TPM Vol. 32, No. S9, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

CONSTRUCTION AND STANDARDIZATION OF AN ACHIEVEMENT TEST FOR ENGLISH LANGUAGE SKILLS OF ELEMENTARY LEVEL STUDENTS

BHARTI GARG
RESEARCH SCHOLAR DEPARTMENT OF EDUCATION CHITKARA UNIVERSITY RAJPURA (PB) 140401 INDIA,
EMAIL: gargbharti2010@gmail.com

DR. SANGEETA PANT
PROFESSOR & DEAN DEPARTMENT OF EDUCATION CHITKARA UNIVERSITY RAJPURA (PB) 140401 INDIA,
EMAIL: Sangeeta.pant@chitkara.edu.in

Abstract: The proposed study will develop, build, and standardise a valid and reliable Achievement Test in English Language Skills among elementary level students. In accordance with the importance of English proficiency in the development of literacy, the research is conducted with systematic test-development process encompassing planning, preparation of successive drafts, pilot testing, item analysis and standardization. The first draft of the questionnaires consisted of 76 items ready in four abilities, i.e., listening, speaking, reading and writing, and filtered with the help of the professionals into a second draft of 55 items. A sample of 200 students took this draft and item analysis was done on the basis of upper and lower 27% groups on the basis of Difficulty Value (DV) and Discriminating Power (DP). Following the application of DV and DP, 10 unworthy items were eliminated leaving an English Language Skills Test (ELST) of 45 items. To determine the reliability, the test-retest method was used on another sample of 100 students and the correlation coefficient was 0.87, which indicated that it is very stable. The resulting tool gives a psychometrically reliable measure of English language skill to the teacher and researchers and offers a means of conducting pedagogical improvement at the elementary level.

Keywords: Achievement test, English language skills, elementary students, standardization, item analysis, reliability, validity

1.INTRODUCTION:

English has become one of the most powerful languages of the twenty-first century that may be both a lingua franca in the world, as well as a scholarly lingua franca, and a language of communication between multilingual communities (1). In the Indian context where hundreds of languages exist, English is a very crucial link language between various linguistic groups and access to educational, economic, and technological opportunities (2).

The elementary school level does not only make English a mandatory academic lesson, but also the cornerstone that future literacy, understanding, and formal communication skills are anchored. Early exposure of children to English is extremely important in determining their cognitive growth, their academic confidence, and their engagements in learning tasks at higher levels (3,4). Since English is the language of textbooks, instructions, and examinations in many state and central schools boards, the future success in the school system and the academic future becomes impossible without proficiency in English (5). Hence, acquisition of English language competency at the early level is not only a learning necessity but a crucial factor in academic equity and life-long learning (6).

The four interdependent skills included in English language learning at the elementary level comprise of listening, speaking, reading and writing which combined with each other lead to communicative competence in children (7). Listening provides the basis of vocabulary development, phonological awareness, and understanding; speaking allows expressiveness and classroom issues; reading helps in the acquisition of knowledge and development of a critical thinker; and writing fosters the organization of ideas, the use of grammar, and expression of literacy (8). Combined, these competencies will affect the overall academic achievement of children in all subjects since their skills in comprehending instruction, comprehending textual data and the ability to respond to questions will greatly rely on their mastery of the skills of the English language (9).

In studies, it has always been shown that students who have better language development in the early-grade levels have better learning achievements in mathematics, science, and social studies implying how the base of language proficiency in English plays a significant role in holistic academic performance (10,11). With the integration of competency-based curricula in educational systems based on national frameworks, proper measurement of these four domains of language can be even more critical (12). The achievement tests take center stage to determine the level of mastery of the students in terms of academic skills, areas of deficiency in learning and the pedagogical decisions (13). Achievement tests, which are well designed, can be used in the field of learning English language as a way of determining whether learners have mastered the understanding at the grade level about comprehension, vocabulary, grammar and communication (14).

In standardized achievement tests, especially, objectivity, reliability and comparability of the results among schools and groups of the population are guaranteed (15). They give instructors some diagnostic data very important in developing remedial teaching, assist policy makers to assess the efficacy of curricula and aid researchers in analyzing literacy patterns by region (16). In the elementary level, standardized tests provide an empirical insight into the effectiveness with which the initial years of language education are performing and particularly in multilingual situations with a wide range of exposure to English (17,18).

Nevertheless, to be effective in fulfilling these purposes achievement tests have to be designed in accordance with the systematic processes which involve blueprinting, item writing, expert validation, pilot testing, and item analysis with statistics (19). Although the assessment of English language is crucial, their current tools in elementary level have limited scope and standard. A good percentage of the available evaluation tool are limited to specific grammar elements, vocabulary awareness or isolated reading comprehension exercises without the representation of the entire range of language skills needed to be considered communicatively competent (20). Majority of the standardized achievement tests that have been created in India and other countries are tailored towards secondary or higher-secondary students to portray the examination expectations at these levels instead of the developmental requirements of young learners (21).

In addition to this, most school level tests are not psychometrically rigorous and are not developed in a systematic manner, which compromises their reliability and fairness (22). This has led to teachers often using subjective, teacher-designed tests that differ in the level of difficulty, the distribution of content, and the accuracy of scoring (23). These inconsistencies tend to make teachers fail to see a clear image of the level of learning of students and confuse the discovery of real learning problems (24). This discrepancy is more alarming in situations that involve the second or third language of English and where students must be provided with a systematic scaffold in order to reach grade-level competence (25). The turn towards evidence-based educational practice across the globe has solidified the need to have powerful, standardised instruments, to evaluate the result of learning language at a lower level (26). The competency-based education, the foundational literacy and numeracy (FLN), and the practices of continuous assessment, based on the principles of objective measurement, are prioritized by the international or national curricular frameworks, including NEP-2020 (27).

These policy guidelines require developmentally appropriate, psychometrically sound, and measurement tools that can track the progress of the learners over time. The development of such tools is to be done in accordance with the demonstrated principles of test development that encompass the correspondence of the items to the learning objectives, the content validity being secured by the means of the expert review, and the empirical validation of the test in terms of the difficulty and discrimination indices (28). Standardized English Language Skills Tests that are aimed at elementary students should thus evaluate listening, speaking, reading and writing abilities within a balanced approach as they should be able to reflect the benchmarking of the curriculum and language developmental patterns (29).

Based on these requirements, the current paper engages in the development and standardization of an English Language Skills Test (ELST) that will be focused on elementary-level students. Also in contrast to the current assessment methods, which emphasize individual grammar elements and advanced grade proficiency standards, the ELST is structured to assess all four key language skills with psychometrically acceptable items. The test development process involves paying attention to the creation of the first set of items according to curriculum specifications, the further refinement of this set in the course of the expert assessment, the testing of the final version on a representative sample, and the item analysis in terms of Difficulty Value (DV) and Discrimination Power (DP) to understand that only high-quality items remained (30). Insisting on matching the evaluation to the developmental stage of elementary learners and using the strict methods of statistical processing, the ELST should provide educators, administrators, and researchers with the reliable instrument to help them diagnose the level of language proficiency, enhance instructional practices, and increase the overall quality of early English language education.

2. REVIEW OF LITERATURE

2.1 Achievement Tests in Language Education

Achievement tests are widely known as important devices in assessing how learners have acquired instructional goals, especially in language learning where comprehension and communication as well as expression is to be assessed in a methodical manner (31). The achievement tests are curriculum-based and unlike proficiency tests which examine the general communicative ability, achievement tests are interested in establishing whether the learners have mastered certain skills they were taught during a given time (32). When applied in the context of English language learning, the tests can assess the progress in listening, speaking, reading, and writing, and help teachers to understand the strengths, learning gaps, and introduce specific pedagogical interventions (33).

According to scholars, achievement tests do not only measure the result of learning, but also cause changes in instructional planning, curriculum development and educational accountability (34). An effective language achievement test will also involve various areas of linguistic competence such as vocabulary, grammar, comprehension, phonological awareness and written expression. It is in line with grade-level outcomes, which makes the content developmentally engaging in terms of cognitive and linguistic development in the learners (35). In multilingual environments like India, in which English is a second or third language, objective achievement test is even more important to provide equitable assessment since exposure to English differs greatly among social-economic and geographic groups (36). Some researches underline that the timely and proper evaluation of the language skills leads to future success in school as students who have good background language skills are likely to succeed in all their subject areas (37). Hence, achievement tests are inevitable tools of enhancing language proficiency, quality education, and effective classroom teaching (38).

2.2 Construction and Standardization of Achievement Tests Procedures:

The construction and standardization of a reliable language achievement test require adherence to globally accepted psychometric principles. Test development typically involves sequential stages, including:
(a) identification of content and objectives,
(b) blueprinting,
(c) expert validation,
(d) item writing and editing,
(e) pilot testing,
(f) item analysis using Difficulty Value (DV) and Discrimination Power (DP), and
(g) establishing reliability and validity (39).

In this step, the identification of content and objectives is carried out. Development of language achievement test begins with clear learning objectives that represent the standards of curriculum and learner expectations. Such objectives should be quantifiable, behavioural and they should be in tandem with instructional objectives in the various fields of language (40). The scholars also note that it is critical to make test items connected with curricular competencies so that the same content is relevant and equal to the test-takers (41). The identification of content usually includes going through text books, syllabi, competency frameworks, and past examination papers in order to identify critical skills and language configurations that should be used in the target grade level (42).

2.2.2 Blueprinting and Test Specifications:

Preparation of a test specification, also known as blueprinting, is done to assure that test items are effectively distributed in terms of language abilities, cognitive abilities and content area. This will ensure that over-representation of some aspects (e.g., grammar) and under-representation of others (e.g., listening or speaking) is prevented (43). The blueprint is of high quality and it balances the items in the taxonomy of Bloom; knowledge, understanding, application, and analysis, so that the test measures the development of lower-order and higher-order thinking (44). Blueprints have been indicated to promote content validity and the coherence of structure of standardized tests (45).

2.2.3 Expert Validation:

Expert review is an important process in getting rid of ambiguity, biasness, redundancy of content, and complex language that can be detrimental to the learners (46). Within panels, there are typically language educators, psychometricians and subject specialists who assess the items in terms of being clear, relevant, representativeness and correspondence to the cognitive goals (47). Research proves that, when subjected to expert validation, the item content validity ratio (CVR) is enhanced as well as the overall quality of the test is strengthened before pilot administration (48).

2.2.4 Writing, editing and piloting of Items:

When a certain item has been proven, it is revised according to the recommendations of specialists and gathered into a working version. On a small scale, pilot testing is useful in recognizing unexpected problems during the wording of items, time allocation, or understanding by the student (49). The pre-test administration also serves as source of item difficulty and discrimination which are vital in perfecting the draft (50).

2.2.5 Analysis of the Items in terms of Difficulty Value (DV) and Discrimination Power (DP):

Item analysis forms the backbone of test standardization. Two classical indices are used:

- **Difficulty Value (DV)** indicates how many students answered the item correctly. Items that are too easy or too difficult fail to differentiate levels of ability (51).
- **Discrimination Power (DP)** measures how well an item distinguishes between high and low achievers, typically using the 27% Kelley method (52).

Items with DP values above 0.30 are generally considered acceptable, while those below 0.19 are usually discarded (53). Likewise, DV values within the moderate range (0.25–0.75) are preferred as they contribute to the test's reliability and discriminatory efficiency (54).

2.2.6 Establishing Reliability and Validity:

A standardized achievement test must demonstrate reliability—consistency of measurement across time, items, and populations—and validity—accuracy in measuring what it intends to measure (55). Reliability may be estimated through test–retest, split-half, Kuder-Richardson (KR-20/21), or Cronbach's alpha methods. Validity involves content, criterion, and construct validation procedures (56). Literature shows that language achievement tests achieving reliability coefficients above 0.70 and demonstrating strong correlations with external criteria are considered psychometrically robust (57). Furthermore, validity evidence ensures that interpretations drawn from test scores are meaningful and appropriate (58).

2.3 Previous Standardized Tests in English / Language Skills:

The methodological precedents of standardized language achievement tests are found in previous studies on the topic. Several scholars have come up with grammar, vocabulary or reading-based achievement tests to school students, which is often followed by systematic processes of blueprint creation, professional assessment and systematic analysis of items (59). An example is that standardized grammar and comprehension tests designed to test middle-school students usually start with large item pools of over 70100 items then are refined through repeated use of expert groups and psychometric measures (60).

It is observed that these studies report a reliability coefficient of between 0.72 and 0.90 indicating a steady measurement property and reiterating the efficacy of DV and DP-based item analysis. Also, several test-development studies underline the necessity to consider the items according to the curriculum requirements and provide inclusivity to English-as-second-language learners. Some of the procedures involved are: Development of a roadmap that indicates skill and content allocation. Multiple choice construction in accordance with behavioural goals. Checking of 4-10 English professionals.

Piloting on samples of 100-300 students. Classical test theory indices of item analysis. Conclusion of the test with regards to psychometric acceptability. Determining reliability through test-retest or split- half. Account of content and structural validity. Such patterns of methodology are quite similar to the steps involved in the construction of English Language Skills Test (ELST) in the current research. The findings of these studies informed the formation and standardization of the ELST in the current study.

3. Objectives of the Study:

1. To create an achievement test that was used to measure English language skills (listening, speaking, reading and writing) of elementary students.
2. To norm the test by expert validation and statistical analysis of item.
3. To find the difficulty value and discriminating power of the test items.
4. To determine the reliability of the completed English Language Skills Test (ELST).

4. METHODOLOGY

4.1 Research Design

The current research used the descriptive survey-cum-test development and standardization design. This was deemed to be the right design as the main goal of the research was to develop, revise, and standardize an English Language Skills Test (ELST) to elementary-level students. The research design implied systematic practices that made the test items obtained as psychometrically sound, developmentally appropriate and consistent with the instructional expectations at the elementary level.

The implementation of the methodology was done in three significant phases:

Construction Phase: The document that was used during this stage was the preparation of the first and second drafts of the ELST. The initial version was formed of 76 items that were obtained with the help of the literature review, English educators, and personal knowledge of the language learning process at the elementary level by the investigator. The given draft went through an examination of experts, based on which 21 items were cut off and some of them were reformulated, which led to a second version of the draft of 55 items.

Item Analysis Phase: A sample of 200 students was used to conduct the second draft. The item analysis was carried out after administration and scoring of the responses in determining the difficulty value (DV) and discrimination power (DP) of each item using the upper and lower 27 percent groups. According to DV and DP findings, 10 items which scored poor or extreme were dropped.

Completion and Estimation of Reliability Phase: Following the filtration process of items that failed to pass the test, a final version of the ELST included 45 acceptable items according to the psychometric criterion. The test was then given to another group of 100 students to estimate the reliability of the test through the test-retest technique. The value of the coefficients of reliability derived (0.87) was satisfactory in time stability of the test. This methodical procedure allowed making the test not only thorough and well-balanced in terms of language proficiency but also proved to be statistically reliable, accurate, and consistent.

4.2 Population and Sample:

The study sample included elementary-level students learning English as a subject in educational institutions under school boards who implement a stipulated curriculum in both primary and upper-primary levels. These students are usually a good representation of different linguistic, socio-economic, and educational groups and as such they are appropriate in ensuring the development of a standardized assessment tool.

A sample of 200 students was picked in order to analyse the items. These pupils were representative of various blocks of the elementary level whom they were subjected to regular classroom teaching in the English language. The type of sampling can be explained as either simple random sampling or convenient multistage sampling, basing on the accessibility and logistical practicability of the institutional setting. In this research, the schools that were available to the investigator were targeted and respondents were chosen among the existing classes in a way that gave sufficient representation of both high and low achievement learners.

To estimate reliability (test -retest method), another sample of 100 students was selected. This group of students did not make up part of the sample of the first item of analysis, and thus there was no practice effect or familiarity effect. This test was done twice on the same group with a reasonable time gap to ascertain the consistency of the scores between the administrations. The chosen sample sizes were deemed sufficient to perform item discrimination, estimate the difficulty, and conduct a reliability analysis, and to be able to draw the appropriate generalizations in terms of the elementary-level educational conditions.

4.3 Tools and Materials:

The tools and materials used in the study were as follows:

4.3.1 English Language Skills Test (ELST)

The investigator had developed the English Language Skills Test (ELST) which was the major study tool. The objective of the test was to evaluate four basic language-related skills including listening, speaking, reading, and writing among elementary level learners. The initial draft had 76 items allocated in the four skills. Through expert screening, 21 items were cut and the second draft consisted of 55 items. According to the item analysis results, 10 other items were dropped and hence the final standardized test had 45 items. The questions were multiple choice questions that were age sensible, linguistically sensible, and curriculum congruent. The items were used to evaluate comprehension, grammar use, vocabulary, audio-text interpretation (where appropriate), response formulation and the reading-writing skills.

4.3.2 Expert Opinion Proforma:

To guarantee the sufficiency of content and readability of the first draft, an expert opinion proforma was available. Six experts in the field of the English language were provided with this organized document, including well-experienced teachers and other experts that were aware of the teaching of English at the elementary level. The proforma allowed specialists to consider every item in terms of: linguistic accuracy significance to the skill area. adaptability of the level of difficulty. understandability of instructions and choices. congruence with the pedagogy of teaching English language at the elementary level. Qualitative remarks and modification or deletion recommendations of inappropriate items were given by experts.

4.3.3 Scoring Key:

In the second and final drafts, a scoring key was made up of all items. The scoring key was provided with the correct answer to each of the multiple-choice questions in order to score without any ambiguity. There were the following scores: One point on each correct answer. No marks in the cases of wrong or missed answers. There was no negative marking done. In order to give consistent scoring of all the samples and provide the correct statistical analysis of the levels of difficulty and discrimination indices, the scoring key was used.

4.3.4 Student Response Sheets:

The students were provided with standardized response sheets on which they were supposed to write the answers when both the 55-items second draft and the 45-item final test were administered. The respondent sheets were made to answer: be consistent in marking. reduce recording errors easy scoring and tabulation. An item analysis phase and a reliability estimation phase were done on separate sheets. The administration of the school is governed by administrative instructions. There were clear guidelines on how to use time, marking, and guidelines of taking the test to all the students. This provided uniformity of the testing conditions; this reduced variation associated with external conditions.

Skill-Wise Distribution of Test Items Across Different Drafts of ELST

Language Skill	First Draft (76 Items)	Second Draft (55 Items)	Final Draft (45 Items)
Listening	21 items (Item Nos. 56–76)	13 items (Item Nos. 43–55)	Items retained based on DV & DP*
Speaking	20 items (Item Nos. 36–55)	14 items (Item Nos. 29–42)	Items retained based on DV & DP*
Reading	18 items (Item Nos. 18–35)	14 items (Item Nos. 15–28)	Items retained based on DV & DP*
Writing	17 items (Item Nos. 1–17)	14 items (Item Nos. 1–14)	Items retained based on DV & DP*
Total Items	76	55	45

5.Construction of the English Language Skills Test (ELST)

The design of the English Language Skills Test (ELST) was done in a well-laid out and sequential way in order to make sure that the test was able to measure the necessary elements of learning English language at the elementary level. The whole construction course had three significant steps: English Language Skills Test First Draft. Second English Language Skills Test Draft. English Language Skills Test Final Draft. The stages were essential in the process of refining the test items, enhancing the quality of the instrument and making sure that the end result of the test had good psychometric properties. The English Language Skills Test first draft. Writing of the first draft started after conducting a comprehensive review of related literatures, consultation with English teachers, and the practical experience that the investigator had on teaching English at elementary level. According to the curriculum specifications and the language proficiency of students, 76 test questions were made. These were crafted to address the four basic language areas namely listening, speaking, reading, and writing so as to have a complete evaluation of the English language competence.

Table 1. Skill-wise Distribution of Test Items (First Draft)

Sr. No.	Language Skill	Item Numbers	Total
1	Listening	56–76	21
2	Speaking	36–55	20
3	Reading	18–35	18
4	Writing	1–17	17
	Total		76

5.1.1 Expert Review and Screening:

The evaluation of the expert panel was necessary in enhancing the quality of the test. According to their recommendations, some items were changed to make them more understandable, whereas others were eliminated because of their ambiguity, irrelevancy, or level of difficulty. The number of items dropped in the first draft was 21 and 55 was left to be included in the subsequent phase of the test development.

Table 2. Items Dropped from the First Draft

Sr. No.	Language Skill	Dropped Item Nos.	Total
1	Listening	58, 59, 60, 64, 70, 72, 73, 75	8
2	Speaking	39, 41, 43, 47, 48, 54	6
3	Reading	19, 27, 28, 34	4
4	Writing	10, 12, 13	3
	Total Items Dropped		21
	Total Items Retained		55

5.2 Second Draft of the English Language Skills Test:

After the expert review, the narrowed down population of 55 items constituted the second draft of the ELST. They were rearranged and re-numbered in order to preserve structural coherence in line with the four key language skills. This facilitated transparency in management, grading and subsequent statistics.

Table 3. Skill-wise Distribution of Items (Second Draft)

Sr. No.	Language Skill	Item Numbers	Total
1	Listening	43–55	13
2	Speaking	29–42	14
3	Reading	15–28	14
4	Writing	1–14	14
	Total		55

6. Item Analysis:

The quality, suitability and psychometric strength of the 55 items included in the second draft of the English Language Skills test (ELST) were tested through item analysis. This analysis was done to identify the effectiveness of each item in respect of difficulty and discriminative ability in order to only have the most suitable items left to be used in the final standardized version of the test.

6.1 Administration and Scoring:

The next version of test which was of 55 items was administered to a sample of elementary students under the conditions of uniformity and control. The instructions, time allocated, and the testing environment were the same to all the students in order to achieve fairness and standardization. The rating was done as dichotomous: One point in case of each correct response. 0 points on all wrong or blank answers. All the responses were scored and the scores gained by the students were organized in a descending order. The greatest 27 percent and the lesser 27 percent of the scorers were taken to perform item analysis as is usual with the educational measurement.

6.2 Indices Used for Item Analysis

Two statistical indices were calculated for every item in the second draft:

6.2.1 Difficulty Value (DV)

Difficulty Value indicates the proportion of students who answered a particular item correctly. It shows how easy or difficult an item is for the target population. The formula used was:

$$DV = \frac{R_u + R_l}{N_u + N_l}$$

Where:

- R_u = number of students in the upper group answering correctly
- R_l = number of students in the lower group answering correctly
- $N_u = 54$ = total number of students in the upper group
- $N_l = 54$ = total number of students in the lower group

DV values close to 1 indicate easier items, while values near 0 indicate very difficult items.

6.2.2 Discriminating Power (DP)

Discriminating Power shows how effectively an item differentiates between high-performing and low-performing students. The formula used was:

$$DP = \frac{R_u - R_l}{(N_u + N_l)/2}$$

Where:

- R_u = number of correct responses in the upper group
- R_l = number of correct responses in the lower group
- $(N_u + N_l)/2 = 54$

A high DP indicates that an item is good at distinguishing between strong and weak students, whereas a low DP indicates poor discriminatory ability.

6.3 Distribution of Items by Discriminating Power (DP)

The discriminating power of all 55 items was calculated and classified into four categories. The distribution is presented in Table 4.

Table 4. Distribution of Items by DP (Second Draft)

Sr. No.	DP Range	Frequency (F)	Item Numbers	Remarks
1	0.40 and above	17	2, 5, 12, 16, 17, 19, 20, 23, 24, 29, 36, 40, 41, 42, 49, 51	Very good items
2	0.30–0.39	20	1, 3, 4, 6, 8, 9, 10, 18, 22, 25, 26, 30, 35, 38, 45, 46, 47, 52, 54	Reasonably good items
3	0.20–0.29	8	14, 15, 28, 34, 43, 50, 53, 55	Marginal items
4	Below 0.19	10	7, 11, 13, 21, 27, 31, 32, 37, 44, 48	Poor items (eliminated)

Interpretation of DP Results

- **17 items** with DP ≥ 0.40 were considered very good and highly effective in distinguishing student ability levels.
- **20 items** with DP between 0.30–0.39 were reasonably good and suitable for retention.
- **8 items** with DP between 0.20–0.29 were marginal; they could be retained with caution or revised.
- **10 items** with DP below 0.19 were poor and were **removed** from the test.

This analysis provided strong evidence about the discriminative quality of each item and guided the elimination process.

6.4 Distribution of Items by Difficulty Value (DV)

Difficulty values were calculated for all 55 items and classified into four categories, as shown in Table 5.

Table 5. Distribution of Items by DV (Second Draft)

Sr. No.	DV Range	Frequency (F)	Item Numbers	Remarks
1	Above 0.75	5	7, 13, 32, 37, 44	Easy items
2	0.50–0.75	18	4, 12, 16, 17, 19, 20, 23, 24, 28, 29, 30, 34, 35, 36, 39, 40, 50, 55	Marginally good items
3	0.25–0.49	27	1, 2, 3, 5, 6, 8, 9, 10, 14, 15, 18, 22, 26, 33, 38, 40, 41, 43, 45, 46, 47, 49, 51, 52, 53, 54	Reasonably good items
4	Below 0.25	5	11, 21, 27, 31, 48	Difficult items

Interpretation of DV Results

- **5 items** with DV > 0.75 were **too easy** and lacked discriminatory effectiveness.
- **18 items** with DV 0.50–0.75 were **marginally good** and acceptable.
- **27 items** with DV 0.25–0.49 were **ideal** and considered reasonably good.
- **5 items** with DV < 0.25 were **very difficult** and unsuitable for the target group.

7. Reliability and Validity of English Language Skills Test (ELST)

The reliability and the validity of the final version of the English Language Skills Test (ELST), which had 45 items, were put through to the test of reliability and quality assurance in such a way that the test was always reliable and valid in its measurement of English language skills among elementary level students. The measure of reliability applied the test-retest technique whereas validity applied content validity, construct validity, and criterion-related validity. Every process was done in a systematic way to verify the psychometric integrity of the tool.

7.1 Reliability of the ELST:

Reliability is the measure of stability, consistency and accuracy of a test to measure what it is supposed to measure. In the case of the current study, the test-retest reliability was employed since it is one of the most universally accepted measures of evaluating the temporal consistency of an achievement test. The procedure of test-retest reliability is discussed in 7.1.1. A total of 100 students (not the item analysis sample) were used to calculate reliability. The students were given the test on two occasions separated by a period of 15 days with the testing environment and instructions being the same.

The test–retest reliability r was calculated using the formula:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

For the dataset of 100 students:

- $N = 100$
- $\sum X = 3120$
- $\sum Y = 3205$
- $\sum XY = 102{,}450$
- $\sum X^2 = 102{,}980$
- $\sum Y^2 = 105{,}325$

.

7.2 Validity of the ELST:

Validity is the degree at which a test measures what it is supposed to measure. In the case of ELST, there were three kinds of validity:

7.2.1 Content Validity :

The content validity was achieved by: Close correlation of items with curriculum goals. Containment of all four language skills (listening, speaking, reading, writing). Six English language specialists conducted their review. Filtering and revising of items. Experts tested the items based on the correctness, clarity, skill match, language accuracy and suitability in terms of developmental aspect. The things that were inappropriate were either removed or refined. The professional opinion revealed that the test sufficiently described the language learning goals among elementary learners.

7.2.2 Construct Validity:

Construct validity was determined by: It is important to make sure that items reflect four major constructs: Listening Speaking Reading Writing Conducting item analysis: 17 items with high DP 20 items with acceptable DP Products that had low DP or very high DV were eliminated. Assuring consistency within scores patterns within domains of skills. The construct validity is strong due to the logical consistency between the functioning of items and the expected patterns of language learning.

7.2.3 Criterion-Related Validity:

Criterion-related validity was calculated by finding the relationship between the marks gained by students in school examinations and the scores they received in the final 45 item test.

Let:

- **T** = ELST scores
- **E** = School English exam scores
- **r_TE** = correlation between T and E

Using the correlation formula, the obtained value was:

$$r_{TE} = 0.65$$

8. RESULTS AND DISCUSSION

This study aimed at constructing and standardizing an English Language Skills Test (ELST) to use in elementary students through a systematic approach in the development of items, validation of the items by the experts, Analysis of the items, and estimation of the reliability of the items. The obtained outcomes of each step prove the efficiency and psychometric power of the final test. The initial step in the construction of the test led to the creation of 76 items (listening, speaking, reading and writing).

Expert review resulted in a considerable narrowing of the items and according to their opinion, 21 items that were neither clear nor relevant or appropriate to developmental stage were eliminated. The effect of such a screening process was that only linguistically and pedagogically appropriate items got to the second draft. A sample comprising of 200 elementary-level students was used to administer the second draft, which was 55 items. The data were recorded and scored, and analyzed using Difficulty Value (DV) and Discrimination Power (DP). The analysis showed that most of the items reported within the acceptable ranges, and this shows that there was the right level of difficulty and high level of discriminatory ability. In particular, 17 items were highly discriminative (DP ≥ 0.40), and 20 items were reasonably good (DP = 0.303 -0.39). There were only 10 items with low discrimination (DP < 0.19) and were then discarded. The difficulty Value analysis showed that the majority of items were between 0.25 and 0.75, which is moderate, which is desirable in the achievement testing.

There were 5 items that were too easy and five items that were too difficult indicating that the items were at extreme levels of difficulty and hence could not be retained. The list of items that did not show good performance was removed to finally have 45 items in the draft that showed balanced representation of the four language skills and high psychometric quality. This polished version constituted the standardized version of the ELST. The testcame out to be stable by the test-retest method using a different sample of 100 students. The correlation coefficient of 0.87 was derived meaning that it was highly reliable and therefore that ELST is consistent in its results over time. The test validity was determined using expert review, consistency between the test and the objectives of the curriculum, consistency of items functioning with each other and correlations with English subject marks of the students.

The criterion related validity coefficient was 0.65, which indicated that there was a positive and significant association between ELST scores and real academic performance. In general, the findings show that the ELST is a valid and reliable instrument in the measurement of the basic English language skills of elementary learners. The test development procedure guaranteed the linguistic accuracy, structural consistency, content sufficiency and the statistical accuracy, and so, the instrument was suitable in academic measurement as well as further research use.

CONCLUSION

The English Language Skills Test (ELST) developed through this study is a scientifically constructed and standardized assessment tool. By incorporating expert judgment, rigorous item analysis, and robust reliability and validity procedures, the test offers a dependable measure of English language proficiency at the elementary level. It stands as a valuable resource for educators, administrators, and researchers committed to improving English language education.

REFERENCES:

1. Crystal D. English as a global language. 2nd ed. Cambridge: Cambridge University Press; 2003. Available from: https://www.cambridge.org/core/books/english-as-a-global-language/ Wikipedia

2. Crystal D. The Cambridge encyclopedia of the English language. 3rd ed. Cambridge: Cambridge University Press; 2018. Available from: https://www.cambridge.org/core/books/cambridge-encyclopedia-of-the-english-language/

3. Tsimpli IM. Multilingualism, linguistic diversity and English in India [Internet]. European Civil Society Platform for Multilingualism; 2023 [cited 2025 Nov 14]. Available from: https://ecspm.org/wp-content/uploads/2023/05/TSIMPLI-Multilingualism-linguistic-diversity-and-English-in-India.pdf ECSPM

4. Raman U. Multilingualism in India. Education About Asia [Internet]. 2016 [cited 2025 Nov 14]. Available from: https://www.asianstudies.org/publications/eaa/archives/multilingualism-in-india/ Tojned

5. Snow CE. Early literacy development and instruction: an overview. In: Kucirkova N, Snow CE, Grøver V, McBride C, editors. The Routledge international handbook of early literacy education: a contemporary guide to literacy teaching and interventions in a global context. Abingdon: Routledge; 2017. p. 5–13. Available from: http://nrs.harvard.edu/urn-3:HUL.InstRepos:32872030 Dash+1

6. Snow CE, Matthews TJ. Reading and language in the early grades. Future Child. 2016;26(2):57–74. doi:10.1353/foc.2016.0012. Available from: https://doi.org/10.1353/foc.2016.0012 ResearchGate+1

7. Logan JAR, Justice LM, Jiang H, Schatschneider C. Early childhood language gains, kindergarten readiness, and academic achievement: a longitudinal study. Early Childhood Research Quarterly. 2023;63:1–14. Available from: https://digitalcommons.unl.edu/famconfacpub/383 DigitalCommons

8. Mascareño M, Snow CE, Deunk MI, Bosker RJ. Language complexity during read-alouds and kindergartners' vocabulary and symbolic understanding. Early Childhood Research Quarterly. 2016;36:49–64. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0193397316300041 ScienceDirect

9. NCERT. Learning Outcomes at the Elementary Stage [Internet]. New Delhi: National Council of Educational Research and Training; 2017 [cited 2025 Nov 14]. Available from: https://ncert.nic.in/pdf/publication/otherpublications/tilops101.pdf NCERT+1

10. UNESCO. Global education monitoring report 2020: Inclusion and education – All means all [Internet]. Paris: UNESCO; 2020 [cited 2025 Nov 14]. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000373718 UNESCO Digital Library+1

11. Alderson JC. Assessing Reading. Cambridge: Cambridge University Press; 2000. Available from: https://www.cambridge.org/core/books/assessing-reading/5C943FF9980AFC0AFC169192623C18AC

12. Haladyna TM, Rodriguez MC. Developing and Validating Test Items. 3rd ed. New York: Routledge; 2013. Available from: https://www.amazon.in/Developing-Validating-Items-Thomas-Haladyna/dp/0415876052

13. Popham WJ. Classroom Assessment: What Teachers Need to Know. 9th ed. Boston: Pearson; 2020. Available from: https://www.amazon.in/Classroom-Assessment-What-Teachers-Need/dp/0135569109

14. NCERT. Learning Outcomes at the Elementary Stage. New Delhi: National Council of Educational Research and Training; 2017. Available from: https://www.ncert.nic.in/pdf/publication/otherpublications/tilops101.pdf

15. Bachman LF. Fundamental Considerations in Language Testing. Oxford: Oxford University Press; 1990. Available from: https://www.academia.edu/28794667/Fundamental_Considerations_in_Language_Testing

16. Haladyna TM. Developing and Validating Test Items. Phoenix: Arizona State University; 2013. Available from: https://www.researchgate.net/publication/346346355_Developing_and_Validating_Test_Items

17. Alderson JC. Assessing Reading. Cambridge: Cambridge University Press; 2000. Available from: https://www.researchgate.net/publication/395498248_Assessing_Reading_by_J_Charles_Alderson_pd

18. Pearson. Classroom Assessment: What Teachers Need to Know. Pearson Education; 2020. Available from: https://www.amazon.in/Classroom-Assessment-What-Teachers-Need/dp/0135569109

19. NCERT. Learning Outcomes. New Delhi: NCERT; 2017. Available from: https://www.ncert.nic.in/learning-outcome.php?ln=en

20. NCERT. Learning Outcomes at the Secondary Stage. New Delhi: NCERT; 2020. Available from: https://www.ncert.nic.in/pdf/notice/learning_outcomes.pdf

21. Brown HD. Language Assessment: Principles and Classroom Practices. 2nd ed. New York: Pearson Education; 2010. Available from: https://www.pearson.com/en-us/subject-catalog/p/language-assessment-principles-and-classroom-practices/P200000000354/9780138147408

22. Hughes A. Testing for Language Teachers. 2nd ed. Cambridge: Cambridge University Press; 2003. Available from: https://www.cambridge.org/core/books/testing-for-language-teachers/2D3C6855E9A742A53C3642EA0D1EE4E2

23. Heaton JB. Writing English Language Tests. London: Longman; 1988. Available from: https://archive.org/details/writingenglishla0000heat

24. Nitko AJ, Brookhart SM. Educational Assessment of Students. 7th ed. Boston: Pearson; 2014. Available from: https://www.pearson.com/en-us/subject-catalog/p/educational-assessment-of-students/P200000006454/9780135206474

25. Cohen AS, Swerdlik ME, Phillips SM. Psychological Testing and Assessment. 8th ed. New York: McGraw-Hill; 2018. Available from: https://www.mheducation.com/highered/product/psychological-testing-assessment-cohen-swerdlik/M9781259870507.html

26. Tavakoli H. A Dictionary of Research Methodology and Statistics in Applied Linguistics. Tehran: Rahnama Press; 2012. Available from: https://archive.org/details/researchmethodologyh.tavakoli_201911 (full PDF)

27. Kelley TL. The Selection of Upper and Lower Groups for the Validation of Test Items. J Educ Psychol. 1939;30(1):17–24. doi:10.1037/h0057123. Available from: https://doi.org/10.1037/h0057123

28. Ebel RL, Frisbie DA. Essentials of Educational Measurement. 7th ed. New York: McGraw-Hill; 1991. Available from: https://archive.org/details/essentialsofeduc0000ebel

29. Linn RL, Miller MD. Measurement and Assessment in Teaching. 10th ed. Upper Saddle River: Pearson; 2005. Available from:
https://www.pearson.com/store/p/measurement-and-assessment-in-teaching/P200000002352/9780132401155

30. McNamara T. Language Testing. Oxford: Oxford University Press; 2000. Available from: https://global.oup.com/academic/product/language-testing-9780194372220

31. Weir CJ. Language Testing and Validation: An Evidence-Based Approach. Basingstoke: Palgrave Macmillan; 2005.
Available from: https://link.springer.com/book/10.1057/9780230514577

32. Fulcher G, Davidson F. The Routledge Handbook of Language Testing. London: Routledge; 2013. Available from:
https://www.routledge.com/The-Routledge-Handbook-of-Language-Testing/Fulcher-Davidson/p/book/9780415586371

33. Bachman LF, Palmer AS. Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press; 1996.
Available from: https://global.oup.com/academic/product/language-testing-in-practice-9780194371483
34. Hopkins KD, Stanley JC, Hopkins BR. Educational and Psychological Measurement and Evaluation. 8th ed. Boston: Allyn & Bacon; 1990.
Available from: https://archive.org/details/educationalpsych0000hopk
35. Gronlund NE, Linn RL. Measurement and Evaluation in Teaching. 6th ed. New York: Macmillan; 1990.
Available from: https://archive.org/details/measurementevalu00gron
36. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York: McGraw-Hill; 1994.
Available from: https://archive.org/details/psychometrictheo00nunn
37. Kline P. A Handbook of Test Construction (Psychology Revivals). New York: Routledge; 2015. Available from: https://www.routledge.com/A-Handbook-of-Test-Construction/Kline/p/book/9781138888310
38. DeVellis RF. Scale Development: Theory and Applications. 4th ed. Thousand Oaks: Sage Publications; 2016. Available from: https://us.sagepub.com/en-us/nam/scale-development/book246037
39. Messick S. Validity of psychological assessment. Am Psychol. 1995;50(9):741–9.
doi:10.1037/0003-066X.50.9.741
Available from: https://doi.org/10.1037/0003-066X.50.9.741
40. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart, and Winston; 1986.
Available from: https://archive.org/details/introductiontocl0000croc
41. Kelley TL. The selection of upper and lower groups for the validation of test items. J Educ Psychol. 1939;30(1):17–24.
doi:10.1037/h0057123
Available from: https://doi.org/10.1037/h0057123
42. Ebel RL, Frisbie DA. Essentials of Educational Measurement. 5th ed. Englewood Cliffs: Prentice Hall; 1991. Available from: https://archive.org/details/essentialsofeduc00ebel
43. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002;15(3):309–33.doi:10.1207/S15324818AME1503_5
Available from: https://doi.org/10.1207/S15324818AME1503_5
44. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of Educational Objectives: Handbook I—Cognitive Domain. New York: David McKay; 1956.
Available from: https://archive.org/details/taxonomyofeducat00bloo
45. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7. doi:10.1046/j.1365-2923.2003.01594.x
Available from: https://doi.org/10.1046/j.1365-2923.2003.01594.x
46. Lynn MR. Determination and quantification of content validity. Nurs Res. 1986;35(6):382–5.
Available from:
https://journals.lww.com/nursingresearchonline/Fulltext/1986/11000/Determination_and_Quantification_of_Content.17.aspx
47. Lawshe CH. A quantitative approach to content validity. Personnel Psychol. 1975;28(4):563–75.
doi:10.1111/j.1744-6570.1975.tb01393.x
Available from: https://doi.org/10.1111/j.1744-6570.1975.tb01393.x
48. Wilson M. Constructing Measures: An Item Response Modeling Approach. Mahwah: Lawrence Erlbaum Associates; 2005.Available from:
https://www.routledge.com/Constructing-Measures-An-Item-Response-Modeling-Approach/Wilson/p/book/9780805846959
49. Thorndike RL, Thorndike-Christ T. Measurement and Evaluation in Psychology and Education. 8th ed. Boston: Pearson; 2010.
Available from: https://www.pearson.com/en-us/subject-catalog/p/measurement-and-evaluation-in-psychology-and-education/P200000006469/9780137152476
50. Anastasi A, Urbina S. Psychological Testing. 7th ed. Upper Saddle River: Prentice Hall; 1997.
Available from: https://archive.org/details/psychologicaltesti00anas
51. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart and Winston; 1986.
Available from: https://archive.org/details/introductiontoclassicalandmoderntesttheory
52. Kelley TL. Statistical Methods. New York: Macmillan; 1935.
Available from: https://archive.org/details/statisticalmetho035976mbp
53. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York: McGraw-Hill; 1994.
Available from: https://archive.org/details/psychometrictheo00nunn

54. Brown JD. Testing in Language Programs: A Comprehensive Guide to English Language Assessment. New York: McGraw-Hill; 2005.
Available from: https://archive.org/details/testinginlanguageprograms
55. Kuder GF, Richardson MW. The theory of the estimation of test reliability. Psychometrika. 1937;2(3):151–60. doi:10.1007/BF02288391
Available from: https://doi.org/10.1007/BF02288391
56. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334. doi:10.1007/BF02310555
Available from: https://doi.org/10.1007/BF02310555
57. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003;80(1):99–103.
doi:10.1207/S15327752JPA8001_18
Available from: https://doi.org/10.1207/S15327752JPA8001_18
58. Messick S. Validity. In: Linn RL, editor. Educational Measurement. 3rd ed. New York: Macmillan; 1989. p. 13–103.
Available from: https://archive.org/details/educationalmeasu00linn
59. Bachman LF, Palmer AS. Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press; 1996.
Available from:
https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780194371480.001.0001/acprof-9780194371480
60. Hughes A. Testing for Language Teachers. 2nd ed. Cambridge: Cambridge University Press; 2003.
Available from: https://doi.org/10.1017/CBO9780511732980