitivity and specificity of a test are completely dependent on the cutoff point score which is used to determine a line between normal and impaired individuals. To confirm the existence of language disorder in client, many clinicians use from arbitrary cutoff score (e.g., –1.5 or 2 SD under the mean) for any language test. However, now, there are substantial data demonstrating that this practice could not lead to accurate diagnoses because children with impaired language frequently do not obtain scores that fall below these commonly applied cutoff scores.[10] The cutoff score derived for one test can differ significantly from that of another test even when these tests were validated on the same sample of children.[13]

The purposes of this study are reviewing published accuracy studies in the last 15 years (until August 2015) which have focused on determining the sensitivity and specificity of specific language tests/measures as inclusionary criteria for the diagnosis of preschool monolingual children with SLI from TD children. Since preschool period is the most important period in the diagnostic process of SLI and regarding the long-standing nature and variable clinical manifestations of this disorder during the development, only the accuracy studies on preschool period have been selected for reviewing in this study. In this study, research on the sensitivity and specificity of the language measures in languages other than English also included examining whether there are shared language behaviors with good diagnostic accuracy for the identification of children with SLI in various languages and whether they can be introduced as universal clinical markers for this disorder. The intention of this review mainly is to specify the linguistic test/measure with acceptable sensitivity and specificity, without necessarily making clinical recommendations for the use of a particular test. Moreover, no attempt has been made to summarize data from test manuals or to evaluate the validity or reliability of the diagnostic procedures.

MATERIALS AND METHODS

Literature search strategies

A systematic computerized search was conducted in electronic databases including MEDLINE via PubMed, Google Scholar, and Web of Science and publisher databases (Springer, Oxford, Thieme, ProQuest, and ScienceDirect) from 2000 through July 30, 2015. For the electronic search, we used the following keywords or MeSH subject headings: "Specific language impairment" or "SLI" and "primary language impairment" or "PLI" with at least one of the followings: "diagnosis," "identification," "accuracy," "sensitivity," and "specificity;" we used identical search items in all resources. Our search strategies moreover included the tracking of references lists of all searched article and searches by hand in books. E-mail for more information was made to professionals and authors. Studies identified in this ways also incorporated into the decision-making process.

Inclusion and exclusion criteria

Based on structured guidelines of systematic reviews, we reviewed the last 15 years (until August 2015) for accuracy studies on the diagnosis of preschool children with SLI from their TD peers, which published in English-language journals. Accuracy studies on the diagnosis of SLI from other language impairments and studies on adults, toddlers (under 3 years of age), and school-aged children or bilingual subjects with SLI were excluded from the study.

Study selection and eligibility criteria

The literature retrieval processes and screenings of articles are illustrated in details in Figure 1. After screening titles and abstracts and applying inclusion and exclusion criteria, 28 potentially relevant articles to our questions were selected and 261 articles were removed. Studies that could not be excluded with certainty were then examined in detail in full text. In cases of doubt, a second investigator was consulted. The quality of studies meeting the inclusion criteria was

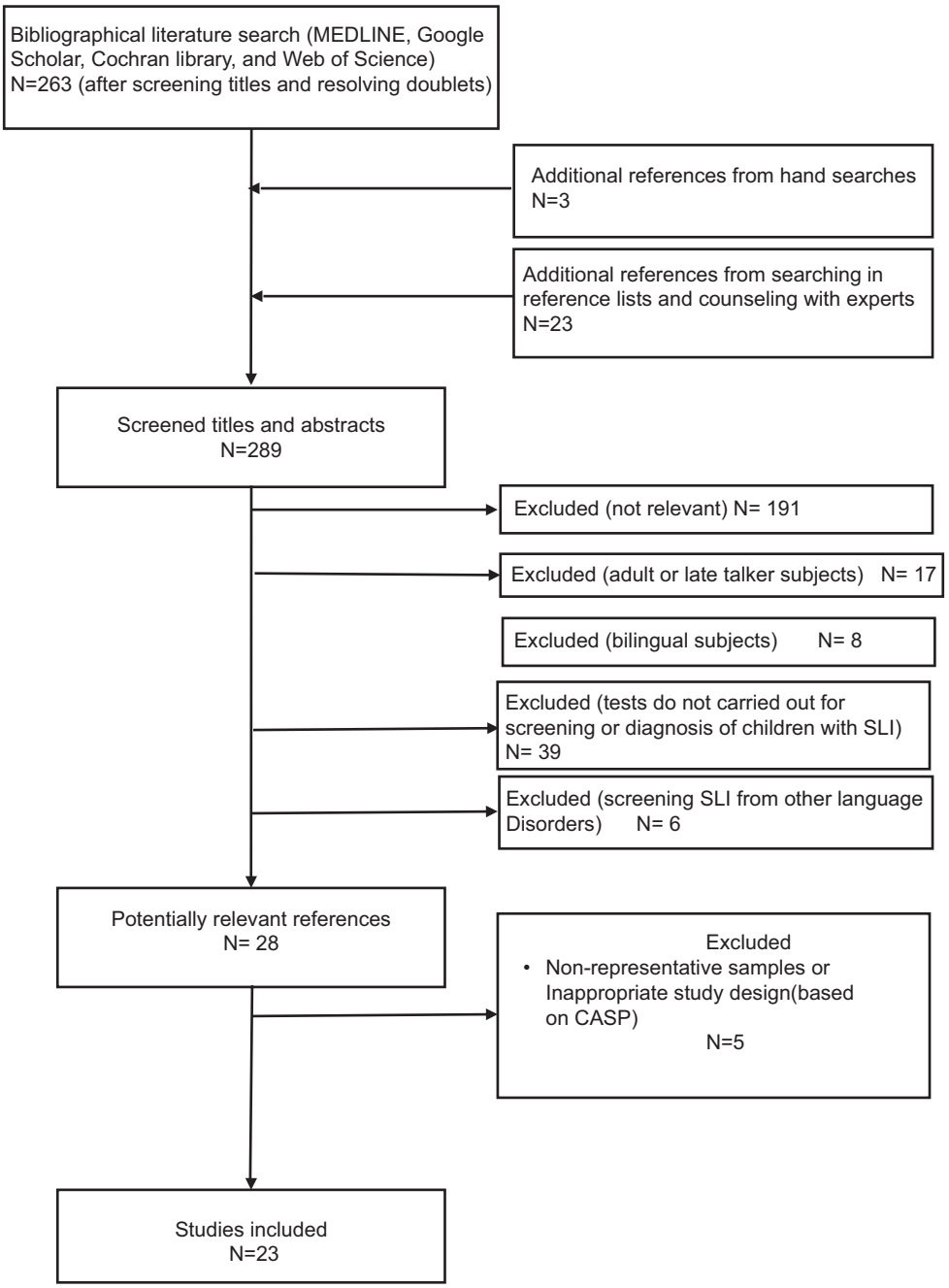

Figure 1: The literature retrieval processes and screenings of articles

appraised by critical appraisal skills program, diagnostic test study form.[14] This form involves 12 questions that consider following three broad issues for appraising a diagnostic test study:

- Are the results of the study valid?
- What are the results?
- Will the results help me and my patient/population?

Nine questions of this form have a three-level scale (yes/do not tell/no) and 3 questions (7, 8, and 12) should be described. The first two questions are "screening questions" and can be answered fast. Even if the answer to one of them is "no" or "cannot tell," it is not worth continuing to the remaining questions. We appraised all potential relevant articles to this systematic review with these criteria. Studies that obtain "yes" answer to the first two question of appraisal form, and in sum, more than eight yes answer ("cannot tell" answer to 2 questions was tolerable), with reported or at least calculable results (question 7) and acceptable results (question 8) categorized as studies with high or moderate quality and remained in the pool of articles. Hence, five articles excluded via this criterion and 23 articles remained for studying in this systematic review. These steps, which are demonstrated in Figure 1,

were conducted by the first author and reviewed twice. The author tried to prevent any bias regarding the author's professional field or celebrity and selective reporting among studies.

Data extraction and abstraction

Tables 1-3 show the original data elicited from included studies. Tables 1 and 2 describe the characteristics of subjects, index measure/s, and reference standard used in every included study. The sensitivity and specificity of the index tests/measures used for differential diagnosis of preschool SLI children from their TD peers are shown in Table 3. Because of the heterogeneity of the tests/measures and differences in the way in which similar tests/measures were implemented in different groups of children, statistical pooling was not possible.

Table 1: Characteristics of included studies on English-speaking subjects

References	Subjects	Index test	Reference test
Conti-Ramsden, 2003[15]	64 English-speaking children: 32 children with SLI and 32 age-matched TD children	2 processing (nonword repetition and digits recall) and 2 linguistic tasks (past tense and plural marking)	Judgment of referral SLPs or specialist teachers for SLI group and report of classroom teacher for TD group
Conti-Ramsden and Hesketh, 2003[16]	64 English-speaking children: 32 SLI and 32 younger language-matched TD children	Past tense task; noun plural task; nonword repetition (CNRep); recall of digits	A combination of expert's judgment and standardized tests
Gray, 2003[17]	44 English-speaking children: 22 with SLI and 22 age- and gender-matched TD children	Nonword repetition (40 nonwords from CNRep in two 20 word lists) and digit span tasks	Judgment of certified SLPs (with the help of standardized language tests certified SLPs (with the help of standardized language tests and intervention status)
Perona *et al*., 2005[18]	85 English-speaking children including 42 SLI and 43 TD children	SPELT-P3	Clinical judgment
Oetting and Cleveland, 2006[19]	83 children including 16 children with SLI, 36 age-matched and 31 younger language-matched TD children who lived in the rural south of the United States. All children were speaker of a nonmainstream dialect of English	NRT (Dollaghan and Campbell's 16 nonwords repetition test)	Judgment of experts (classroom teacher/SLP) and standardized language and IQ tests
Pankratz *et al*., 2007[20]	64 English-speaking children: 32 SLI and 32 TD age-matched children	American version of Renfrew Bus Story	Experts judgments
Greenslade *et al*., 2009[21]	96 English-speaking children in two different groups Exploratory group: 32 SLI and 32 aged-matched TD children Confirmatory group: 10 children with SLI and 22 aged-matched TD children	SPELT-P2	A combination of clinical judgment and the results of formal testing
Deevy *et al*., 2010[22]	29 SLI and 47 age-matched TD children	NRT consists of 16 nonwords	SPELT-P2
van der Lely *et al*., 2011[6]	51 English-speaking children in two different SLI groups (11 young - SLI children, aged from 3;6 to 6:6 years and 10 older - SLI children aged from 6;9 to 8;11 years) and one TD group (including 30 children aged from 3;6 to 6:6 years)	GAPS test	Judgment of SLP and educational psychologists as well as discrepancy-based criteria
Gladfelter and Leonard, 2013[23]	55 English speakers in two different age-domains: 4 years-olds: 12 SLI and 15 TD children 5 years olds: 13 SLI and 15 TD children	Three measures from spontaneous speech TMT; PS; FVMC	Expert judgment based on tests results and parents reports. SPELT-P2 employed only for SLI children
Spaulding *et al*., 2013[24]	80 English speakers; 40 SLI and 40 TD peers	PPVT-III PPVT-IV	Certified SLP's judgment
Souto *et al*., 2014[25]	112 4- and 5-year-old children Exploratory group 4 years old: 14 SLI and 16 TD children 5 years old: 11 SLI and 15 TD children Confirmatory group 4 years old: 14 SLI and 16 TD children 5 years old: 11 SLI and 15 TD children	Spontaneous speech measures concentrated on tense/agreement morphology including Measures of developmental level of tense/agreement Measures of tense/agreement consistency	SPELT-P2

PPVT-III=Peabody Picture Vocabulary Test; Third Edition; PPVT-IV=Peabody Picture Vocabulary Test; Fourth Edition; TMT=Tense marker total; PS=Productivity score; FVMC=Finite verb morphology composite; GAPS=Grammar and phonology screening; NRT=Nonword repetition task; SPELT-P2=Structured Photographic Expressive Language Test-Preschool=Second Edition; SPELT-P3=Structured Photographic Expressive Language Test-Preschool=Third Edition; TD=Typically developing; CNRep=Children's Test of Nonword Repetition; SLPs=Speech-language pathologists

Table 2: Characteristics of included studies on subjects speaking languages other than English

References	Subjects	Index test	Reference test
Bortolini *et al*., 2002[26]	Study 1: 12 Italian-speaking SLI and 12 age-matched TD children	The children's production of grammatical morphology including articles and third-plurals	Judgment of clinicians based on parent/ teacher reports and discrepancy-based criteria
	Study 2: 15 Italian-speaking children with SLI and 15 TD age-matched children	assessment of children's use of articles, third-plural inflection and direct object clitics (replicating and extending the findings of Study 1)	
Klee *et al*., 2004[27]	45 Cantonese-speaking children: 15 SLI children, 15 TD aged-matched and 15 TD-language matched children	A composite score (including MLU and lexical diversity [D]) elicited from free-play-based language samples	Judgments of a qualified SLP and discrepancy-based criteria
Bortolini *et al*., 2006[28]	Expriment 1: 33 Italian speaking children including 11 children with SLI, 11 age-matched TD peers and 11 MLU- matched TD children	Children's production of three morpheme types: third-person plural inflection and direct object clitics were of primary interest. Third-person singular inflection was included as a control	Expert judgment and discrepancy-based criteria
	Experiment 2: SLI and TD-AM children participated in Experiment 1	Nonword repetition	
Stokes *et al*., 2006[29]	44 Cantonese children: 14 SLI children, 15 TD-aged matched and 15 younger TD-peers matched on receptive grammar scores	Nonword repetition Sentence repetition	Judgment of SLP based on standardized language and IQ tests
Wong *et al*., 2010[30]	29 Cantonese-speaking children including 14 TD and 15 SLI children	A composite variable made up of MLU+lexical diversity (D) + age obtained from conversationally-based language samples	Clinical judgment of an experienced SLP
Thordardottir *et al*., 2011[2]	92 I French-speaking children from Quebec: 14 SLI and 78 TD children	EVIP (French version of the PPVT) TACL (a test of receptive language) A French test of nonword repetition A test of sentence repetition (adopted from CELF-p) The RAN The digit span subtests of the CELF-4 The ENNI: Micro and macro structures MLU in words and morphemes from a spontaneous language sample	Expert judgment
Dispaldro *et al*., 2013[31]	34 Italian-speaking children: 17 SLI and 17 TD-age-matched children	Nonword and real word repetition tasks with two different scoring methods Scoring method 1: Percentage of phonemes correct with allowances for developmental phonological errors Scoring method 2: Percentage of whole-words correct, with no allowances for developmental phonological errors	Expert judgment (based on test results and intervention status) and performance on a set of probes designed to assess children's use of third person direct object clitics (Bortolini *et al*., 2002, 2006)
Grinstead, *et al*., 2013[32]	54 Spanish-speaking children: 26 SLI and 29 TD children	Six distinct indices of grammatical development extracted from spontaneous speech samples MLUw MLU-m MLTU; which differs from MLU in only counting clausal utterances ETU; the mean number of errors in a child's clausal utterances NDW SUB-I: The ratio of the total number of clauses to the total number of T-units Two experimental measures EP GCT	Expert judgment based on standardized tests and parents'/teachers' reports
Kapalková *et al*., 2013[33]	46 Slovak-speaking children 16 SLI, 16 age-matched and 14 MLU-matched TD children	Nonword repetition test with two different scoring method Whole-item scoring method Vowel scoring method	Expert judgment

Contd...

Table 2: Contd...

References	Subjects	Index test	Reference test
Katzenberger and Meilijson, 2014[34]	454 (383 TD and 71 SLI) Hebrew-speaking children	KHLA for preschool children	Expert judgment
Kazemi *et al*. 2015[35]	27 TD and 24 SLI Persian speaking children	Measures from spontaneous speech samples	Professional judgment

ENNI=Edmonton Narrative Norms Instrument; KHLA=Katzenberger Hebrew Language Assessment; SLI=Specific language impairment; CELF=Clinical Evaluation of Language Fundamentals; PPVT=Peabody Picture Vocabulary Test; TACL=Test for Auditory Comprehension of Language; TD=Typically developing; AM=Age-matched; MLU=Mean length of utterance; MLUw=Mean length of utterance by words; MLUm=Mean length of utterance by morphemes; NDW=Number of different words; EP=Elicited production; GCT=Grammaticality choice task; ETU=Errors per terminable unit; MLTU=Mean length of terminable unit; SLPs=Speech-language pathologists

Table 3: Diagnostic performance of the various tests/language measures in included studies

Index test/language feature	Language	Cutoff	Sensitivity	Specificity	Acceptability	Reference
PPVT-III	English	103	80	75	No	Spaulding *et al*., 2013[24]
PPVT-IV	English	103	80	70	No	
SPELT-P2	English	Not reported	82	95	Yes	Gray, 2003[17]
SPELT-P2						
Exploratory group	English	Scores >87: TD; score <87: SLI	90.6	100	Yes	Greenslade *et al*., 2009[21]
Confirmatory group	English	Scores >87: TD; score <87: SLI	100	95.6	Yes	
SPELT-P3						
Exploratory group	English	Score >95: TD; score <95: SLI	90.6	100	Yes	Perona *et al*., 2005[18]
Confirmatory group	English	Score >95: TD; score <95: SLI	90	100	Yes	
Renfrew bus story (American-version)						
Information score		Scores <87: SLI; scores <99: SLI	81.3	75	No	Pankratz *et al*., 2007[20]
Length scores			75	75	No	
Information and length scores in combination		A range of information (82-90) and length (90-112) scores	84.4	78.1	No	
GAPS test (overall)	English	15th percentile	100	93.3	Yes	van der Lely *et al*., 2011[6]
GAPS grammar		15th percentile	90.9	100	Yes	
GAPS phonology		15th percentile	90.9	93.3	Yes	
Nonword repetition test						
Scoring method 1 (no allowances for out-of-inventory phonemes)	English	TPPC scores ≤66%: SLI and TPPC scores >66%: TD	86	91	Yes	Deevy *et al*., 2010[22]
Scoring method 2 (excluding out-of-inventory phonemes)		TPPC scores of ≤68%: SLI and TPPC scores of >68%:TD	79	89	No	
Nonword repetition	English	Not reported	95	100	Yes	Gray, 2003[17]
Digit recall	English	Not reported	91	77	No	
Nonword repetition	English	Not reported	56	92	No	Oetting and Cleveland, 2006[19]
NWR and CSSB	English	Not reported	81	94	Yes	
Digit recall	English	Scores up to 25th percentile: TD; Scores under 25th percentile: SLI	53	94	No	Conti-Ramsden, 2003[15]
Test of nonword repetition (CNRep)	English		66	100	No	
Past tense task	English		71	91	No	
Noun plural task	English		16	100	No	
CNRep + past tense marking	English		81	91	Yes	

Contd...

Table 3: Contd...

Index test/language feature	Language	Cutoff	Sensitivity	Specificity	Acceptability	Reference
Digit recall	English	Scores up to 25th percentile: TD; Scores under 25th percentile: SLI	53	90	No	Conti-Ramsden and Hesketh, 2003[16]
Test of nonword repetition (CNRep)	English		66	85	No	
Past tense task	English		71	14	No	
Noun plural task	English		16	77	No	
Finite verb morphology composite						
In 4-years-old children	English	Not reported	100	100	Yes	Gladfelter and Leonard, 2013[23]
In 5-years-old children		Not reported	92.31	93.33	Yes	
TMT						
In 4-year-old children		Not reported	83.33	86.67	Yes	
In 5-year-old children		Not reported	76.92	80.00	No	
PS						
In 4-year-old children		Not reported	66.67	86.67	No	
In 5-year-old children		Not reported	84.62	80.00	Yes	
TMT and PS (in combination)						
In 4-year-old children		Not reported	83.33	86.67	Yes	
In 5-year-old children		Not reported	84.62	80.00	Yes	
The mean tense/agreement developmental scores						
In 4-year-old children	English	Not reported	79	81	No	Original study Souto *et al.*, 2014[25]
In 5-year-old children		Not reported	64	80	No	
Mean of the five highest tense/agreement developmental scores						
In 4-year-old children		Not reported	71	69	No	
In 5-year-old children		Not reported	73	87	No	
Finite verb morphology composites						
In 4-year-old children		Not reported	93	94	Yes	
In 5-year-old children		Not reported	91	93	Yes	
The mean sentence points						
In 4-year-old children		Not reported	93	91	Yes	
In 5-year-old children		Not reported	100	100	Yes	
The overall DSS scores						
In 4-year-old children		Not reported	79	94	No	
In 5-year-old children		Not reported	72	87	No	
The mean tense/agreement developmental scores						
In 4-year-old children		Not reported	50	75	No	Replication study
In 5-year-old children		Not reported	82	80	Yes	
Mean of the five highest tense/agreement developmental scores						
In 4-year-old children		Not reported	14	94	No	
In 5-year-old children		Not reported	82	87	Yes	
Finite verb morphology composites						
In 4-year-old children		Not reported	93	100	Yes	
In 5-year-old children		Not reported	82	93	Yes	
The mean sentence points						
In 4-year-old children		Not reported	100	100	Yes	
In 5-year-old children		Not reported	100	100	Yes	
The overall DSS scores						
In 4-year-old children		Not reported	93	94	Yes	
In 5-year-old children		Not reported	82	87	Yes	

Contd...

Table 3: Contd...

Index test/language feature	Language	Cutoff	Sensitivity	Specificity	Acceptability	Reference
Article	Italian	Not reported	58.33	91.67	No	Experiment 1 Bortolini *et al.*, 2002[26]
Third person plural	Italian	Not reported	100	91.67	Yes	
Article + third plural	Italian	Not reported	100	100	Yes	
ATP composite	Italian	Composite scores >81%: TD; Composite scores <81%: SLI	83.33	91.67	Yes	
Article	Italian	Not reported	100	100	Yes	Experiment 2
Third person plural	Italian	Not reported	86.67	86.67	Yes	
Clitics	Italian	Not reported	86.67	93.33	Yes	
Article + third plural	Italian	Not reported	100	100	Yes	
Article + clitics	Italian	Not reported	100	100	Yes	
Third plural+clitics	Italian	Not reported	100	100	Yes	
Article + third plural + clitics	Italian	Not reported	100	100	Yes	
ATP composite	Italian	Composite scores >77%: TD; Composite scores <77%: SLI	93.33	93.33	Yes	
Third person plural	Italian	Not reported	72.73	90.91	No	Bortolini *et al.*, 2006[28]
Direct object clitics	Italian	Not reported	90.91	100	Yes	
Third plural + direct object clitics	Italian	Not reported	90.91	100	Yes	
Nonword repetition	Italian	Not reported	81.82	81.82	Yes	
Nonword repetition+third plural	Italian	Not reported	81.82	100	Yes	
NWR + direct-object clitics	Italian	Not reported	90.91	100	Yes	
NWR + third plural + direct object clitics	Italian	Not reported	90.91	90.91	Yes	
Nonword and real word repetition task	Italian					Dispaldro *et al.*, 2013[31]
Scoring method 1 (see Table 1)						
Nonword repetition		93% correctly repeated phonemes	94.1	94.1	Yes	
Real word repetition		96.5% correctly repeated phonemes	94.1	94.1	Yes	
Nonword and real word R		93.75% correctly repeated phonemes	94.1	100	Yes	
Scoring method 2 (see Table 1)						
Nonword repetition		65% correctly repeated target	100%	100	Yes	
Real word repetition		79% correctly repeated target	100%	100	Yes	
Nonword and real word R		72% correctly repeated target	100%	100	Yes	
Composite score (made up from age, MLU, and D)	Cantonese	Not reported	100	96.5	Yes	Klee *et al.*, 2004[27]
Composite variable (made up of MLU, lexical diversity, and age)	Cantonese	0 indiscriminant function equation [(−0.037×age) + (0.931×MLU) + (0.099×D) − 7.269]	73.3	57.1	No	Wong *et al.*, 2010[30]
Sentence repetition	Cantonese	Not reported	77	97	No	Stokes *et al.*, 2006[29]

Contd...

Table 3: Contd...

Index test/language feature	Language	Cutoff	Sensitivity	Specificity	Acceptability	Reference
EVIPss	French	−1 SD	88	85	Yes	Thordardottir *et al*., 2011[2]
Carrow/TACL		−1 SD	71	86	No	
MLUw		−1 SD	40	85	No	
MLUm		−1 SD	36	87	No	
ENNI story grammar		−1.28 SD	46	87	No	
ENNI first mentions		−1 SD	31	88	No	
NWR		−1.28 SD	85	88	Yes	
Sentence imitation		−1 SD	86	92	Yes	
Following directions		−1.28 SD	93	92	Yes	
RAN error		−1.28 SD	71	91	No	
RAN time		−1 SD	43	86	No	
Forward digit span		−1 SD	54	89	No	
Nonword repetition						
Whole-item scoring method	Slovakian	8 (8/16 correct repetitions)	93.75	100	Yes	Kapalková, *et al*., 2013[33]
Vowel scoring method		15 (15/16 correct repetitions)	75	100	No	
MLUw	Spanish	Not reported	81	76		Grinstead *et al*., 2013 [32]
NDW		Not reported	85	72		
EP		Not reported	89	89		
MLUw, EP		Not reported	89	83		
MLUm, GCT		Not reported	100	84		
MLUw, GCT		Not reported	100	79		
EP, MLTU		Not reported	72	100		
EP, MLUw, SUB-I, GCT		Not reported	93	87		
MLUw, GCT, SUB- I		Not reported	86	84		
KHLA	Hebrew	−1.25 SD	98.4	82.2		Katzenberger and Meilijson, 2014[34]
Grammaticality	Persian	94.25	98	84	Yes	Kazemi *et al*., 2015[35]
MLUw-excluding one word utterances		2.96	82	98	Yes	
Semantic errors		2.50	92	96	Yes	
MLUw		2.37	92	78	No	
Total errors		15.5	74	98	No	
MLUm-excluding one word utterances		4.08	66	98	No	
MLUm		3.39	92	78	No	
Nonsense string of words		0.5	83	78	No	
Missing verbs		0.5	83	74	No	
Number of different words		132	97	78	No	
Total number of one-word utterances		83	75	85	No	
Wrong responses		0.5	72	96	No	

PPVT=Peabody Picture Vocabulary Test; SPELT=Structured Photographic Expressive Language Test; SLI=Specific language impairment; TD=Typically developing; NWR=Nonword repetition; CSSB=Comprehension subtest VI of the Stanford-Binet; CNRep=Children's Test of Nonword Repetition; TACL=Test for Auditory Comprehension of Language; MLUw=Mean length of utterance by words; MLUm=Mean length of utterance by morphemes; NDW=Number of different words; ENNI=Edmonton Narrative Norms Instrument; EP=Elicited production; GCT=Grammaticality choice task; ETU=Errors per terminable unit; MLTU=Mean length of terminable unit; KHLA=Katzenberger Hebrew Language Assessment; GAPS=Grammar and Phonology Screening test; TMT=Tense marker total; PS=Productivity score; RAN=Rapid automatized naming; TPPC=Total Percentage of Phonemes Correct; DSS=Developmental sentence score; SD=Standard deviation; ATP=Auditory temporal processing; EVIPs=Échelle de vocabulaire en images Peabody

RESULTS

Among included studies, 12 studies have been conducted on English or American – English-speaking populations – and the remaining 11 studies have been conducted on non-English speakers; among them, three studies carried out on Cantonese- and three on Italian-speaking children, and one study have been carried out on each of the French-, Spanish-, Slovakian-, Hebrew-, and Persian-speaking populations.

The sample size of studies ranged from 29 to 454 children. Participants of all studies are SLI children and their aged-matched TD peers. In 6 studies, younger

language-matched TD peers also included. The numbers of subjects in SLI and TD groups are not the same in 9 studies; however, only in two studies (2 and 34), these differences in number are very obvious. Gender was not a significant factor in studies, and no separate analysis was done on girls and boys although in some studies there were a matching between SLI and TD control group according to the gender. All children with SLI in studies were receiving clinical services for their problem or were eligible for registration in speech-language services.

For primary categorization of children as impaired or normal and determine the case status, authors need a reference standard. The reviewed studies show considerable variety in how these case statuses are defined. Expert judgment has been the most popular reference standard among included studies. Only in two studies, a standardized test with predefined sensitivity and specificity in the diagnosis of SLI has been used for samples categorization. The lists of reference tests used in articles are seen in Tables 1 and 2.

The index tests/measures vary among studies. Among included studies, eight studies used from one or more standardized tests as the index measure, in which seven studies from these eight carried out on English-speaking populations. Five studies focused on language measures elicited from spontaneous language and nine articles concentrated on linguistic or processing features extracted via language probes. In one study (2), a collection of all these speech extraction methods has been used as index. Moreover, in one study (32), indexes have been extracted from spontaneous language and experimental measures separately with the aim of comparing these two methods of extracting.

Evidence from included studies indicates that the majority of studies compare the performance of two or more diagnostic procedures when applied to a single population; this provides easier state to make judgments about the relative value of different procedures/measures. Moreover, three of included studies (21, 25, and 26) conducted their study on two separate populations of children with SLI with the aim of increasing the reliability of the estimated sensitivity and specificity of the index tests/measures.

The sensitivity and specificity of behavioral psycholinguistic measures/tests for the diagnosis of preschooler children with SLI from their TD peers are shown in Table 3. The sensitivities of tests or linguistic/processing measures have a range between 16% to100% and specificities vary from 14% to 100%.

The cutoff score used for index tests are demonstrated in Table 3. From 23 papers reviewed, the cutoff score is not reported in eight articles. Some authors used more than one cutoff point for one index test, but we reported only the score which defined as the optimum cutoff point by authors.

DISCUSSION

Tests with more sensitivity and specificity rates can lead to increased reliability of detection rates for true positives and true negatives.[36] Moreover, since there is no single widely accepted "reference standard" for subject identification in the field of SLI,[37] introducing the tests or measures with empirical evidence of an acceptable sensitivity and specificity is of the high importance because they can then be used as reference test in future studies or clinical practices.

As demonstrated in the result section, the index tests used in included studies can be generally divided into two main categories: Standardized language tests which target different areas of language and psycholinguistic features elicited from the children's linguistic or processing system via speech sample analysis or psycholinguistic probes.

A survey in included studies shows that far more research is available on the diagnostic accuracy of standardized test in English than any other language. Hence, the preference of much of included studies that have been carried out on other languages (including Italian, Cantonese, Slovakian, Spanish, and Persian) is finding linguistic or processing measures that can be introduced as clinical markers for SLI. These tendencies may be related to excessive studies carried out on SLI in English language from the first time this concept emerged, so the linguistic characteristics and deficits of English-speaking children with this disorder are more explicit than SLI children speaking other languages, where this field of study is nearly new. Furthermore, the availability of various well-standardized language assessments on English-speaking populations could be another factor. However, as much as standardized tests, English-language investigators focused on the diagnostic performances of psycholinguistic markers for differential diagnosis of preschool children with SLI from TD children; moreover, among included studies on subjects who speak languages other than English, two studies used from standardized language or language processing tests or subtests of them as index tests.[2,34]

Regarding Plante and Vance's (1994) criteria for acceptable sensitivity and specificity, among the English standardized tests used as the index test in included studies, Renfrew bus story had adequate sensitivity but weak specificity. Hence, its application to identifying preschool children with SLI can results in over-identification of TD children as SLI. Grammar and phonology screening, Structured Photographic Expressive Language Test (SPELT) – P2, and SPELT-3 are tests of grammatical production and all of

them have good sensitivity and specificity for diagnosis of preschool children with SLI. Vocabulary tests including Peabody Picture Vocabulary Test (PPVT-III) and PPVT-IV had unacceptable sensitivity and specificity levels which made them inappropriate tools for identifying SLI children. These results are consistent with the results of the previous study by Gray *et al*. on diagnostic accuracy of four vocabulary tests (including PPVT-Ill) that show none of vocabulary tests is accurate measure for differential diagnosis of preschool English children with SLI.[38] It is notable that PPVT-IV is the newest version of PPVT, and regarding the Betz *et al*., is the third most commonly employed norm-referenced test used by clinicians for the diagnosis of children with SLI in the United States.[39] However, the results of these two studies not only show that despite known deficits of children with SLI in the area of vocabulary, these children are unlikely to score low on these commonly used vocabulary tests but also show that the newer test version is not superior to older in the diagnostic process. Hence, these results again confirm the importance of investigating the diagnostic performances of every linguistic test before its application for diagnostic purposes. It should be noted here that contrary to the results of these two English studies, the results of Thordardottir *et al*.'s study on the diagnostic power of Échelle de vocabulaire en images Peabody (EVIP), French version of PPVT, shows that this vocabulary test has acceptable sensitivity and specificity for differential diagnosis of preschool French children with SLI from their TD peers.[2] However, due to these inconsistencies between studies' results, it seems that clinician should be cautious about the application of EVIP as the only diagnostic tool for detecting French children with SLI and it is ideal if the results of Thordardottir *et al*. are repeated in another independent sample of French-speaking children.

Over the last two decades, research has consistently shown that English-speaking children with SLI score significantly lower than their age-matched TD peers and even than younger language-matched TD peers on tests of working memory such as nonword repetition (NWR), digit recall, and sentence repetition (SR).[40-43] Bishop based on a twin study proposed that NWR can be served as a phenotypic marker of heritable language impairment.[44] Then, Dollaghan and Campbell suggested that tasks such as NWR may serve as a method of identifying children with language impairments.[40] After that, many studies were conducted to investigate this suggestion (e.g., Conti-Ramsden, 2003; Conti-Ramsden, Botting and Faragher, 2001; Archibald and Gathercole 2007).[15,45,46] The results of this study show that NWR is one of the tasks which have received much attention in included studies.

The majority of studies conducted in English which have investigated the potential of NWR as a clinical marker for SLI have used from one of these two tests: The children's test of nonword repetition (CNRep) and the nonword repetition test (NRT)[44] (these tests have been compared in detail elsewhere - see 46 for review). NRT has been used as an index test in two of included studies.[19,22] The results of Deevy *et al*.'s study imply good sensitivity and adequate specificity of this test.[22] In spite of that, the results of Oetting and Cleveland demonstrate that NRT, alone, could not be used as an accurate diagnostic tool because of low sensitivity although it is diagnostic power increases in combined with scores from one other nonbiased assessment (comprehension subtest VI of the Stanford–Binet).[19] The causes of the difference between the results of these two studies are not clear, but it could be attributed to different cutoff points, the difference between reference standard employed, the differences of age, cognitive characteristics and severity of impairment among participants, and the sample size of studies.

Conti-Ramsden and Conti-Ramsden and Hesket used CNRep in their studies to evaluate the performance of phonological working memory (pWM) in preschool children with SLI and to determine the CNRep's accuracy indistinguish these children from their TD peers.[15,16] As demonstrated in Table 3, in both studies at the optimum cutoff point, the specificity of CNRep was fair but the sensitivity was low. Although the result of these two studies does not allow us to introduce the CNRep as an appropriate screening test to identify preschool children with SLI, Gray's study shows that CNRep had an excellent sensitivity and specificity. Interestingly, Gray's study also shows that while CNRep can be used as a diagnostic tool for SLI, the digit span task cannot.[17] Gray used the previous version of CNRep in her study. Moreover, the reasons expressed above could contribute to the variability of the results obtained in these studies.

Although it proposed that the children's performance on NWR task permits accurate classification of children with SLI and same-age peers even when the children spoke a nonstandard dialect of American-English,[47] regarding these incommensurable results, it seems that NWR cannot be introduced confidently as an adequate measure for diagnosis of English preschool children with SLI by itself. It seems necessary therefore to carrying out more studies with larger samples. It is worth to note here that the results of Stokes *et al*.'s study on diagnostic power of NWR and SR as language processing markers for SLI in Cantonese show that unlike English children, Cantonese preschool children with SLI do not score significantly lower than their age-matched peers on NWR task. Moreover, although SR was able to show group differences, at the individual level, this task has good specificity but unacceptable sensitivity.[29] Hence, the results of Stokes *et al*.'s study suggest that may

be no limitation in pWM in Cantonese-speaking children with SLI or may be the executed tasks need and essay skills other than pWM. Stokes *et al*. proposed that poorer NWR capacity for English-speaking children with SLI might be related to weaker use of the red-integration strategy in word repetition.[29] These results imply that the clinical accuracy of NWR tasks may be related not only to individual subjects differences in language use and exposure but also to the language(s) tested. Hence, it is clear that further cross-linguistic investigations of language processing strategies including NWR are required.

Thordardottir *et al*. used tests of NWR, SR, following directions and digit span as linguistic processing markers for differential diagnosis of 5-year-old French-speaking children with SLI. Their results show that although the digit span test is not sensitive enough to detect Italian SLI children, the diagnostic powers of other tests are adequate.[2]

Kapalková *et al*. developed a fast and easily-administered NWR task and determined the performances of this task and its different scoring methods in distinguishing between Slovak-speaking children with SLI and TD children. The NWR task used in their study differs from English NWR tasks in number of items per length and scoring methods.[33] As could be seen in Table 3, whole-item scoring method (number of correctly repeated consonants) has good sensitivity and specificity, but the diagnostic performance of vowel scoring method (number of correctly repeated vowels in addition consonants) is not fair due to the sensitivity of 75%. As Archibald and Gathercole *et al*. and Kapalková *et al*. found in their study that children repeat high word-like nonwords better than low word-likes;[33,46] this finding implies the influence of accumulated language knowledge on the performance of item repetition. The results of Dispaldro *et al*.'s study on real word repetition and NWR in normally developing children confirm this finding too. What was interesting in the results of Dispaldro *et al*.'s study was the strength of real word repetition in predicting the grammatical ability of children rather than NWR.[48] Hence, Dispaldro *et al*. appraised the diagnostic performance of both real word repetition and NWR in differentiating between Italian-speaking children with or without SLI and propound the question whether real word repetition could be as effective as NWR as a clinical marker for Italian-speaking children with SLI.[31] The high diagnostic value of NWR for identification of Italian preschool children with SLI had been marked in previous study by Bortolini *et al*.[28] As can be seen in Table 3, not only nonwords but also real words show good to excellent sensitivity and specificity with both the two scoring methods.

Besides NWR and the other language processing measures, some linguistic features also have been surveyed and introduced by researchers as potential clinical markers for SLI in a variety of languages. For example, Rice and Wexler suggested that certain aspects of verb morphology, such as tense-marking, are especially difficult for SLI children and may constitute clinical marker which can improve the identification of SLI.[49]

From the 23 included studies, 12 studies evaluated the diagnostic performances of some linguistic measures elicited from spontaneous speech samples or linguistic probes as potential clinical markers. Among them, 4 studies have been performed on English, 3 on Italian, 2 on Chinese (Cantonese), and one on each of French, Spanish, and Persian samples. Since there are many dissimilarities in the linguistic characteristics of different languages, the results of these studies cannot be assimilated.

To finding potential clinical markers for SLI, researchers chiefly focused on those linguistic features that the previous studies consistently shown that are problematic for this group of language-impaired children[50] such as many tense/agreement morphemes in English. These morphemes include third-person singular – s, past tense – ed, both copula and auxiliary – is, are, am, and auxiliary – do, did, and does. Among included English studies, 2 studies (23 and 25) investigated the diagnostic values of specific combinations of these morphemes extracted from spontaneous speech.

Gladfelter and Leonard evaluated the diagnostic accuracy of two composite measures of tense/agreement from spontaneous speech (tense marker total and productivity score developed by Hadley and Short, 2005[51]) besides the diagnostic accuracy of more traditional measure of finite verb morphology composite (FVMC) adapted from Leonard *et al*.[52] to determine whether these new composite measures could be serve as better identifiers for SLI children. The actual difference between these measures is in the number of obligatory contexts found for each morpheme. The FVMC is a combination of the number of obligatory contexts for all tense/agreement morphemes that divided into the total number of tense/agreement morphemes actually produced. In contrast, Hadley and Short's measures of spontaneous tense/agreement morpheme emphasize on the diversity of contexts in which these morphemes are used by diverse scoring and excluding contexts that are often associated with nonanalyzed productions.[23]

The results of Bedore and Leonard's study on diagnostic performances of FVMC have been shown that this measure has acceptable sensitivity and specificity.[53] The results of Gladfelter and Leonard show such as FVMC, these newly introduced measures seem largely successful in distinguishing 4- and 5-year-old children with SLI from their

TD age-mates, but their power of diagnosis is not beyond the FVMC's power. Furthermore, their results imply that the combination of the FVMC measure and the measures of Hadley *et al*. would seem to be most informative.[23]

Souto *et al*. studied the diagnostic value of measures of global and developmental level of a child's tense/agreement morpheme use. Their results show although the diagnostic values of the two types of measures that provided developmental levels of tense/agreement morpheme use are not satisfactory, the diagnostic accuracy of traditional FVMC that involves a smaller collection of tense/agreement morphemes but treats all of these morphemes equally can be considered acceptable. Furthermore, their results show that among other studied global measure of grammatical accuracy, sentence point, and overall developmental sentence score, sentence point could be introduced as a suitable tool for identifying 4- and 5-year-old children with SLI, but the diagnostic accuracy of the overall DSS is not acceptable. Hence, the results of this study indicate that different grammatical measures do not yield equivalent results for children with SLI.[25]

The results of Conti-Ramsden and Conti-Ramsden and Hesketh studies on the diagnostic performances of grammatical marking (include tense-marking and plural-marking extracted via language probes) in distinguishing preschool children with SLI show neither past tense marking nor noun plural has acceptable diagnostic.[15,16] Furthermore, the results of Conti-Ramsden (2003) show that although the combination of past tense task and CNRep could be served as a diagnostic tool for differential diagnosis of preschool children with SLI, neither of them has acceptable sensitivity in separation.[15]

As mentioned previously, the most important trait of variables that can be labeled as clinical markers is low within-group variability in performance of SLI children as a group and the total absence of an overlap of scores for the SLI and TD groups in these measures.[29] However, the results of included studies on diagnostic performances of potential clinical markers for SLI in English again confirm the previously suggested idea that many measures yield significant group differences do not necessarily meet the higher standard of reliable identification of language impaired children individually.

Studies carried out to determine the diagnostic accuracy of linguistic indexes in distinguishing Italian children with SLI from normal children mainly focused on articles, clitics, and third-person plural inflections, separately or jointly, as more problematic aspects of language in Italian-speaking children with SLI. Table 3 provides a summary of the results. The results show that there are some disagreements between the findings of Bortolini *et al*. (2002) and Bortolini *et al*. (2006). For example, third-person plural inflections, when considered alone, have acceptable diagnostic performances in one study while do not have sufficiently high sensitivity in another although specificity is quite good.[26,28] Since the two studies carried out on the same status and the subjects of two studies were similar in age and severity of language impairment, Bortolini *et al*. mentioned that the origin of these inconsistencies between the results is not clear,[28] but differences in IQ level may be a determinant factor. Clitics have acceptable sensitivity and specificity in the two studies. An outcome which can be seen in Table 3 is the improvement of diagnostic accuracy when two or more measures are considered together; in the other word, the values improved or stayed without considerable changes (did not become poorer) when the measures were used jointly. Therefore, these results suggest the value of considering measures together.

Fair to good discriminant accuracy has also been reported for grammatical markers in one study carried out in another Romance language, Spanish.[32] This study concerned with the utility of tense as a clinical marker of SLI and authors used two different methods of data extraction including experimental methods (elicited production and grammaticality choice task) and spontaneous speech sample analysis (to extract six distinct indices of grammatical development). Their results show that Spanish-speaking children with SLI have problem with tense, and tense marking could be introduced as a potential clinical marker for SLI. Moreover, their results indicate that elicited production test has the most balanced accuracy for both sensitivity and specificity. Furthermore, some combined functions of experimental and spontaneous measures such as mean length of utterance by morphemes (MLU-m) + grammaticality choice task or elicited production task + mean length of terminable unit have good diagnostic performances.[32]

Gross indexes from spontaneous speech (including MLU by words and MLU-m) did not achieve acceptable discriminant accuracy in the Thordardottir *et al*.'s study on children speaking the other romance language, French. Moreover, this study examined the diagnostic power of a range of French standardized measures of language (including receptive vocabulary [by EVIP], receptive grammar [by Test for Auditory Comprehension of Language], and narrative production [by Edmonton Narrative Norms Instrument]) and language processing.[2] As could be seen in Table 3, except narrative production indexes, other standardized measures of language and language processing provide accurate diagnostic tools for SLI in French.

Among 3 included studies on Cantonese language, Klee *et al*. and Wong *et al*. used a composite variable made up of MLU,

lexical diversity (D), and age in their study as the index measure.[27,30] The results of Klee *et al*.'s study show this composite variable has excellent discriminative potential.[27] In spite of good diagnostic performances, because of the wide confidence intervals for sensitivity and specificity due in part to the sample size, Klee *et al*. cautioned that before recommending this measure for clinical use, its accuracy must be re-examined in another independent sample of Cantonese-speaking children.[27] The aim of Wong *et al*.'s study was replicating Klee's study in a second, independent sample of Cantonese-speaking children with or without SLI. Unlike the findings of the original study, the results of Wong *et al*. demonstrate that this measure cannot be used as an accurate instrument for the diagnosis of SLI because neither the sensitivity nor specificity values were acceptable.[30] Hence, regarding the results of these two studies, to ensure about the clinical usefulness of a diagnostic test or measure, it is helpful or even necessary to evaluate its diagnostic values in different studies on the target populations.

Finally, among included studies, one study is about the performances of language measures derived from play-based, conversational language samples in diagnosis of Persian preschool children with SLI. The results of this study show that although the majority of measures extracted from language samples were capable in differentiating children with or without SLI at the group level, only three of these measures exhibited good diagnostic performances at the individual level [Table 3].[35]

CONCLUSION

The results of this study demonstrate that any test/measure that initially shows acceptable diagnostic power should subsequently be put to the test of replication in other accuracy studies on different samples. Among included studies, only a few studies compared a single diagnostic measure across different groups of samples. Moreover, the numbers of studies that compare the performance of more than one diagnostic test/measure on a single sample of children are limited across studies. If more than one test done simultaneously on one population, comparative information can be obtained and then the relative performance of the tests can be described. Hence, an important outcome of this study is the value of considering measures together to improve the diagnostic accuracy.

In addition, the results particularly encourage cross-linguistic research. Tests that have been standardized on specific population are not suitable for other populations, and specific linguistic or even processing measures are not applicable as diagnostic markers in different languages.

The results of this review also reveal that standardized tests vary in how sensitive they are to language impairment and also there is no single cutoff point which is appropriate across tests. It is notable that in most studies, the empirically derived cutoff score which provides the highest discriminative capacity is not the same as statically estimated cutoff point.

The final point then must be emphasis is the construction of the subjects. In all of included studies, the number of SLI subject is nearly equal to the number of normally developing subjects and the SLI group mostly constituted from clinically referred sample. Hence, it is clear that obtained values are not necessarily generable to general population of preschool-aged children, where the prevalence of SLI is nearly 7%.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

AUTHORS' CONTRIBUTIONS

- TM contributed in the conception and design of the work, conducting the study, drafting and revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work
- ShJ contributed in the conception and design of the work, conducting the study, revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work
- FH contributed in the conception of the work, revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work
- ZS contributed in the conception of the work, revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work
- PN contributed in the conception of the work, revising the draft, approval of the final version of the manuscript, and agreed for all aspects of the work.

REFERENCES

1. Leonard LB. Children with Specific Language Impairment. 1st ed. Massachusetts: MIT Press; 2000. p. 3-25.
2. Thordardottir E, Kehayia E, Mazer B, Lessard N, Majnemer A, Sutton A, *et al.* Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. J Speech Lang Hear Res 2011;54:580-97.
3. Tomblin JB, Records NL, Buckwalter P, Zhang X, Smith E, O'Brien M. Prevalence of specific language impairment in kindergarten children. J Speech Lang Hear Res 1997;40:1245-60.
4. Maleki Shahmahmood T, Soleymani Z, Jalaei S. A comparison study in test of language development (TOLD) and speech samples between children with specific language impairment and their

MLU matched group. Mod Rehabil 2009;2:25-33.
5. Maleki Shahmahmood T, Soleymani Z, Faghihzade S. The study of language performances of Persian children with specific language impairment. Audiology 2011;20:11-21.
6. van der Lely HK, 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:e22432.
7. van der Lely HK, Marshall CR. Assessing component language deficits in the early detection of reading difficulty risk. J Learn Disabil 2010;43:357-68.
8. Maleki Shahmahmood T, Nakhostin Ansari N, Soleymani Z. Methods for identification of specific language impairment. Audiology 2014;23:1-18.
9. Tomblin JB, Records NL, Zhang X. A system for the diagnosis of specific language impairment in kindergarten children. J Speech Hear Res 1996;39:1284-94.
10. Spaulding TJ, Plante E, Farinella KA. Eligibility criteria for language impairment: Is the low end of normal always appropriate? Lang Speech Hear Serv Sch 2006;37:61-72.
11. Sackett DL, Haynes RB. Evidence base of clinical diagnosis; the architecture of diagnostic research. Br Med J 2002;324:539-41.
12. Kazemi Y, Stringer H, Klee T. Study of child language development and disorders in Iran: A systematic review of the literature. J Res Med Sci 2015;20:66-77.
13. Plante E, Vance R. Selection of preschool language tests: A data-based approach. Lang Speech Hear Serv Sch 1994;25:15-24.
14. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1994;271:389-91.
15. Conti-Ramsden G. Processing and linguistic markers in young children with specific language impairment (SLI). J Speech Lang Hear Res 2003;46:1029-37.
16. Conti-Ramsden G, Hesketh A. Risk markers for SLI: A study of young language-learning children. Int J Lang Commun Disord 2003;38:251-63.
17. Gray S. Diagnostic accuracy and test-retest reliability of non-word repetition and digit span tasks administered to preschool children with specific language impairment. J Lang Commun Disord 2003;36:129-51.
18. Perona K, Plante E, Vance R. Diagnostic accuracy of the structured photographic expressive language test: Third edition (SPELT–3). Lang Speech Hear Serv Sch 2005;36:103-15.
19. Oetting JB, Cleveland LH. The clinical utility of nonword repetition for children living in the rural south of the US. Clin Linguist Phon 2006;20:553-61.
20. Pankratz ME, Plante E, Vance R, Insalaco DM. The diagnostic and predictive validity of the Renfrew Bus Story. Lang Speech Hear Serv Sch 2007;38:390-9.
21. Greenslade KJ, Plante E, Vance R. The diagnostic accuracy and construct validity of the structured photographic expressive language test – preschool: Second edition. Lang Speech Hear Serv Sch 2009;40:150-60.
22. Deevy P, Weil LW, Leonard LB, Goffman L. Extending use of the NRT to preschool-age children with and without specific language impairment. Lang Speech Hear Serv Sch 2010;41:277-88.
23. Gladfelter A, Leonard LB. Alternative tense and agreement morpheme measures for assessing grammatical deficits during the preschool period. J Speech Lang Hear Res 2013;56:542-52.
24. Spaulding TJ, Hosmer S, Schechtman C. Investigating the interchangeability and diagnostic utility of the PPVT-III and PPVT-IV for children with and without SLI. Int J Speech Lang Pathol 2013;15:453-62.
25. Souto SM, Leonard LB, Deevy P. Identifying risk for specific language impairment with narrow and global measures of grammar. Clin Linguist Phon 2014;28:741-56.
26. Bortolini U, Caselli MC, Deevy P, Leonard LB. Specific language impairment in Italian: The first steps in the search for a clinical marker. Int J Lang Commun Disord 2002;37:77-93.
27. Klee T, Stokes SF, Wong AM, Fletcher P, Gavin WJ. Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. J Speech Lang Hear Res 2004;47:1396-410.
28. Bortolini U, Arfé B, Caselli CM, Degasperi L, Deevy P, Leonard LB. Clinical markers for specific language impairment in Italian: The contribution of clitics and non-word repetition. Int J Lang Commun Disord 2006;41:695-712.
29. Stokes SF, Wong AM, Fletcher P, Leonard LB. Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. J Speech Lang Hear Res 2006;49:219-36.
30. Wong AM, Klee T, Stokes SF, Fletcher P, Leonard LB. Differentiating Cantonese-speaking preschool children with and without SLI using MLU and lexical diversity (D). J Speech Lang Hear Res 2010;53:794-9.
31. Dispaldro M, Leonard LB, Deevy P. Real-word and nonword repetition in Italian-speaking children with specific language impairment: A study of diagnostic accuracy. J Speech Lang Hear Res 2013;56:323-36.
32. Grinstead J, Baron A, Vega-Mendoza M, De la Mora J, Cantú-Sánchez M, Flores B. Tense marking and spontaneous speech measures in Spanish specific language impairment: A discriminant function analysis. J Speech Hear Res 2013;56:352-63.
33. Kapalková S, Polišenská K, Vicenová 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.
34. Katzenberger I, Meilijson S. Hebrew language assessment measure for preschool children: A comparison between typically developing children and children with specific language impairment. Lang Test 2014;31:19-38.
35. Kazemi Y, Klee T, Stringer H. Diagnostic accuracy of language sample measures with Persian-speaking preschool children. Clin Linguist Phon 2015;29:304-18.
36. Merrell AW, Plante E. Norm-referenced test interpretation in the diagnostic process. Lang Speech Hear Serv Sch 1997;28:50-8.
37. Crestani AH, Oliveira LD, Vendruscolo JF, Ramos-Souza AP. Specific language impairment: The relevance of the initial diagnosis. Rev CEFAC 2012;15:228-36.
38. Gray S, Plante E, Vance R, Henrichsen M. The diagnostic accuracy of four vocabulary tests administered to preschool-age children. Lang Speech Hear Serv Sch 1999;30:196-206.
39. Betz SK, Eickhoff JR, Sullivan SF. Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. Lang Speech Hear Serv Sch 2013;44:133-46.
40. Dollaghan C, Campbell TF. Nonword repetition and child language impairment. J Speech Lang Hear Res 1998;41:1136-46.
41. Gathercole SE, Willis CS, Baddeley AD, Emslie H. The children's test of nonword repetition: A test of phonological working memory. Memory 1994;2:103-27.
42. Montgomery JW. Working memory and comprehension in children with specific language impairment: What we know so far. J Commun Disord 2003;36:221-31.
43. Montgomery JW, Magimairaj BM, Finney MC. Working memory and specific language impairment: An update on the relation and perspectives on assessment and treatment. Am J Speech Lang Pathol 2010;19:78-94.
44. Archibald LM, Joanisse MF. On the sensitivity and specificity of

non-word repetition and sentence recall to language and memory impairments in children. J Speech Hear Res 2009;52:899-914.
45. Conti-Ramsden G, Botting N, Faragher B. Psycholinguistic markers for specific language impairment (SLI). J Child Psychol Psychiatry 2001;42:741-8.
46. Archibald LM, Gathercole SE. Nonword repetition in specific language impairment: More than a phonological short-term memory deficit. Psychon Bull Rev 2007;14:919-24.
47. Campbell T, Dollaghan C, Needleman H, Janosky J. Reducing bias in language assessment: Processing-dependent measures. J Speech Lang Hear Res 1997;40:519-25.
48. Dispaldro M, Deevy P, Altoé G, Benelli B, Leonard LB. A cross-linguistic study of real-word and non-word repetition as predictors of grammatical competence in children with typical language development. Int J Lang Commun Disord 2011;46:564-78.
49. Rice ML, Wexler K. Toward tense as a clinical marker of specific language impairment in English-speaking children. J Speech Hear Res 1996;39:1239-57.
50. Simon-Cereijido G, Gutierrez-Clellen VF. Spontaneous language markers of Spanish language impairment. Appl Psycholinguist 2007;28:317-39.
51. Hadley PA, Short H. The onset of tense marking in children at risk for specific language impairment. J Speech Lang Hear Res 2005;48:1344-62.
52. Leonard LB, Miller C, Gerber E. Grammatical morphology and the lexicon in children with specific language impairment. J Speech Lang Hear Res 1999;42:678-89.
53. Bedore LM, Leonard LB. Specific language impairment and grammatical morphology: A discriminant function analysis. J Speech Lang Hear Res 1998;41:1185-92.