The Promise Foundation | UNIVERSITY OF OXFORD | Oxford Policy Management | UKaid from the British people

Assessing literacy in developing countries

EVIDENCE BRIEF: What is available and what is required?

About this brief

This brief summarises a rigorous literature review examining the quality and range of the measurement tools that are used to assess literacy and foundation learning in developing countries. The review was written by Sonali Nag and a team at The Promise Foundation and funded by DFID. It benefits from substantial support from the Principal Investigator and consultant from an earlier review: Maggie Snowling and Shaher Banu Vagh.

How to use this brief

The brief starts with background information including the theory underpinning the review and an outline of what assessments are used for. It then explores two issues that cut across all foundational learning assessments: the importance of context and problems with how results are reported. The main literacy assessments are then summarized, one sub-skill at a time. Factors that should be considered whenever assessing each sub-skill are identified. A short section outlining the main gaps in evidence is followed by a table showing how the main assessment instruments have performed.

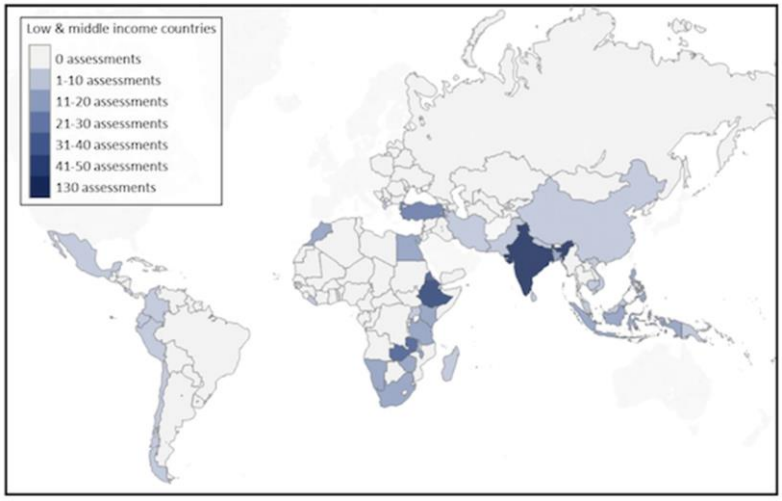
Methodology

The review covered assessments of language and literacy skills in children aged 3-14 conducted in a developing country between 1990 and 2014.

Multiple electronic databases (ERIC, PsycInfo, Web of Science) and websites were screened.

Only studies rated as moderate or high in methodological quality were included.

The final set covers 55 languages and 53 countries.



Recommendations

For policy makers and funders

- Assessments should be designed to emphasise the importance of reading with understanding.
- Assessments should include items that cover the range of literacy skills. The difficulty should be appropriate for the ability of the group being tested.
- Greater attention needs to be paid to adapting assessments to the local context, recognizing the different language characteristics and socio-cultural environment.
- Supplementary information should be reported in every study to enable readers to interpret results properly. This should include contextual factors and the psychometric properties of tests.
- Assessment data need to be communicated well and key stakeholders need to make good use of the information they provide in order to improve education quality.
- Innovations are required, including using technology, to support the scale-up of testing.
- Assessments should be free to use and adapt if we are to encourage more countries to measure children's learning. This will have greatest impact on the most marginalized.
- A resource bank of robust tests should be established to make it easier for researchers to identify useful pre-existing assessments that can be adapted to fit language, writing system, culture and other contextual factors.

For researchers

- Researchers should report results in a way that reflects the inter-play between different skills. Granular reporting of items one at a time can be misleading.
- More data should be collected on the affordability of tests and the conditions needed to enable those within the education system to implement them reliably.
- Further research is required to better understand the mechanisms by which teacher-led assessments can lead to improved learning outcomes. Policy makers should act on the results.
- More assessments should be designed to enable cross-country comparisons. E.g. through the use of link items.

Reading with understanding

It is important that children learn to read with understanding. The skills and knowledge that children require to do this (symbol knowledge, oral language skills, emergent literacy skills, decoding and language comprehension) inform each other and develop together. Complementary skills develop in tandem, rather than sequentially. Assessments must reflect this view of reading development.

Similarly, children need to write to convey meaning. Assessment should include this area of literacy development (emergent writing, spelling and narrative writing).

Background

In the past two decades, there has been a huge expansion in the range of measurement tools available to assess foundation learning and literacy in developing countries. The vast majority of the measures are researcher-developed tools, which are generally used in smaller samples, and in response to a specific research aim. However, there is an increasing demand for learning assessments that can be applied at-scale to monitor education quality and inform teaching practice. Commercial tests with demonstrated good psychometric properties in Western contexts, are increasingly being adapted for use in developing countries at large and small scales. Unfortunately, the process of adapting the assessments to new contexts and testing how they perform has often been carried out without sufficient rigour and care.

Why assess?

There are two reasons why it is useful to assess children's learning and underlying skills:

1. Assessment can monitor educational quality. Communicating test results about what children can do (or cannot do) can improve decision making at every level of the education system. This improves educational quality and thereby lifts children's attainment.

2. Assessment can inform teaching practice. Teachers who assess well and use test information well, teach better. Towards this aim, the synthesis collates measures that potentially could be part of a teacher's toolkit.

The recommendations for what should be assessed at scale and in a teachers' toolkit are shown below:

	Principles	Skills to assess
At Scale 	Assessments should focus on reading accuracy and reading comprehension. Supplementary skills should be chosen depending on three criteria: • Assessments do not require too much resource • Assessments must provide additional useful information about the education system • It is likely that the assessment can be conducted at scale.	• Reading accuracy • Reading comprehension • Emergent writing • Symbol knowledge • Spelling • Narrative writing • Vocabulary • Listening comprehension • Grammatical awareness
Teachers' toolkit 	Three core principles guide the choice of tests to be included: • The tests should align with what children must learn in order to read and write well. Otherwise, teachers' responses to the data may not represent an improvement in teaching practice. • The toolkit should only include tests the teachers are able to use properly. • Data from assessments should be useful to inform teachers and aid their practice in the classroom.	• Vocabulary and spoken language • Concepts About Print • Symbol knowledge • Reading and spelling accuracy • Reading comprehension and narrative writing

Importance of context

Language characteristics and cultural factors significantly affect how pupils respond to test items. It is therefore important that assessments should always suit the context in which they are administered in order to ensure that the data they produce is valid and reliable.

Test localisation is not only a matter of literal translation of the test; it involves ensuring that the test format, its content and the testing process are familiar and meaningful for test-takers.

Piloting is essential to ensure that test design decisions do not contribute to costly errors of measurement. However, for a large proportion of reviewed studies, pilots were either not carried out or not reported. The consequence of such oversight is to limit the usefulness of the resulting data. For example, in languages like English and Spanish, letter naming tasks generally discriminate poorly between children, as most children master this skill by Grade 2. Unless an appropriate combination of single letters (e.g. “a”) and more complex letter combinations (e.g. “-tion”) are included, this task may not provide useful information.

The Young Lives international longitudinal study of children and youth [http://www.younglives.org.uk]

The Young Lives study provides an example of a best practice in transparent reporting of both localisation procedures and of statistical test performance across cultural groups.

The survey tracks 3,000 children over a 15-year period in each of its study countries – Ethiopia, India (Andhra Pradesh), Peru and Vietnam. The data structure includes a range of measures of children’s literacy, mathematics and cognitive skills as well a broad range of information about the children’s contexts.

The researchers demonstrate:

- Sensitivity to establishing the fairness of tests (e.g. comparing the behavior of the assessment in different groups).
- Use of methods to establish fairness. These include impressionistic procedures (e.g. cultural relevance and appropriateness, adequate conditions for test taking) and empirical procedures (e.g. Differential Item Functioning).

Cueto, S. and Leon, J. (2012) Psychometric characteristics of cognitive development and achievement instruments in Round 3 of Young Lives. Oxford, UK: Young Lives.

Cueto, S., Leon, J., Guerrero, G. and Munoz, I. (2009) Psychometric characteristics of cognitive development and achievement instruments in Round 2 of Young Lives. Oxford, UK: Young Lives.

Seager, J. and de Wet, T. (2003) Establishing large panel studies in developing countries: the importance of the Young Lives pilot phase, working paper 9, Oxford: Young Lives.

Issues in reporting

Reporting of results tends to look at each sub-skill separately even though some assessments appear to be designed using an approach that acknowledges the importance of considering multiple skills simultaneously.

This can lead to misleading conclusions. For example, it can impose strict assumptions about the order in which skills are learnt, which do not hold in reality. Identifying reading fluency rates of children without a focus on reading with meaning pushes towards a teaching approach for reading speed before comprehension. In reality, after a certain level, the two can develop simultaneously, each supporting the other.

Important information about assessments is also often missing:

- The profile of the assessor – interpersonal processes may influence outcomes
- Processes for increasing contextual relevance including use of culturally-embedded material and translation
- Contextual factors that can influence results. E.g. familiarity with printed materials and other task demands
- Psychometric properties

Assessments by literacy skill

Emergent Literacy

What? A rudimentary understanding of how written language connects with spoken language and knowledge of how to handle printed materials.

Why measure? These skills help pupils when they later come to learn literacy skills.

How many? 22 measures in 15 studies

Most common assessment: Concepts about print - basic questions about a book like "I will read this book. Show me where to read."

Alternatives:

- Emergent writing tasks - writing their name, symbols (e.g. letters, akshara) or common words
- Moving word task - children recognising when the wrong word is used to label a picture
- Emergent orthographic knowledge - Multiple choice questions asking children to select a symbol, symbol string or word from a set of pseudo-print distracters

Symbol knowledge

What? Understanding the relationship between symbols and units of sound as well as how to write symbols.

Why measure? It is a building block toward accurate reading and spelling.

How many? 58 measures in 30 studies

Most common assessment: Knowledge of letter sounds or letter names. This develops from individual symbols (e.g. ‘r’) to multi-symbol strings (e.g. ‘th’).

Alternatives:

- Asking children to name or write as many symbols (e.g. letters, akshara) as they know
- Discriminating between visually confusable symbols
- Symbols in context - matching words with symbols or "say a word that starts with the letter _"
- Mixed symbols lists

Considerations: The total number of symbols and their frequency of use influence performance.

To improve the ability of assessments to distinguish between children at similar levels of attainment, assessors could add clusters and affixes and frequent and uncommon symbols.

Reading accuracy

What? The knowledge necessary to recognise words by decoding or using other strategies.

Why measure? Decoding is both critical for literacy development and sensitive to instruction and opportunity. The association between being able to decode words and being able to comprehend written text is seen among monolingual, bilingual and biliterate readers.

How many? 90 measures in 37 studies

Most common assessment: Children are asked to read lists of words, nonwords or connected texts, typically chosen to fit the curriculum.

Alternatives:

- Lexical judgement tasks – children are asked to distinguish between words and pseudo-words
- Word chain – children are asked to mark word boundaries in a continuously printed word chain

Considerations: The pace at which children develop decoding skills depends on the consistency, familiarity and complexity of symbol-sound mapping. It is also influenced by access to varied books and printed materials, which provide opportunities to practice these skills.

The use of nonwords is not advised as it detracts from a focus on reading for meaning, while showing no advantage over using words.

Spelling

What? Writing words accurately.

Why measure? Spelling is a decoding skill like reading accuracy. There is a strong correlation between the two, but spelling tends to be more difficult. The fine motor skills required to write are normally included within this skill too.

How many? 35 measures in 17 studies

Most common assessment: Spelling a dictated list of words.

Alternatives:

- Spelling nonwords and multimorphemic words (e.g. compound words, inflections)
- Assessing spelling from free writing samples
- Recognition tasks: Tasks that do not require writing (e.g. pick from multiple spelling options)

Considerations: Spelling skills develop faster when languages have consistent symbol-sound mapping.

The expression of spelling skills can be obscured by limitation in transcription skills.

Reading fluency

What? Reading connected text accurately at a speed similar to a conversational rate, with appropriate expression and intonation.

Why measure? Higher speed and accuracy suggests automaticity in word level decoding and signals that more attentional resources are available for reading comprehension processes. Prosody is the ability to reflect understanding of what is being read through the use of expression and intonation. Assessing prosody is rare, but it adds an indication of the child's ability to read with understanding.

How many? 52 measures in 16 studies.

Most common assessment: The number of words read per minute (usually measured in a 1-minute window). Most use connected text, but some use lists of words or nonwords.

Considerations: There is the risk that these assessments can send out the wrong signals and lead to poorer practice in the classroom. A focus on reading fluency may unwittingly encourage a shift of focus away from meaning-based instruction and towards the mechanics of speed and accuracy. Including prosody in assessments can reduce this risk

Reading comprehension

What? Extracting meaning from written text. This combines two skills: deducing the correct words from written symbols and extracting meaning from words. Skilled comprehenders may use multiple strategies to understand the text (e.g. looking back at the text).

Why measure? It provides direct evidence of how well a child can read and how well a teaching programme is working. It encompasses all sub-tasks of reading and is related unambiguously with end benefit.

How many? 66 measures in 27 studies.

Most common assessment: (i) Question and answer – children read a passage and answer questions on it.

(ii) Cloze tests – texts with some words replaced by blank spaces. Children demonstrate an understanding of the text by correctly filling in the gaps.

Alternatives:

- Modified Cloze (or Maze) tests, choosing from a list of suggested words to fill in gaps
- Matching a sentence with a picture

Narrative writing

What? Multiple cognitive-linguistic processes underpin narrative writing. Of these, the mechanics of writing, narrative generation and memory are key. At higher levels, it includes planning skills and writing for an audience.

Why measure? These assessments have the potential to provide direct inferences about what the child can do in the area of writing, what they need to write better and what the teacher can do to help them.

How many? 17 measures in 9 studies.

Most common measure: Writing in response to prompts. Children may be asked to complete an unfinished story, re-write a story or collate information for a factual piece.

Measures can cover multiple aspects of the written responses, including:

- transcription skills - handwriting, punctuation, spelling
- narrative generation skills - vocabulary, style-related details (signalling time and chronology, tone of the story, etc.), cohesiveness of the narrative, awareness of the reader
- working memory
- writing fluency - words written per minute
- the quality of language and detail in the narrative
- creative (which is not well defined)

Considerations: Performance is sensitive to opportunity: Children perform less well when their instruction does not cover a broad range of writing skills.

Content generation can be constrained if a child needs to give too much attention to the physical task of writing. So these tasks are a better indication of a child's language skills once their transcription skills have reached a certain level.

Vocabulary

What? The breadth and depth of knowledge about words, either expressing them or understanding them.

Why measure? It aids reading comprehension and the decoding of words that are difficult to decode (e.g. multi-morphemic, written with an uncommon symbol or with an exceptional spelling).

How many? 63 measures in 33 studies

Most common measure: Picture vocabulary test - the child is asked to point to one of four pictures that match a just-heard word

Alternatives:

- Identify a target word from a set containing distractor words; identify a synonym
- Semantic fluency- "name as many ___ as you can"
- Definitions - define target word
- Focus on parts of a word (e.g. drop or change inflections in words)

Considerations: Vocabulary development is exceptionally sensitive to ambient language.

Other areas of spoken language assessment

What? Listening comprehension – obtaining meaning from spoken language; understanding of grammar and language structure.

Why measure? Spoken language assessment can inform the question: What does the child need in order to read and write well? It provides information about the skills that children bring to the task of reading and writing from their home and culture.

How many? 36 measures in 17 studies

Most common measure: Comprehension questions following a just-heard message; grammatical awareness (repeating a message, judging the appropriateness of sentence construction); retelling a short story or proverb.

Considerations: These measures are sensitive to language and context due to the relationship with oral traditions. Using instruments adapted from measures developed in other contexts is problematic.

Gaps in the evidence

The review summarises four key gaps in the evidence:

- **The profile of the assessor**. The identity of the assessor influences how they relate to the child and, in turn, how the child performs. Gender, socio-economic status, urban or rural background, ethnicity, religion and linguistic affiliation should all be considered and reported on.
- **How the assessment results relate to the context**. Researchers need to consider the degree to which their assessments observe contextual factors rather than pupils' skill. Apparent improvements or variations in performance may not be caused by improvements in students' skills as we tend to assume. Instead, they could be caused by teachers coaching students on the particular tasks in the tests or variations in familiarity with the printed text used in the assessment. Bilingual contexts add further complications in understanding pupils' performance in literacy tests. This needs to be better understood.
- **Dissemination of assessment results.** The flow of information from assessments to decision makers and stakeholders is crucial to ensure that they are useful. However, studies did not report on their dissemination plans so it is not clear how far this was considered.
- **Reporting standards.** It is important for researchers to build confidence in their assessment tool among decision makers. However, the review found that reporting of the contextual relevance and psychometric properties of measures was poor.

The reliability and appropriateness of the most popular measures

Skill	Most common measure	Reliability	Appropriateness
Emergent literacy	**Concepts About Print**: Basic questions about a book like "I will read this book. Show me where to read." **Example reference** Chinyama, A. et al. (2012*). Literacy boost Zimbabwe: Baseline report*. Zimbabwe: Save the Children.	There is no reliability information for 13 of the 22 measures. However, where reported, reliability estimates (Cronbach's alpha) are typically moderate to excellent for measures of concepts about print, emergent writing and emergent orthographic knowledge where children have to select a symbol or word from a list of non-symbol or non-word distractors.	Does not capture much variation for narrow SES groups in poor areas so would not be appropriate to evaluate educational quality, but may still be useful in a teacher's toolkit.
Symbol knowledge	**Knowledge of letter names or letter sounds**: Correctly identifying letter names or sounds. **Example reference** Alcock, K. J., et al. (2000). The development of reading tests for use in a regularly spelled language. *Applied Psycholinguistics*, 21(4), 525-555.	39 of the 58 measures do not report reliability information. However, reliability estimates are typically high (above 0.8). These estimates are reported also in second language settings and bi-scriptal contexts.	Positively correlated with print experience at home and sensitive to differences in instruction quality. The stage at which these tasks best differentiate between children depends on the language. They distinguish better in earlier grades for languages that are have simple symbol-sound mapping and/or have small symbol sets. They are useful in later grades for languages with large symbol sets or when complexities in symbol-sound links are included in the task. The teaching strategy has a significant effect on results. When letter names are explicitly taught, almost all children learn them and the level of attainment across the group becomes uniform.

Reading accuracy	**Decoding words**: Reading lists of words (10-200 items). **Example reference** Babayigit, S., & Stainthorp, R. (2010). Component processes of early reading, spelling, and narrative writing skills in Turkish: A longitudinal study. *Reading and Writing: An Interdisciplinary Journal*, 23(5), 539-568.	There is no reliability information for 64 of the 90 measures. However, reliability estimates (Cronbach's alpha, split half or test-retest) are typically excellent (above 0.9). There is also evidence of convergent divergent validity (correlation with appropriate measures). The shortest test with a reliability estimate of 0.95 was 20 words for Grade 1 and 30 words for Grade 2.	The ability of the assessment to distinguish between children at different levels depends on: (i) Language characteristics: Good distributions are found well into middle school with irregularly spelled languages, while performance reaches a ceiling within the initial school years for regularly spelled languages. Error analysis or speed measures can capture variations in regular languages. For languages with simple symbol-sound mapping (transparent orthographies), the tests differentiate well through into middle school when there is a large number of symbols or some low frequency symbols. (ii) Word selection: Distributional properties of scores appear to be better when item selection is based on psycholinguistic and orthographic characteristics or a random selection from a dictionary list.
Spelling	**Spelling dictated words**: spell a list of words dictated to them and the accuracy is measured. **Example reference** Nag, S., Treiman, R., & Snowling, M. (2010). Learning to spell in an alphasyllabary: The case of Kannada. *Writing Systems Research*, 2(1), 1-12.	Spelling measures show some of the highest reliability indices in the review.	An important innovation is to remove the writing component in the task (e.g., ask children to sequence symbol cards to show spelling, identify correct spelling in a multiple choice format, etc.). These innovations need to be evaluated across contexts (e.g., first first-generation learners, with visually complex orthographies, etc.).

Reading Fluency	**Words read per minute:** The number of words read per minute. Most (25) use connected text, but some use lists of word (17). **Example reference** Asfaha, Y. M., Kurvers, J., & Kroon, S. (2009). Grain size in script and teaching: Literacy acquisition in Ge'ez and Latin. *Applied Psycholinguistics*, 30(4), 709-724.	It is easier to achieve high levels of reliability for fluency assessments than for comprehension assessments. Word reading fluency assessments have produced internal consistency estimates between 0.70 and 0.96 (8 studies). Only 1 nonword fluency measure reported reliability - a rest-retest reliability of 0.74. 50% of connected text measures reported reliability estimates above 0.90. Reliability estimates for measures applied in a second language were significantly lower - between 0.68 and 0.87.	In languages with simple symbol-sound mapping (transparent languages), reading fluency tends to be the measure of choice because reading accuracy stops differentiating between children at a fairly early stage of education. Reading fluency is less important for languages with more complex symbol-sound mapping and spelling system because reading accuracy differentiates well to a more advanced educational level. It is not possible to compare reading fluency results across languages. There are concerns about whether measuring reading fluency shifts the focus of teaching away from meaning-based instructions towards the mechanics of speed and accuracy.
Reading comprehension	"**Question and answer** (38 measures): Children answer questions on a text to demonstrate direct understand and/or accurate inference. **Cloze** (20 measures): Children fill in gaps deliberately left in some text. **Example reference** **Q&A:** Nag, S., & Snowling, M. J. (2011). Cognitive profiles of poor readers of Kannada. *Reading and Writing: An Interdisciplinary Journal*, 24(6), 657-676. **Cloze:** Williams, E. (1998). *Investigating bilingual literacy: Evidence from Malawi and Zambia* (Education Research Paper, p. 110). London, UK: Department for International Development (DFID).	Q & A tests: 0.73 for 1 question per passage; 0.62 for 2 questions; 0.81-0.85 for 3-6 questions in 1st language, 0.70-0.80 in 2nd language; 0.73-0.79 for 7-12 questions in 1st language, 0.60-0.82 in 2nd language.	Reading comprehension is closely linked to the end benefit for children. Assessing comprehension rather than fluency moves attention away from the mechanics of speed and accuracy towards the importance of extracting meaning from text.

Narrative writing	**Written responses to trigger material (prompts)**: Measures can cover multiple aspects of the written responses, including transcription skills, working memory, writing fluency, the quality of language and detail in the narrative and creativity. **Example reference** Johnson, D., Hayter, J., & Broadfoot, P. (2000). *The quality of learning and teaching in developing countries: Assessing literacy and numeracy in Malawi and Sri Lanka* (Education Research Paper No 41). Kent, UK: Department for International Development (DfID).	Reliability may arguably be higher for transcription skills and accuracy of recall than for narrative generation skills at the level of content and structure. One study partially supports this - inter-rater reliability between 2 primary school teachers was 0.97 and 0.99 for transcription skills and narrative content, but 0.74 for narrative structure.	Scaffolding has been added in some tests to support children with contextual information and memory prompts. These can help facilitate greater variability in written outputs because those who would not have written much may be encouraged to write more. The skills required to rate children's written work may not be easily available among some teachers.
Vocabulary	**Picture Vocabulary Test**: Child has to point to one of four pictures that matches a just heard word **Example reference** Cueto, S., Leon, J., Guerrero, G., & Munoz, I. (2009). *Psychometric characteristics of cognitive development and achievement instruments in Round 2 of Young Lives* (Young Lives Technical Note 15). Retrieved from www.younglives.org.uk	Picture identification activities have the largest proportion of measures with reliability estimates above 0.80 (approx. 50%).	Second language learners tend to know fewer words and learn new words more slowly than native learners. Localisation is important in order to accommodate linguistic and cultural considerations. The full complexity of vocabulary is not assessed - abstract words and words in multiple contexts are under-represented. Different types of vocabulary measures are not comparable because they assess different elements of the same construct.