Adolescent language development being a poorly researched area requires an in-depth retrospection on the various constructs which should be identified inorder to assess language abilities in this particular group of individuals. The pattern of language development in this group is subtle, thus requiring a carefully planned and well executed assessment protocol. The motivation behind the development of this language tool was the lack of focus observed in language development in older children and adolescents especially in a multilingual country like India. The neglect observed in this population has led to a lacuna of adolescent language assessment tools to identify persisting language impairments which may impede the individual’s communication and academic performance when confronted by new educational, social and vocational demands during secondary school. Hence, the present research was an attempt to develop an assessment tool to assess the language development in adolescents. The tool aimed to incorporate tasks to identify language-learning disabilities, specific to each group.

A two stage cluster random sampling research design was adopted for the present research. The research protocol was approved by Institutional Ethical Committee of Kasturba Medical College (Manipal University), Mangalore. The samples required for the research were taken from schools within Mangalore, having English as their medium of instruction. The following method was adopted for the conduction of research.

**Participants**

The participants in the current research included typically developing school going adolescents aged between 10–15.11 years. The Registrar General and Census Commissioner of the Government of India in 1991 revealed that one-fifth of India’s population is in the adolescent age group of 10 – 19 years. According to Gupta (2003), India comprises of almost 200 million adolescents, and it is expected that this group will continue to grow reaching over 214 million by 2020. In the lower age limit was 10 years with an upper age limit as 15.11 years. This limit was for considering children attending secondary high school till Xth standard. Moreover, education until high school is often the minimum level of education that is received in most social communities, with only 27% of students getting into class XI (Berlia, 2007).
Selection Criteria

The participants were recruited from schools which followed the state board syllabus with English as the medium of instruction. Prior to the conduction of research, school authorities were explained about the purpose of the research and a written permission (Appendix A) was obtained from them. The school teachers were provided with a checklist (Appendix B) adopted for the research to recruit children based on selection criteria. The selected children’s parents were contacted and given a checklist (Appendix C) inorder to obtain details regarding the child’s schooling. The participants who fitted the selection criteria based on the school teachers and parents report were recruited for the research. An informed consent (Appendix D) was obtained from all the participants involved in the research prior to their inclusion in the study.

Inclusion criteria.

- Those who fitted the age and standard criteria.
- Those selected by the teachers based on the checklist adopted for the research.
- Those selected by the parents based on the checklist adopted for the research.

Exclusion criteria.

- Those with a history/complaint of any speech and/or language deficits.
- Those with a history/complaint of any reading and/or writing problems.
- Those who had any history/complaint of acquired hearing loss.
- Those with complaints of cognitive deficits such as poor memory, attentional deficit, organizational and/or sequencing issues.
- Those with a history of any transfer from more than one school
- Those with a history of any shift in the medium of instruction.
- Those with a history of any academic failures.

The current research comprised of 6 groups of participants based on an age and standard criteria.
• The participants in the first pilot study included $N = 90$ individuals between $10 – 15.11$ years, with 15 participants in each group.

• The participants in the second pilot study included $N = 60$ individuals between $10 – 15.11$ years, with 10 participants in each group.

• The participants for the final data collection included $N = 432$ individuals between $10 – 15.11$ years of age.

The sample size was calculated using the following formulae.

$$n = \frac{N}{1 + N(e)^2}$$

$n$ = sample size

$N$ = population size

$e$ = level of precision at 0.05

The sample size was estimated based on the population of adolescents between 10-19 years in India (Census of India, 2001). A 10% of additional samples were included in each group to accommodate dropouts if any during the course of the study.

The participants were allocated into six groups based on their age and standard for the data collection as follows:

**Group I.** Group I consisted of participants ($n = 72$) (36 males and 36 females) within the age range of $10.0 – 10.11$ years studying in Standard V.

**Group II.** Group II consisted of participants ($n = 72$) (36 males and 36 females) within the age range of $11.0 – 11.11$ years studying in Standard VI.

**Group III.** Group III consisted of participants ($n = 72$) (36 males and 36 females) within the age range of $12.0 – 12.11$ years studying in Standard VII.
**Group IV.** Group IV consisted of participants \( n = 72 \) (36 males and 36 females) within the age range of \( >13.0 – \leq 13.11 \) years studying in Standard VIII.

**Group V.** Group V consisted of participants \( n = 72 \) (36 males and 36 females) within the age range of \( >14.0 – \leq 14.11 \) years studying in Standard IX.

**Group VI.** Group VI consisted of participants \( n = 72 \) (36 males and 36 females) within the age range of \( >15.0 – \leq 15.11 \) years studying in Standard X.

**Procedure**

The development of the language assessment tool followed 3 phases. Phase 1 comprised of the designing of the tool along with pilot study, Phase 2 included the administration of the developed language assessment tool, and Phase 3 focused on establishing test validity and reliability measures.

**Phase 1: Tool design and construction.** The designing and construction of the language assessment tool followed a systematic approach. Figure 1 illustrates the process followed for the designing and construction of the adolescent language assessment tool.

**Planning the language constructs, tasks, and test items.** The planning level of the language assessment tool included the identification of the relevant language constructs to provide an insight about the adolescent language development. The preparation of the language constructs to be used in the tool were decided based upon an extensive review of the research done on adolescent language development, as well as studying the available adolescent language assessment tools used by the western population. The constructs that were selected for the research aimed to provide an insight about the adolescent language development. The three language constructs that were selected were semantics, morphology and syntax. The semantic language construct consisted of ten tasks. The ‘contrastive relations task’ measures an individual’s knowledge of antonyms which is an important semantic ability (Soifer, 2005). The ‘multiple meanings task’ measures an individual’s knowledge of synonyms which is an important semantic ability for the understanding of the meaning of a term (Soifer, 2005).
The ‘associated relations task’ measures the semantic knowledge of an individual which includes information about a word and its relations to other words and its possible meanings. The ‘convergent naming task’ measures the comprehension of word definitions which required an understanding of words and knowledge, a combination of linguistic, metalinguistic and cognitive competencies. The ‘double-function words task’ measures the understanding of two statements which had two different meanings, but

Figure 1. Tool design and construction. This flowchart illustrates the procedure followed for the designing and construction of the language assessment tool.
sharing the same orthographic representation. The ‘homophones task’ assesses the individual’s ability to identify words that shared the same pronunciation but with different orthographic representations. The ‘compare/contrast task’ measures the individual’s ability to identify words that commonly face confusion during usage. The ‘analogical reasoning task’ primarily focuses on the linguistic and cognitive competence, requiring a controlled semantic structure of analogies. The ‘proverbs/idioms task’ and ‘similes task’ had figurative expressions which are often considered to be giant lexical units that are used to assess for the higher-linguistic skills. Though such expressions are used to enhance the pragmatic use of language, it is considered to be a semantic aspect.

The morphologic language construct consisted of a ‘morphological derivations task’, which incorporated the use of derivational suffixes to deliver information pertaining to the syntactic and semantic aspects, in turn conveying a subtle distinction. The syntactic language construct consisted of a ‘sentence combining task’, which assesses the ability to integrate sentences using appropriate grammatical markers.

The selection of the vocabulary to be used for the preparation of the items for each of the constructs was taken from the core curriculum (IV$^{\text{th}}$ standard - X$^{\text{th}}$ standard). This contained vocabulary from different disciplines - English Literature, Social Sciences, Sciences (Physics/Chemistry/Biology), Mathematics (Algebra/Geometry) and Environmental Sciences. Each of the school texts was carefully scrutinized and words were extracted from them and allocated specifically for each group. The selected words were from most of the word classes in English (verbs, nouns, adjectives, adverbs, and preposition) comprising of both concrete and abstract words.

The vocabulary allocations of the tasks under the semantic language construct consisted of selecting words for the item which were present in the prescribed curriculum. For the allocation of figurative expressions under the proverbs/idioms and similes tasks, the occurrence of such expressions in the curriculum was not consistent throughout the selected academic standards. Hence, the test items comprising the proverbs/idioms and simile task were pooled in together from school textbooks and children’s fictional books, forming a common stimuli for all age groups. The construction of the test items under the simile task included the inclusion of simile based statements.
from the school textbooks and children’s fictional books. Each item included the standard presentation of a simile in a sentence format with a missing referent.

The test items that were created for the morphological derivations task followed the selection of words that were formed on the basis of including suitable suffixes to a root word. Either the root words and/or the derived words were required to be present in the vocabulary list of the corresponding group. For the construction of the items in the sentence combining task, the sentences were selected from the prescribed curriculum. Each item comprised of a minimum of two sentences and a maximum of four sentences which included functional markers.

**Deciding the presentation level.** The tasks that were selected from each language construct required to be presented in a standard format. Table 1 shows the tasks and their types of presentation levels (word/sentence/both). The word level tasks were planned so as to avoid any contextual bias from the sentence and also for providing objectivity in scoring. The word and sentence level tasks included the presentation of the items in a sentence format. Such a presentation method was planned to elicit a single word response which also aided in the objective scoring of the items.

**Modality based assignment of tasks.** Language being multimodal in nature indicated the need to assign the tasks under both auditory and visual modalities. Though most of the tasks that were prepared for the current research consisted of items tapping upon two separate modalities, there were few tasks which inevitably required to be presented in specific modalities. For example, when considering the homophones task, producing the polysemous item verbally will result in identifying the appropriate choice as the pair is phonologically similar. Therefore this task had to be presented in the visual modality itself. In order to maintain homogeneity across modalities in the usage of polysemous words, the double-function words task (homographs) also polysemous in nature were presented in the auditory modality. Similarly, when considering the proverb/idiom task, the items which consisted of a figurative expression, accompanied by three other sentences which served as multiple choices was presented in the visual modality. To balance the figurative tasks used in both modalities, the simile task, also
figurative in nature was presented in the auditory modality. Table 1 shows the tasks along with the modalities it taps upon.

**Deciding the response format.** The test items of the language tasks followed either a selected response format or a constructed response format. The selected response format required the examinee to select their response from a variety of response choices that were given. These are also referred to as forced choice formats and included multiple choice items, which are less affected by guessing. In the constructed-response format, the items required the examinee to create and organize a response, which is also a free-response, required the examinee to complete or fill in one or more blanks of a statement with the correct word or phrase. Table 1 shows the response formats for each of the tasks that were selected for the present research.

Table 1

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Modalities tested</th>
<th>Presentation levels</th>
<th>Response formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrastive relations</td>
<td>Auditory + Visual</td>
<td>Word</td>
<td>Constructed response</td>
</tr>
<tr>
<td>Multiple meanings</td>
<td>Auditory + Visual</td>
<td>Word</td>
<td>Constructed response</td>
</tr>
<tr>
<td>Associated relations</td>
<td>Auditory + Visual</td>
<td>Word</td>
<td>Selected response – multiple choices</td>
</tr>
<tr>
<td>Convergent naming</td>
<td>Auditory + Visual</td>
<td>Word + Sentence</td>
<td>Constructed response</td>
</tr>
<tr>
<td>Double-function words</td>
<td>Auditory</td>
<td>Word + Sentence</td>
<td>Constructed response</td>
</tr>
<tr>
<td>Homophones</td>
<td>Visual</td>
<td>Word + Sentence</td>
<td>Selected response – two choice</td>
</tr>
<tr>
<td>Compare/Contrast</td>
<td>Auditory</td>
<td>Word + Sentence</td>
<td>Selected response - two choice</td>
</tr>
<tr>
<td>Analogical reasoning</td>
<td>Auditory + Visual</td>
<td>Word</td>
<td>Selected response - multiple choices</td>
</tr>
<tr>
<td>Proverbs/Idioms</td>
<td>Visual</td>
<td>Sentence/Phrase</td>
<td>Selected response - multiple choices</td>
</tr>
<tr>
<td>Similes</td>
<td>Auditory</td>
<td>Phrase</td>
<td>Constructed response</td>
</tr>
<tr>
<td>Morphological derivations</td>
<td>Auditory + Visual</td>
<td>Word + Sentence</td>
<td>Constructed response</td>
</tr>
<tr>
<td>Sentence combining</td>
<td>Visual</td>
<td>Word + Sentence</td>
<td>Constructed response</td>
</tr>
</tbody>
</table>

**Preparing the number of test items under each task.** A large number of test items were initially devised under each of the groups. Individual sets of stimuli were prepared for each of the six groups (Group I - VI). A total of 5 test items were retained under each of the tasks (except for the proverbs/idioms and similes task) for each group,
specific to each modality, totaling to 30 items (5 items X 6 groups) under each task. The proverbs/idioms and similes task consisted of a common stimuli of 30 items each.

**Planning for the instructions and scoring.** Appropriate instructions were framed to suit each task. The scoring was aimed to attain objective responses. Except for the proverbs/idioms task, all other tasks followed a standard biserial scoring system, with a score of 1 for a correct response and a score of 0 for an incorrect response/unattempted test item. There was no difference in the scoring system of the auditory and visual based tasks. Since the responses for the proverbs/idioms task were to be rated according to the figurative, literal or incorrect interpretations, the scoring followed a categorical system. Here the figurative interpretations received a score of 2, the literal interpretations received a score of 1, and the incorrect interpretations received a score of 0. The instruction which was used for the presentation of an auditory modality based task followed a different pattern from its counter-part, the visual modality based task. Table 2 shows the instructions used under each of the tasks specific to the auditory and visual modalities.

<table>
<thead>
<tr>
<th>The Instructions created for each Task based on the Modality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td><strong>Contrastive relations</strong></td>
</tr>
<tr>
<td>Auditory</td>
</tr>
<tr>
<td>Visual</td>
</tr>
<tr>
<td><strong>Multiple meanings</strong></td>
</tr>
<tr>
<td>Auditory</td>
</tr>
<tr>
<td>Visual</td>
</tr>
<tr>
<td><strong>Associated relations</strong></td>
</tr>
<tr>
<td>Auditory</td>
</tr>
<tr>
<td>Visual</td>
</tr>
<tr>
<td>Task</td>
</tr>
<tr>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Convergent naming</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Analogical reasoning</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Morphological derivations</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Double-function words</td>
</tr>
<tr>
<td>Homophones</td>
</tr>
<tr>
<td>Compare/Contrast</td>
</tr>
<tr>
<td>Proverbs/Idioms</td>
</tr>
<tr>
<td>Simile</td>
</tr>
</tbody>
</table>

*Cautions taken during the preparation of test items.* The test items were prepared in a clear, precise, and grammatically correct manner by adopting certain caution. Firstly, each of these tasks was designed to have items which were age specific,
fitting within the curriculum prescribed. For example, the test items created for the Vth standard group (Group I) were different from the test items created for the VIth standard group (Group II), which in turn was different from the test items created for its subsequent group (Group III), and so on. This manner of allocation of the test items for the stimuli was increasing in complexity with age. Secondly, care was taken to avoid inter-related items; that is wording of one item providing a clue to the answer of other item. Thirdly, caution was taken to avoid any inter-locked items; that is when it is necessary to know the answer to one of them to get the other one right. Fourthly, since certain tasks were to be presented in both auditory and visual modalities, caution was taken to not repeat the items within the task across the modalities. Lastly, it was seen to it that there was no repetition of any vocabulary across any of the tasks and modalities.

Initial validation and revision. The initial validation of the developed tool was conducted before the commencement of the pilot study. The scrutinization was done by two judges who were experts in the field of speech language pathology, to examine the validity of the constructs, tasks and items. Both judges had a teaching and work experience of more than 5 years in handling children with language disorders. The validity of the language constructs and tasks was scrutinized on the basis of two parameters – ‘appropriate to assess adolescent language’ and ‘inappropriate to assess adolescent language’. The validity of the test items under each task was scrutinized on the basis of three parameters – ‘retaining the item’, ‘modifying the item’, and ‘replacing the item’. The validity of the scoring system under each task was scrutinized on the basis of two parameters – ‘appropriate scoring’ or ‘inappropriate scoring’. A section was also included to convey any comments or suggestions by the experts. The tool was also scrutinized by the respective English teachers of the six standards. They scrutinized the tool to examine the validity of the items of the respective curriculum level. The validity of the test items under each task was scrutinized on the basis of three parameters – ‘items that were very easy for the student’, ‘items that were appropriate for the student’, and ‘items that were difficult for the student’. A section was also present to convey any comments or suggestions by the teachers regarding the stimuli. After the retrieval of the
scrutinized language tool from the subject experts and teachers, necessary modifications were incorporated, and the tool was used for the pilot study.

**First pilot study.** Fifteen students were recruited based on the subject selection criteria, with a total of $N = 90$ students participating in the pilot study. The participants of a particular standard received three sets of stimuli – easy, medium, and difficult levels. The easy level comprised of the set which was a level below the target standard; the medium level comprised of the set which was at the level of the same target standard; and the difficult level comprised of the set which was at a level above the target standard.

**Revision of tasks and test items.** The obtained data was scored accordingly and subjected to analysis. The items which attained less than 30% scores or more than 80% scores were eliminated from the stimuli set in order to prevent a floor and ceiling effect of the chosen stimulus. The ‘sentence combining task’ had to be eliminated, in view of the varied responses generated by the participants. With the objective of the task being to integrate the sentences using an appropriate grammatical marker, the aim was to achieve a pattern of grammatical markers which increased in their usage with age. Due to the inability to derive a developmental pattern of grammatical markers used by the participants across the six groups, this task was found to be unacceptable and hence eliminated.

**Second pilot study.** The second pilot study was done after the item selection based on the first pilot. A total of $N = 60$ students participated in the second pilot study with 10 in each group. The individuals who served as participants in the first pilot study were not included in the second pilot study. For the stimuli presentation, the task level of difficulty was retained which was at par with the first pilot study. With the completion of the data collection, the responses obtained from the participants were subjected to analysis. Figure 2 shows the number of tasks and items that were selected for the first pilot study, second pilot study, and the final data collection.

**Content validation of the test items.** The items of each task of every group were subjected to validity and reliability related measures, inorder to select the good items. For the validity related measure, point-biserial correlation technique was done to assess what
the examinee knows and doesn’t know. This was done by correlating the scores of the item with scores of the criterion measure. This method of item analysis provided the item’s discriminating power and helped in the identification of the poor items of the test. Though an item’s discriminating power can range from -1.00 to +1.00, items having a higher coefficient value were preferred. An item’s inclusion or elimination depended on the size of the coefficient. Not only were the items dependent on the item correlation coefficient, but also on the correlation of the item with the other items in the task. Items that had low correlations with the other items but high correlations with the criterion were considered best because those items made an independent contribution to the prediction of criterion scores. Nunnally and Bernstein (1994) and Garrett (1965) suggested that items with a discriminating power of 0.2 or more will mostly be satisfactory, especially if the test is long. Garrett (1965) reported that a discriminating power of 0.3 can be considered for shorter tests. Similarly Hammill, Brown, Larsen, and Wiederholt (2007) considered a discriminating power of 0.3 for the content validity of the Test of Adolescent and Adult Language – 4. However, Anastasi (1976) opined that a discriminating power of 0.2 or 0.3 coefficients may be acceptable. For the current research, an item discrimination power of >0.3 was considered for the inclusion of the item, which was also at par with the Test of Adolescent and Adult Language – 4 and the Fullerton Language Test for Adolescents – 2. Items which received a discrimination power of <0.3 were eliminated. Item difficulty index referred to the percentage of examinees that gave correct responses to a given item, helped to identify items that were too easy or too difficult. A difficulty index can range from 0%–100%. For an item having a low difficulty value can indicate the item to have been mis-represented, too challenging, ambiguous, or the item may have only one correct answer. Hammill, Brown, Larsen, and Wiederholt (2007) considered an item difficulty range of 15 – 85% for the consideration of items to obtain content validity for the Test of Adolescent and Adult Language – 4. The present research followed a more stringent item difficulty index range of 40 – 70% to maintain an intermediate level of difficulty for the inclusion of the items. Items which received a difficulty index of <40% and >70% were eliminated. The decision to consider point biserial correlation coefficient and item difficulty index as a measure to identify the good items was done for every task, except for the
proverbs/idioms and similes task which were analyzed only on the basis of item difficulty index. The reason to analyze these tasks in terms of the item difficulty index was because age specific items could be derived from these tasks, as both the tasks follow a different developmental pattern compared to other tasks. For the reliability related measure, Cronbach’s alpha method was used to assess how reliable are the items. A high coefficient alpha indicates a high degree of internal consistency. Achieving a good internal consistency for a test is important. Such measures define the consistency of the results in a test. Cronbach’s alpha measures how well a set of items measured a single skill of individuals. Researchers have recommended alpha values ranging between 0.70 and 0.95 to be considered as acceptable (DeVellis, 2003). Language tests such as the Test of Adolescent and Adult Language – 4 have used internal consistency measures such as the Cronbach’s alpha in the construction of the items of the test. For the purpose of the present research a Cronbach’s alpha value of >0.70 was considered as a criteria for the inclusion of an item. Items which received a Cronbach’s alpha value of <0.70 were eliminated.

The scrutinization of the tool was conducted by two experts in the field of speech language pathology. The validity of the test items under each task was scrutinized on the basis of three parameters, which was at par with the process adopted during the pilot study. The tool was also subjected to scrutiny by the respective English teachers of the six standards. The teachers who served as judges for the initial validation were not considered for the final validation. After the retrieval of the scrutinized language tool from the subject experts and teachers, necessary modifications were incorporated, and the tool was ready to be used for the final data collection.

**Phase 2: Test administration.** This phase of the research began with the identification of the participants who were not included in the pilot studies. A total of n = 72 participants were recruited under each group, with a total of N = 432 samples. The final stimuli for the data collection included 11 tasks. The details regarding each of the tasks are explained below.
Figure 2. Total number of tasks and items selected for data collection. This flowchart illustrates the number of tasks and items that were prepared and presented under the auditory (A) and visual (V) modalities for the first pilot study, second pilot study, and the final data collection.
• **Contrastive relations.** The test items under the contrastive relations task consisted of having antonymous relations on the basis of being mutually exclusive, mutually dependent, and based on the relative degree of comparison.

• **Multiple meanings.** The test items under the multiple meanings task consisted of having synonymous relations that belonged to the use of different styles, words that differed only in their evaluative or emotive meaning, and words that overlapped in meaning or were near in meaning.

• **Associated relations.** The test items under the associated relations task consisted of four words that were presented as a set, from which the examinee had to select the unrelated or the odd word from it.

• **Convergent naming.** The test items under the convergent naming task consisted of a standard definition which followed an Aristotelian style (A is a B that C). For example, ‘A curtain (A) is a cloth (B) which is used to cover the windows from light (C). In the present task, the definition was presented in a similar manner without the presence of ‘A’, which had to be identified by the examinee.

• **Analogical reasoning.** The test items under the analogical reasoning task was formatted in a A: B :: C : __? (D) pattern; wherein each position indicated a word. Three multiple choices (words) were also included along with the test item. The items consisted of analogies based on characteristic property, part-whole relationships, functional relationship, sequential relationship, causal relationship, and superordinate-subordinate category.

• **Morphological derivations.** The test items under the morphological derivations task consisted of a statement with a missing word, and a root word.

• **Double-function words.** The test items under the double-function words task consisted of two sentence meanings wherein the examinee was required to identify the respective homograph.

• **Homophones.** The test items under the homophones task consisted of a statement with a missing word, along with a choice of a pair of homophones from which an appropriate homophone had to be selected by the examinee.

• **Compare/contrast.** The test items under the compare/contrast task consisted of a statement with a missing word, along with a choice of a pair of confusable words.
• **Proverbs/Idioms.** The test items under the proverbs/idioms task consisted of figurative expressions (proverbs or idioms) along with a three multiple choices (literal interpretation, figurative interpretation and an incorrect interpretation).

• **Simile.** The test items under the similes task consisted of figurative phrases with a missing word.

The examples for each of the tasks are shown in Appendix E.

**Setting up testing environment and guidelines for stimuli presentation.** The research was conducted in a well-lit distraction free room within the school premises. Each individual was made to sit comfortably on a chair facing the examiner who was also seated at a table to record the responses. The testing room consisted of only the examiner and the examinee. The devised language tool was individually administered on each individual from all six groups. For the auditory based tasks the examiner verbally presented the items with its respective instructions and noted down the examinee’s responses in a response sheet. Whereas for the visual based tasks, the examinee was given a booklet containing all the instructions and the target items. Each participant was given a pencil to complete the visual section of the tool. A practice item was provided on need basis. The similar testing and stimulus guidelines were followed for the first pilot study, second pilot study, and the final data collection. Figure 3 shows a photo of the examinee undergoing the adolescent language assessment.
**Administration of the language assessment tool.** The test administration included the presentation of a single stimuli set targeting its age equivalent group. This method of presenting the stimuli was adopted as each age-based stimuli was inclusive of items of different complexity levels tapping upon respective age level. The examinee’s preparedness and motivation level was considered before the commencement of the test. The participant was free to withdraw from the test or was allowed to take breaks during the course of the testing. The test administration began with the presentation of the instruction pertaining to the task of interest. The auditory based tasks were presented first, followed by the visual based tasks. The total time taken to complete the assessment tool was between 1 – 1½ hours. Figure 4 illustrates the order of presentation of the tasks of the adolescent language assessment tool for each participant.

*Figure 3: Language assessment in progress. This figure shows a participant being tested under a testing environment.*
Figure 4. Presentation order of the tasks of the adolescent language assessment tool. This flowchart illustrates the order of the presentation of the tasks on the target sample size.

The data collection was carried out from August 2013 – March 2014. The data was then entered in SPSS version 16 and subjected to statistical analysis. Descriptive statistics was done to analyze the scores that were obtained under each of the tasks under every modality across the groups. Test of normality was performed using Kolmogorov-Smirnov and Shapiro-Wilk Test and was found to be significant. Descriptive statistics was done to determine the mean and standard deviation of the scores of the auditory and visual tasks of the typically developing adolescents.
and age matched adolescents with language disorder. One way ANOVA was done to determine the level of significance (p<0.001) across the groups for the proverbs/idioms task (visual) and similes task (auditory). Post hoc analysis was done to determine the level of significance (p<0.001) between the groups.

**Phase 3: Test validity and reliability.** Standardization is an important step in designing and scrutinizing a language test. Before using a test, the information concerning the validity and reliability of the test for its specific purposes must be acquired. Both of these measures have traditionally been considered as the basic criteria which any language test should fulfill.

**Test Validity.** The content validity of the language tool was established through a process of validation. Such measures should ideally be done before the results can be used for any specific purpose. The main objective of such measures was to arrive at a conclusion as to whether the interpretations and uses of the test are valid. Measures such as analyzing the content of the test and computing the correlation between scores of the normative sample with those on the criterion of interest were performed to ascertain the test validity.

**Content validity.** The content validity of a test is concerned with whether the content of a test elicits a range of responses that were representative of the entire domain of skills that the test is intended to measure. This is evaluated by analyzing the composition of the test to determine the extent to which it represents the objectives of the constructs and tasks. The process of evaluating the content validity was done during the test construction phase itself. The validity was attained with the point biserial correlation which was done to obtain the discrimination power of the items, and by calculating the difficulty index of the items. This was done to improve the test items by discarding the poor items. Further validation was also obtained by the subject experts and teachers examining if the test measures what it is intended to measure. When considering its diagnostic implication, item analysis provides information regarding what examinees know and doesn’t know. Since one of the objectives of the present research was to develop a criterion-referenced test, item analysis was performed to assess how much an examinee knows about the content of each of the tasks.

**Construct validity.** Construct validity is the ability to accurately reflect upon the conceptual foundation upon which a test has been developed. If a measure has good construct
validity, then the test will appear to measure the abilities that it is designed to measure. The clinical accuracy which is the ability to discriminate between the language skills is a fundamental property of any diagnostic test. In order to satisfy this, the typically developing adolescents were compared with 60 age matched adolescents with language disorders (10 adolescents with language disorder X 6 groups) using Mann Whitney test to compare their responses to items and the total scores in every task. A similar construct validation method was also observed in the construction of other adolescent language tests (Hammill et al., 2007; Thorum, 1986). The adolescents with language disorders for the present research portrayed poor language and academic performances. This disordered population was identified by the teachers using the inclusion/exclusion checklist, and were found to demonstrate a relatively poor linguistic competence in the Linguistic Profile Test. Receiver Operating Characteristic (ROC) analysis was also performed to attain the cut-off scores, sensitivity, specificity, and area under the curve. The cut-off scores provide the criterion to be considered to delineate individuals with a disorder. The sensitivity provides the probability of an individual who has a disorder to test positive for it; while specificity provides the probability of an individual who does not have a disorder to test negative. The area under the curve provides the concentration of the individuals identified with a disorder.

**Test Reliability.** The development of any language has to be a consistent measure, to measure what it was designed to measure.

**Cronbach’s alpha.** A sampling error is associated with the degree of homogeneity among the test items. Items which measure the same quality require to be related to each other. If so the amount of test error due to content sampling would be less. In contrast, if the items are unrelated to each other, then the test may be measuring different qualities. This may indicate the test to have a content sampling error which is high. To account for this sampling error, and to indicate the internal consistency of the tasks, the Cronbach’s (1951) coefficient alpha method was used. The process of evaluating the reliability based on the Cronbach’s alpha value was done during the test construction phase itself.

**Test-retest reliability.** The test-retest reliability measure which is another method to examine the internal consistency of a test is also important to determine the consistency of a test which was administered on two occasions. Language tests such as the Test of Adolescent and
Adult Language – 4, and the Fullerton Language Test for Adolescents – 2 have used such measures to determine the reliability of their respective tests. The present research has also adopted a similar method to examine the internal consistency of the language tool. Therefore inorder to test the reliability of the tool, the test was re-administered on 10% of the total sample size (7 participants from each group), after 2 weeks of the initial administration. A similar proportion was used by Thorum (1986) for the construction of the Fullerton Language Test for Adolescents – 2. The coefficient of reliability was found by correlating the scores of the responses of all tasks which were obtained at the two instances using Kappa statistics (agreement between each item in every task) and Intra-class correlation coefficient (agreement between total scores for every task).