Middle East J Rehabil Health Stud. 2022 October; 9(4):e129345. doi: 10.5812/mejrh-129345.

Published online 2022 September 16. **Systematic Review**

A Systematic Review of Psychometric Properties of the Vocabulary Tests for Iranian Persian-Speaking Children: Current Status and Future Directions

Farhad Sakhai [1], Mozhgan Asadi [1], Kowsar Baghban [2], Golnoosh Golmohamadi [1,*] and Talieh Zarifian [3]

[1] Department of Speech Therapy, School of Rehabilitation, Semnan University of Medical Sciences, Semnan, Iran
[2] Department of Speech Therapy, School of Rehabilitation, Hamadan University of Medical Sciences, Hamadan, Iran
[3] Department of Speech Therapy, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran

[*] *Corresponding author*: Department of Speech Therapy, School of Rehabilitation, Semnan University of Medical Sciences, Semnan, Iran. Email: go.golmohammadi@semums.ac.ir

Received 2022 June 26; **Revised** 2022 August 20; **Accepted** 2022 August 21.

Abstract

Context: One aspect of spoken language skills is vocabulary, which provides a basis for acquiring other language aspects. Assessing a child's vocabulary knowledge aids in identifying the child's language strengths and weaknesses and predicts reading ability and academic success. Speech-language pathologists frequently employ various procedures in clinical and research settings to assess the children's language skills and help make decisions about diagnosis, eligibility for services, and intervention.
Objectives: This systematic review investigated currently available vocabulary tests developed or adapted for Iranian Persian-speaking children.
Data Sources: Based on the PRISMA guideline, electronic searches of three national (SID, Irandoc, and Magiran) and four international (ScienceDirect, ProQuest, PubMed, and Google scholar) databases were carried out from 2000 to 2022 to identify Persian vocabulary assessment tools.
Study Selection: Search in the reference lists of papers, unpublished theses, and content of related journals also supplemented the database searches.
Data Extraction: The psychometric properties of these tests were reviewed based on specific criteria used in the literature. The papers and test manuals were examined according to these criteria.
Results: Eight tools have been developed or adapted for assessing vocabulary knowledge in Iranian Persian-speaking children. Reviewing the content and psychometric properties of the included tools indicated that the Test of Language Development- Primary:3 (TOLD-P:3) is the only accessible published tool with the most reported psychometric evidence. It measures language development in children; however, it is a multi-modal test that includes vocabulary subtests.
Conclusions: This review revealed that most of the reviewed tools were in the primitive stages of test development or adaptation procedures and did not examine many psychometric properties. As a result, vocabulary is a field that requires more attention because there is no accessible, standardized tool with adequate psychometric properties.

Keywords: Vocabulary, Psychometric, Validity, Reliability, Assessment, Systematic Review, Children

1. Context

Language acquisition is one of the most critical domains in child development (1). Children should acquire various language aspects in typical development. The words are small semantic units and make the basic building blocks of language. Production of first words is considered one of the early developmental language milestones (2). Expressive vocabulary development begins with the emergence of the first words around 12 months (3, 4), but the pace of vocabulary acquisition is usually slow to 18 months (5). The rapid increase in vocabulary development occurs when the lexicon is close to the border of about 50 words, and at this time, many children experience a vocabulary spurt (3, 4, 6, 7). It seems that verbs are more complicated than nouns, so they are acquired later in typical development (8-10).

Vocabulary has a vital role in developing information exchangeability (11). Therefore, the richness of vocabulary knowledge guarantees successful and appropriate com-

munication (12). On the other hand, the early lexicon has a predictive value for later language and literacy skills (13-15), and vocabulary development difficulties are one of the primary indicators of language impairment (14). The complex construct of lexical knowledge has been studied regarding the distinction between receptive versus expressive vocabulary and breadth versus depth of vocabulary (16, 17). In oral language, receptive vocabulary reflects the words children recognize by hearing, and expressive vocabulary reflects the words they produce (18). Receptive and expressive vocabularies have different growth rates (19, 20). The extent or how many words one knows represents a breadth, and the conceptual familiarity of the word denotes the depth dimension of vocabulary knowledge. However, the distinctions mentioned above are helpful for research and education purposes. It is important to remember that these components are separate but interrelated (19-21).

Given the above distinctions in vocabulary knowledge, various tools are designed to address these different domains. Vocabulary skills are often assessed in children as a method of screening, diagnosing a possible impairment, setting intervention goals, measuring change following an intervention, and tracking change over time (22-24). Several assessment approaches have been used to study vocabulary skills, including standardized vocabulary tests, parent/caregiver report measures, spontaneous speech sample analysis, and researcher-made assessments (25). Each of these approaches has advantages and disadvantages. Nonetheless, standardized tests are popular for assessing vocabulary (25). These standardized tests include the Peabody Picture Vocabulary Test (26), the expressive one-word picture vocabulary test (27), and the receptive one-word picture vocabulary test (28), all of which have been adapted to various languages. The MacArthur-Bates communicative development inventories (CDIs) are the most commonly used parent reporting tool (29). The CDIs are divided into three forms for three age groups (8 to 16 months, 16 to 30 months, and 30 to 36 months), each equipped with a vocabulary section (comprehension and production). The CDIs' vocabulary includes words from various semantic and grammatical classes. Adaptations of the CDIs have been developed in some languages other than English, such as French (30), Arabic (31), Danish (32), and Italian (33). The decision-making process of the assessment approach is influenced by factors such as the purpose of assessment, vocabulary dimension (receptive-expressive and breadth-depth), and age of the participants (25). Furthermore, the psychometric properties of a tool must be determined whether it is intended for clinical or research use (24, 34, 35). When there is no evidence of the tool's psychometric properties, the evaluation results are unreliable and raise concerns (36).

2. Objectives

This systematic review aimed to (1) identify, describe, and appraise the psychometric quality of available vocabulary assessment tools for Persian-speaking children; and (2) determine knowledge gaps in evidence and identify areas that need further research.

3. Data Sources

The preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines (37) were used to guide the literature search and review process in the present systematic psychometric review.

3.1. Eligibility Criteria

The tools were included for review if they (1) were adapted or developed in the Persian language; (2) evaluated expressive, receptive vocabularies or both; (3) reported at least one aspect of validity or reliability; (4) included children up to middle childhood (i.e., children to 12 years old); (5) reported in the field of speech and language pathology, linguistics, and child psychology; (6) were published between 2000 and 2022; and (7) were accessible by published manuals, full-text research papers, or documents. The tools were excluded if they (1) were adapted or developed on bilingual or multilingual children; (2) examined other aspects of lexical processing such as word association and word finding; (3) reported the use of a researcher-made tool without providing validity or reliability; and (4) examined the depth of vocabulary knowledge.

The hallmark of a systematic review is to reduce bias at all phases of the review process (38). To help identify and avoid selection bias, the review protocol was written (but not registered due to time constraints of the research team), and the selection criteria of the studies were defined before the start of the study. Two authors independently reviewed the protocol for ensuring that the study design, that is, the procedures for study selection, were appropriate for addressing the study hypotheses (selection bias).

3.2. Information Sources

Few publishers publish tests related to speech and language in Iran. Moreover, when a test is developed, the articles are usually published at the beginning, followed by

the test booklet. Therefore, firstly, databases were searched to find available tools. The candidate tools were identified by search in electronic databases and hand searching using a third-stage approach. National and international electronic databases were identified during the first stage of electronic searching. Persian national databases in which academic publications are recorded include (1) Scientific Information Database (http://www.sid.ir); (2) Iranian research institute for information sciences and technologies (http://irandoc.ac.ir); and (3) Journal information bank (www.magiran.com). The international electronic databases included ScienceDirect, ProQuest, PubMed, and Google Scholar search engine.

During the second stage, a manual search was conducted in the libraries of three universities (The University of Social Welfare and Rehabilitation Science, Rehabilitation Faculty of Iran University of Medical Sciences, and Rehabilitation Faculty of Tehran University of Medical Sciences) in Tehran, Iran. These universities educate graduate and postgraduate students in speech and language pathology and are pioneers in the fields of language and test design. In this stage, the lists of master's and doctoral dissertations that were not published as articles were screened. Thus, all of the related theses and research projects were retrieved.

During the third stage, the table of contents of most relevant Iranian peer-reviewed journals centered on publishing studies related to the topic of this review was examined in the manual search. The searched journals included journal of Audiology, journal of paramedical sciences and rehabilitation, journal of modern rehabilitation, journal of Rehabilitation, journal of research in Rehabilitation sciences, Koomesh journal, and Pajouhan scientific journal. Only national journals were manually reviewed for two reasons: first, Iranian researchers are more likely to publish their papers in national journals, and second, because of a lack of specialized Persian terminology. To reduce the selection bias, searches for relevant literature were conducted in multiple commercial databases, grey literature sources, citations, and via hand searching to identify as many relevant studies as possible.

4. Study Selection

4.1. Search Strategy

Due to a lack of unified keywords for professional terminology among Iranian researchers, different search terms were used to retrieve all related tests in the abovementioned databases. The following search terms were used for Iranian databases in the Persian language: "/vɑegan/ for vocabulary," "/bæyɑni/ for expressive," "/dærki/ for receptive," "/ʔærzjɑbi/ for assessment," "/ʔɑzmun/ for test," "/rævɑji/ for validity," "/ʔeʔtebɑr/ for reliability," "/tæklif/ for task," and /kudækɑn/ for children." English databases were searched using search terms (MeSH and text words) related to (1) construct (vocabulary); (2) population (children); (3) instrument (task, test); and (4) designing and measurement properties (development, adaptation, validity, reliability, and psychometric) based on the search filters guide suggested by Terwee (39). To limit the search results to publications in Persian, all the combinations included the "Persian" term. The only Boolean operator used in the search queries was "AND." The restriction on the publication date was from 2000 to 2022. The database searches were limited to titles and abstracts. We used the snowballing technique to maximize the identification of relevant studies in this stage (40). In other words, the reference lists for retrieved documents were manually searched.

4.2. Study Records

All found references were imported into EndNote (EndNote X8, Thomson Reuters), and duplicate studies were removed. Next, two reviewers (T.Z. and K.B.) independently reviewed the studies identified based on research questions and assessed the eligibility criteria in two phases. First, the titles and abstracts of all retrieved studies were reviewed, and studies that did not meet the inclusion criteria were excluded. Second, if the eligibility of studies was not clear from abstracts, the full-text versions were examined. Then, the reference lists of relevant studies were searched. In the end, the final set of identified studies was reviewed. Figure 1 illustrates the flow diagram of the search process for all phases of the study, from identification to selection according to the eligibility criteria.

5. Data Extraction

5.1. Data Collection Process

Two authors (F.S. and G.G.) independently evaluated and extracted study characteristics and data from included studies to reduce information bias. All included studies were evaluated qualitatively in terms of psychometric properties that largely followed the criteria used in McCauley and Swisher (41), Bogue, DeThorne (42), and McCauley and Strand (43) studies. These criteria are summarized below. The strategy used for resolving disagreements between the two authors was discussion.

The definition of standardization sample: The clear definition of normative sample in the test manual in terms of (1) geographic residence; (2) socioeconomic status; (3)

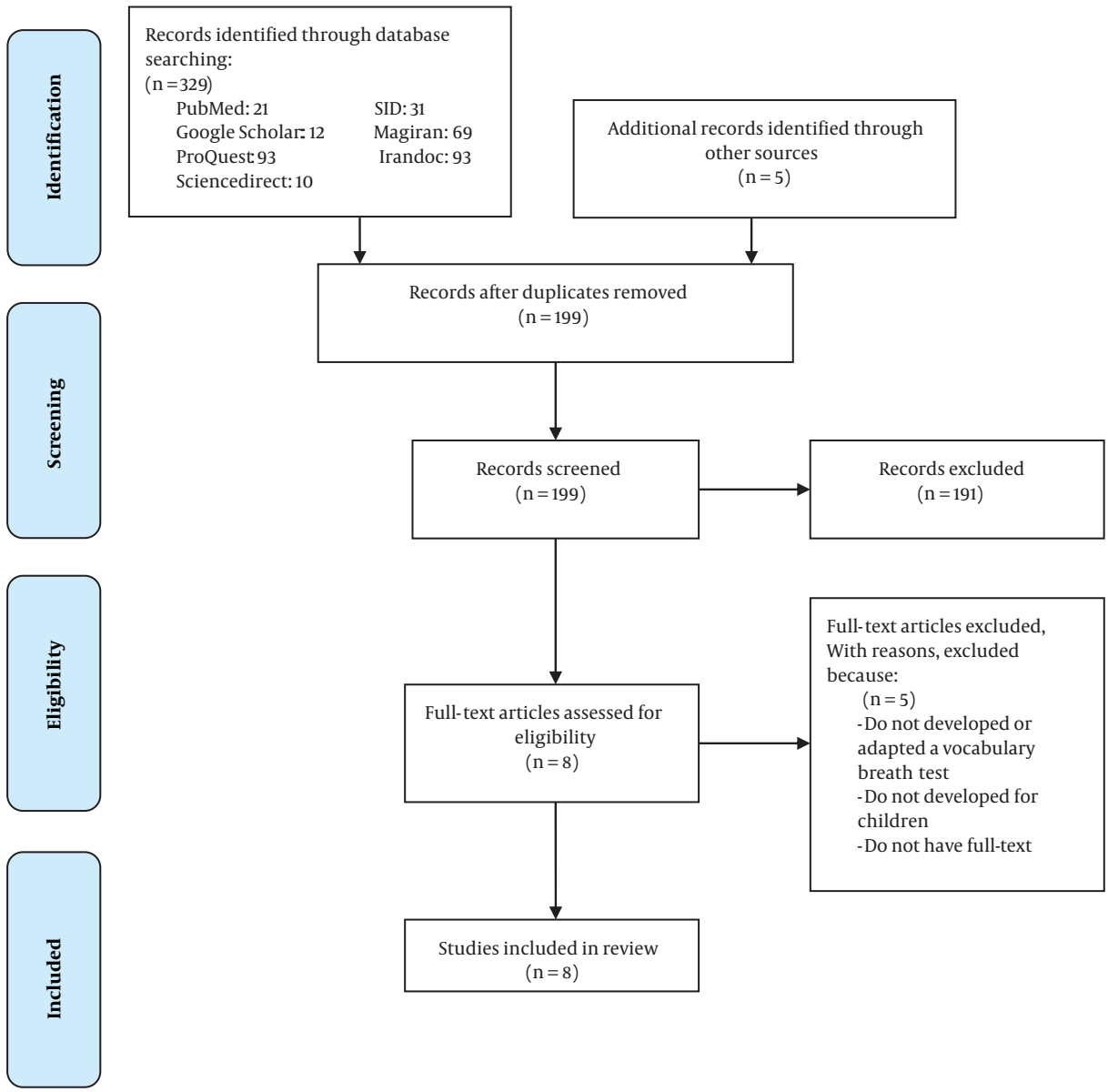

Figure 1. Flow diagram of study selection

the "normalcy" of subjects in the sample, including the number and cause of excluding children; and (4) recency.

The adequate sample size for each subgroup examined during standardization: Subgroups with sample sizes of 100 or more.

- The content validity-item-analysis: The quantitative methods used to study and control item difficulty, item validity, or both.

- Mean and standard deviation: For total raw scores of all relevant subgroups.

- Concurrent validity: The agreement results obtained from other valid methods of categorizing children as normal or impaired.

- Predictive validity: The test's ability to predict later performance on another valid criterion of speech or language behavior addressed by the same test in question.

- Construct validity: (1) evidence from confirmatory factor analysis study; (2) evidence from the test performance improvement with increasing age (developmental trends); and (3) evidence of expected group differences in the test

performance (group comparison).

- Internal consistency reliability: 0.70 - 0.90 as an acceptable correlation coefficient for each age group.

- Test-retest reliability: Correlation coefficient of 0.90 or above at a significance level of 0.05 or lower.

- Inter-rater reliability: Correlation coefficient of 0.90 or above at a significance level of 0.05 or lower.

5.2. Detailed Descriptions of Test Administration and Scoring Procedures

- The description of special qualifications: required for the test administrator or scorer.

The psychometric characteristics of the vocabulary assessment tools were extracted using the above-mentioned criteria. The second author prepared the data collection form (M.A.), and explanations about the items were provided to data extractors (F.S. and G.G.) at the beginning of the data extraction process and regularly throughout the project. These criteria were used with caution because they have been supposed for purely normative studies, but more reviewed studies were in the primitive phases of test development or adaptation.

5.3. Reliability

Before progressing to the full-text review phase, the second reviewer re-examined a random sample of 20% abstracts to evaluate reliability. Inter-rater reliability index (kappa Cohen) was 89% for the abstract review process. For the inter-rater reliability of full-text screening, a random sample of 20% full-texts by the same reviewer was re-examined to assess eligibility. Inter-rater reliability index (kappa Cohen) was 96% for the full-text review process.

6. Results

A total of 329 studies were identified through searches in databases. After removing duplicated studies, 199 articles remained. In the next step, 191 studies that did not meet inclusion criteria based on the abstracts were excluded. The full text of another study was not available. The eligibility of another five studies was not confirmed based on a full-text review. Three studies were excluded following a review of the full text of retrieved studies because they were related to developing a test for assessing word association, rapid automatic naming, and word finding. Two studies were appropriate for adults that did not meet the inclusion criteria. Reference lists for relevant documents were manually searched to maximize the identification of eligible studies, and five articles were identified. In the second and third stages of the search, no new studies were identified. Finally, eight studies were identified as eligible that developed a tool to assess vocabulary for Iranian Persian-speaking children. A summary of information about each tool is presented in Table 1. These studies were related to developing or adapting and determining psychometric properties of vocabulary tests. Five studies were related to the adaptation of the original version: peabody picture vocabulary test (PPVT) (44), test of language development- primary: 3 (TOLD-P: 3) (45), British Picture Vocabulary Scale (BPVS-II) (46), MacArthur-Bates communication development inventory: Persian version (CDI-I: P) (47), and short form of Persian picture vocabulary scale (PPVS) (48). Three other studies were related to the development of a new vocabulary test in the Persian language: receptive picture vocabulary test (RPVT-I) (49), receptive picture vocabulary test (RPVT-II) (50), and picture verb test (PVT) (51). All of these tests provide assessments of vocabulary breadth knowledge. Five tests (RPVT-I, RPVT-II, PPVT, BPVS-II, and PPVS) only assess receptive vocabulary; the PVT assesses expressive vocabulary, and two others (TOLD-P: 3 and CDI-I: P) target both receptive and expressive vocabulary knowledge. The TOLD-P: 3 and CDI-I: P are not unimodal, but only the vocabulary section is examined in the present study. In other words, whenever we talk about a CDI-I: P, we mean the word section, and whenever we talk about the TOLD-P: 3 test, we mean the two oral and picture vocabulary sub-tests.

Regarding administration and response elicitation, participants should indicate their responses by selecting the target picture from an array of four receptive ones or naming the picture in expressive ones. However, the CDI-I: P is the only tool that includes a parent report checklist. The RPVT-II and PPVS are not timed. The CDI-I: P takes between 20 and 40 minutes to administer, depending on the mother's level of education. The administration time for other tests is not specified. Except for the PVT and CDI-I: P, the remaining vocabulary tests assess lexical noun categories. All reviewed studies were reported in journal articles, except for the TOLD-P: 3, which was also published as an accessible vocabulary assessment tool. As a result, the tests' information was obtained from their subsequent articles. Most of them were carried out as master's thesis projects, and they could only be accessed with the permission of test developers. Some of them (TOLD-P: 3 and PPVT) have provided norm scores in large groups, and researchers have claimed that they can be used as reference (norm-reference) for identifying clinical samples. Three tools (TOLD-P: 3, PPVT, and CDI-I: P) included some items related to noun, verb, and adjective, whereas two others had no items related to the verb: the RPVT-I and RPVT-II. The grammatical category of the items is not mentioned in the PPVS and BPVS-II. The PVT was designed exclusively

Table 1. Summary of Iranian Studies on Test Development for Vocabulary Assessment in Persian-Speaking Children

Test	Study	Subtests	Number of Items	Age Range	Sample Size	Picture Plate Description
Peabody picture vocabulary test (PPVT)	Razavieh and Shahim (44)	Receptive vocabulary	97	3 - 11 years	1010	Full-color drawings; One picture per plate
Test of language development-primary: 3 (TOLD-P: 3)	Hasanzadeh and Minaei (45)	Picture vocabulary and oral vocabulary subtests	30; 28	4 - 8 years	1235	Color drawings; One picture per plate for picture words subtest; Four pictures per plate for the spoken word subtest
MacArthur-Bates Communication Development Inventory: Persian version (CDI-I: P)	Kazemi et al. (47)	Words and Gestures:	680	8 - 16 months	30	Checklists
Picture verb test (PVT)	Soltaninejad et al. (51)	Expressive verb vocabulary	55	36 - 54 months	106	One picture per plate
British picture vocabulary scale; (BPVS-II)	Kazemi et al. (46)	Receptive vocabulary	168	5 - 11 years	180	Four white and black pictures per plate
Receptive picture vocabulary test (RPVT-I)	Hassanpour et al. (49)	Receptive vocabulary	240	30 - 71 months	91	Color drawings; Four pictures per plate
Receptive picture vocabulary test (RPVT-II)	Salehi Zahabi et al. (50)	Receptive vocabulary	240	6 - 13 years	118	Color drawings; Four pictures per plate
Short form of Persian Picture Vocabulary Scale (PPVS)	Pouretemad et al. (48)	Receptive vocabulary	38	5 - 6 years	Pilot study: 100; Original study: 410	White and Black line drawings

to assess the expression of verbs. The familiarization process and practice items were explicitly mentioned in the PVT and the TOLD-P: 3.

6.1. Criterion 1 (Standardization Sample)

Our investigation of this criterion focused primarily on the TOLD-P: 3, the PPVT, and the BPVS-II, all claiming to have attempted to standardize a vocabulary test. However, the sample's quality was examined in other studies. The TOLD-P: 3 standardization sample represented the population proportion in different geographic regions of Tehran city and the socioeconomic status proportion in different regions based on census data. This test's normative sample included both typical and atypical children (e.g., intellectual disabilities and learning disorders) and did not exclude them. No consensus exists on whether atypical individuals should be included in the normative sample. Proponents of including atypical individuals believe that such a sample would more accurately represent the full range of language abilities (34, 52).

For the PPVT, the distribution of socioeconomic status in the standardization sample was close to its proportion in the target province's population based on census data. However, no explanation has been provided for the geographical residence, normalcy violation, and the PPVT's approach to dealing with it. The BPVS-II study failed to meet any of the properties of criterion 1. In this study, only the sampling method (simple random) is mentioned without considering geographical residence and population proportion. Similarly, in other studies, only the sampling method has been mentioned.

Another consideration is the recency of the standardization sample. Because the population can change over time, children should not be measured against an out-of-date sample. The vocabulary area of language development, especially in children, is sensitive to changes over time. Another issue to consider is the appearance of depicted objects and the visual attractiveness of test pictures. The TOLD-P: 3 test (i.e., the only test with a published manual) was developed in 2002 and has not been revised since then. Furthermore, there is no mention of the year in which the sample data were collected.

6.2. Criterion 2 (Sample Size)

Some studies (e.g., the PPVT) grouped subjects by age at six-month intervals, while others (e.g., the TOLD-P: 3)

grouped subjects by year. The TOLD-P: 3 test only provided an adequate sample size in each of the five age sub-groups ranging from four to eight years. The PPVT provided an adequate sample size only in each of the six age sub-groups from six to 11 years Subjects were grouped into 6-months intervals from 36 to age 72 months and yearly intervals from 6 through age 11 years. In BPVT, none of subgroups met the age criterion: ages five (n = 35), seven (n = 46), nine (n = 50), and 11 (n = 49). The PPVS provided an adequate sample size in each age group. The remaining five tests (PVT, RPVT-I, RPVT-II, and CDI-I: P) failed to meet this criterion. The CDI-I: P, similar to other tools, did not have any normative data due to the small sample size and was at a preliminary level of adaptation. However, it was found through email contact with the corresponding author that the determining psychometric characteristics of the newest edition of CDI has been done, and standardization is scheduled to take place soon.

6.3. Criterion 3 (Content Validity)

In the PPVS, TOLD-P: 3, and PVT, the classical approach to item analysis (i.e., estimating the difficulty level and the discriminate power of items) was used. The content validity was estimated using the Content Validity Index (CVI) in the PVT, RPVT-I, and RPVT-II and the content validity rating (CVR) in the RPVT-I and PVT. The BPVS-II did not use any method to evaluate content validity.

6.4. Criterion 4 (Mean and Standard Deviation)

Central tendency indices scores are reported in the PPVT, BPVS-II, PVT, TOLD-P: 3, RPVT-I, and RPVT-II for age subgroups. In addition, three studies, PPVT, RPVT-I, and RPVT-II, also reported these indices for gender subgroups. These indices were not reported in the CDI-I: P.

6.5. Criterion 5 (Concurrent Validity)

The PPVT provided a weak correlation between test scores and academic achievement in one age group. Evidence for this criterion has been provided for the picture vocabulary and oral vocabulary subtests of TOLD-P: 3 by examining the correlation between the similarities and vocabulary subtests of the Wechsler test. In BPVS-II, concurrent validity has been investigated by examining correlations between vocabulary scores and the verbal, practical, and general intelligence scores of the Wechsler test. Five reviewed tests did not provide evidence of concurrent validity (47-51).

6.6. Criterion 6 (Predictive Validity)

None of the reviewed tools examined the predictive validity evidence.

6.7. Criterion 7 (Construct Validity)

Of the eight reviewed tests, only the TOLD-P: 3 used the exploratory factor analysis to assess the construct validity. Developmental trends, increasing the mean raw scores in the form of one-year age ranges, were considered the second evidence to assess the construct validity. An increase was observed in the mean of raw scores among age subgroups of BPVS-II, TOLD-P: 3, and PPVS. Mean scores increased across age subgroups in the PVT (three age groups: 36 to 42 (M = 37/35), 42-48 (M = 39/06), and 48-54 (M = 42/97)) and RPVT-I (means presented in six-months intervals). In the RPVT-II, mean scores increased with age, except for 8 to 9 and 9 to 10 years, where mean scores remained constant. There was no information about the subjects' mean scores in the CDI-I: P paper. The mean scores of the PPVT revealed developmental trends across all age groups.

Only the TOLD-P: 3 provided evidence for group comparisons. Children with learning disorders, speech and language development delay, mental retardation, and attention deficit hyperactivity disorder were assessed. The mean differences between the groups with disorder and the normative or control groups were more than 2 SD in picture vocabulary and oral vocabulary subtests. Gender differences in test scores were compared in two reviewed tests (PPVS and RPVT-I), but the findings were insignificant.

6.8. Criterion 8 (Internal Consistency)

One method of estimating the reliability of a test or scale is to calculate the correlation coefficient among items (Cronbach's alpha coefficient). The RPVT-II (0.83), PVT (0.71), TOLD-P: 3 (with Cronbach's alpha for oral and picture vocabulary subtests of 0.89 and 0.76, respectively), PPVS (0.84), CDI-I: P and BPVS-II (0.84) passed the 0.70-0.90 criterion. The vocabulary production (0.87) and vocabulary comprehension (0.98) subscales of CDI-I: P showed the highest values but these Cronbach's alpha coefficients, especially in the case of the second subscale, which is above 0.9, probably indicate the presence of highly related and redundant items. The internal consistency of two tests (RPVT-I and PPVS) was assessed by measuring the split-half reliability.

6.9. Criterion 9 (Test-Retest Reliability)

Four tests (TOLD-P: 3, PPVS, CDI-I: P, and BPVS-II) did not meet this criterion because they did not assess the test-retest correlation. In four reviewed tests (RPVT-I, RPVT-II, PVT, and PPVT) with reported test-retest reliability, the values were above 0.70.

6.10. Criterion 10 (Inter-rater Reliability)

None of the reviewed tools examined the inter-rater reliability evidence.

6.11. Criterion 11 (Administration and Scoring)

Four tests (PVT, TOLD-P: 3, RPVT-I, and RPVT-II) provided brief descriptions of administration and scoring procedure, but administration procedure was described by only two: PPVS and CDI-I: P. The TOLD-P: 3 was the only reviewed test that provided standard score and percentile cut-off point information.

6.12. Criterion 12 (Qualification)

None of the reviewed tools described the required qualifications for the administration and scoring.

The results of the review of the psychometric characteristics of vocabulary assessment tools are provided in Table 2.

7. Discussion

This systematic review aimed to investigate the psychometric characteristics of the available vocabulary assessment tools in the Persian language. Eight tools were identified. This review revealed that vocabulary is one of the language areas in Iran, we have no available standardized assessment tool with enough psychometric properties, and many present tests failed to meet psychometric expectations. Similar studies on language assessment tools in English (36, 41, 53, 54) and other languages also demonstrate that many reviewed tools did not meet all of the psychometric criteria (55). In terms of validity, quantitative evidence of content validity has not been reported in only three tools (49-51). More evidence of construct validity has been reported for the multi-dimensional TOLD-P: 3 test. The evidence for criterion validity, especially concurrent, has been reported for only three tools (44-46), but predictive validity has not been examined in any tool. Regarding reliability, internal consistency was reported by almost all tools, and only four tools examined test-retest reliability (44, 45, 50, 51). Any of the tools did not meet the predictive validity, inter-examiner reliability, and description of examiner qualification.

The surprising issue is that, despite much research in recent years in test adaptation and development in the speech and language pathology profession in Iran, there is only one commercially published standardized multidimensional test with vocabulary subtests (45) that has not been revised after two decades of development. There appears to have been no improvement in the overall quality of the tests since 2000. A further issue that should be considered is that the objectives of the development of these tools have not been explicitly mentioned, and this issue may call into question the validity and reliability of these tools for diagnostic accuracy. If the results of evaluating children with poor tests are used to diagnose and determine intervention priorities, this can harm clinical and research activities. It appears that the researchers did not address the diagnostic accuracy of the tests, although this aspect is critical in psychometric studies today. This is one of the disadvantages of studies. Another disadvantage is that research studies, including test adaptation or development, do not complement each other. As a result, the researchers and clinicians are confronted with multiple vocabulary tests at a primitive level of test development that may not even be available.

In the meantime, clinicians are positioned to have the CDI-I: P contained in the child language assessment package for monitoring progress or only accessible standardized language assessment tool (TOLD-P: 3) to measure vocabulary in Persian-speaking children. Alternatively, they may design their informal measurements. In any case, the use and interpretation of the results of each of these should be cautious. Nevertheless, it seems that these psychometric inadequacies of vocabulary assessment tools in the Persian language have some probable reasons. One possible explanation for the current situation is that graduate students lack appropriate tools for evaluating various speaking and language skills when deciding on a topic for their thesis. On the one hand, using researcher-made tasks and assessment tools makes it more difficult to publish papers from their theses in international journals. As a result, most test development research is done as a thesis by master's students. Also, issues such as time constraints for graduation and a lack of experience and knowledge in this field can lead to current test problems. Future studies in the research field of test development or adaptation, particularly in the vocabulary area, appear to require more attention and broader collaboration of researchers in related fields in order to attract financial support. In this case, we can finally hope that future tools will provide a gold standard for assessing vocabulary. However, the work of the researchers of these eight tests is commendable, as they have taken the first steps in providing vocabulary assessment tools, and their work will serve as the foundation for future improvements in developing vocabulary tests.

7.1. Limitations

Our study has some limitations. First, because some abstracts and full-text articles were only available in Persian, the readers of the present review may be unable to appraise the study results fully. Second, a limited number of included studies met eligibility criteria. Third, a manual or electronic search was not taken in a limited number of rehabilitation universities in Iran. It would be better to include the libraries of all universities that research the de-

Table 2. Psychometric Criteria Reported by Vocabulary Assessment Tools in Persian Speaking Children

Psychometric Criteria	Tests							
	PPVT	TOLD-p: 3	CDI-I: P	PVT	BPVS: I	RPVT-	RPVT-II	PPVS
Standardization sample								
a	-	√	-	-	-	-	-	-
b	√	√	-	-	-	-	-	-
c	-	-	-	-	-	-	-	-
d	-	-	-	-	-	-	-	-
Adequate sample size	√	√	-	-	-	-	-	√
Content validity	-	√	-	√	-	√	√	√
Mean and standard deviation	√	√	√	√	√	√	√	√
Concurrent validity	√	√	-	-	√	-	-	-
Predictive validity	-	-	-	-	-	-	-	-
Construct validity								
a	-	√	-	-	-	-	-	-
b	√	√	-	√	√	-	-	√
c	-	√	-	-	-	√	-	√
Internal consistency reliability	√	√	√	√	√	√	-	√
Test-retest reliability	√	-	-	-	-	√	√	-
Inter-rater reliability	-	-	-	-	-	-	-	-
Test administration and scoring procedures	-	√	√	√	-	√	√	√
Qualifications for the test administrator or scorer	-	-	-	-	-	-	-	-

Abbreviation: PPVT, peabody picture vocabulary test; TOLD-P: 3, test of language development- primary: 3 (TOLD-P: 3); BPVS-II, British Picture Vocabulary Scale; CDI-I:P, MacArthur-Bates Communication Development Inventory: Persian version; PPVS, short form of Persian picture vocabulary receptive; RPVT-I, picture vocabulary test; RPVT-II, receptive picture vocabulary test; PVT, picture verb test.

velopment and adaptation of language tests for children in Iran.

Footnotes

Authors' Contribution: Study concept and design, F. S., G. G., M. A., and T. Z.; Analysis and interpretation of data, F. S. and G. G.; Drafting of the manuscript, M. S.; Critical revision of the manuscript for important intellectual content, F. S., M. A., and G. G.; Statistical analysis, K. B., T.Z., and F.S.

Conflict of Interests: The first and fourth authors are family members. The second author is a member of the editorial board of the Middle East Journal of Rehabilitation and Health Studies, but she was not involved in the review process.

Data Reproducibility: The data presented in this study are uploaded during submission as a supplementary file and are openly available for readers upon request.

Funding/Support: The authors did not receive any funding for this study.

References

1. Kazak Berument S, Guven AG. [Turkish expressive and receptive language test: I. Standardization, reliability and validity study of the receptive vocabulary sub-scale]. *Turk Psikiyatri Derg*. 2013;**24**(3):192–201. Turkish. [PubMed: 24049009].
2. Hoff E. *Language development*. 5th ed. USA: Cengage Learning; 2013. 480 p.
3. Bornstein MH, Cote LR, Maital S, Painter K, Park SY, Pascual L, et al. Cross-linguistic analysis of vocabulary in young children: spanish, dutch, French, hebrew, italian, korean, and american english. *Child Dev*. 2004;**75**(4):1115–39. doi: 10.1111/j.1467-8624.2004.00729.x. [PubMed: 15260868].
4. Bloom P. *How children learn the meanings of words*. USA: MIT press; 2002.
5. Fenson L, Dale PS, Reznick J, Bates E, Thal DJ, Pethick SJ, et al. Variability in Early Communicative Development. *Monogr Soc Res Child Dev*. 1994;**59**(5). i185. doi: 10.2307/1166093.
6. Bates E, Dale PS, Thal D. Individual differences and their implications for theories of language development. In: Fletcher P, MacWhinney B, editors. *The handbook of child language*. **30**. New Jersey, USA: Blackwell Publishing Ltd; 1995. p. 96–151.
7. Goldfield BA, Reznick JS. Early lexical acquisition: rate, content, and the vocabulary spurt. *J Child Lang*. 1990;**17**(1):171–83. doi: 10.1017/s0305000900013167. [PubMed: 2312640].

8. Gentner D. *Why Nouns Are Learned before Verbs: Linguistic Relativity Versus Natural Partitioning. Technical Report No. 257*. Champaign: University of Illinois; 1982. Report No.: 257.
9. Gleitman LR, Cassidy K, Nappa R, Papafragou A, Trueswell JC. Hard Words. *Lang Learn Dev*. 2005;**1**(1):23–64. doi: 10.1207/s15473341lld0101_4.
10. Gleitman L. The Structural Sources of Verb Meanings. *Lang Acquis*. 1990;**1**(1):3–55. doi: 10.1207/s15327817la0101_2.
11. Paul R. *Language disorders from infancy through adolescence: Assessment and intervention*. **324**. Netherlands: Elsevier Health Sciences; 2007.
12. McGregor KK, Oleson J, Bahnsen A, Duff D. Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. *Int J Lang Commun Disord*. 2013;**48**(3):307–19. doi: 10.1111/1460-6984.12008. [PubMed: 23650887]. [PubMed Central: PMC3648877].
13. Paul R. Predicting outcomes of early expressive language delay: Ethical implications. In: Bishop DVM, Leonard LB, editors. *Speech and language impairments in children: Causes, characteristics, intervention and outcome*. 1st ed. USA: Psychology Press; 2000. p. 195–209.
14. Lee J. Size matters: Early vocabulary as a predictor of language and literacy competence. *Applied Psycholinguistics*. 2010;**32**(1):69–92. doi: 10.1017/s0142716410000299.
15. Libertus ME, Odic D, Feigenson L, Halberda J. A Developmental Vocabulary Assessment for Parents (DVAP): Validating parental report of vocabulary size in 2-to 7-year-old children. *J Cogn Dev*. 2015;**16**(3):442–54. doi: 10.1080/15248372.2013.835312.
16. Christ T. Moving Past "Right" or "Wrong" Toward a Continuum of Young Children's Semantic Knowledge. *J Lit Res*. 2011;**43**(2):130–58. doi: 10.1177/1086296x11403267.
17. Schoonen R, Verhallen M. The assessment of deep word knowledge in young first and second language learners. *Lang Test*. 2008;**25**(2):211–36. doi: 10.1177/0265532207086782.
18. Nation ISP. Teaching and Learning Vocabulary. In: Hinkel E, editor. *Handbook of Research in Second Language Teaching and Learning*. 1st ed. UK: Routledge; 2005. p. 605–20. doi: 10.4324/9781410612700-44.
19. Tomasello M. *Constructing a Language*. Massachusetts, USA: Harvard University Press; 2005. doi: 10.2307/j.ctv26070v8.
20. Webb S. Receptive and Productive Vocabulary Sizes of L2 Learners. *Stud Second Lang Acquis*. 2008;**30**(1). doi: 10.1017/s0272263108080042.
21. Webb S. Depth of Vocabulary Knowledge. *The Encyclopedia of Applied Linguistics*. USA: Wiley; 2012. doi: 10.1002/9781405198431.wbeal1325.
22. Tomblin JB, Records NL, Zhang X. A system for the diagnosis of specific language impairment in kindergarten children. *J Speech Hear Res*. 1996;**39**(6):1284–94. doi: 10.1044/jshr.3906.1284. [PubMed: 8959613].
23. Paul R, Norbury C. *Language disorders from infancy through adolescence-E-Book: Listening, speaking, reading, Writing, and Communicating*. 2nd ed. Netherlands: Elsevier Health Sciences; 2012.
24. Dockrell JE, Marshall CR. Measurement Issues: Assessing language skills in young children. *Child Adolesc Ment Health*. 2015;**20**(2):116–25. doi: 10.1111/camh.12072. [PubMed: 32680388].
25. Pan BA. Assessing Vocabulary Skills. In: Hoff E, editor. *Research Methods in Child Language: A Practical Guide*. USA: Wiley; 2011. doi: 10.1002/9781444344035.ch7.
26. Dunn LM, Dunn DM. *Peabody Picture Vocabulary Test*. 4th ed. USA: APA PsycTests; 2007. doi: 10.1037/t15144-000.
27. Brownell R. *Expressive one-word picture vocabulary test: Manual*. Novato, CA: Academic Therapy Publications; 2000.
28. tafiadis D, Karagianni E, Tafiadi M. The Expressive and the Receptive One Word Picture Vocabulary test (EOWPVT & ROWPVT). (A combine pilot study and validation of the tests' in normal Greek population - aged from 6 years till 6 years and 11 months). *Ann Gen Psychiatry*. 2010;**9**(1). S105. doi: 10.1186/1744-859X-9-S1-S105.
29. Fenson L, Marchman VA, Thal DJ, Dale PS, Reznick JS, Bates E. *MacArthur-Bates Communicative Development Inventories*. USA: APA PsycTests; 2006. doi: 10.1037/t11538-000.
30. Kern S. Lexicon development in French-speaking infants. *First Lang*. 2007;**27**(3):227–50. doi: 10.1177/0142723706075789.
31. Abdelwahab AGS, Forbes S, Cattani A, Goslin J, Floccia C. An Adaptation of the MacArthur-Bates CDI in 17 Arabic Dialects for Children Aged 8 to 30 Months. *Lang Learn Dev*. 2021;**17**(4):425–46. doi: 10.1080/15475441.2021.1916502.
32. Bleses D, Vach W, Slott M, Wehberg S, Thomsen P, Madsen TO, et al. The Danish Communicative Developmental Inventories: validity and main developmental trends. *J Child Lang*. 2008;**35**(3):651–69. doi: 10.1017/S0305000907008574. [PubMed: 18588718].
33. Rinaldi P, Pasqualetti P, Stefanini S, Bello A, Caselli MC. The Italian Words and Sentences MB-CDI: normative data and concordance between complete and short forms. *J Child Lang*. 2019;**46**(3):546–66. doi: 10.1017/S0305000919000011. [PubMed: 30773152].
34. Andersson L. Determining the adequacy of tests of children's language. *Commun Disord Q*. 2005;**26**(4):207–25. doi: 10.1177/15257401050260040301.
35. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;**21**(4):651–7. doi: 10.1007/s11136-011-9960-1. [PubMed: 21732199]. [PubMed Central: PMC3323819].
36. Friberg JC. Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Lang Teach Ther*. 2010;**26**(1):77–92. doi: 10.1177/0265659009349972.
37. Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*. 2009;**151**(4):264–9. W64. doi: 10.7326/0003-4819-151-4-200908180-00135. [PubMed: 19622511].
38. Mueller M, D'Addario M, Egger M, Cevallos M, Dekkers O, Mugglin C, et al. Methods to systematically review and meta-analyse observational studies: a systematic scoping review of recommendations. *BMC Med Res Methodol*. 2018;**18**(1):44. doi: 10.1186/s12874-018-0495-9. [PubMed: 29783954]. [PubMed Central: PMC5963098].
39. Terwee CB, Jansma EP, Riphagen ,I, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res*. 2009;**18**(8):1115–23. doi: 10.1007/s11136-009-9528-5. [PubMed: 19711195]. [PubMed Central: PMC2744791].
40. Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *J Clin Epidemiol*. 2005;**58**(5):444–9. doi: 10.1016/j.jclinepi.2004.09.011. [PubMed: 15845330].
41. McCauley RJ, Swisher L. Psychometric review of language and articulation tests for preschool children. *J Speech Hear Disord*. 1984;**49**(1):34–42. doi: 10.1044/jshd.4901.34. [PubMed: 6700200].
42. Bogue E-L, DeThorne LS, Schaefer BA. A Psychometric Analysis of Childhood Vocabulary Tests. *Contemporary Issues in Communication Science and Disorders*. 2014;**41**(Spring):55–69. doi: 10.1044/CICSD_41_S_55.
43. McCauley RJ, Strand EA. A review of standardized tests of nonverbal oral and speech motor performance in children. *Am J Speech Lang Pathol*. 2008;**17**(1):81–91. doi: 10.1044/1058-0360(2008/007). [PubMed: 18230815].
44. Razavieh A, Shahim S. [Adaptation and standardization of Peabody picture vocabulary test]. *Journal of Psychology and Educational Sciences*. 1998;**57**(3):25–40.
45. Hasanzadeh S, Minaei A. [Adaptation and Standardization of the Test of TOLD-P: 3 for Farsi - Speaking Children of Tehran]. *Journal of Exceptional Children*. 2002;**1**(2):119–34. Persian.
46. Kazemi MS, amiri S, Malekpour M, Molavi H. [Psychometric properties (standardization, validity, and reliability) of british picture vocabulary scale (bpvs_ii)]. *Training Management*. 2012;**9**(3):123–44. Per-

sian.
47. Kazemi Y, Nematzadeh S, Hajian T, Heidari M, Daneshpajouh T, Mirmoeini A. [The validity and reliability coefficient of Persian translated McArthur-Bates Communicative Development Inventory]. *Journal of Research in Rehabilitation Sciences*. 2008;**4**(1):45–51. Persian.
48. Pouretemad H, Mosa-Kazemi M, Hooman A, Sadeghi MS, Hasanzada-Tavakoli MR. [Psychometric Properties of the Short Form of Persian Picture Vocabulary Scale]. *Adv Cogn Psychol*. 2011;**13**(3):1–8. Persian.
49. Hassanpour N, Jalilevand N, Masumi E, Ghorbani A, Kamali M. Development of a picture receptive vocabulary test and evaluation of its validity & reliability for normal 36-71 months Persian children. *Journal of Paramedical Sciences & Rehabilitation*. 2015;**4**(3):34–43.
50. Salehi Zahabi S, Ghorbani A, Jalilehvand N, Kamali M. [Development and determine the Psychometric properties of Picture perceptive Objective Vocabulary Test for normal Persian-speaking 6-13 years-old children]. *Modern Rehabilitation*. 2016;**9**(6):159–67. Persian.
51. Soltaninejad N, Ghorbani A, Salehi M, Fakhrrahimi S. Development of picture verb test for 36-54 month-old normal Persian-speaking children and determination of its validity and reliability. *Aud Vestib Res*. 2012;**21**(3):70–6.
52. DeThorne LS, Schaefer BA. A guide to child nonverbal IQ measures. *Am J Speech Lang Pathol*. 2004;**13**:275–90. doi: 10.1044/1058-0360(2004/029).
53. Eisenberg SL, Hitchcock ER. Using standardized tests to inventory consonant and vowel production: A comparison of 11 tests of articulation and phonology. *Lang Speech Hear Serv Sch*. 2010;**41**(4):488–503. doi: 10.1044/0161-1461(2009/08-0125).
54. Plante E, Vance R. Selection of Preschool Language Tests. *Lang Speech Hear Serv Sch*. 1994;**25**(1):15–24. doi: 10.1044/0161-1461.2501.15.
55. McLeod S, Verdon S. A review of 30 speech assessments in 19 languages other than English. *Am J Speech Lang Pathol*. 2014;**23**(4):708–23. doi: 10.1044/2014_AJSLP-13-0066. [PubMed: 24700105].