Content validation and rubric development of language tools tested in multi-state multi-language project: Challenges Faced

Gayatri K. Vaidya
Educational Initiatives, Ahmedabad. Gujarat. INDIA.

Email: gayatri.vaidya@ei-india.com

Validity of tools is a key issue in large scale assessments. For any assessment to have its valid impact, the inferences and decisions made on the basis of assessments should be well-founded. To ensure this, the content, construct, testing procedures and analysis procedures have to be validated. As such a complex feature, validity of tools enters a different paradigm when the tools have to be tested in multiple states of India in various vernacular languages. While mathematical tools require simpler adaptation and precise translation in other languages, for language testing, the tools developed should be validated for cultural and linguistic purposes as well. Questions where the construct includes skills like identifying letters and sounds, word usage, grammar and comprehension, mere translation in vernacular language may not suffice. More than translation it is trans-creation and adaptation of tools in that language, while maintaining the difficulty levels. This requires various validity checks. The paper focuses on a case study of a project funded by an international organization working on children education and welfare. As part of the project, the tools were developed and administered for grades 2 & 3 in 7 different states of India in 5 different vernacular languages. Various validity checks including harmonization of tools, constructs, difficulty levels and pilot testing were employed to ensure uniformity and efficacy of the assessment. The paper lists out challenges faced and remedies worked out to ensure that the objective of the assessment was achieved.

Introduction:

Educational assessments should always have a very clear testing objective. Assessment will not give insights about learning unless there are validity and reliability checks. Hence, validity is the single most important attribute of a good test. Validity can be defined as the extent to which it measures what it was designed to measure, without contamination from other characteristics (Darr, 2005).

More recently, the assessment specialists also propose that validity is better understood as an evaluation of the quality of the interpretations and decisions that are made on the basis of an assessment result—that is, how well the inferences we make or actions we take on the basis of an assessment result can be justified (Kuzek & Rist, 2004). It is crucial that the inferences and decisions made on the basis of assessment results are well-founded. Determining validity involves amassing evidence that supports these interpretations and decisions. The strength of that evidence will lead us to a strong,

moderate, or weak case for validity. What evidence we collect will depend on the kind of interpretations and decisions we want to make.

The paper focuses on the processes followed to develop tools in 7 languages administered for grades 2 & 3 in 7 different states of India in 5 different languages. Schools from seven states including Gujarat, Madhya Pradesh, Tamil Nadu, Karnataka, Andhra Pradesh, Rajasthan & Jharkhand were assessed in five different languages i.e. Gujarati, Hindi, Tamil, Kannada and Telugu respectively. The validity checks were ensured and uniformity and efficacy were checked.

Methodology:

This project, which was funded by an international agency working in the field of child care and education, aimed at finding out the effectiveness of a learning methodology in primary grades. The tools thus developed for classes 2 and 3 were based on age appropriate competencies.

Fig.1: Overall process of tool development

To make tools relevant to the content, the first step was to map the textbooks of all 7 states to check for the coverage of curriculum in various states. Based on the mapping of the textbooks, skills that were tested across all 7 states were identified and the blueprint design was created to develop tools. The blueprint defined the skills to be tested number of questions under each skill to ensure balanced question paper having acceptable representation of various types of questions that can be asked under the same skill. These question types included direct or familiar questions that are seen and worked on by the students in their textbooks and workbooks and some non-familiar type of questions which were based on HOTS (Higher Order Thinking Skill) defined by Bloom (1964). National and international benchmark items were also included in the tools to have a benchmarking reference.

The tools assessed skills like students' ability to read, speak, recognize letters and sounds, vocabulary, knowledge of grammar, sentence construction and reading and comprehension to retrieve information and make inferences by analyzing texts. The tools included questions from other large scale studies conducted in India and certain international items were also included to benchmark the performance. The assessments were low-stake assessments at the student's level, but high-stake at the decision making levels.

The base paper was developed in English. It was first approved for its content validity by language and maths experts, following the recommendations of the funding agency. The tools were also commented upon by the education department of each state and appropriate suggestions were incorporated to ensure a balanced paper which could test similarly across the states. They were then translated in 5 different languages- Gujarati, Hindi, Telugu, Tamil and Kannada.

The process and guidelines followed for translation (and verification of translation) were similar to the processes followed by international assessment agencies like PIRLS and TIMSS (Yu and Ebbs, 2011).

Sometimes when the text from one language is translated to another language, due to different syntactic and grammatical rules in different language, the question either becomes invalid or becomes easier than the source question. At times, the testing objective of the question changes, which can impact the overall balance of the tool and also the findings we arrive at. Since translation is a very restrictive term and doesn't always help in upholding the testing objective of the item (Joldersma, 2004), the base papers were translated and adapted in 5 different languages- Gujarati, Hindi, Telugu, Tamil, Kannada.

While translators were oriented and trained to translate the items staying faithful to the source paper, adaptations typically required inputs from domain experts, as and when it

was essential to deviate from the source version to maintain psychometric variance which focuses on the students' mental capacities to answer the question (Linn *et al., 1991)*.

Since the tools were to be tested in 5 different languages, there was an additional step required- *harmonization of tools*. It was essential that the translated and adapted tools also maintained identical testing objective and difficulty level across languages. This was carefully carried out with the help of experts. The free response questions that were included in the tools to assess students' writing abilities required a very well thought-through harmonization. Starting from letters to words to sentences, the items were harmonized to ensure that they maintained the difficulty level and age appropriateness. Similar to what is defined in the Cambridge Assessment Processes, the internal structure of the assessments was kept consistent with the content domain across languages to ensure credibility of the test (Cambridge Press, 2009).

Every translated, adapted and harmonized version was then submitted for a validity check. Experts not involved in translations verified these papers for the quality of translation, usage of correct and age appropriate terms in translation and equivalence of the translated version against the source version.

The validated and thus commented tools were then approved via Independent Paper Solving (IPS). IPS was done by a person not involved in the test development and validation processes. This process ensured that the correct answers in the system were marked correctly, and no item had construct validity issues. If there were items that could not convey the testing objective to the IPS person or where the person doing IPS could not answer the correct question, the question was further scrutinized and either corrected, modified or dropped.

To support analysis, rubrics were designed once the questions were harmonized. This applied mainly to the Free Response questions.

Table 1: Validity checks for translated versions

Validation Checks	**Explanation**
Added information	Any information given in the vernacular version which is not present in the base version. This information can be a word or a group of word or additional information giving explanation.
Missing information	Any information given in the vernacular version which is not present in the base version.
Layouts/	The lay outs, fonts used, visuals and information in visuals should be

Visuals	identical to the base version.
Grammar/ Syntax	Any such grammar or syntactical mistake that may invalidate the question: can be due to inflection, conjugation error (wrong mood, aspect, tense or voice used) or any such literal translation in vernacular language that can deviate question from its testing objective.
Consistency	Patterns to be followed for synonyms or literal matches across the translated version. For Maths, the units should be translated uniformly across the paper.
Terminology	Scientific terminology may differ from the word in use in language. This should be taken care in the translated version.
Adaptation	Adaptation should be done especially when the literal translation would change the testing objective and would put the assessed at a disadvantage.
Mistranslation	Wrong translation of terms/ phrases or sentences resulting in misconstruction.

When test instruments are piloted, it happens routinely that some items do not work the way the test developers expected them to work. Analysis of field test results using, for example, the Item Response Theory model (IRT) may reveal cases of Differential Item Functioning (DIF) for certain test questions in certain countries, and this may (or may not) lead to identifying residual translation issues or ambiguities that verifiers may (or may not) have overlooked. The case analysis has to be carried out post hoc. When a verifier checks equivalence *ex ante*, s/he can check whether all the information contained in the source is also present in the target version, whether the term remains the same, whether the level of difficulty of the test items is likely to have been affected by linguistic or syntactical artefacts, whether hints for the correct responses have been added or removed, etc. The verifiers follow the verification checklist and try to assess as many aspects of equivalence as possible, but this process has limitations.

The post-assessment analysis was also carried out to find out whether the validity checks employed at various stages worked or not. To check whether the students were able to do certain types of questions in a similar difficulty level range, overall difficulty levels of papers, difficulty values of the tool across states and languages were compared to understand content validity of the papers.

PBC (Point Biserial Coefficient- rpbi) value for each question was compared with its equivalent in other languages to see if there were any major/drastic differences in the findings. Point Biserial Coefficient (PBC) values give the correlation between the students' overall scores and score for any particular question. The PBC ranges from a low of **-1.0** to a high of **+1.0**. The more positive values of PBC are preferable as it shows that an item discriminates well among the students who did well in the test, compared to those who did not. Positive PBC of any options for a question indicates that high

performing students are selecting that option, while negative PBC indicates that low performing students are selecting that option (Brown, 2001)

Item–wise difficulty levels for each vernacular version were calculated and plotted against each other for a comparative view. Item difficulty is simply the percentage of students taking the test who answered the item correctly. The larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be (Wood, 1960). To compute the item difficulty, average performance is taken as the difficulty value. So if the average performance is 75%, difficulty value is 0.75, difficulty value is 1-0.75 = 0.25. The proportion for the item is usually denoted is called item difficulty (Crocker & Algina, 1986; Hetzel, 1997).

Results:

Translation and harmonization of free responses:

The free responses in Language papers required thinking through and arriving at a common set of letters/words which could be tested at the same difficulty level. Some alphabets which are simple to write in one language, were found to be difficult to write in the other language. Through a detailed analysis of writing skill, shapes and curves of alphabets and students' capacity to do such questions, a final list of dictation letters and free response questions was designed.

Despite the efforts to come to a letter that can be common across language, as can be seen in table 2, for class 2, different letters had to be tested for Telugu and Kannada as the equivalent of 'h' in those two languages was very simple to write i.e. the number of curves and strokes required in writing the letter were lesser as compared to their equivalents in other languages. and thus was changing the difficulty level of the question. Similarly, for class 3, depending on the letter and the curves involved in writing while in Gujarati and Hindi equivalent letters of 'x' were tested, in Tamil it was 'h' and in Kannada and Telugu equivalent of 'sh' were tested.

Similarly, for the free response questions assessing words, such questions were identified from the pool for which the word across languages were harmonized to maintain difficulty levels as well as the testing objective. Table 2 narrates the free responses with their pronunciations and their equivalents in those languages.

Rubrics were designed for each Free Response question to ensure that the responses in such questions were scored identically across districts and states. In absence of such a defined structure scoring can get affected depending on subjective understanding of the person evaluating the responses. Table 3 shows some examples of rubrics designed to ensure identical scoring across languages.

Table 2. Harmonization in FRQs to ensure validity of tools

Question	English (Base paper)	Hindi	Gujarati	Tamil	Kannada	Telugu
Class: 2						
Letter dictation	T	ग (g)	ગ (g)	க (k)	ಗ (g)	గ (g)
Letter dictation	H	ह (h)	હ (h)	ஹ (y)	ಯ (y)	య (ya)
Letter dictation	B	कि (Ki)	પ (p/ pu)	ப (pu)	ಪು (pu)	పు (pu)
Picture identification	Rose	गुलाब (gulab)	ગુલાબ (gulab)	ரோஜா (roja)	ಗುಲಾಬಿ(Gulabi)	గులాబీ (gulabi)
Picture identification	Leg	पैर (Paer)	પગ (pug)	கால் (kaal)	ಕಾಲು (Kaalu)	కాలు(Kallu)
Class: 3						
Word completion	B	के (ke)	કે (ke)	வ (V)	ಬ (ba)	అ (aa)
Picture identification	Lamp	दीप (deep)	દિયા (diya)	தீபம் (dheepam)	ದೀಪ (dee-pa)	దీపం (deepam)
Picture identification	Eye	आंख (Aankh)	આંખ (Aankh)	கண் (kann)	ಕಣ್ಣು (kaNNu)	కన్ను (Kannu)
Picture identification	Empty	खाली (Khali)	ખાલી (Khali)	படத்திற்கு (padaththirku)	ಖಾಲಿಯಾಗಿದೆ (Kaaliyaagidee)	బొమ్మకు (bhomaku)
Letter dictation	X	क्ष (ksh)	ક્ષ (ksh)	ஹ (ha)	ಶ (sh)	శ్ (sh)
Picture identification	Tree	झाड (jhaad)	પેડ (ped)	மரம் (maram)	ಮರ (Mara)	చెట్టు (chattu)

As can be seen, for the second question, since the equivalent of H was becoming difficult for Tamil, Telugu and Kannada students, in place of H, vernacular equivalent of Y was tested.

This process of harmonization required multiple rounds of revisions even in the base paper. Since harmonizing of papers across 5 languages and completely different scripts had to be done with the help of language experts, such picture identification questions were selected, wherein the words students had to write were of the same difficulty level. The base paper was hence revised to come up with questions which could be tested at almost the same difficulty level for all the languages.

Table. 3 Rubrics designed in various languages to ensure identical scoring pattern across languages

<table>
<tr><th>Testing Objective</th><th>Rubric/ scorecard example</th><th>Adaptation examples</th></tr>
<tr><td>Letter writing</td><td>
<table>
<tr><td colspan="2">Score Card</td></tr>
<tr><td>Writes T or t.</td><td>01</td></tr>
<tr><td>Any other answer</td><td>85</td></tr>
<tr><td>Invalid Answer Code/More Than One Option Ticked</td><td>86</td></tr>
<tr><td>Not Attempted</td><td>88</td></tr>
</table>
</td><td>
<table>
<tr><td colspan="2">Score Card</td></tr>
<tr><td>ग लिखा हो</td><td>01</td></tr>
<tr><td>अन्य उत्तर</td><td>85</td></tr>
<tr><td>Invalid Answer Code/More Than One Option Ticked</td><td>86</td></tr>
<tr><td>Not Attempted</td><td>88</td></tr>
</table>
</td></tr>
<tr><td>Identifying word for picture</td><td>
<table>
<tr><td colspan="2">Score Card</td></tr>
<tr><td>Writes 'eye' or any other appropriate answer with correct spelling.</td><td>01</td></tr>
<tr><td>Writes 'eye' or any other appropriate answer with spelling errors</td><td>21</td></tr>
<tr><td>Any other answer</td><td>85</td></tr>
<tr><td>Invalid Answer Code/More Than One Option Ticked</td><td>86</td></tr>
<tr><td>Not Attempted</td><td>88</td></tr>
</table>
</td><td>
<table>
<tr><td colspan="2">Score Card</td></tr>
<tr><td>'આંખ'/'નયન'/'નેત્ર'/'ચક્ષુ' શબ્દ સાચી જોડણી સાથે લખે છે.</td><td>01</td></tr>
<tr><td>આંખના કોઈ ભાગનું નામ સાચી રીતે લખે છે અથવા જોડણીની ભૂલ સાથે 'આંખ' લખે છે.</td><td>21</td></tr>
<tr><td>અન્ય કોઈપણ જવાબ</td><td>85</td></tr>
<tr><td>Invalid Answer Code/More Than One Option Ticked</td><td>86</td></tr>
<tr><td>Not Attempted</td><td>88</td></tr>
</table>
</td></tr>
<tr><td>Writing correct sentence using given jumbled words</td><td>
<table>
<tr><td colspan="2">Score Card</td></tr>
<tr><td>Kanchan cooked food today. (or) any other correct variant. (Do not check for spellings.)</td><td>01</td></tr>
<tr><td>Any other answer</td><td>85</td></tr>
<tr><td>Invalid Answer Code/More Than One Option Ticked</td><td>86</td></tr>
<tr><td>Not Attempted</td><td>88</td></tr>
</table>
</td><td>
<table>
<tr><td colspan="2">Score Card</td></tr>
<tr><td>'ಖಾಲಿಯಾಗಿದೆ' ಎಂದು ಬರೆದರೆ</td><td>01</td></tr>
<tr><td>'ತುಂಬಿದೆ' ಎಂದು ಬರೆದರೆ</td><td>81</td></tr>
<tr><td>ಇತರ ಯಾವುದೇ ಉತ್ತರ</td><td>85</td></tr>
<tr><td>Invalid Answer Code/More Than One Option Ticked</td><td>86</td></tr>
<tr><td>Not Attempted</td><td>88</td></tr>
</table>
</td></tr>
<tr><td>Writing correct sentence for the given picture</td><td>
<table>
<tr><td colspan="2">Score Card</td></tr>
<tr><td>Writes at least 1 sentence that is relevant to the picture and has no spelling or grammatical errors</td><td>01</td></tr>
<tr><td>Writes at least 1 sentence that is relevant to the picture but has minor spelling or grammatical errors</td><td>21</td></tr>
<tr><td>Any other answer</td><td>85</td></tr>
<tr><td>Invalid Answer Code/More Than One Option Ticked</td><td>86</td></tr>
<tr><td>Not Attempted</td><td>88</td></tr>
</table>
</td><td>
<table>
<tr><td colspan="2">Score Card</td></tr>
<tr><td>படத்திற்கு பொருத்தமான குறைந்தது ஒரு வாக்கியமாவது எழுத்துப் பிழைகள் இல்லாமல் எழுதுகிறார்.</td><td>01</td></tr>
<tr><td>படத்திற்கு பொருத்தமான குறைந்தது ஒரு வாக்கியமாவது எழுதுகிறார். ஆனால் வாக்கியத்தில் எழுத்துப் பிழைகள் உள்ளன.</td><td>21</td></tr>
<tr><td>வேறு பதில்</td><td>85</td></tr>
<tr><td>Invalid Answer Code/More Than One Option Ticked</td><td>86</td></tr>
<tr><td>Not Attempted</td><td>88</td></tr>
</table>
</td></tr>
</table>

As can be seen in Table.3, for each free response questions, where the answers were expected to be differently correct or wrong, score cards or scoring rubrics were pre-designed as a part of tool development process. Specific guidelines against each possible answer were given and a code was fixed for each possible answer. The answers included the putative answers like fully correct answer and wrong answer in case of

letter writing skill; fully correct answer, partially correct answer due to wrong spelling and a fully wrong answer in the word writing skill; Sentence writing skill was tested by writing correct variant for the given jumbled words where spellings were not given weightage while in class 3, in sentence writing skill where students had to write a sentence on the basis of the given picture, relevance to the picture, grammatical correctness, spellings were all given weightage. Codes like 01, 21, 81 85, 86, 88 were decided for the backend analysis programming. Also, by not featuring the actual scores in the score cards, the score cards were made more objective in nature where the focus of the evaluator was taken away from the marks to be given to the codes to be assigned.

Confirmation of validity post assessment:

1. **Content validity by PBC comparison of vernacular versions:**

As discussed above, for language testers, PBC can give an insightful understanding for the existing data. While performance can vary with varying capacity of students in different states, PBC can be indicative of the validity of question across mediums. PBC can also give insights about the quality of an individual item and hence can also help in identifying a flawed item which can affect the overall outcome of the assessments.

PBC values for all the five languages were calculated and plotted on the graph to understand the effect of harmonization on the performance.

As can be seen from figures 2 & 3, the overall pattern of the PBC values are in the same range. The PBC values obtained were significant at >0.5 levels. Thus, having adapted tools for each vernacular version against the base versions showed content validity.

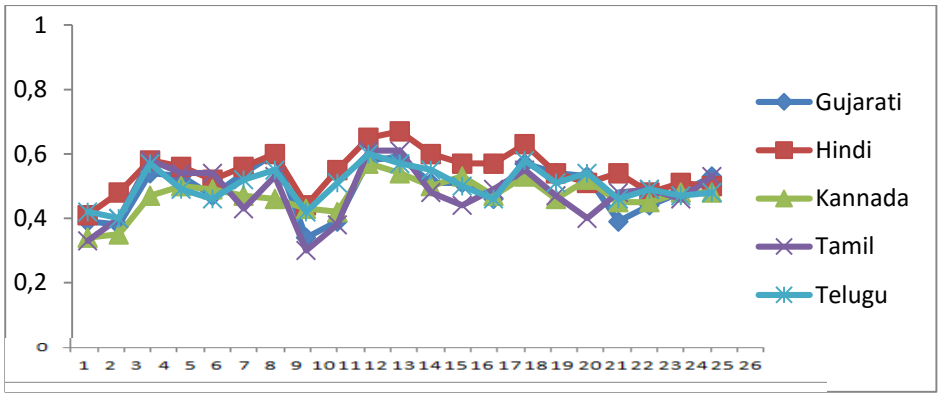

Fig.2: PBC distribution in each language version after harmonization in class 2 tools

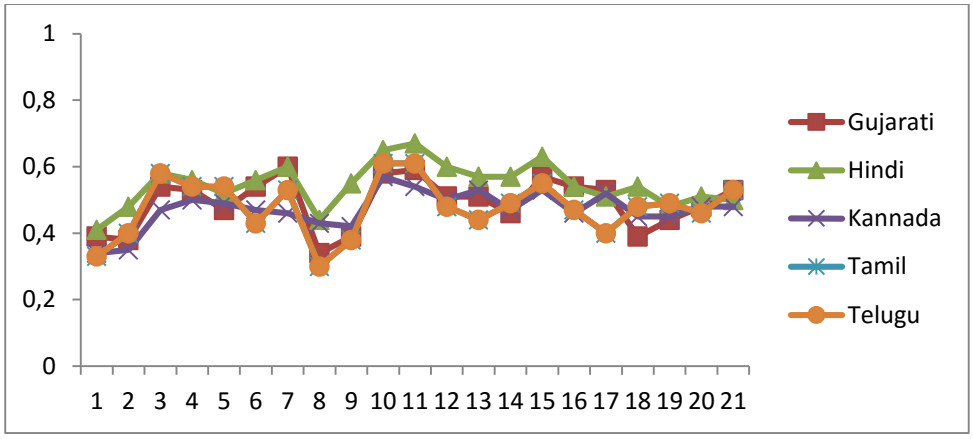

Fig.3: PBC distribution in each language version after harmonization in class 3 tools

2. Content validity by difficulty level comparison of vernacular versions:

The difficulty levels for each item were also calculated for each vernacular version. These difficulty values were significant at >0.5 and were plotted on the same graph to compare and analyze how the difficulty levels of questions changed when they were translated in other languages. Due to extensive and careful harmonization of tools, the range was found to be significantly close. Figures 4 & 5 show difficulty level comparison across vernacular versions for class 2 and 3 respectively.

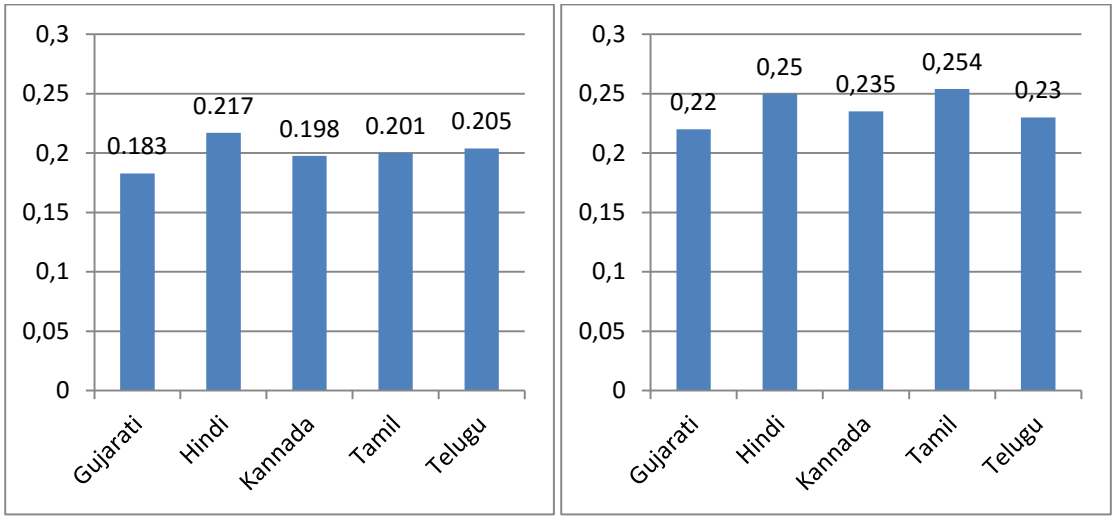

Fig. 4 & 5: Graphs showing average difficulty levels of each final vernacular version classes 2 & 3

As can be seen from the data, for class 2, while the overall performance in for each item varies, the difficulty levels for different paper range from 0.18 – 0.23. Similarly for class 3, the difficulty levels range from 0.18 to 0.25. However, for class 3, as compared to other classes, Kannada students find the items less difficult. This can also be attributed

to the higher learning levels in the region. However, overall, the difficulty levels seem to be comparable.

Conclusion:

Content validation is an essential aspect of any assessment, especially when the items are required to be translated in various vernacular versions. The given research paper discusses the processes followed in translation of the items. An additional step after translation and adaptation, called 'harmonization of tools', was carried out to ensure that the items were tested for the same objective. Since the tools were to be developed in multiple languages, the base paper was designed in such a way that while it tested the competencies and skills with age appropriateness, it also maintained the difficulty levels and testing objectives. The post assessment data was analysed for content validity and acceptable range of PBC values of items across papers were observed. A similar trend was observed in difficulty levels as well.

References:

1. Bloom B., B. Mesia, and D. Krathwohl (1964). Taxonomy of Educational Objectives (two vols: The Affective Domain & the Cognitive Domain). New York. David McKay.
2. Brown, J D (2001). Statistics Corner: Questions and answers about language testing statistics: Point-biserial correlation coefficients biserial correlation coefficients. Shiken: JLT Testing & Evlution SIG Newsletter. 5 (3) October 2001 (pp. 13 - 17)
3. Darr, Charles (2005). A hitchkicker's guide to validity, Assessment News.
4. Joldersma, KB (2004). Cros- linguistic instrument comparability. In: The meeting of the American Educational Research Association, San Diego, CA. http://www.ed-web3.educ.msu.edu/CEPSE/mqm/forms/KJ.doc.
5. Kusek, JZ & Rist RC (2004). Ten Steps to a result based monitoring and evaluation system. World Bank publication.
6. Linn, RL; Baker, EL & Dunbar, SB (1991). Complex performance based assessment: Expectations and validation criteria. Educational Researcher, 20 (8), 15- 21.
7. Matlock, Hetzel S (9970). Basic Concepts in Item and Test Analysis. In: The annual meeting of the Southwest Educational Research Association, Austin, January, 1997.
8. Oates, T. (2009) The Cambridge Approach Principles for designing, administering and evaluating assessment. Cambridge Press.
9. Wood, D.A. (1960). Test construction: Development and interpretation of achievement tests. Columbus, OH: Charles E. Merrill Books, Inc.
10. Yu, A. & Ebbs, D. (2011) Methods and procedures| Translation and translation verification, TIIMS and PIRLS International Study Centre.