HEART
HEALTH & EDUCATION ADVICE & RESOURCE TEAM

UKaid
from the British people

ASSESSMENT OF LITERACY AND FOUNDATIONAL LEARNING IN DEVELOPING COUNTRIES

Final Report Sonali Nag

This assessment is being carried out by HEART (Health & Education Advice & Resource Team).

The project manager/Team leader is Sonali Nag. She was supported by the Promise Foundation. For further information contact HeartforPeaks@opml.co.uk.
The contact point for the client is 'Goli Whittaker' (g-whittaker@dfid.gov.uk)The client reference number for the project is 7825-A1218.

HEART
Level 3, Clarendon House
52 Cornmarket Street
Oxford, OX1 3HJ
United Kingdom

Tel +44 (0) 1865 207 333
Email info@heart-resources.org
consultancy@heart-resources.org
Web www.heart-resources.org

Acknowledgements

This report has benefitted from the inputs of several colleagues who asked searching questions as we synthesised the data. They are Gideon Arulmani (Promise Foundation, India), Shaher Banu Vagh (ASER Centre, India), Yvonne Griffiths (University of Leeds, UK), Gloria Ramirez (Thompson Rivers University, Canada), Rhona Stainthorp (University of Reading, UK) and Margaret Snowling (University of Oxford, UK). We would also like to thank our peer reviewers Reg Allen and Rachel Hinton, as well as participants at a Round Table in July 2016, who discussed a first draft of the synthesis. I also gratefully acknowledge the meticulous work of the research assistants on this project.

The review team

Name	Initials	Name	Initials
Principal Investigator		**Consultant**	
Sonali Nag	SN	Gideon Arulmani	GA
Research assistants			
Sajma Aravind	SA	Khyati Sampat	KS
Rayan Miranda	RM	Gurpreet Reen	GR
Kalyani Sadekar	KS	Emily Reeves	ER
Mini Krishna	MK	Prerna Menon	PM
Riona Lall	RL	Sudha Vijay	SV

Review team for the Nag *et al.* (2014) review

Name	Initials	Name	Initials
Core team		**Panellists**	
Margaret Snowling	MJS	Monica Melby-Lervag	MML
Sonali Nag	SN	Shaher Banu Vagh	SBV
Shula Chiat	SC	Terezinha Nunes	TN
Carole Torgerson	CT	Yonas Asfaha	YA
Research assistants			
Dominique Shure	DS	Gurpreet Reen	GR
Prerna Menon	PM	Yvonne Griffiths	YG
Kamila Polisenska	KP	Marina Puglisi	MLP
Angshuman Phukan	AP	Meenakshi Parameshwaran	MP
Sundas Ali	SA	Emily Reeves	ER
Mini Krishna	MK	Mili Kalia	MKa
Centre for Reviews and Dissemination			
Steven Duffy	DF	Lisa Stirk	LS

Executive summary

Context, scope and framework

This review examines the quality and range of tools used to measure literacy and foundational learning in developing countries. It covers the assessment of language and literacy skills in children from age three to 14 (or preschool to Grade 8) and includes assessment tools from studies published between 1990 and 2014, rated as 'Moderate' or 'High' in methodological quality.

There are two main reasons 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 review is underpinned by a Systems View of Reading, which sets literacy in the context of other language, cognitive and social skills. This view highlights the importance of developing complementary skills together with each benefitting from the development of the other. The various skills and knowledge that children require to read with meaning inform each other and develop together. They do not necessarily develop sequentially.

With this theory providing an underlying framework, the review identifies the most common assessments used to measure all subskills: emergent literacy, symbol knowledge, reading accuracy, spelling, reading fluency, reading comprehension, narrative writing, vocabulary and other areas of spoken language assessment.

Each section gives an overview of why it is useful to assess the skill, the approaches that have been taken to assessing it and current innovations and challenges. The validity and reliability of assessments are discussed wherever these are reported in sufficient detail.

Implementation a scale and teachers' toolkit

There are several assessments that have been successfully implemented at scale, including tests of symbol knowledge, reading accuracy, reading fluency and reading comprehension. These should be augmented by further skills that could feasibly be implemented at scale and would provide information that is useful for strengthening the education system: emergent writing, spelling, narrative writing, vocabulary, listening comprehension and grammatical awareness.

The review sets out criteria for identifying the assessments that teachers could use to gain information to improve their practice in the classroom. Importantly, these include ensuring that the assessments align with what children must learn in order to read and write well and ensuring that teachers have the skills to use the assessment well. Teachers can be led by data (indeed, that is often our hope) so it is important to avoid diverting their attention away from important aspects of the learning process by over-emphasising others. The review recommends that this toolkit could include assessments of vocabulary and spoken language, concepts about print, reading and spelling accuracy, reading comprehension and narrative writing.

Localisation

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 result in measurement errors. 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. “sh” and “-tion”) are included, this task may not provide useful information.

Ensuring that an assessments are appropriate for the context in which they are administered requires care and resource. Good practice to establish fairness includes assessing the cultural and linguistic appropriateness and relevance qualitatively and empirically.

In general, the review shows that there is significantly better reporting of localisation among researcher-developed tools, but that commercial test adaptation processes are poorly documented. A large proportion of the studies using adapted versions mention translation and the use of consultations with local experts to select and translate items, but do not detail the actual translation procedures.

Communicating results

The way that results are communicated influences the actions that will be taken as a result. Reporting is therefore an important stage in the assessment process. The Systems View of Reading highlights the importance of assessing multiple skills simultaneously to reflect the way that children learn. However, reporting of results tends to look at each sub-skill separately.

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 the letters that are misidentified pushes towards a teaching approach that ensures that all letters must be mastered before any word decoding is introduced. In reality, after a certain level, the two can develop simultaneously, with each supporting the other.

Transparency around the robustness of measures is often poor with important information about assessments often missing:

- The profile of the assessor as interpersonal processes often influence outcomes
- Processes for localizing such a translation and the use of culturally-embedded material
- Contextual factors that can influence results, such as familiarity with printed materials
- Contextual relevance
- Psychometric properties

Reporting these details helps to build trust in the instrument. It also influence how results should be interpreted. For example, it can help to identify where a result is the result of school experience, rather than learnt skill or whether questions could have been misinterpreted by some test-takers.

Future directions

It is important that learning measures are designed in a way that ensures that the results they produce reflect the skills that they pertain to measure in a reliable manner. To do this, careful consideration needs to be given to whether theories and approaches developed for other languages, school systems and socio-cultural contexts can be applied to the local population.

Efforts should be made to develop affordable tests so that the benefits of assessment can be felt by a broader group, including the poor and marginalised.

Further to this, assessment tools should be placed in the hands of teachers to enable them to develop a better understanding of what is happening in their classroom and what they can do to improve it. This direct feedback can help in a way that at-scale assessment cannot. This being said, the mechanisms by which teacher-led assessment can lead to better practice and improved learning needs to be better understood.

Further rigorous reviews should be commissioned to supplement this study. These include: a review of multi-country assessments, such as PIRLS, ASER and EGRA; and review of how to measure the contextual factors that play such a key role in determining both how much pupils learn and how assessments should be interpreted.

Finally, a free-to-use resource bank of robust and useful tests should be developed. This would help to further the use of high-quality tests whilst also pooling a larger volume of information about how tests perform in different settings.

References

For quick reference, examples of each type of assessment are listed in the table on the following pages.

Emergent literacy	**Most common**	**Concepts About Print** Chinyama, A., Svesve, B., Gambiza, B., Guajardo, J., Onunda, D., & Dowd, A. J. (2012*). Literacy boost Zimbabwe: Baseline report*. Zimbabwe: Save the Children. Vagh, S. B. (2009). Learning at home and at school: A longitudinal study of Hindi language and emergent literacy skills of young children from low-income families in India. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 70(11-A), 4183.
	Alternatives	**Emergent writing tasks - writing their name, symbols or common words** Strasser, K., & Lissi, M. R. (2009). Home and instruction effects on emergent literacy in a sample of Chilean kindergarten children. *Scientific Studies of Reading*, 13(2), 175-204. **Moving word task - recognising when the wrong word is used to label a picture** Rochdi, A. (2010). Developing pre-literacy skills via shared book reading: The effect of linguistic distance in a diglossic context. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 70(8-B), 4801. **Emergent orthographic knowledge – selecting a symbol, symbol string or word from a set of distractors** Jere-Folotiya, J., Chansa-Kabali, T., Munachaka, J. C., Sampa, F., Yalukanda, C., Westerholm, J., …, Lyytinen, H. (2014). The effect of using a mobile literacy game to improve literacy levels of grade one students in Zambian schools. *Education Tech Research Dev*, 62, 417-436.
Symbol knowledge	**Most common**	**Symbol name, Symbol sound** Alcock, K. J., Nokes, K., Ngowi, F., Musabi, C., Mbise, A., Mandali, R., . . . Baddeley, A. (2000). The development of reading tests for use in a regularly spelled language. *Applied Psycholinguistics*, 21(4), 525-555. Asfaha, Y. M., Beckman, D., Kurvers, J., & Kroon, S. (2009). L2 reading in multilingual Eritrea: The influences of L1 reading and English proficiency. *Journal of Research in Reading*, 32(4), 351-365. Nag, S., & Snowling, M. J. (2012). Reading in an alphasyllabary: Implications for a language-universal theory of learning to read. *Scientific Studies of Reading*, 16(5), 404-423.
	Alternatives	**Asking children to name or write as many symbols as they know** Piper, B. (2010). *Ethiopia early grade reading assessment (Data analysis report)*. Research Triangle Park, NC: RTI. **Discriminating between visually confusable symbols** Elbeheri, G., & Everrett, J. (2007). Literacy ability and phonological processing skills amongst dyslexic and non-dyslexic speakers of Arabic*. Reading and Writing: An Interdisciplinary Journal*, 20(3), 273-294. **Symbols in context - matching words with letters or "say a word that starts with _"** Oktay, A., & Aktan, E. (2002). A cross-linguistic comparison of phonological awareness and word recognition in Turkish and English*. International Journal of Early Years Education*, 10(1), 37-48. **Mixed symbols lists** Nag, S. (2007). Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading*, 30(1), 7-22.

Reading accuracy	**Most common**	**Read lists of words or connected texts** 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. Veii, K., & Everatt, J. (2005). Predictors of reading among Herero–English bilingual Namibian school children. *Bilingualism: Language and Cognition*, 8(3), 239-254. Winskel, H., & Widjaja, V. (2007). Phonological awareness, letter knowledge and literacy development in Indonesian beginner readers and spellers. *Applied Psycholinguistics*, 28(1), 23-45.
	Alternatives	**Lexical judgement tasks – distinguish between words and pseudo-words** Jukes, M., Vagh, S., & Kim, Y. (2006). *Development of assessments of reading ability and classroom behaviour.* Washington, DC: World Bank. **Word chain – mark word boundaries in a continuously printed word chain** Nakamura, P. (2014). *Facilitating Reading Acquisition in Multilingual Environments in India (FRAME-India): Final report*. American Institutes for Research.
Spelling	**Most common**	**Dictated list of words** Ledesma, H. M. L. (2002). Language factors influencing early reading development in bilingual (Filipino-English) boys. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 63(6-A), 2096. Nag, S., Treiman, R., & Snowling, M. (2010). Learning to spell in an alphasyllabary: The case of Kannada. *Writing Systems Research*, 2(1), 1-12.
	Alternatives	**Spelling nonwords and multimorphemic words (e.g. compound words, inflections)** Winskel, H., & Widjaja, V. (2007). Phonological awareness, letter knowledge and literacy development in Indonesian beginner readers and spellers. *Applied Psycholinguistics*, 28(1), 23-45. **Assessing spelling from free writing samples** 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. **Recognition tasks: Tasks that do not require writing (e.g. pick from multiple spelling options)** Test by Ojanen et al. 2013, reported in Jere-Folotiya, J., Chansa-Kabali, T., Munachaka, J. C., Sampa, F., Yalukanda, C., Westerholm, J., . . . Lyytinen, H. (2014). The effect of using a mobile literacy game to improve literacy levels of grade one students in Zambian schools. *Education Tech Research Dev*, 62, 417-436.
Reading fluency	**Most common**	**Connected text** Piper, B., Zuilkowski, S. S., & Mugenda, A. (2014). Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. *International Journal of Educational Development*, 37, 11-21. **Words** 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. **Nonwords** Mohamed, W., Elbert, T., & Landerl, K. (2011). The development of reading and spelling abilities in the first 3 years of learning Arabic. *Reading and Writing: An Interdisciplinary Journal,* 24(9), 1043-1060.
	Alternatives	**Exception to the one minute rule (five minutes with sentences)** Alcock, K. J., Nokes, K., Ngowi, F., Musabi, C., Mbise, A., Mandali, R., . . . Baddeley, A. (2000). The development of reading tests for use in a regularly spelled language. *Applied Psycholinguistics*, 21(4), 525-555.

Reading comprehension	**Most common**	**Question and answer – read a passage and answer questions on it** Davidson, M., & Hobbs, J. (2013). Delivering reading intervention to the poorest children: The case of Liberia and EGRA-Plus, a primary grade reading assessment and intervention. *International Journal of Educational Development*, 33, 283-293. Nag, S., & Snowling, M. J. (2011). Cognitive profiles of poor readers of Kannada. *Reading and Writing: An Interdisciplinary Journal*, 24(6), 657-676. **Cloze tests – texts with some words replaced by blank spaces. Children demonstrate an understanding of the text by correctly filling in the gaps** Williams, E. (1998). *Investigating bilingual literacy: Evidence from Malawi and Zambia* (Education Research Paper, p. 110). London, UK: Department for International Development (DFID).
	Alternatives	**Modified Cloze (or Maze) tests, choosing from a list of suggested words to fill in gaps** Jukes, M., Vagh, S., & Kim, Y. (2006). *Development of assessments of reading ability and classroom behaviour*. Washington, DC: World Bank. **Matching a sentence with a picture** Spratt, J., Seckinger, B., & Wagner, D. (1991). Functional literacy in Moroccan school children. *Reading Research Quarterly*, 26(2), 178-195.
Narrative writing	**Most common**	**Writing in response to prompts** 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. 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).
Vocabulary	**Most common**	**Picture vocabulary test – point to on of four pictures that match a just-heard word** 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 Vagh, S. B. (2009). Learning at home and at school: A longitudinal study of Hindi language and emergent literacy skills of young children from low-income families in India. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 70(11-A), 4183.
	Alternatives	**Identify a target word from a set containing distractor words; identify synonym** Alcock, K. J., Ngorosho, D., Deus, C., & Jukes, M. C. H. (2010). We don't have language at our house: Disentangling the relationship between phonological awareness, schooling, and literacy. *British Journal of Educational Psychology*, 80(1), 55-76. **Semantic fluency- "name as many ___ as you can"** Jukes, M. C. H., & Grigorenko, E. L. (2010). Assessment of cognitive abilities in multiethnic countries: The case of the Wolof and Mandinka in the Gambia. *British Journal of Educational Psychology*, 80(1), 77-97. **Definitions – define a target word** Nag, S., & Snowling, M. J. (2011). Cognitive profiles of poor readers of Kannada. *Reading and Writing: An Interdisciplinary Journal*, 24(6), 657-676. **Focus on parts of a word (e.g. drop or change inflections in words)** Winskel, H., & Widjaja, V. (2007). Phonological awareness, letter knowledge and literacy development in Indonesian beginner readers and spellers. *Applied Psycholinguistics*, 28(1), 23-45.

Other spoken language	**Most common**	**Comprehension questions following a just-heard message** Davidson, M., & Hobbs, J. (2013). Delivering reading intervention to the poorest children: The case of Liberia and EGRA-Plus, a primary grade reading assessment and intervention. *International Journal of Educational Development*, 33, 283-293. Jukes, M. C. H., & Grigorenko, E. L. (2010). Assessment of cognitive abilities in multiethnic countries: The case of the Wolof and Mandinka in the Gambia. *British Journal of Educational Psychology*, 80(1), 77-97. **Grammatical awareness (repeating a message, judging the appropriateness of sentence construction)** Fernald, L. C. H., Kariger, P., Engle, P., & Raikes, A. (2009). *Examining early child development in low-income countries: A toolkit for the assessment of children in the first five years of life*. Washington, DC: The World Bank. **Retelling a short story** Castilla, A. P. (2008). Developmental measures of morphosyntactic acquisition in Monolingual 3-, 4-, and 5-year-old Spanish-speaking children. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 71(4-B), 2362.

Table of contents

List of abbreviations

ASER	Annual Status of Education Report
CRD	Centre for Reviews and Dissemination
CTT	Classical Test Theory
DFID	Department for International Development
IRT	Item Response Theory
MUW	Most Used Words
PIRLS	Progress in International Reading Literacy Study
PPVT	Peabody Picture Vocabulary Test
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SSCI	Social Science Citation Index
WRAT	Wide Range Achievement Test

1 Context, scope and framework

1.1 Background to the current review

In 2013, a team of researchers undertook a rigorous review of the literature on literacy and foundational learning in developing countries (Nag *et al.*, 2014). During the course of the review it was found that the measurement of children's learning and underlying skills has been approached from a wide range of perspectives and methodologies. In parallel, the variety in assessment methods suggested that innovations were occurring in developing countries and established methods were going through contextual variations. In this current rigorous review on assessment we return to the Nag *et al.* (2014) review to examine the quality and range of measurement tools. We focus on measures of individual differences in children's literacy, language and learning. We consider, in the light of theory, the types of assessment that are both available and required to assess literacy development.

- Under **literacy** we review measures of emergent literacy, symbol knowledge, reading accuracy, spelling, reading fluency, reading comprehension, narrative writing and grade-level tests.
- Under **language** we cover vocabulary and other areas of language assessment such as listening comprehension and grammar knowledge.[1]

An area of particular interest is assessment within exceptionally low levels of achievement and when children have limited test-taking experience. The economic cost of one-on-one testing is daunting, particularly in low-resource, large population contexts. We therefore also draw attention to tests that show potential for a group-testing format.

1.2 The purposes of assessment

One approach to understanding measurement tools is to specify the purposes of assessment that are of interest. For example, a measure developed for comparison across contexts may not yield sufficient information to initiate change within any one context. In a similar vein, assessments that are useful for a teacher may be of limited value to a researcher. Against this background of multiple contexts and purposes of assessment, we address two themes:

- **Assessment that can monitor educational quality**. Here, a strong assumption is that communicating test results about what children can do (or cannot do) can inform educational quality and thereby lift children's attainments. Toward this end, the synthesis is structured around skills and sub-skills. This level of granularity gives information that can complement the use of composite measures to inform policy-level discussions.
- **Assessment that can inform teaching practice**. Here, a strong assumption is that teachers who assess well and use test information well teach better. Toward this aim, the synthesis reviews measures for their potential as a teacher-led assessment tool. The level of fine-grained information gathered about children's literacy and language skills and knowledge can complement curriculum-focused testing to inform pedagogical decisions.

[1] The list of measures included in the review is available in Annex A. Note that we do not review tools used in multi-country comparisons such as the Progress in International Reading Literacy Study (PIRLS) and citizen-led surveys such as the Annual Status of Education Report (ASER). We recommend a separate rigorous review of the rapidly growing literature on tests developed for these purposes.

1.3 The linguistic landscape of developing countries

Developing countries present a diverse linguistic landscape and what it is useful to assess must be considered carefully. Bilingualism and multilingualism are common. The distance between spoken and written language forms often vary (e.g. for Arabic, Tamil and Spanish, the distance to the standard written form differs across the multiple countries where the language is in use). In addition, it is common for children to acquire literacy in more than one language and the potential challenges of biliteracy may differ depending on the particular combination of language and orthography (e.g. compare Arabic–English, Filipino–English and Kannada–English learning). It follows that the nature of the relationship between literacy development and oral language skills in such linguistic contexts is complex. This is an important reason why we have chosen to widen the review beyond literacy to include spoken language skills as a foundation for literacy learning.

Throughout this review we identify the language of assessment as either being in the child's home language (referred to as L1) or in another language (referred to as either L2 or other than L1).

1.4 Synthesis framework for the review

This review is broadly structured around the Simple View of Reading (Gough and Tunmer, 1986). Within this view, reading with understanding depends upon two critical skills: decoding and linguistic comprehension. In turn, decoding depends on symbol knowledge, phonological awareness, the foundations of emergent literacy skills (e.g. early print awareness) and language proficiency. We also take a systems view of reading (and writing) development (see, for example, Perfetti and Stafura, 2014). Within this view, component skills in literacy involve multiple knowledge bases including knowledge about how the symbol system works for transcribing the language at hand (orthography, phonology), vocabulary knowledge (semantics) and knowledge of grammar (morphology, morpho-syntax). We examine the assessment of such within-child skills and knowledge, broadly calling them written and spoken language skills.[2]

The causal relationships between spoken and written language skills remain widely debated, although the emerging consensus is that: a) language skills are an important foundation for literacy skills; and b) the language–literacy relationship from novice through to expert levels of proficiency is reciprocal. In other words, pairs of sub-skills (e.g. phonological awareness–reading accuracy, vocabulary–reading comprehension, or listening comprehension–reading comprehension) show a two-way influence; the first in each pair is not only a predictor but also itself changes as a consequence of accumulating skills with reading. A similar reciprocity can also be assumed between the component skills of language and writing, but this is an understudied area.

This synthesis framework also lends itself to being **an assessment framework.** This framework, based on current understandings of the literacy learning process, can potentially inform priorities for teaching and support assessment design (e.g. what to assess, how to interpret assessment results, etc.). At the written language level, the framework covers emergent literacy, symbol knowledge, reading accuracy, reading fluency, reading comprehension, spelling and narrative writing. At the spoken language level, it covers vocabulary and broader knowledge about language related to listening comprehension and grammatical awareness. Working together, skills and knowledge in these areas of written and spoken language allow for effective engagement in daily uses of literacy as well as more complex (but often academic) literacy tasks.

[2] An alternative to our focus on within-child factors is an approach that is inclusive of contextual factors (e.g. assessment of the political economy, home environments, etc.). While we see such an ecologically broad approach as being of maximum value, we were unable to include contextual factors within the scope of our review.

1.5 The quantity and quality of the evidence available

The scope of this review is the assessment of language and literacy skills in children from age three to 14 (or preschool to Grade 8, with preschool referring to both preschool and kindergarten years). All studies have been conducted in a developing country between 1990 and 2014 and were identified in the earlier review by Nag *et al.* (2014). Details of the search strategy used in the Nag *et al.* (2014) review are given in Annex C. Briefly, studies from multiple electronic databases (ERIC, PsycInfo and Web of Science) and websites (What Works clearinghouse) were screened. The procurement rate for papers and documents in the original review was average (about 80%). Non-procurements were mainly of doctoral theses and papers in technical journals. The current review focuses on those studies that were rated as Moderate or High in methodological quality. The final set covers 55 languages and 53 countries. Table 1 below summarises the number of studies available for each area of assessment, with examples of tools.

Table 1: A summary of the review database given by area of assessment

Serial no.	Area of assessment	Example of tools	Number of tools [number of studies]
Written language			
1	Emergent literacy	CAP, word concept task, the book task	22 [15 studies]
2	Symbol knowledge	Identification, discrimination between pairs, symbol usage, symbol writing fluency	58 [30 studies]
3	Reading accuracy	Familiar and/or unfamiliar word reading, non-word reading, words in connected texts	90 [37 studies]
4	Spelling	Single word spelling, accuracy of words in sentences, non-word spelling	35 [18 studies]
5	Reading fluency	1 min. reading, 3 min. reading, speed of nonsense word reading	52 [16 studies]
6	Reading comprehension	Sentence and/or passage comprehension, the gap test, Cloze (maze) test	66 [27 studies]
7	Narrative writing	Composition structure, write short story/factual writing/letter of complaint	15 [9 studies]
8	Grade-level tests	End of term language tests, composite of continuous classroom evaluations	16 [15 studies]
Spoken language			
9	Vocabulary	Picture vocabulary, expressive vocabulary, semantic fluency	63 [32 studies]
10	Other language measures	Listening comprehension, story comprehension, grammatical awareness, sentence repetition	36 [16 studies]
11	Phonological skills	Rhyme generation, segmentation, blending, syllable and/or phoneme deletion	83 [27 studies]

Tests were allotted to the area that they were judged to assess. This assignment was based on the content of test items and sometimes was different from the stated title of the test. The tests assigned under phonological processing are listed (Annex A11) but are not part of this review. Grade-level achievement tests are discussed only briefly (see Section 2.8); the most obvious examples of such tests are found in multi-country comparisons (e.g. PIRLS and SECMEQ) but these are outside the scope of this review.

The numbers per area of assessment are unequal but, despite this, the available evidence within each area has a geographic and linguistic spread that allows for a rigorous review for the purposes of assessment of interest. Annex D gives the guidance note for data extraction to support such a review.

A statistical synthesis is an important objective for a rigorous review. The Nag *et al.* (2014) review, however, found a meta-analysis untenable because of the variety in the measures used on every parameter of interest. In addition, the review showed that there are either too few studies within each grade/age band or that the measures are not equivalent in the cognitive-linguistic processes they assess. A statistical synthesis is therefore not provided.

1.6 The CTT and IRT perspectives

A first extraction of data from the review set showed that most the studies had followed the CTT perspective. IRT is an alternative to the CTT. IRT analysis allows for an estimate of the child's ability and thereby confirms that the measure is reliably discriminating between different levels of skills (e.g. lower and higher attainments). A probabilistic model is used (the Rasch method) to evaluate the likelihood of individual items in a test capturing a child's performance as a function of stated characteristics of the item (e.g. its complexity level) and characteristics of an individual child taking the test (e.g. skill level). This review does not provide a separate synthesis on the robustness of tests using IRT analyses since very few of the studies examined in this review used this approach.[3]

The consistency of a test is called its reliability. Here again the overwhelming presence of the CTT perspective is evident, with the following estimates of reliability most common in the review set: Stability of performance over repeated testing is offered as evidence of test-re-test reliability. High inter-rater reliability is evidence of consistency in decisions made by two assessors about the child's performance on a task. A third estimate is based on the internal consistency of items in a test.

The validity of a test is demonstrated by accumulating evidence from multiple sources. Within the CTT perspective, four sources of validity are discussed. **Content validity** is related to items of a measure being seen as equivalent to other material intended for the same or related skill. There are two common ways to establish the content validity of literacy measures: seeking equivalence with grade-level textbooks or with content that appears in school exams. **Concurrent validity** is inferred when there is a strong and positive correlation between a child's performance on the index test and another test that examines approximately similar cognitive-linguistic processes. **Convergent-discriminate validity** refers to a pattern of correlations where the index test shows a stronger correlation with theoretically related measures relative to measures that are known to have only distant connections. **Predictive validity** of a test is inferred when there is a strong association between concurrent performance on the task and another theoretically related construct or between earlier and later performance on the task or a theoretically related construct.

An important target within this review has been to establish the validity of a test for the following three inferences: (i) What can the child do with regard to reading and writing? (ii) What does the child need in order to read and write well? and (iii) What does the teacher need to know in order to support children's reading and writing?

[3] An example of the use of IRT is for the measures reported under the Young Lives Project, i.e. Crookston *et al.* (2014).

1.7 Structure of the report

The next two sections give a summary of each area of assessment of written and spoken language skills. The synthesis includes an overview, a map[4,5] and a summary graph of cohort characteristics (i.e. grade levels assessed, socioeconomic characteristics, and information on whether the assessment was in the child's home language). Each section gives an overview of why it is useful to assess a particular construct, how the construct has been assessed, and the innovations and challenges in the assessment of this area. The sections focus on only those psychometric, administrative and contextual properties of tests that have been reported to a sufficient level of detail. As mentioned in footnote 1, the list of measures is given in Annex A. Annex B lists the number of studies reporting: a reliability of above .80 (and numbers with a lower estimate), procurement details (free-to-use tests, commercial tests and researcher-developed tests), mode of data gathering (individual or group testing; performance, reported or observed data) and contextualisation (pilots and localisation effort).

The final sections address lessons learnt, gaps in evidence and future directions for the field of assessment of literacy and foundational learning in developing countries.

[4] Maps in this report give an indication of the geographical spread of measures reviewed in each area of assessment. When the number of measures from a particular country is between 1 and 3 this is indicated with a small flag, between 4 and 6 studies by a medium-sized flag, 6 and 12 studies by a larger flag, and 13 or more studies by a star. Note that a given study may have between 1 and 10 measures in a particular area (e.g. multiple linguistic units of assessment under phonological processing, fluency measures for multiple types of lists and texts, etc.).

[5] All maps are produced using the 1-Page Mapmaker from National Geographic Education (http://nationalgeographic.org/education/mapping/outline-map/).

2 Assessment of written language skills

2.1 Emergent literacy

Why is it useful to assess this construct?

Well before children 'read' and 'write' they already demonstrate print awareness in their rudimentary understanding of how written language connects with spoken language and represents meaning, and the way they handle literacy artefacts (see Clay, 2000 for details). During these early years, children's emergent writing demonstrates their concepts about the outward form of print and how symbols represent sounds and meaning. The international evidence base on emergent writing is smaller than on reading but steadily growing (see Treiman and Kessler, 2014 for an overview). Beyond awareness of print and writing, the construct of emergent literacy includes oral language proficiency and attitudes toward reading.

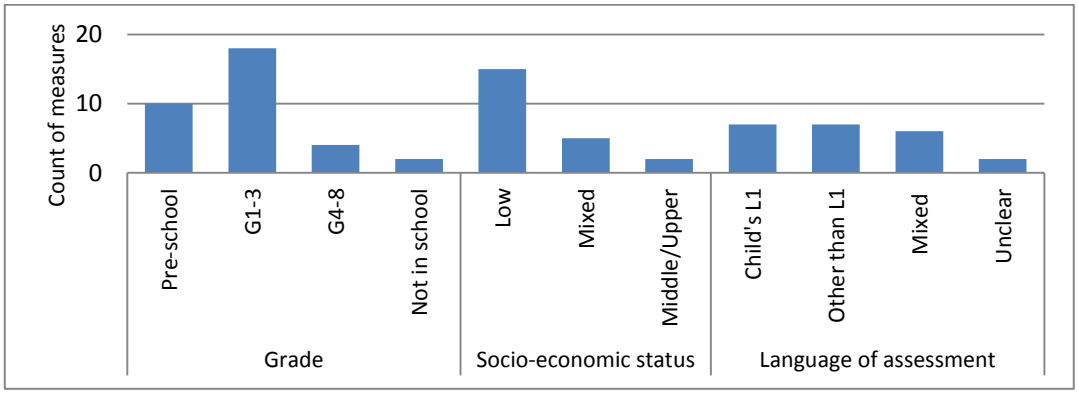
Children who come to the task of literacy learning with higher levels of emergent literacy do better. Assessment of emergent literacy therefore allows for an estimate of the foundational skills available to a child for literacy learning.

What is available and how is this area assessed?

We reviewed 22 measures from 15 studies conducted in 10 countries. The studies were predominantly in low-income and bi- or multilingual contexts. The locations are shown in the adjacent map (legend in footnote 4) and the cohort characteristics are given in Figure 1 below.

This section focuses on measures of print awareness and emergent writing. Oral language measures are reviewed in Section 3. We did not find measures on attitudes for these early years.

Figure 1: Number of measures shown by cohort characteristics (total measures = 22)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

A direct and practical approach to assessing print awareness is using a CAP assessment. A book is brought to the child and the child's concepts about print become evident from responses to questions such as '*Show me the front of the book*' and *'I will read this book. Show me where to read*.' Sixteen of the 22 measures are adaptations of Clay's (2000) measure. While the 16

measures have a common core, they do not cover all components of the original task, which leads to differences in what they each assess.[6] Other measures include:

- Emergent writing tasks (e.g. Chile: Strasser and Lissi, 2009; India: Vagh, 2009). In this task, children are variously asked to write their own name, orthographic symbols and common words.
- The moving word task, where a word – as if by accident – is moved to sit under the wrong picture and a note is made if the child recognises the picture–label mismatch (Morocco: Rochdi, 2010).
- 'Readiness' composites covering multiple domains, including concept knowledge and perceptual skills (e.g. Philippines: Ocampo, 1996).

What are the innovations and challenges?

CAP assessment. There is a small body of evidence for the potential of CAP assessments at scale (e.g. Nepal: Pinto, 2010; Zambia: Friedlander *et al.*, 2014; Zimbabwe: Chinyama *et al.*, 2012). The task is administered one child at a time, making it a time- and resource-intensive task. Given the potential of the test, two points need consideration:

- CAP tests cover multiple components of print experience and test results can directly translate into priorities in an early years teaching programme. A CAP measure may, however, not capture variability within a narrow SES band, particularly in contexts of high poverty. Because of this, CAP tasks may contribute more within a teacher toolkit than in large-scale surveys for monitoring educational quality.
- A well-illustrated book with simple text has become the tool of choice for CAP tests. Use of printed artefacts from the neighbourhood of the child is an appealing but understudied alternative. Adding variety with local printed materials to a teacher's toolkit, for example, broadens the focus from shop-bought (or supplied) books to contextually embedded material. To build on innovations with print readily available in the environment, evidence has to be built at two levels: Are the data from such material more reliable? and; Does the test capture greater variability than a book task? The latter would be of particular interest for research purposes.

Emergent writing tasks. A small body of evidence shows that in environments with low exposure to print and limited instruction, emergent writing tasks capture greater individual differences compared to CAP tasks. Emergent writing tasks may be done in small groups and hence are less time- and resource-intensive.

- To exploit the full potential of the task, clarity is needed for the type of item – free form, symbols, own name, common words, open-ended writing – that is most sensitive to individual differences and is portable across contexts. Writing one's own name has face validity and is culturally universal but makes for unfair comparison across children (consider these names: Ali, Rana, Yonas, Nesrin, Meenakshi, Lyabwene). Single-study evidence suggests that symbol writing may have greater reliability, particularly in the first months into school instruction, and that distributional properties of scores on both symbol writing and word writing improve as instruction effects begin to show (Cronbach's alpha of .98 for Hindi symbols; Vagh, 2009; and.77 at the start and.81 six months later in the school year for Spanish words; Strasser and Lissi, 2009). In other words, once formal literacy instruction begins, some children pull ahead of

[6] Concepts assessed include parts of a book, meaning conveyed through print, the direction of print, tracking text, punctuation, concepts about symbol units and words, and book handling. Our item analysis of the CAP measures is based on the *Books Inside Out* subtest from the Assessment of Foundation Learning (see Nag, 2013).

other children in both symbol and word writing, and these individual differences are captured by the test.

- Scoring schemes need attention. Both dichotomous scoring (1 for correct, 0 for errors) as well as a scale for approximations (0 = no output or pseudo-letters, 5 = perfect rendition) have been used. A challenge in both schemes is achieving high inter-rater reliability.
- Tablet-based assessment (and computer-based assessment) is gaining popularity. The current state of the science is some distance away from use of this medium for assessment of emergent writing. A key issue is the availability of handwriting recognition software that functions within the constraints of the tablets typically used in large-scale surveys.

Emergent orthographic knowledge. This cluster of tasks uses pseudo-print to assess children's orthographic knowledge. These tasks use a multiple-choice format where children have to select a target symbol, symbol string or word from a set of distracters. The distracters may be visually close, phonologically close, or completely unrelated items including pseudo-symbols and pseudo-words. These tests have been used across orthographies: Hindi (Vagh, 2009; Cronbach's alpha reliability was .93 for symbol choice and .85 for the word choice tasks), Arabic (Rochdi, 2010, reliability information not available), and CiNyanja (Jere-Folotiya *et al*., 2014; test–retest reliability of.50). In addition, these tests show promise for small group administration.

2.2 Symbol knowledge

The symbol units of interest to this review are the *akshara*, *fidel* and letter.[7,8] *Akshara* and *fidel* units map on to syllables (e.g. /ka/, /ki/, ku/) and when *akshara* are in strings, they may also map on to values smaller than a syllable (e.g. to represent the initial and final sound values in 'rain' and 'sun'). Letters represent phonemic values, and letter clusters represent other sound values (e.g. the phonemic values of /s/-/ʌ/-/n/ in 'sun' but the letter cluster 'ai' in 'rain' to represent /eɪ/).

Why is it useful to assess this construct?

Assessing symbol knowledge is useful because of the association between symbol knowledge and attainments in literacy. The evidence is currently dominated by research on letter knowledge, and this shows that letter knowledge in preschool and Grade 1 is a predictor of lower- and higher-order processes such as eye movements during reading and accuracy during word reading (for a review, see Grainger *et al.*, 2016). To confirm that these trends apply to other languages and symbol systems, we examined the association between symbol knowledge and reading accuracy (data available from 12/23 studies). Irrespective of language and orthography, a moderate to strong correlation is typically found between symbol knowledge and reading accuracy. An association between symbol knowledge and word reading efficiency into the middle school years is evident when the assessment of symbol and word learning uses uncommon symbols and more complex words (i.e. test items go beyond common symbols and common words).

Growing symbol knowledge is characterised by shifts toward greater efficiency and these shifts are at several levels. Three of these are discussed here. First, when individual symbols have many visual details or represent more than one sound, children move from treating symbols as global wholes to becoming analytic about component parts (e.g. phases in *akshara* learning and

[7] An important fourth symbol set is the character of the Chinese orthographies. We did not find any paper in the review set that assessed character knowledge exclusively, although character knowledge is found as part of a school readiness composite in one study (Rao *et al.*, 2012).

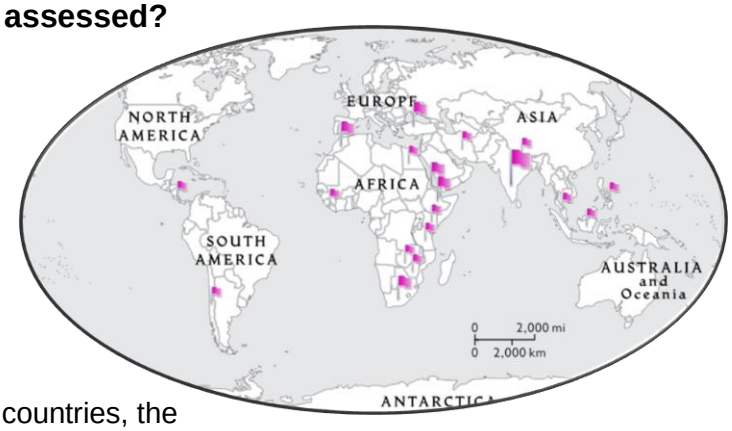
[8] Example of languages using the *akshara* are Nepali, Bengali, Hindi, Gujarati, Tamil, Thai and Kannada. Examples of languages using the *fidel* are Tigrinya and Tigre. Examples of languages using Latin-based letters are English, Spanish, Bahasa Indonesian, Malay, Swahili and CiNyanja.

disambiguation in Arabic symbols). Second, symbols and symbol clusters are learnt at a faster pace when they are encountered more frequently (e.g. compare 'r' and 'th' to 'w' and 'lm' in English). Third, symbol–sound mapping moves from singleton units to larger orthographic chunks (e.g. multi-letter representations like the blends 'sw' and 'pl' and word endings such as '-tion', '-sion' and '-cion' in English). Thus, signs of a maturing orthographic processing system include increased accuracy for a wider range of symbols and symbol combinations, and a facility with mapping sounds beyond singleton symbols. Together these orthographic competencies converge to provide word-level information.

The number of symbols to be learnt differs by writing system. The number of symbols in the *fidel* and *akshara* systems is far larger, with instruction planned across the primary school years. The small set of letters is typically taught by the end of the first year of instruction. Thus, while it is clear that symbol knowledge is useful to assess, the contexts of symbol learning decide which symbols give the most meaningful information about educational quality and for teaching practice.

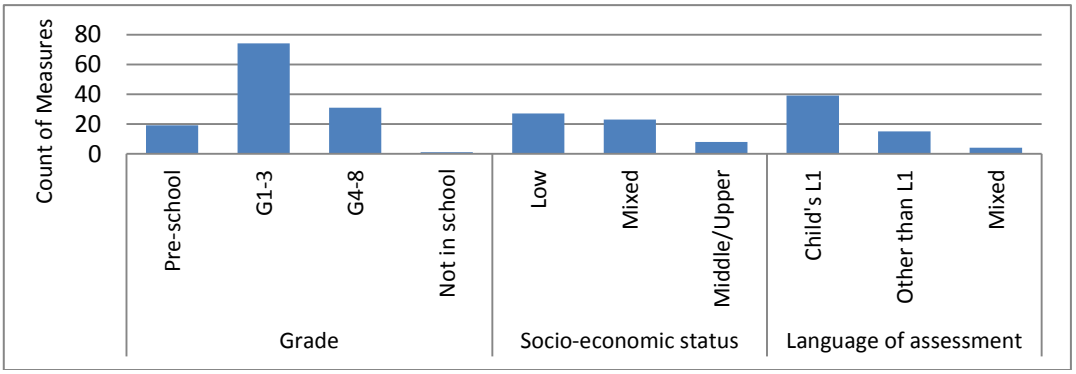
What is available and how is this area assessed?

Our review found 58 measures of symbol knowledge from 30 studies, covering *akshara*, *fidel*, Arabic letters and Latin-based letters[9] (the adjacent map gives the countries covered; legend in footnote 4). Measures typically assessed children in Grades 1 to 3, in low or mixed income surveys and university research projects.



Because of the linguistic landscape of some countries, the assessment was not in the home language for a sub-sample (e.g. in the Philippines, India, Nepal, Zambia and Iran). The composition of the cohort is given in Figure 2.

Figure 2: Number of measures shown by cohort characteristics (total measures = 58)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories

Second language assessment is mainly in English (18 measures, 10 studies, eight countries), covering:

- Knowledge of letter names, letter sounds, or allowing responses of either;

[9] A further six studies included symbol knowledge as the final items in tests assessing concepts about print (e.g. India: Kalia, 2009) and the initial items in tests assessing word reading accuracy (e.g. Costa Rica: Rolla San Francisco *et al.*, 2006; Tanzania: Alcock *et al*., 2000) or spelling accuracy (e.g. Zambia: by Ojanen *et al*. 2013, reported in Jere-Folotiya *et al*., 2014).

- Letter-naming fluency or the number of letters read correctly in one minute;
- Symbol recognition within visually confusable symbol pairs where the distracter is either a visually close symbol or a pseudo-symbol; and
- Letters-in-context either as a letter–word association task or a letter usage task with the instruction: '*say a word that starts with letter __'*.

Of these, letter naming is by far the most popular, closely followed by letter sound. The common item is a singleton letter; inclusion of blends and letter strings is rare. Letters are typically presented in the lower case, although some tests only used the upper case letter and others include both. The dominance of single letters appears to be linked to a narrow definition of *knowledge about the alphabet as taught in the first year of reading instruction*. Such a focus ignores the growth of symbol knowledge that is linked with multi-letter representations.

The legacy of this alphabetic testing tradition is evident in assessments across developing countries, particularly in Latinised orthographies such as Bahasa Indonesia (Indonesia) and Kunama (Eritrea). Assessment of *akshara* (and to some extent the *fidel*) is more nuanced, with lists covering simple-to-complex and more-to-less frequently used symbols (India: Nag, 2007; Eritrea: Asfaha *et al*., 2009), although some initiatives (e.g. EGRA and ASER) apply the Latin-based logic and limit assessment to symbols taught in the first year of school. Other measures are:

- Symbol usage: a simpler version of the letter–sound task with potential to show better distributions at the earliest stages of literacy acquisition (Turkish: Oktay and Aktan, 2002).
- Symbol writing fluency: the task to '*write as many as you know'* shows stable correlation across the preschool year with name–sound knowledge (Hindi: Vagh, 2009).
- Visual form recognition: allows for group administration, is sensitive to differences across the attainments continuum (CiNyanja: Jere-Folotiya *et al*., 2014, Arabic: Elbeheri and Everett, 2007; Tahan *et al.*, 2011), and shows promise with Latin-based symbols in contexts of exceptionally low exposure to print (Tanzania: Alcock *et al*., 2000).
- Mixed symbols: the task covers early-to-later learnt symbols and can potentially include multi-letter strings (Kannada: Nag, 2007).

What are the innovations and challenges?

Poor variability of scores and therefore an inability to predict individual differences is an important issue when designing measures of symbol knowledge. Based on the languages and contexts in the review, the following reasons emerge for skew in the data:

- Characteristics of the writing system. When the symbol set is small (e.g. the 26 letters of English), the test does not pick individual differences after Grade 2 because children already have mastery on all test items. When the symbol set is large (e.g. the 700+ symbols of Kannada) then symbol knowledge tasks register individual differences into middle school as long as there are test items with different frequency of occurrence in children's texts.
- Type of measure. Symbol measures behave differently across orthographies because of the nature of sound–symbol linkages and the size of the symbol set.[10] In name–sound tasks, performance in transparent languages with contained symbol sets shows higher variability in

[10] Orthographies may be characterised along the dimensions of 'orthographic depth' and 'orthographic breadth'. Orthographic depth refers to the extent to which a language has a predictable linkage between sounds and symbols; those that are predictable are called shallow orthographies (also called transparent, consistent), and those with many ambiguous linkages are deep orthographies (opaque, inconsistent). Orthographic breadth refers to the size of the symbol set; those with a small number of symbols are called contained orthographies and those with many symbols are called extensive orthographies.

the early rather than the later grades. The pattern reverses in transparent languages that have extensive symbol sets (for a brief definition of these technical terms, see footnote 10). Visual discrimination measures show promise at two levels: among the visually simple symbol sets, the measure appears to be sensitive at the symbol learning phase of literacy acquisition and for the visually complex sets the measure becomes sensitive in the later stages of mastery. There is limited evidence for symbol fluency and in-context measures.

- Instruction effect. It appears that when letter names are explicitly taught almost all children learn these and since the symbol task limits item selection to within this curriculum, the level of attainment across the group becomes indistinguishable. In parallel, because sounding out is often ignored in instruction (see Nag *et al.*, 2016), almost all children fail on the letter–sound task. However, when teaching has not homogenised group performance, the letter–sound task is more effective than the letter–name task at capturing variations in attainment. In languages where the name–sound of symbols is one, teaching may prioritise teaching of some symbols and this again homogenises performance across the group.
- Learning environment. A predictor of individual differences in symbol knowledge is print experience at home and instruction in school (e.g. India: Sen and Blatchford, 2001). The task becomes particularly confusing for simultaneous biliterates learning to read in two languages that share a symbol set (e.g. Filipino–English or Swahili–English). In short, if the group of children being assessed gain experience from more-or-less similar learning environments then all children manifest similar advantage (or disadvantage) and by extension show similar profiles on symbol tasks.

Given such homogenising effects, researchers have turned to adding information from symbol tests to other tests (typically CAP or reading accuracy) to improve distribution of scores. Another alternative is to relax item selection and move beyond the name–sound identity of singletons to include clusters and affixes for all orthographies and for extensive orthographies include simple, complex, common and uncommon symbols.

2.3 Reading accuracy

Why is it useful to assess this construct?

A critical component skill for reading comprehension is accurate recognition of individual words. A word may be read using at least one of two approaches (the 'two routes to reading'). One uses the lexical-semantic route to match the written word with a word known to the child and thus already available to the child. The second is the phonological route, where the child systematically decodes the sound sequence to identify the word. The lexical-semantic route draws upon the stored meanings of words (the semantic lexicon) while the phonological route draws upon stored sounds (the phonological lexicon). Words with ambiguous sound–symbol linkage (i.e. irregular words such as 'knee') will require the lexical-semantic route while words with no ambiguity in sound–symbol mapping (i.e. regular words such as 'tree') can use either route. Languages differ in the number of words with ambiguous sound–symbol linkages. English has several irregular words[11] and languages like Turkish, Kannada and Swahili have only a few and hence are called consistent languages.

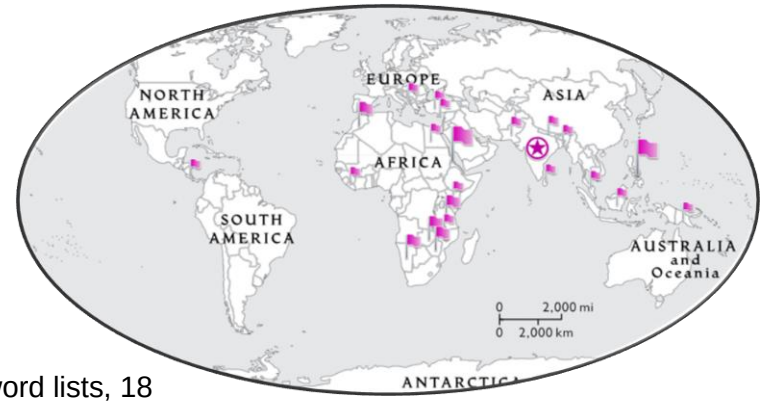
Reading accuracy and reading comprehension show significant association both among monolingual readers as well as bilingual and biliterate readers. There are differences across languages in the pace of development from being a novice to an expert decoder, with children gaining mastery quickly in the consistent languages because letter–sound linkages are simple. In

[11] For a comprehensive discussion on word reading development in English see Stuart and Stainthorp (2016).

addition, decoding skills are particularly sensitive to opportunity for practice. The quality of instruction in school and contextual factors such as parents' SES and home literacy environment predict individual differences in decoding attainments as well as the rate of growth in decoding skills.

Given the critical nature of single word decoding for literacy learning and its sensitivity to instruction and opportunity, an assessment of reading accuracy is informative for monitoring educational quality as well as to inform teaching practice.

What is available and how is this area assessed?



We reviewed 90 measures from 37 studies covering 23 countries and a wide range of languages[12] (for details see map; footnote 4 gives the legend for the map). The composition of the cohort is given in Figure 3.

Fifty-one measures used word lists and 25 non-word lists, 18 used connected texts and six mixed two or more of the above item types.

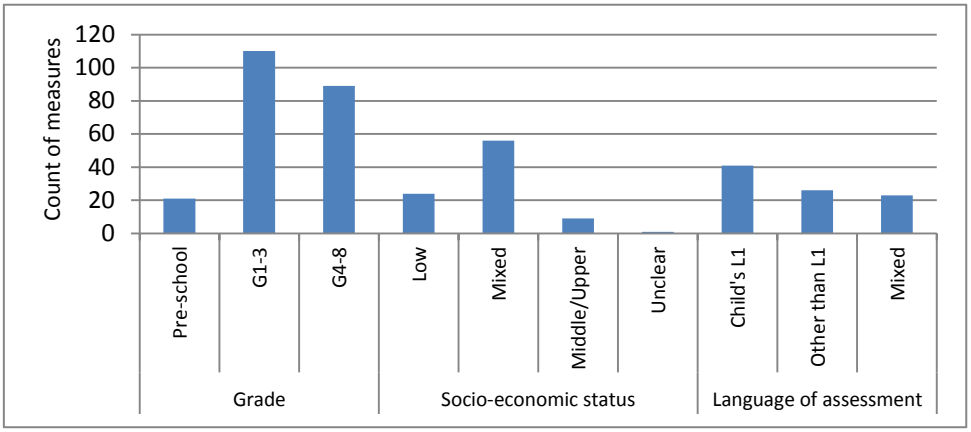
The connected texts are at the length of phrases (Kannada: Ramchandra and Karanth, 2007), sentences (Bengali: Chowdhury *et al.*, 1994) and up to 30-word narratives (Turkish: Babayigit and Stainthorp, 2010).

Two departures from the use of a list or connected text format are:

- a lexical judgement task (Kiswahili, Tanzania: Alcock *et al*., 2000; Kenya: Jukes *et al.*, 2006), which requires recognition of a target from a set of distracters; and
- the word chain task, where the child must mark word boundaries in a continuously printed word chain (Arabic, Egypt: Elbeheri and Everett, 2007; Kannada, Telugu and English, India: Nakamura, 2014).

[12] Languages covered are Albanian, Arabic, Bahasa Indonesia, Bemba, Bengali, Chichewa, English, Filipino, Herero, Kannada, Kiswahili, Kunama, Malay, Nepali, Oriya, Saho, Shona, Sinhala, Spanish, Tamil, Telugu, Tigre, Tigrinya, Turkish and Urdu.

Figure 3: Number of measures shown by cohort characteristics (total measures = 90)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

Longer tests appear more often in research studies. There is greater variability in the length of word lists compared to non-word lists (10 to 200 items vs. 10 to 54 items). The connected texts, when all phrase/sentence/narrative items in the measures are calculated together, range from 30 to 50 words for the earlier grades and between 74 and 160 for Grade 3 and above. From the available information, the shortest test with a reliability estimate of .95 has 20 and 30 words for Grades 1 and 2 respectively (Bahasa Indonesia: Winskel and Widjaja, 2007).

What are the innovations and challenges?

Item selection for word lists. Word lists are typically constructed to fit the curriculum (i.e. are criterion referenced). Though some studies simply include words because they appear in a standardised test developed elsewhere, there are other selection criteria in use:

- words taken from primers (early grades: Kiswahili: Alcock *et al*., 2000; Bahasa Indonesia: Winskel and Widjaja, 2007; later grades: Herero: Veii and Everett, 2005);
- words randomly selected from a dictionary list of all words in the textbook or a selection of children's literature (e.g. Eritrean languages: Asfaha *et al*., 2009; Kannada: Nag, 2007);
- words with the highest token frequency within a textbook (e.g. Most Used Words (MUW) in Zambia: Friedlander *et al*., 2014; and Zimbabwe: Chinyama *et al*., 2012);
- words that reflect the psycholinguistic properties of a language (e.g. words with digraph accents and/or stop sounds in Filipino: Ledesma, 2002; affixes in Malay: Lee and Wheldall, 2011); and
- words that reflect the orthographic properties of the written language (e.g. two- and three-letter words in Turkish: Oktay and Aktan, 2002; four- and five-letter words in Albanian: Hoxhallari *et al.*, 2004; words with 'joint symbols' in Bengali: Chowdhury *et al*., 1994; words with various symbol types in Kannada: Nag, 2007).

A direct comparison of word lists developed based on different criteria is not available but distributional properties of scores appear to be better, especially for use with Grades 3 and above, when item selection is based on psycholinguistic and orthographic characteristics or is a random selection from a dictionary list. These measures need evaluation for robustness across languages.

Distributional properties of scores. Word recognition assessments will show language- and orthography-specific variations, and this general trend is confirmed for the review set (for a brief definition of the technical terms used here, see footnote 10):

- There is a systematic difference in the distributions of test scores in consistent and inconsistent languages. Good distributions well into middle school are found with irregularly spelled languages such as English while performance reaches a ceiling within the initial school years for regularly spelled languages such as Spanish. In such instances, error analysis or speed measures capture the variations in individual attainments (e.g. demonstrated in Turkish).
- Among transparent orthographies there is a systematic difference when the symbol register varies in size (e.g. Turkish has few symbols but Bengali has many). Word lists in the transparent but extensive orthographies register good score distributions when words are selected for frequency of the symbols they contain – words with low-frequency symbols pick out individual differences well into middle school (on Kannada, see Nag, 2007).

Reporting of psychometric properties. Researcher-developed measures form the bulk of the reading accuracy dataset (60/90), reflecting the popularity of decoding assessment in literacy research in developing countries. Despite the active research interest, reporting of psychometric details is very poor (e.g. no information was provided on reliability for 64 of 90 measures, while reporting on collecting evidence for the validity argument is rare). Where information is available:

- Reliability estimates are typically excellent (above .9). These estimates are based on Cronbach's alpha and split half reliability, or test–retest reliability.
- Scores vary in the expected direction across grades and age bands.
- Concurrent validity can be inferred for some measures (15 measures report associations with another measure, while seven report with two other measures).
- Convergent–divergent validity is inferred from 22 measures that report correlations in the expected direction with three or more measures of interest.

In summary, while several measures in this dataset appear to be of robust psychometric quality, reporting standards have been compromised in a majority of the publications.

Words are sufficient. Non-word testing has gained popularity, particularly in EGRA-style test batteries. This section will consider the rationale for why assessment of reading accuracy with words is sufficient to monitor educational quality and for making pedagogical decisions.

- Word lists assess the use of both semantic-lexical and phonological-decoding routes to word recognition while non-words, by the very nature of their absence from the language, focus on only the phonological-decoding route. Each can contribute valuable information but their use must depend on the purpose of assessment. Testing with non-words has been useful for diagnostic assessment when there are concerns about literacy learning difficulties in a child. Non-word-based tests have been particularly useful in psycholinguistic research to uncover the cognitive-linguistic underpinnings of the decoding process. Such a level of detail is arguably useful for clinical practice but not necessary for the twin pedagogical goals of monitoring quality and informing practice.
- A direct comparison of performance on word and non-word accuracy measures is available from five studies and supports the conclusion that the non-word reading task does not add meaningful and pedagogically useful information:
    - children achieve closely similar ranks within a group whether they read a word list or a non-word list. The correlations are moderate to high in the home language (Bahasa Indonesia: .89; Turkish: .92) and among biliterates reading in two languages (Oriya–, Herero–, Filipino– and English: .61 to .92).
    - for both measures, association with other decoding-focused component skills of literacy (e.g. symbol knowledge and spelling) is between .73 and .88 across all languages.

  - predictors of word and non-word reading are similar across languages (Oriya–English, Filipino–English) although in some languages, as expected in light of theory, greater weight of explanation for individual differences in non-word reading comes from phonological skills or speed of processing (Bahasa Indonesia, Turkish).
- Word lists have the advantage of face validity and can be shown to have criterion validity against school textbooks and the stated curriculum. By contrast, interpretation of test results from a non-word test is not immediately clear.
- In second (and third) language contexts some items in a word list will appear as non-words because they are not as yet known to the child; these are words that are neither in the child's vocabulary nor orthographic lexicon ('sight vocabulary'). In such instances, children will draw upon a phonological-decoding approach to read the word. Even so, the experience with word lists is fundamentally different from decoding of non-words. Word lists have the advantage of prompting new vocabulary learning.
- Developing non-words is a specialist task requiring items to reflect the phonology of the target language. By contrast, developing word lists is simpler.
- Non-word tests can unwittingly suggest that instruction time must be taken away from meaning-focused activities to practice phonological decoding with non-words.

Taken together, a word reading test is a simple, direct and sufficient way to monitor educational quality and inform teaching practice. A well-structured word list can support several direct inferences from the test results: what a child can do in reading single words, what a child needs in order to do read well, and how well a teacher is teaching reading.

2.4 Spelling

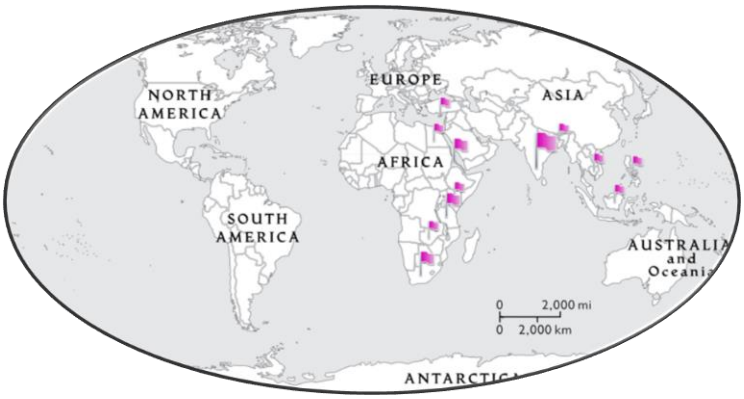
Why is it useful to assess this construct?

Spelling is the skill of writing words accurately. Foundational skills for spelling include knowledge about individual symbols and their sounds, knowledge about the rules of sound-to-symbol mapping in the language and, when writing down, skills in the mechanics of writing (transcription). Taken together, a spelling task may be seen as assessing the child's decoding competence.

There is a high correlation between spelling accuracy and reading accuracy because both skills are directly linked to decoding competence. That said, spelling tends to be harder than reading. One reason for this is that reading a word allows for guessing from context; children can guess at the word by using pictures on the page or the sentence in which the word appears, or even some of the symbols in the word. In spelling, each appropriate symbol has to be recalled. The dissociation between spelling and reading is particularly distinct in languages where there is more transparency in the symbol–sound linkage compared to sound–symbol linkage.

The pace of spelling development depends on the nature of the orthography and the transparency of sound–symbol associations. More consistent or transparent languages allow for a faster pace of spelling development because children quickly gain insight into the mapping of sounds and symbols. Opaque languages take longer. In languages with many long words that are also multi-morphemic (e.g. compound words or words with inflections) knowledge about the morphological structure of words also help in spelling development (e.g. in Turkish). Contextual factors such as specific attributes of the home literacy environment also influence spelling development.

What is available and how is this area assessed?



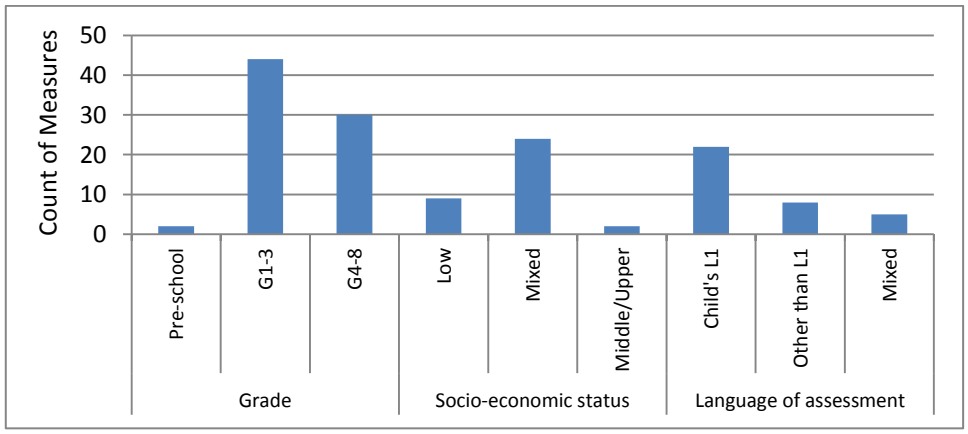
Thirty-five measures[13] from 17 studies conducted in 12 countries comprise this set (legend for the adjacent map in footnote 4, with the cohort details given in Figure 4 below).

Item characteristics are as follows:

- words selected from children's textbooks;
- words selected for their morphological and phonological characteristics;
- phoneme and syllable units common in the language, and non-words;
- words embedded in sentences (fill in the blanks) and words in dictated sentences; and
- personally meaningful words ('write your name', 'write your village name', 'write a thank you letter to your teacher').

Item lists are between three and 50 words and the lengths of connected texts range from two to 10 words.

Figure 4: Number of measures shown by cohort characteristics (total measures = 35)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

Spelling measures show some of the highest reliability indices in the review. Details are presented by the size of the symbol set:

Contained orthographies (fewer symbols to learn)

- Latin-based languages (16 measures, L1). All tests in this set are transparent (for a list of languages, see footnote 13). Spelling accuracy on dictated words is the most common measure (10 measures), with a smaller set using non-words (1), connected sentences (1), free writing samples (2) and a mix of words and non-words (3). One study required children to pick

[13] In our review set, the Latin-based orthographies are transparent (Bahasa Indonesia, CiNyanja, Filipino, Kiswahili, Kunama, Turkish, Saho and Zulu), as are the *fidel*-based languages (Tigre, Tigrinya). Some *akshara*-based languages are more transparent than others (more transparent: Kannada, Telugu; less transparent: Bengali, Khmer). Similarly, there are variations within Arabic (transparent: vowelised, opaque: unvowelised). Assessment of second language spelling is in the Latin-based and opaque language, English.

out the correct spelling from a set of four words. Estimates of internal consistency on word spelling tests (using alpha coefficients, split half) are typically above .90 (four measures) with one instance of an estimate of .76. Inter-rater reliability for performance on connected text is high in the single study using this format (.99). Test–retest reliability is moderate to high (one week, Kiswahili, Grades 2–5: .62; two weeks, CiNyanja, Grade 1 = .82, greater than two weeks (time unspecified) = .65).

- Arabic languages (two measures, L1). One measures accuracy with the vowelised form of words dictated along with a sentence to give context (test–retest, Egypt, Grades 1–3: 'grapheme accuracy' = .92). The second measure assesses accuracy with the unvowelised form on all words of a dictated passage (reliability estimates not available).
- English (seven measures, L2). These assessments are with biliterate children (Filipino–English, Zulu–English, multiple Indian languages–English). Estimates of internal reliability (alpha coefficients, split half) range between .71 and .91.

Extensive orthographies (more symbols to learn)

- *Akshara*-based languages (nine measures, L1). This set of measures cover different levels of orthographic transparency (see footnote 10). Word tests are common (seven), with one measure each using non-words and a mix of words and non-words. Estimates of internal consistency (using Cronbach's alpha) for four measures are above .90.

- *Fidel*-based languages (two measures, L1). Both measures assess accuracy in word spelling and report internal consistency of tests as greater than .90.

What are the innovations and challenges?

Removing the writing component in spelling. In the early grades, the transcription component of a spelling dictation task can be substantial, particularly in the visually complex Arabic and some *akshara* orthographies. Two innovations to reduce the transcription component in the assessment of spelling are available:

- Children sequence symbol cards to show the spelling of a word (Khmer, Grade 1, Cambodia, Nonoyama-Tarumi and Brendenberg, 2009). Reliability information is not available.
- Children identify the correct spelling in a multiple-choice format (CiNyanja, Grade 1, Zambia, test by Ojanen *et al*. 2013, reported in Jere-Folotoya *et al*. 2014). Test–retest reliability with a two-week lag = .82 and for greater than two weeks (time unspecified) = .65.

Both innovations need evaluation for usability and robustness across languages.

Readiness for scale-up. It is surprising that spelling tests have not found a place in large-scale surveys but perhaps this is because there has been a focus on reading rather than all aspects of literacy learning. Spelling measures show high reliability across languages, writing systems, with first and second language learners and across the school years. Spelling measures are relatively easy to administer and inter-rater reliability is high on the task. Findings from a spelling test are easy to interpret because a focus on spelling is already common in many schools (Nag *et al*., 2016). There is a high correlation between spelling skills and reading accuracy so both could potentially give information about the quality of literacy instruction. However, an advantage with spelling tests is that they readily lend themselves to group administration. The available evidence suggests that spelling assessment is ready to go to scale for the twin purposes of monitoring educational quality as well as in a toolkit for teachers.

2.5 Reading fluency

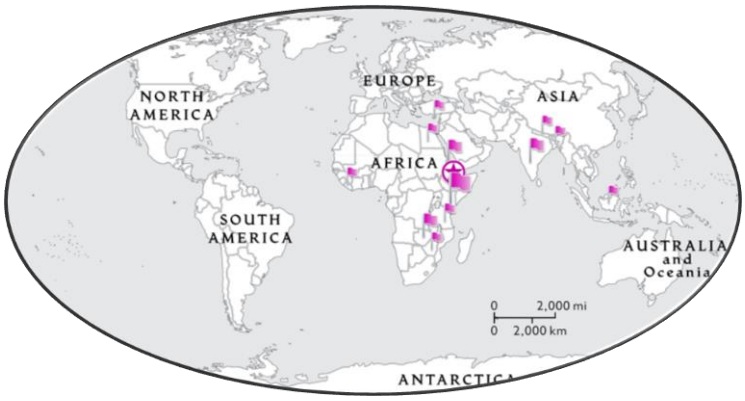
Why is it useful to assess this construct?

Reading fluency is the ability to accurately read connected text at a speed akin to a conversational rate along with appropriate expression and intonation. The sub-components of the construct are therefore speed, accuracy and prosody, although this last nuance of reading is often lost in reading fluency assessment. Higher speed and accuracy suggest automaticity in word-level decoding and signals that limited attentional resources are available to a reader for reading comprehension processes. Prosody is seen as mirroring one's understanding of what is being read (i.e. you cannot read with expression if you do not understand what you have read) but also as helping to understand what is being read (i.e. reading with appropriate emphasis allows the narrative to become clearer).

Reading fluency and reading comprehension show strong associations across the school years. There is robust evidence for reading fluency being a predictor of reading comprehension across multiple languages, with the dominance of reading fluency as a predictor reducing somewhat in the later school years, where spoken language skills become more dominant. Perhaps because of its strong and continuous association, reading fluency has come to be a key component in the assessment of reading development at scale (e.g. Dubeck and Gove, 2015).

What is available and how is this area assessed?

We review 52 measures from 16 studies conducted in 14 countries[14] (legend to adjacent map in footnote 4, with cohort details in Figure 5 below). Seventeen use word stimuli, 10 non-word stimuli and the rest connected text. Details of the measures are as follows:



- Reading fluency is typically assessed on performance within a one-minute time window, but there are instances of three- and five-minute time windows.
- Word lists for one-minute tests range from 50 to 136 words. The length appears to be dictated by language (some languages have a bigger proportion of longer words and thus shorter lists) and grade of children being assessed (younger children have shorter lists).
- Word lists typically contain simple words, with rare instances of multi-morphemic words (e.g. Turkish, an agglutinating language: Babayigit and Stainthorp, 2010).
- When items are passages, word lengths vary substantially: measures for Grades 1 to 3 comprise 14 to 154 words, the most common measure being either a 30-word length or a 60-word length. The largest variability in passage length is seen in assessment of children in Grade 3 level. Measures for Grades 4 and 5 generally comprise passages of between 14 and 26 words, with one measure spanning two pages and another consisting of 125 sentences. A similar pattern recurs in tests for the older grades.
- Some connected text is presented with accompanying pictures as would be seen naturally in an illustrated book (Bengali: Johnson, 2003). Typically, however, the connected text is simply a printed passage.

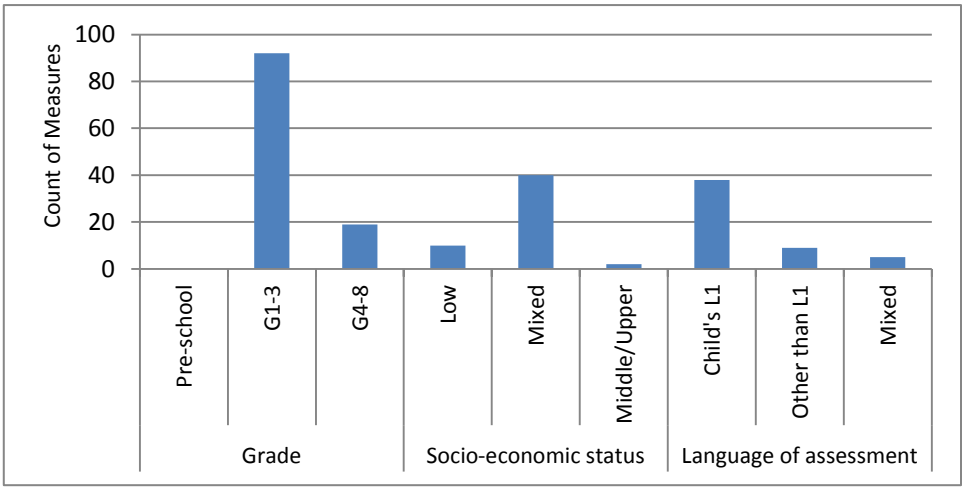
[14] Languages covered are Afan Oromo, Amharic, Arabic, Bengali, Bemba, English, Hararigna, Kannada, Kiswahili, Kunama Lozi, Mbunda, Nepali, Saho, Shona, Sidaamu Afoo, Somaligna, Telugu, Tigre, Tigrinya, and Turkish.

- Most measures focus on quantifying fluency through a 'words-per-minute' or equivalent measure. Some studies use an impressionistic approach asking raters to assess if children read 'smoothly' or 'haltingly'.

Reports of robustness of the reading fluency measures are as follows:

- On word-reading fluency, internal consistency estimates from eight measures range from .70 to .96. Test–retest reliability estimates are between .75 (Northern Cyprus, Turkish, Grade 1, 11-month lag: Babayigit and Stainthorp, 2010) and .95 (Egypt, Arabic, Grades 1–3, time lag unspecified: Mohammed *et al*., 2011).
- Reporting of psychometric properties on the non-word fluency task is missing for all but one measure (test–retest = .74: Northern Cyprus, Turkish, Grade 1, 11-month lag: Babayigit and Stainthorp, 2010).
- Fifty percent of measures using connected text report reliability estimates of greater than .90, and two more with estimates between .80 and .89.
- Only one study reports assessing reading with intonation along with reading speed and reading accuracy (Nakamura, 2014). Estimates of inter-rater reliability for intonation are not reported.
- Among the measures of L2 reading fluency, the reliability estimates are between .68 and .87.
- Assessors need to be skilled in recording the accuracy of reading and simultaneously keeping time (and assessing prosody if this is included). Training for this level of skilled recording of children's responses has been reported to be challenging. Despite this, reports on estimates of inter-rater consistency are sparse. Reporting of assessor training is also poor.

Figure 5: Number of measures shown by cohort characteristics (total measures = 52)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

What are the innovations and challenges?

Age- and grade-appropriate texts. This is an important concern in reading fluency assessment. Reporting is poor on how age and grade appropriateness has been established. This is an area that needs transparent reporting.

Equivalence across languages. This remains a challenge because of the inherent differences across languages. Some languages are multi-morphemic and thus may have a lower word count but communicate the same message as another language that has very few inflections (e.g. compare Turkish and English). Variations are also introduced by differences in the principles of

writing systems and specific orthographies. Orthographic complexity differs, as do symbol–sound mapping ambiguities. An analysis of cross-linguistic differences in words and passages chosen for the task is beyond the scope of this review, but future work must address this.

Reading fluency as the measure of choice in transparent languages. In languages likes Spanish, children reach the ceiling early on reading accuracy tasks because the language is transparent and decoding is an easy skill to acquire. It is on a reading fluency task that the individual differences become evident. Hence, reading fluency is the measure of choice for transparent languages, especially in the older grades. There is, however, one group of transparent languages where reading accuracy will continue to capture variability in attainments well into middle school: the *akshara*-based languages, with their extensive orthography. Reading fluency measurement therefore does not need to be an alternative measure in this set of languages.

When measurement sets undesirable pedagogical targets. There are several debates around this area of assessment (for recent points of view, see UNESCO Institute for Statistics (UIS), 2016). In defence of this area of assessment is the relative ease with which high standards of reliability can be achieved on this measure when compared to reading comprehension. A defence of reading fluency assessment also invokes the strong association between reading fluency and reading comprehension to argue that reading fluency is a good window into what the child can potentially do in the area of reading compression. Nonetheless, keeping aside the measurement argument there is the question of educational quality. Does reading fluency testing send out the wrong message? Within this view, a focus on the assessment of reading fluency will (unwittingly) shift focus from meaning-based instruction to the mechanics of speed and accuracy. Such a focus would be undesirable in those classrooms where teachers are light on explanation and instruction for reading comprehension (e.g. see Nag *et al*., 2016). The tension is between a measurement approach to developing a quality test and a pedagogy approach to what is worthwhile to assess.

2.6 Reading comprehension

Why is it useful to assess this construct?

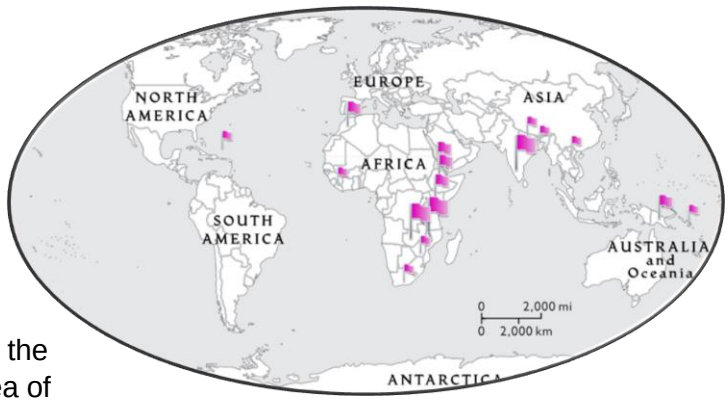
Reading comprehension is the skill of extracting meaning from written text. Within the ‘simple view’, variation in reading comprehension is related to word decoding and oral language comprehension, and this is true for both first and second language learners.

Multiple strategies are used by skilled comprehenders to extract meaning. For example, as texts become more difficult, inferential skills and a range of reading strategies (such as looking back at the text) are increasingly used. Predictors of rate of change in reading comprehension are the rate of growth in children’s vocabulary and decoding skills. In older children, individual differences in reading comprehension are associated with differences in morphological processing (specifically, inflection knowledge), syntactic processing (grammar awareness), and vocabulary knowledge. These findings point to the critical role of oral language proficiency as a foundation for reading comprehension.

Among biliterates, the associations between oral language, decoding skills and reading comprehension are more complex. Much needs to be still clarified in the area of biliteracy development, but the available literature suggests that the unique predictors of reading comprehension in the second language include L2 oral language proficiency and L1 reading comprehension.

Assessment of reading comprehension gives direct evidence of how well a child can read and how well a teaching programme is working.

What is available and how is this area assessed?



Sixty-six measures from 27 studies and 16 countries comprise this set (see adjacent map, with its legend in footnote 4; for composition of cohort, see Figure 6).

Use of a group format is more common in assessing reading comprehension than in any other area within our synthesis framework (23 of the 66 measures). The types of measures in this area of assessment include:

- Question and answer. This is the most common format. Children read a passage and then answer questions on it. Questions may be direct or require inference. There may be one or more questions for each passage. Answers may be in spoken form, written form or chosen from multiple options (38 measures).
- Cloze. These are tests where a selection of words is left blank and the child has to comprehend the available text to supply the missing words (20 measures).
- Modified Cloze (also called Maze task). This test supplies potential words to choose from to fill in the blank: *'Which of these words best completes the sentence?'* (Morocco: Wagner, 1993) (five measures).
- Matching. In this test the child matches a sentence to a picture (one measure).

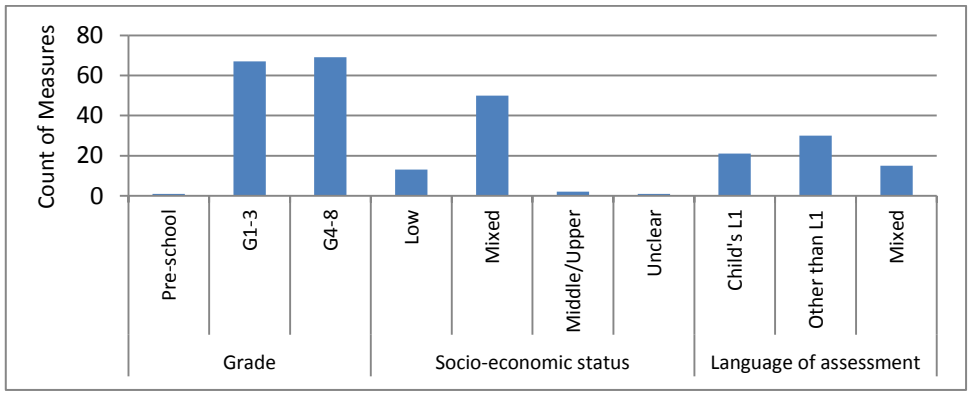
There is insufficient reporting of task details for two measures.

Reliability estimates for internal consistency of measures typically range between .70 and .80 for Cloze tasks (11 measures). Reporting is poor for the modified Cloze measures but the correlation pattern is in the expected direction: high association between performance on the modified Cloze measures and question–answer measures.

On the question–answer tests, we examined measures grouped for the number of questions asked on a given passage. Estimates of internal consistency for one question per passage is .73 (one measure), for two questions is .62 (one measure), for three to six questions between .81 and .85 in the L1 and between .70 and .80 in L2 (18 measures) and for seven to 12 question between .73 and .79 in L1 and .60 and .82 in L2 (seven measures).

Internal consistency estimates for the matching task is at .90 (Hindi: Sharma, 1997).

Figure 6: Number of measures shown by cohort characteristics (total measures = 66)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

What are the innovations and challenges?

Choice of material. Studies are lax in the reporting of how a passage was chosen. A good description of selection criteria is that the material reflects the textbook 'in topic and in style' (Malawi and Zambia: Williams, 1998). There are more sophisticated measures to ensure there is a match between grade-level reading material and a chosen reading comprehension passage but these forms of establishing equivalence were not found in this literature. We did not find guidelines for the levelling of connected text.

2.7 Narrative writing

Why is it useful to assess this construct?

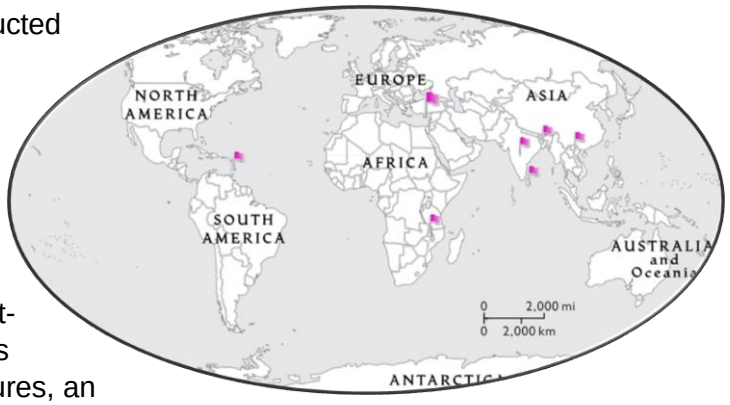
Multiple cognitive-linguistic processes underpin narrative writing. Of these, transcription, narrative generation and memory are three key components (adapted from Berninger, 1996; 1999). Among novice writers, narrative writing is seen as a telling of what they know when given a trigger or a prompt ('writing whatever a prompt brings to their mind', in the words of Babayigit and Stainthorpe, 2010). Higher-order processes of planning and writing for an audience are not yet evident. At this stage, the lack of automaticity with the mechanics of writing (transcription) may take away attention resources from generating the content for the writing. The constraints of transcription skills on narrative-generation skills are, however, closely linked to the nature of instruction. If instruction focuses on good handwriting and spelling, then these may quickly become automatic and no longer constrain content generation. However, when the focus on transcription skills in parallel limits practice in different genres, an equally plausible outcome is slow development of narrative-writing skills (Nag *et al*., 2016). Children may then approach the writing task using taught templates such as writing by rephrasing the question prompt or reproducing taught essays, and some children may not write at all. Thus, written language (e.g. a composition, a letter, etc.) is a window into the child's language skills with the strength of association between written and spoken language stronger when transcription skills have reached a certain level of automaticity *and* instruction has supported the development of a broad range of writing skills.

Narrative writing is useful to assess because it gives specific insights about the component skills of writing and general insights about the quality of the education being provided. The assessment has the potential for direct inferences on what the child can do in the area of writing, what the child needs to write better and what the teacher can do to help the child write better.

A narrative-writing task is useful for monitoring educational quality and in a toolkit for teachers because of the transparent link between assessment results and stated curricular goals. The field is, however, under-researched both internationally and, as will become evident below, within developing countries.

What is available and how is this area assessed?

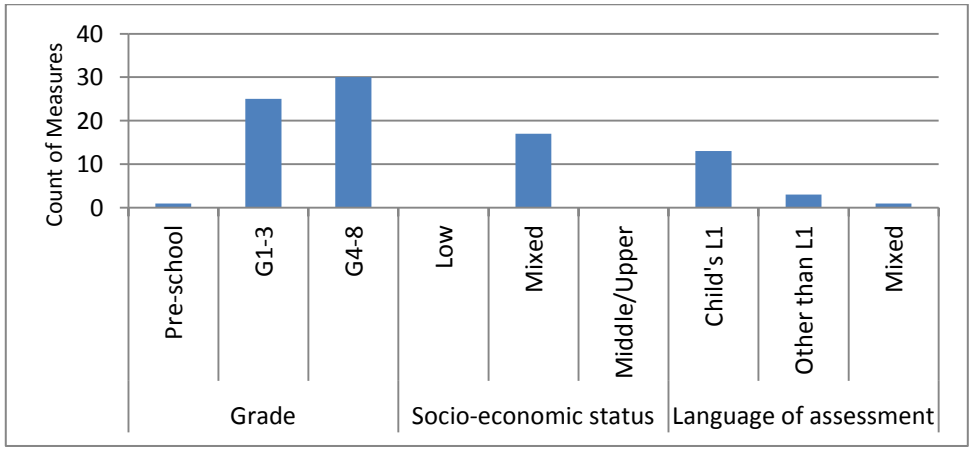
There are 17 measures from nine studies conducted in seven countries (for details see adjacent map, with its legend in footnote 4). The assessment is typically in the child's home language (see Figure 7 for other details about the cohort). All measures are researcher-developed, bespoke measures.



The prompts in these writing tasks include a just-heard story, a just-heard passage from the class textbook, a story narrated using a series of pictures, an unfinished letter and factual information on a specified topic gathered from multiple sources.

The narrative outputs in the writing tasks are in the form of completion of an unfinished story, re-writing of a story in one's own words and collation of information for a factual piece. Outputs at the simplest level include copying from a written model or writing one's name.

Figure 7: Number of measures shown by cohort characteristics (total measures = 17)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

Measures cover multiple component skills of narrative writing:

- Transcription skills, including handwriting, 'appearance', use of punctuations and spelling.
- Narrative-generation skills, including quality of vocabulary, the grammar of the written text (e.g. use of inflections and specific word families such as prepositions, adjectives and adverbs), style-related details (e.g. marking time and chronology in the narrative, tone of the story, etc.), cohesiveness of the narrative, awareness of genre-related details, and 'an awareness of the reader'.
- Working memory as well as short- and long-term memory assessed from the accuracy and comprehensiveness of idea units covered in the written text. For stories, the focus is on details

related to characters, settings and sequence of events. For factual writing, the focus is topic-specific information.

- Measures that draw upon more than one sub-system include writing fluency (the total number of words written per minute), 'vividness' (in quality of language usage and the level of detailing in the narrative) and the ambiguously titled area of 'creativity'.

Assessments of narrative writing in the L2 or in multilingual settings cover the same component skills as L1 assessment, with an additional evaluation for traces of the first language in written expression in another language (e.g. use of dialect and native language words: Nag, 2013).

Scoring schemes in eight of the nine studies reflect the multi-dimensional nature of the construct of narrative writing. The number of dimensions in a scheme range from three to 10, and scoring is based on three-, four- and five-point rating scales or categorical scores. Analysis of children's attainments uses both individual scores for each dimension and a simple composite of some or all dimensions.

Reports from two studies (Johnson *et al.*, 2000; Nag, 2013) suggest that rating children's written narratives requires skills that may not be easily available among some teachers. Estimates of reliability may arguably be higher for transcription skills and accuracy of recall of idea units when compared to scoring for narrative-generation skills at the level of content and structure. One study on early grade writing that allows a direct examination of this issue shows partial confirmation of this hypothesis (Northern Cyprus: Babayigit and Stainthorp, 2010). Here, inter-rater reliability between two primary school teachers is high (between .97 and .99) for transcription skills (writing fluency and spelling accuracy) and narrative content (judgement of relevance, accuracy and vividness of content) but lower (.74) for narrative structure (judgement of completeness of sentences, repetitiveness of sentence structures, number of subordinate clauses in sentences, and use of linking expressions such as 'and' and 'but').

What are the innovations and challenges?

Scaffolding the narrative-writing process. Innovations are seen in the supporting of children through the process of narrative writing. These supports provide contextual and memory prompts.

- 'The children were told to study the pictures carefully and then when they were ready, to go back to the beginning and start writing the story.' The pictures remain visible throughout the writing exercise (Babayigit and Stainthorp, 2010).
- Listen to a story, then participate in a class discussion about a picture depicting the just-heard story, and then use the picture to write the story (Johnson *et al*., 2000).
- Complete a part-written letter filling in an appropriate salutation, one key message and an appropriate ending (Chowdhury *et al*., 1994).
- Use supplied words such as connectors (e.g. 'and'), inflections (e.g. 'to', 'for', 'of') and transitional tags that communicate time sequence (e.g. 'first', 'next', 'afterwards', 'then') to develop the narrative (Nag, 2013).

Scaffolding is perhaps especially meaningful in contexts where there is limited opportunity to practice a broad range of writing skills. It is possible that scaffolding allows for greater variability in written outputs (those who may not have written much now produce longer narratives because of the support), but we did not find a direct evaluation of this hypothesis.

Evidence needed at several levels. To build on the available innovations with narrative-writing assessment, evidence has to be built at several levels. At the outset, it is unclear if some topics are

better than others at capturing the individual differences in narrative-writing skills. This is an important question because topics were always specified in the measures we reviewed. Coding schemes also require evaluation, particularly to operationalise potentially ambiguous parameters of assessment such as 'relevance of the writing', 'vividness' and 'appropriateness of vocabulary'. Related to this is the need to find ways to achieve high inter-rater reliability, particularly for the scoring of narrative-generation skills. Finally, it is unclear if scaffolding (in general, and for each of the types of support described above) captures greater variability than an unsupported writing task, particularly in contexts where narrative-writing practice is limited.

2.8 Grade-level tests

Why is it useful to assess this construct?

Tests of grade-level competencies cover a stated subject domain and draw heavily on local curriculum and assessment frameworks. Such measures therefore can potentially allow for a real-world check of children's attainments even when the approach to assessment does not exactly match testing practices that are popular with local teachers. Moreover, tests of grade-level competencies are assumed to hold high ecological validity because they resonate with the learning outcomes that are sponsored and promoted by the education authority of a region (e.g. ministries of education, etc.).

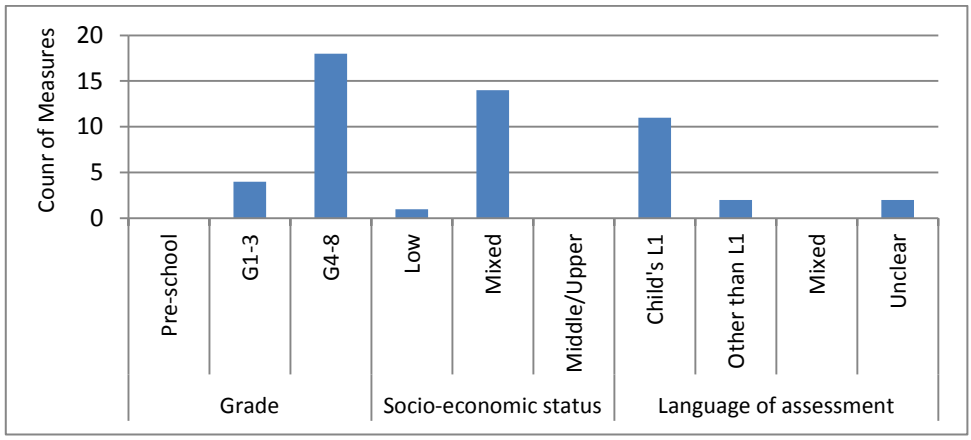
What is available and how is this area assessed?

This set comprises 15 measures from eight studies conducted in seven countries in South and Southeast Asia (the legend for the map can be found in footnote 4). For the composition of this cohort, see Figure 8.

The assessments included in this set were conducted within a single country (with one exception, where trends are examined both within and beyond the individual country). It must be noted that an extensive set of grade-level tests used in large, multi-country surveys has not been included here because such tests fall outside the scope of this review.[15]

[15] Examples of large, multi-country assessments of reading achievement include: a) the PIRLS for fourth graders conducted by the International Association for the Evaluation of Educational Achievement (IEA) every five years, with 2016 being the fourth round; b) the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) for sixth graders conducted by the Consortium, with four rounds completed. A stated mission of these cross-national surveys is to provide meaningful data for informed decision-making to improve the quality of education.

Figure 8: Number of measures shown by cohort characteristics (total measures = 15)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

Grade-level tests in the review set follow two formats:

- They are based on a one-off assessment (e.g. India: Lakshminarayana *et al*., 2013; Philippines: Abeberese *et al.*, 2011; Sri Lanka: Aturupane *et al.*, 2014; Vietnam: Rolleston and Krutikova, 2014).
- They are based on a composite score derived from continuous classroom assessment (Eretria: Asfaha *et al.*, 2009).

The most elaborate tests are styled after the SACMEQ, with six 'developmental' levels and three item types (Vietnam: Griffin and Thanh, 2006).

The local experts who develop these tests are either members of the in-country authority for educational assessment or an informed third party. The field assessors range from teachers and school principals to graduate research assistants and hired independent assessors.

Both criterion-referenced and norm-referenced tests are seen. In other words, grade-level tests have been used both to assess actual achievement and mastery of the 'language' subject domain as well as the relative ranking of test-takers within that domain.

What are the innovations and challenges?

Quality of teacher-led assessment. This is an area of concern in some developing countries. One study found that language and literacy assessments conducted by teachers did not show the expected associations with the tasks of reading comprehension conducted by the research team (Eritrea, multiple languages of instruction: Asfaha *et al*., 2009). This suggests the likely absence of rigour in classroom assessments, meaning that, even when student performance is elicited and recorded, the use of this information by teachers is not informative. The extent of mismatch in classroom and independent assessment is unclear from the reported data but this single study highlights the challenges involved in improving the standards of teacher-led assessment within school systems.

3 Assessment of spoken language skills

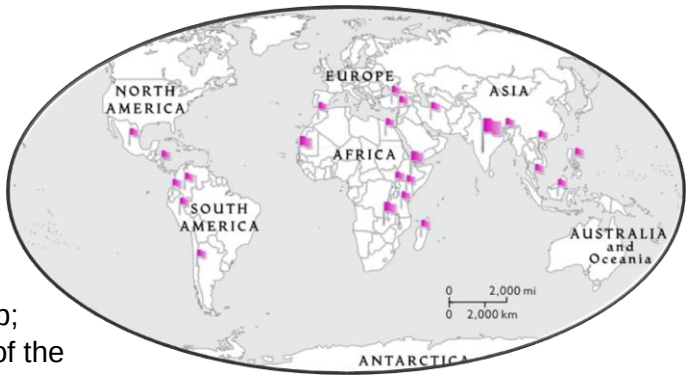
3.1 Vocabulary

Why is it useful to assess this construct?

Vocabulary knowledge is a complex and multi-faceted construct. The construct covers, for example, the breadth and depth of knowledge about words, in either the expressive or receptive mode. Vocabulary knowledge may be demonstrated in multiple ways, including quickness to generate words on a theme and awareness of component parts in words (morphological awareness). Vocabulary assessment is useful because this component of language is related in important ways to individual differences in literacy attainment. Children with better vocabulary pull ahead in their skills for abstracting meaning from texts (reading comprehension). Vocabulary knowledge is also useful for decoding words that are multi-morphemic, written with an uncommon symbol or with an exceptional spelling (reading accuracy).

Children expand their vocabulary at an exponential pace between the ages of three and nine (Biemiller, 2015). However, this vocabulary development is exceptionally sensitive to ambient language and children with exposure to wide-ranging vocabulary are at an advantage. There is also robust evidence to show that second language learners lag behind native learners in their knowledge of word meanings. The vocabulary gap is seen both in a lower number of known words (vocabulary size) and in the speed with which new words are added to the child's lexicon (rate of acquisition). Since many children acquire literacy in a language other than the home language, assessment of vocabulary knowledge becomes an imperative.

What is available and how is this area assessed?



We review 63 measures. The assessments are conducted in 24 countries covering 26 major languages, a small number of 'minor' and 'town' languages, and dialects.[16] The geographical spread of the vocabulary measures is one of the best in this review (see adjacent map; for the legend, see footnote 4). The composition of the cohort is given in Figure 9.

Vocabulary assessment in developing countries is clearly influenced by an approach first made popular in the 1950s with the Peabody Picture Vocabulary Test (PPVT), in which a child must point to one of four pictures that match a just-heard word (50 of 63 measures used this approach). Other measures use a spoken word or a sentence to assess vocabulary knowledge (12 and one respectively). Task details and number of measures in our review set are as follows:

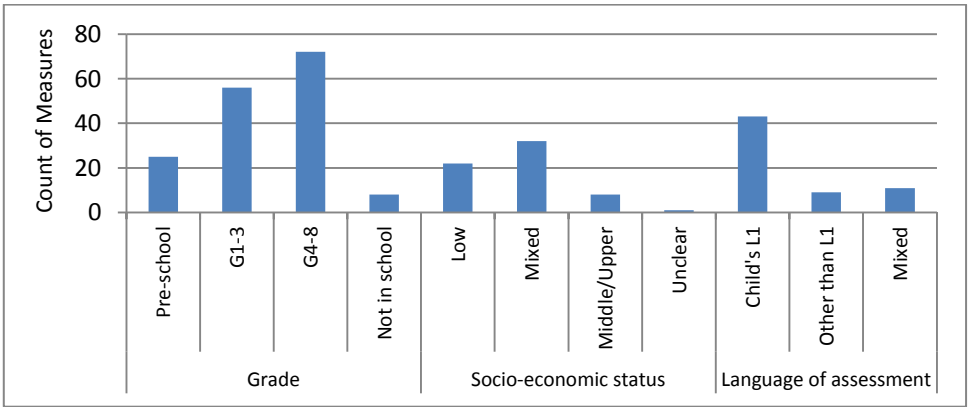
- Identify target word from a set containing distracter words (receptive vocabulary with a focus on vocabulary breadth: 28 measures).
- Name target word (expressive vocabulary with a focus on vocabulary breadth: 13 measures).

[16] Languages covered are Amaringa, Arabic, Bahasa Indonesian, Bangla/Bengali, CiNyanja/ Nyanja/Town Nyanja, English, Filipino, Hindi, H'mong, Kannada, Kiswahili, Lozi, Luganda, Malagasy, Malay, Mandinga, Mbunda, Oromifa, Persian, Quechua, Spanish, Telugu, Tieng Viet Nam, Tigrigna, Turkish and Wolof. This list does not include multiple minor languages covered in one study.

- '*Name as many xxx as you can'* (semantic fluency with a focus on vocabulary breadth: 11 measures).
- Define/describe target word (expressive vocabulary with a focus on vocabulary depth: three measures).
- Identify a synonym for a target word (receptive vocabulary with a focus on vocabulary breadth: three measures).
- Divide a word into its prefix and root (word manipulation with a focus on morphological awareness: one measure).
- Identify number of words in a spoken sentence (awareness of words as a linguistic unit: one measure).

Three measures mix receptive and expressive items and one study turns two independent sub-tasks into a composite.

Figure 9: Number of measures shown by cohort characteristics (total measures = 63)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

What are the innovations and challenges?

Localisation effort. Several studies acknowledge that, after procurement of commercial tests, substantial adaptation was required but the more consistent effort at localisation is with the bespoke measures (33/63). Of these, 10 specify linguistic and cultural considerations, accumulate evidence that can demonstrate validity of the task and report moderate-to-high reliability estimates of the measure: examples for the early years and older grades for vocabulary depth are the measures reported in Opel *et al.* (2009) and Nag and Snowling (2011), and for vocabulary breadth the measures reported in Vagh (2009) and Jukes and Grigorenko (2010).

Item selection. A range of parameters guide the selection of target words and distracters. Of these, examples c) to e) below demonstrate innovation in language-specific item generation:

a) selection from word families particularly focusing on nouns and verbs (styled after PPVT).

b) selection of synonyms.

c) selection of phonologically close, semantically close and semantically unrelated words as distracters (Kiswahili *et al.*, 2010).

d) selection of words that are similar/identical and very different in the home dialect and the language of assessment (Standard and Moroccan Arabic: Rochdi, 2010).

e) selection of words for their prefixes and suffixes (Bahasa Indonesia: Winskel and Widjaja, 2007).

Unlike word lists, guidelines for the selection of pictures are poorly articulated. Most studies report consultation with local experts to check the appropriateness of pictures, but the parameters guiding such judgements are vague. The first challenge is that pictures may be ambiguous and the potential reasons for this are numerous, e.g. pictorial representations may be alien for children with exceptionally low print experience, the pictorial idiom may be difficult for children to understand or the visualisation may be outside their lived experience. A further challenge is to create picture items that appear equivalent, perform equally across subgroups, and do not elicit word knowledge around unintended constructs. Finally, picture-based assessments are constrained by the medium; those words that are not easily captured in pictures need to be set aside. We did not find good examples of innovations to circumvent this limitation (e.g. use of accompanying context sentences as a way to include words with poor picturability).

Reliability estimates for each type of measure. Reliability data are reported for most measures (see Annex B). The most common estimates are for the internal consistency of the measure (using odd–even split half and alpha coefficients).

- Picture identification (receptive vocabulary) has the largest proportion of measures with estimates above .80. This small but consistent body of evidence has one counter-trend from a test–retest analysis among first graders (Zambia: Jere-Folotiya *et al*., 2014), where the reliability estimate is 'rather weak' at .23 (time lag not reported). The finding may be related to the children's age or the extent to which negotiating the task depends on experiences that come with schooling.
- Approximately 50% of measures of picture naming and vocabulary depth have internal consistency estimates of above .8. The reliability of the semantic fluency measures is unclear because of low reporting except for one study where internal consistency is estimated at .40 (computed for fluency scores across three semantic categories). Only one measure each of morphological manipulation and word awareness were available: the first has high internal consistency (>.80) and the second moderate (.60).
- Two studies allow for a direct comparison of internal consistency of tasks: picture identification .74, picture naming .88 (India: Vagh, 2009); semantic fluency .40, picture naming .65 (India: Brouwers *et al.*, 2006).

In summary, there are identifiable innovations in vocabulary assessment, with the bulk of evidence for picture identification measures and a very small body of evidence for synonym identification, picture naming and test of vocabulary depth. Single-study reports suggest that morphological-awareness and word-awareness tasks have the potential to capture variability but evidence needs to be built. Both tasks may be valuable for monitoring educational quality and in a teacher toolkit. Evidence needs to also be built for the claim that semantic fluency meaningfully discriminates different levels of vocabulary knowledge.

Capturing the complexity of vocabulary knowledge. The available measures rely heavily on the assessment of concrete words (particularly object names) and words in a single context such as its association to a specific picture or a specific sentence frame. Abstract words and words in multiple contexts are under-represented. The insights available in current theorising about vocabulary knowledge are yet to find a place in assessment, of which two are highlighted here:

- Vocabulary is a dynamic representation of the world rather than simple and single word entries in the mental lexicon. Within such an interpretation, vocabulary levels may be judged as higher when multiple uses of individual words are known (e.g. 'a fat cat' (plump cat) vs. 'a fat profit' (a

hefty profit)). The next generation of assessments must consider accessing the multiple meanings of words.

- Vocabulary is considered to also comprise knowledge of meanings of multi-word units ('because of', 'period of time', 'except that': examples from Biemiller, 2015). This is another item type that may be considered.

Vocabulary assessment incorporating these kinds of items would be particularly relevant for monitoring educational quality and to inform teaching practice in the older grades.

Limited comparability across vocabulary measures. The range of vocabulary measures available in the developing country literature reflects the multi-dimensional nature of the construct. Some studies use naming, others identification, fluency and manipulation tasks. The measures also differ in task demand. To illustrate, three tasks are ranked by increasing working memory demand:

> *'... name as many animals as possible in 1 minute'* (Philippines: Ledesma, 2002)
>
> *'... say as many words starting with 'M' as they could in 1 min., omitting all proper nouns (names of people, places, etc.)'* (Mexico: Ardila *et al.*, 2005)
>
> *'a chick (kifaranga) is a chigger (funza)? or a lock (kifunguo)? or a chicken (kuku)? or a t-shirt (fulana)?'* (Tanzania: Alcock *et al*., 2010)

These variations limit comparability across measures. It is also fair to note that variations because of differences in instruction formats are found in any area of assessment, but are perhaps more common among spoken language measures.

3.2 Other areas of spoken language assessment

What are the useful constructs to assess?

In line with the framework of this review, several spoken language skills beyond vocabulary knowledge have an abiding influence on attainments in reading and writing. Listening comprehension is associated with reading comprehension. Grammar knowledge and syntactic processing show associations with reading comprehension and narrative writing in multiple languages, and with accuracy in reading and spelling multi-morphemic words in some languages (e.g. the agglutinating languages like Kannada and Turkish). Spoken language assessment in areas such as listening comprehension, grammatical awareness and expressive language can inform the question: What does the child need in order to read and write well?

Another reason to focus on spoken language assessment is to counter the ignoring of the linguistic assets that children bring to the task of reading and writing from their home and culture (see Nag *et al*., 2016). Such disengagement with the foundational skills for literacy is costly for the school system and the teaching–learning process. An important message from these inter-linkages is that spoken language assessment is important even if the stated interest is assessment to improve literacy. These assessments address the question: What does the teacher need to know in order to support children to read and write well?

What is available and how is this area assessed?

We reviewed 36 measures drawn from 17 studies covering 23 home languages and 13 second languages. The geographical spread of the studies is given in the adjacent map (see footnote 4 for the legend and for composition of the cohort see Figure 10).

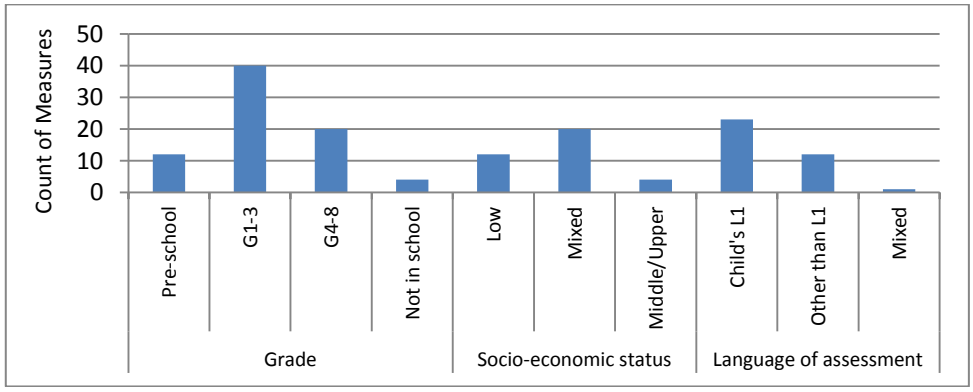
The 36 measures, clustered by the nature of the task, are as follows:

- Comprehension measures that ask for demonstration of understanding through questions about a just-heard message. Items include sentences, local proverbs and short stories. Questions may be factual or demand inferences (26 measures).
- Grammatical-awareness measures give information about the child's grasp of the grammatical forms of the language. Repeating a just-heard message, making a judgement about the appropriateness of sentence construction and accurate reception of grammatical information are in this set (three, one and one measures respectively).
- Retelling measures require the child to narrate back a short story. This is an expressive oral language task that, like the narrative-writing task, must draw upon multiple language subsystems (four measures).

The grammatical-awareness measures use phrases or sentences while the comprehension and retelling measures use longer narratives or a mix of shorter and longer linguistic units. The reporting of psychometric details is low across all measures but where available the estimates are noteworthy. Estimates of internal consistency are moderate to high for listening comprehension measures (Liberia, English (L2), alpha coefficient .88: Piper and Korda, 2011) and grammatical-awareness measures (India, Kannada, alpha coefficient .70: Nag and Snowling, 2012; Northern Cyprus, Turkish, alpha coefficient .78: Babayigit and Stainthorp, 2010).

Ratings of children's retellings are typically elaborate but none of the measures report a robust check of inter-rater consistency.

Figure 10: Number of measures shown by cohort characteristics (total measures = 36)



Note: Grade count is across overlapping categories; SES and Language of Assessment counts are across discrete categories.

Studies vary in regard to the subgroups who receive the test and the purpose of assessment. Some studies assess children's listening comprehension across the attainments continuum while others (e.g. the EGRA-based studies) assess only the non-readers. In intervention studies, these measures have been used to match groups or as outcome measures. In bi- and multilingual settings, the measures help select a sample with a homogenous linguistic profile. The grammatical-awareness measures appear in studies examining the language–literacy relationship.

In summary, several measures are available to assess listening comprehension, grammatical awareness and expressive oral language but the level of reporting is poor. This therefore precludes an analysis of which method is superior for which age band and for assessment in L1 and L2.

What are the innovations and challenges?

The challenges of creating language material for spoken assessment are considerable and arguably reflect some of the most vexing issues in the assessment of children and learning.

Test development. A typical but not desirable process during the development phase of a test is as follows: A test in language X that has the advantage of an existing body of research about its robustness and usefulness is taken as the material to be adapted for language Y. Drafts of the material receive critical attention through a pilot study and/or a consultation with language Y experts. The material goes through iterations and the iterations are seen as strengthening localisation of the linguistic content; this process is assumed to weaken the hold of the problematic direct-translation approach. The evidence from this review shows that such an approach appears to have served well at the level of vocabulary tests but does not work at a level of language testing beyond single words. The problems with using another language as the first reference point for material development are that:

- such material may or may not reflect the inherent properties of the language of interest; and
- such materials may or may not have a deep association with the oral traditions of a region.

Furthermore, there may be several other cultural factors that the material is blind to. Innovations to address these concerns include:

- use of culturally embedded linguistic materials such as proverbs, children's riddles and folk stories (e.g. Jukes and Grigorenko, 2010; Nag, 2007; Veii and Everatt, 2005); and
- detailed psycholinguistic analysis prior to material development (e.g. Castilla, 2008).

There are other reasons why language tasks may not be equivalent. In the area of listening comprehension measurement, for example, contributing processes include vocabulary and background knowledge, inference making and working memory. Individual measures differ in the extent of demand made on vocabulary and background knowledge. Some measures tax working memory more than others. However, the most obvious difference between measures is in the quality of inference expected of a child. Some questions can be answered from explicitly stated ideas but others need connections to be made or gaps to be closed between neighbouring phrases or sentences, or idea units placed far apart. No individual measure in this review set reported systematically manipulating any of these parameters to improve the sensitivity of the measure.

4 Lessons learnt

4.1 Safeguarding against token localisation

Many adaptations of assessment tools report on the translation process and/or contribution of local experts but are short on documentation of the exact processes followed. Where translation is mentioned, the use of back translation is not – it is therefore difficult to know if this important step was bypassed or overlooked in the reporting. Where local experts contribute, they are typically university-based academics in the fields of language, education and linguistics, or they are school teachers. The guidance note for quality checks of items is not reported and again it is unclear if there was consensus among those consulted, and when there was not, how the differences were reconciled.

An excellent example of localisation across multiple countries is the Young Lives measures.[17] For example, the measurement researchers demonstrate:

- **Sensitivity to establishing fairness in test design.** One way this is done is to identify vulnerable groups within a specific context and examine how the test behaves in this group when compared to a more privileged group (e.g. girls vs. boys, major vs. minor languages, etc.). This allows for the development of measures sensitive to socio-demographic and socio-cultural variations and minimises interference from factors unrelated to the construct being assessed.
- **Use of methods to establish fairness.** This includes impressionistic procedures (e.g. cultural relevance, cultural and linguistic appropriateness, adequate conditions for test taking, etc.) and empirical procedures (e.g. using Differential Item Functioning). These procedures improve the chances for inferences from test results to be similarly meaningful for all groups. Another important adherence to fairness is to explicitly limit the comparison of test results if there is doubt that similarly titled tests may not in fact be equivalent.

One widely adopted effort at localisation is to align assessment material to local school textbooks. The methods for doing this include calculating type and token frequencies of words in textbooks and gathering impressionistic data on whether items reflect curricular targets.

4.2 Communicating assessment results

As has happened in the last decade, information from assessments is likely to continue to guide countries toward the new Sustainable Development Goals for education. A quick analysis of themes in public documents about assessment results show two dominant concerns: governance of education systems and quality in education for the common good. This section speaks to the second concern, specifically within the context of young children's literacy learning.

There are two inter-linked points from the foregoing synthesis that are important for the communication of assessment results to improve quality in education: first, a profile approach to understanding foundation learning and literacy; and, second, the use of assessment results to serve the development of grounded education programmes. A profile approach to children's learning appears to be inherent in several studies in this review, although sometimes the rationale is not entirely clear (e.g. why should an early grades test battery include letter-naming fluency, familiar word fluency, unfamiliar word fluency, and fluency with connected texts?). But beyond the

[17] Cueto *et al.* (2009). Retrieved from www.younglives.org.uk

multiple areas of assessment, the use of the data often becomes exceptionally unitised. We take one specific example to illustrate the fractured nature of data interpretation:

> '...students are not able to identify one third of letters... help children learn all of their letters, especially the letters children most struggled with: Q, W, Y, J, I, L, and G (and q, w, y, j, i, l, and g)').[18]

Such a recommendation may be seen as an *All Symbols First* recommendation and such an approach raises several questions. Will a focus on these specific letters make a difference to children's skills for reading and writing? Teachers in countries using extensive orthographies are settled on the reality that some symbols are learnt later than others. Might such an approach be applied to the learning of, for example, q, w and j? Or, should the assumption of *All Symbols First* be applied not just to the English alphabet but also other symbol sets? Clearly, a linguistically sensitive conceptual framework for assessment would not import the *All Symbols First* notion to extensive symbol sets, while a pedagogically sensitive framework for assessment would be more circumspect about what from the profile of attainments captured in an assessment session is to be prioritised in the classroom. Lastly, do such communications pass on a perspective about teaching and learning that is limited to specific tasks and skills, and a perspective about educational outcomes that is restricted by what was assessable?

The purpose of quality in education would be better served when communications about assessment results are alert to such broader linguistic and pedagogical issues.

4.3 What should be assessed at scale?

We have used a systems view of literacy development to structure the data available from developing countries. Within this view, there are multiple component skills of literacy and each involves multiple knowledge bases related to orthography, phonology, semantics, morphology and morpho-syntax, as well as world knowledge or topic-related and 'general' knowledge. It is not necessary to assess all these cognitive-linguistic areas at scale. A subset would suffice.

A direct assessment of literacy is through the already accepted focus on reading accuracy and reading comprehension. Other areas to include are chosen because this review has shown that: a) they are not very resource intensive to assess; b) they add to our understanding of an educational system; and c) they hold promise for being brought to scale. These include:

- emergent writing (for very young children[19] and children in print-starved environments);
- symbol knowledge (across primary school);
- spelling (across the school years);
- narrative writing (across the school years);
- vocabulary (across the school years);
- listening comprehension (across the school years); and
- grammatical awareness (across the school years).

Fluency tests (of both reading and writing fluency) are useful at scale but only if accompanied by tests that directly assess reading comprehension and narrative-writing skills.

[18] From the Executive Summary in Chinyama *et al.* (2012).
[19] For a recent addition of this task at scale, see Pisani *et al.* (2010).

Some assessment tasks are resource intensive in terms of both their development and the establishing of inter-rater reliability. Examples include the CAP tasks, non-word tasks, phonological tasks, expressive language tasks and assessing children's retelling of narratives. These tasks also do not add substantial new information about the quality of an educational system and therefore do not represent good value for money.

4.4 Assessment toolkits for teachers

Three core principles guide the choice of tests for an assessment toolkit for teachers. These principles are that:

- **An assessment framework must align assessment results with actionable responses to pedagogical questions.** The framework of written and spoken language that we have followed in this report is useful to align teacher-led assessment with what children must learn in order to read and write well.
- **The scope of a toolkit must be defined by the proficiency of the teachers who will use it.** The developing country literature suggests that, even though highly skilled teachers are available, many still come to the task of teaching with exceptionally low skill and knowledge.
- **The purpose of assessment must first serve pedagogical decision-making and then, if appropriate, decision-making in additional areas.** We illustrate below one additional purpose of assessment by choosing diagnostic assessment for children at-risk or with a developmental disorder. Examples of other additional purposes of assessment possible for teacher toolkits include assessment for certification of students, for placement of students, and for accountability of colleagues.

The following toolkits draw upon these core principles and address the tension between assessing at a grain-size meaningful for instruction and using tools within the skills of the teacher-assessor:

A full toolkit for pedagogical decision-making. This toolkit focuses on multiple spoken and written language skills and includes assessment tools in the area of:

- vocabulary and spoken language;
- CAP and symbol knowledge;
- reading and spelling accuracy; and
- reading comprehension and narrative writing.

Assessment of reading fluency and writing fluency are also useful but only if accompanied by tests that give information about reading comprehension and narrative writing.

This toolkit may be considered for teachers who are skilled both in assessment processes and in instruction. The tasks in the toolkit allow for some degree of micro-level analysis of children's behaviour, and information from these tasks complements information gathered from routine classroom assessment of grade-level/curricular targets.

A light toolkit for pedagogical decision-making. This toolkit with simple to administer and easy to interpret tests can serve teachers with low skills and proficiencies. Tests from this review that meet these two criteria are related to symbol knowledge, reading and spelling accuracy, and picture vocabulary. Such a toolkit can serve the purpose of pedagogical decision-making but only in a limited manner. The aim would be to upgrade the toolkit to cover other areas of assessment but first provision has to be made for sustained training to meet standards (e.g. in test administration, recording and scoring of children's responses, and interpretation of results).

An advanced toolkit for diagnostic decision-making. Such a toolkit is needed if the purpose of testing is early recognition of developmental disorders or offering tiers of specialist support when there are concerns about a child's literacy progress relative to grade peers. Such a diagnostic toolkit is for teachers who can go beyond the curriculum and can interpret the role of skills that undergird literacy development. Examples of relevant tests for this purpose and level of assessor skills include non-word reading and phonological processing.

A final point is related to the need for smooth uptake of assessment results into classroom practice. Without smooth uptake the full potential of a teacher's toolkit cannot be realised.[20] Although this area is outside the scope of our review, it is clear that, in many developing country contexts, teachers will almost certainly require demonstration of how to translate assessment findings into lessons in the classroom.

[20] For one example of promising practice, see Johnson *et al.* (2000): 'Teachers involved in the study were able to collect information about children as learners, collect evidence of learning and to record the achievements of children' (p. 70).

5 Gaps in evidence

This section focuses on gaps in evidence that are wide-ranging in nature and relevant for all areas of assessment. Gaps in evidence specific to each area are given in the relevant sections above.

5.1 Profile of the assessor

Interpersonal processes linked to the identity of the assessor may influence the outcomes of an assessment. For example, in contexts where gender relations are firmly defined, the gender of the assessor may impact children's performance. SES, urbanicity, ethnicity, religion and linguistic affiliation are other deep and entrenched social stratifiers in some contexts; these too are likely to influence the assessment process. Bringing focus on who is the assessor is quite different from the common interest in how reliable the assessor is. We did not find any systematic examination of the profile of the assessor on children's test performance, although some studies invoke the assessors' identity as an unaccounted-for variable to explain group differences.

5.2 Assessment results as reflecting context

Several tests make task demands that are closely linked with school experiences. In narrative-writing tasks, for example, an assumption is that children will write spontaneously and what they write will showcase their skills. However, if classroom practices favour certain narrowly defined written productions then test performance would capture this coaching. Beyond the individual child, any divergence in attainments across schools could then simply reflect the nature of classroom practices. Similarly, the nature of the printed material may itself become a barrier to assessing literacy-related skills and knowledge. Single-study evidence from Tanzania shows improvement in performance on handwritten words compared to their printed form. This difference was noted even after care was taken to ensure test materials were printed in a font close to that used in textbooks (Alcock *et al*., 2000). Unfamiliarity of printed materials, the procedures of the test and test-like situations are hidden challenges in many contexts. These are neither fully understood nor accounted for in the interpretation of test results.[21] A final area that needs concerted attention is learning in multilingual and biscriptal contexts: our review found attention to this area is growing but much remains unclear.

5.3 Dissemination of assessment information

It is clearly important to ensure that assessment findings are appropriately and in a timely fashion communicated to different audiences. The expected flow of assessment information should match the stated purpose of assessment, with dissemination, for example, to management, policy-makers, teachers and parents. We did not find clear reporting of such types of information flow; it is not clear if a dissemination plan was followed or neglected.

5.4 Reporting standards

Accurate measurement of children's attainments is critical for advancing our understanding of foundational learning and literacy development. It is important for decision-making in the real world whether about individual children or a class, or larger units such as a section, a school, a school district or a school system. An important requirement for all of these aims is to build confidence in

[21] For similar concerns see Sternberg *et al*. (2002).

the measurement tool. However, we found reporting about the contextual relevance of a measure to be poor. Reporting about the psychometric properties of measures was also poor.

6 Future directions

6.1 Prioritising measurement research in developing countries

Psychometrically rigorous and contextually sensitive measurement is an important research agenda for developing countries. This review has given valuable insights into some of the key issues that can improve the quality of assessment. Examples of areas where protocols to ensure quality standards are needed include:[22,23,24]

- The construct, group and purpose. It is desirable that the constructs being assessed are clearly stated at the outset, and there is careful consideration on whether theorising – if drawn from other languages, orthographies, school systems and socio-cultural contexts – can be applied to the local population. Also essential is to explicitly describe who will be the user of the test and for what purpose(s) the test will be used.
- The processes and judgements. The field of assessment acknowledges the need for measures with strong psychometric properties and cultural relevance. Adopting greater transparency regarding what is done before a tool is taken into the field would be an important first step toward improving confidence in individual measures. In other words, one important way to ensure quality standards in assessment is through adopting protocols for documenting the contextualisation of a measure and establishing its robustness.
- The cost and value for money. Assessment often occurs in contexts that are severely resource-constrained. It is essential that protocols for reporting on cost are routinely included. Examples of areas of reporting include cost of test production or test procurement, training, administration, analysis and communication of results. Also important to examine is whether tests otherwise fit for purpose also bring value for money.

Currently there are more assessment tools available in developing countries for research purposes and fewer for use by teachers. Measurement research in developing countries must ensure teacher-led assessment tools are on the research agenda. Also related to this is the need for a systematic evaluation of the usefulness of test-based pedagogical decisions and the outcomes of such decisions on children's literacy attainments.

In some developing countries tests are available through high-quality research studies and small-scale surveys, but normative data are yet to be collected. One priority for the research agenda is to identify high-quality tests for generating further local evidence and for norming studies.

Developing affordable tests is also important for the research agenda. Affordable tests can help to ensure fairness in access to the benefits accrued from the use of assessment results.

6.2 Innovations using group-testing formats

Currently the bulk of testing is one on one. Much needs to be done to clarify if group-testing formats can be used meaningfully, keeping the economic viability of such efforts in view. Group testing certainly is not indicated for the preschool years but more examples are needed for group testing across the rest of the school years. The review has highlighted the skills that may lend themselves to group testing (emergent orthographic knowledge, emergent writing, symbol

[22] A useful reference to develop standards is American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999).
[23] For an example of an explicit and transparent statement of protocols see Education Testing Service (2014).
[24] For a conservative discussion on concerns when importing tests into new contexts, see Fernald *et al.* (2009).

knowledge, spelling, reading comprehension and narrative writing), but other areas also require innovation.

6.3 Multi-country, citizen-led and common-framework assessments

Large-scale assessment initiatives show the level of learning attainments to be extremely low in many contexts. These assessments are often linked to multi-country comparisons (e.g. PIRLS or SACMEQ), citizen-led initiatives (e.g. ASER or UWEZO) or common-framework assessments (e.g. EGRA). Substantial resources are spent on these initiatives and some have become high-stakes testing because of comparison between countries (or smaller units such as districts). Against this background, a social audit of the outcomes of such initiatives and a rigorous review of the theoretical frameworks and assumptions that underpin these assessment tools is called for.

6.4 Assessment of contextual factors

This review has focused on the assessment of within-child factors. As pointed out in footnote 2, however, a similar review should be considered for the assessment of contextual factors. Examples of such tools would be checklists and observation schedules to evaluate classroom, school, home and neighbourhood processes related to literacy and foundational learning, and tools for capturing broader constructs such as SES and socio-cultural influences.

6.5 Teachers as assessors and learning outcomes

A recommended mechanism for improving learning outcomes is teacher-led assessment in the classroom (e.g. Westbrook *et al*., 2013), although the current evidence base for this proposition is descriptive and correlational in nature. Future work must look for direct causal evidence for higher attainments because of assessment by teachers, and uncover what moderates the effects of teacher-led assessment. Examples of potential moderators include characteristics of the child, the area of skill assessment, the nature of the assessment, the skills and proficiencies of the teacher, and the nature of the school environment for supportive responsive teaching based on assessment results.

6.6 The need for free-to-use tests

The field is dominated by a huge variety of researcher-developed, bespoke measures. These are, however, almost always part of small-scale studies and hidden from the attention of key decision-makers who address educational issues at scale. Instead, commercial tests dominate such initiatives either as the test of choice or as the preferred template to be exported into multiple countries. These commercial tests are expensive and even after procurement require a lot of adaptation. Future work must consider a free-to-use resource bank of robust and useful tests. This could be an open-access online library available to researchers and practitioners interested in improving the learning experience for the child.

References

References marked with an asterisk indicate studies included in the rigorous review.

*Abeberese, A. B., Kumler, T. J. and Linden, L. L. (2011). *Improving reading skills by encouraging children to read: A randomized evaluation of the Sa Aklat Sisikat reading program in the Philippines* (NBER Working Paper No. 17185). National Bureau of Economic Research: Cambridge, MA.

*Abeberese, A. B., Kumler, T. J. and Linden, L. L. (2014). Improving reading skills by encouraging children to read in school: A randomized evaluation of the *Sa Aklat Sisikat* reading program in the Philippines. *Journal of Human Resources*, *49*(3), 611–633.

*Alcock, K. J. and Ngorosho, D. (2003). Learning to spell a regularly spelled language is not a trivial task – Patterns of errors in Kiswahili. *Reading and Writing: An Interdisciplinary Journal*,16(7): 635–666. doi:10.1023/A:1025824314378

*Alcock, K. J. and Ngorosho, D. (2007). Learning to spell and learning phonology: The spelling of consonant clusters in Kiswahili. *Reading and Writing: An Interdisciplinary Journal*, 20(7), 643–670.

*Alcock, K. J., Ngorosho, D., Deus, C. and Jukes, M. C. H. (2010). We don't have language at our house: Disentangling the relationship between phonological awareness, schooling, and literacy. *British Journal of Educational Psychology*, 80(1), 55–76.

*Alcock, K. J., Nokes, K., Ngowi, F., Musabi, C., Mbise, A., Mandali, R., . . . Baddeley, A. (2000). The development of reading tests for use in a regularly spelled language. Applied Psycholinguistics, 21(4), 525–555.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standards for Educational and Psychological Testing*, Washington DC: American Educational Research Association

*Ardila, A., Rosselli, M., Matute, E. and Guajardo, S. (2005). The influence of the parents' educational level on the development of executive functions. *Developmental Neuropsychology*, 28(1), 539–560.

*Asfaha, Y. M., Beckman, D., Kurvers, J. and Kroon, S. (2009). L2 reading in multilingual Eritrea: The influences of L1 reading and English proficiency. *Journal of Research in Reading*, 32(4), 351–365.

*Asfaha, Y. M., Kurvers, J. and Kroon, S. (2009). Grain size in script and teaching: Literacy acquisition in Ge'ez and Latin. *Applied Psycholinguistics*, 30(4), 709–724.

*Aturupane, H., Glewwe, P., Ravina, R., Sonnadara, U. and Wisniewski, S. (2014). An assessment of the impacts of Sri Lanka's programme for school improvement and school report card programme on students' academic progress. *The Journal of Development Studies*, 50(12), 1647–1669.

*Babayigit, S. and 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.

*Baydar, N., Küntay, A. C., Yagmurlu, B., Aydemir, N., Cankaya, D., Göksen, F. and Cemalcilar, Z. (2013). 'It takes a village' to support the vocabulary development of children with multiple risk factors. *Developmental Psychology*. doi: 10.1037/a0034785

*Bekman, S., Aksu-Koc, A. and Erguvanli-Taylan, E. (2011). Effectiveness of an intervention program for six-year-olds: A summer-school model. *European Early Childhood Education Research Journal*, 19(4), 409–431.

Berninger, V. W. (1996). Multiple constraints and shared subsystems in writing acquisition. In V. W. Berninger. *Reading and writing acquisition: A developmental neuropsychological approach*. (pp. 129–152). Oxford, UK: Westview Press.

Berninger, V. W. (1999). Coordinating transcription and text generation in working memory during composing: Automatic and constructive processes. *Learning Disability Quarterly*, 22, 99–112.

*Berry, C. (2001). Achievement effects of multigrade and monograde primary schools in the Turks and Caicos Islands. *International Journal of Educational Development*, 21(6), 537–552.

Biemiller, A. (2015, Summer). Which words are worth teaching? Perspectives on Language and Literacy. Retrieved from http://www.onlinedigeditions.com/article/Which+Words+Are+Worth+Teaching%3F/2244530/0/article.html#

*Brouwers, S. A., Mishra, R. C. and van de Vijver, F. J. (2006). Schooling and everyday cognitive development among Kharwar children in India: A natural experiment. *International Journal of Behavioral Development*, 30(6), 559–567.

*Castilla, A. P. (2008). Developmental measures of morphosyntactic acquisition in Monolingual 3-, 4-, and 5-year-old Spanish-speaking children. Dissertation Abstracts International: Section B: The Sciences and Engineering, 71(4-B), 2362.

*Chinyama, A., Svesve, B., Gambiza, B., Guajardo, J., Onunda, D. and Dowd, A. J. (2012). Literacy boost Zimbabwe: Baseline report. Zimbabwe: Save the Children.

*Chowdhury, A. M. R., Ziegahn, L., Haque, N., Shrestha, G. L. and Ahmed, Z. (1994). Assessing basic competences – A practical methodology. *International Review of Education*, 40(6), 437–454.

*Clarkson, P. C. (1993). The effects of bilingualism on examination scores: A different setting. *RELC Journal: A Journal of Language Teaching and Research in Southeast Asia*, 24(1), 109–117. doi: 10.1177/003368829302400107

*Clarkson, P. C. and Galbraith, P. (1992). Bilingualism and mathematics learning: Another perspective. *Journal for Research in Mathematics Education*, 23(1), 34–44.

Clay, M. M. (2000). *Concepts about print*. Auckland, New Zealand: Heinemann.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

*Crookston, B. T., Forste, R., McClellan, C., Georgiadis, A. and Heaton, T. B. (2014). Factors associated with cognitive achievement in late childhood and adolescence: The Young Lives cohort study of children in Ethiopia, India, Peru, and Vietnam. *BMC Pediatrics*, 14(253).

*Cueto, S., Leon, J., Guerrero, G. and 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

*Davidson, M. and Hobbs, J. (2013). Delivering reading intervention to the poorest children: The case of Liberia and EGRA-Plus, a primary grade reading assessment and intervention. *International Journal of Educational Development*, 33, 283–293.

*De Sousa, D. S., Greenop, K. and Fry, J. (2010). The effects of phonological awareness of Zulu-speaking children learning to spell in English: A study of cross-language transfer. *British Journal of Educational Psychology*, 80(4), 517–533.

*Dixon, P., Schagen, I. and Seedhouse, P. (2011). The impact of an intervention on children's reading and spelling ability in low-income schools in India. *School Effectiveness and School Improvement*, 22(4), 461–482.

Dubeck, M. and Gove, A. (2015). The EGRA: Its theoretical foundation, purpose and limitations. *International Journal of Educational Development*, 40.

*Elbeheri, G. and Everrett, J. (2007). Literacy ability and phonological processing skills amongst dyslexic and non-dyslexic speakers of Arabic. *Reading and Writing: An Interdisciplinary Journal*, 20(3), 273–294.

*Elmonayer, R. (2012). Promoting phonological awareness skills of Egyptian kindergarteners through dialogic reading. *Early Child Development and Care*, 183(9), 1229–1241.

Education Testing Service (2014) *ETS Standards for Quality and Fairness* (SQF). www.ets.org/s/about/pdf/standards.pdf

*Farukh, A. and Vulchanova, M. (2014). Predictors of reading in Urdu: Does deep orthography have an impact? *Dyslexia*, 20, 146–166.

*Fedda, O. D. and Oweini, A. (2012). The effect of diglossia on Arabic vocabulary development in Lebanese students. *Educational Research and Reviews*, 7(16), 351–361.

Fernald, L. C. H., Kariger, P., Engle, P. and Raikes, A. (2009). *Examining early child development in low-income countries: A toolkit for the assessment of children in the first five years of life*. Washington, DC: The World Bank. Available at: http://siteresources.worldbank.org/INTCY/Resources/395766-1187899515414/Examining_ECD_Toolkit_FULL.pdf

*Fernald, L., Webber, A., Galasso, E. and Ratsifandrihamanana, L. (2011). Socioeconomic gradients and child development in a very low income population: Evidence from Madagascar. *Developmental Science*, 14(4), 832–847.

*Friedlander, E., Zalila, A., Kasuba, K. and Sichamba, B. (2014). *Literacy boost Zambia: Baseline Report*. Zambia: Save the Children.

Gough, P. B. and Tumner, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7, 6–10.

Grainger, J., Dufau, S. and Zeigler, J. (2016). A Vision of Reading. *Trends in Cognitive Sciences*, 1529, http://dx.doi.org/10.1016/j.tics.2015.12.008

*Griffin, P. and Anh, P. N. (2005). Assessment of creative writing in Vietnamese primary education. *Asia Pacific Education Review*, 6(1), 72–86.

*Griffin, P. and Thanh, M. T. (2006). Reading achievements of Vietnamese grade 5 pupils. *Assessment in Education: Principles, Policy and Practice*, 13(2), 155–177.

*Guild, D. E. (2000). The relationship between early childhood education and primary school academic achievement in Solomon Islands. *International Journal of Early Childhood*, 32(1), 1–8.

*Hopkins, S., Ogle, G. Kaleveld, L., Maurice, J., Keria, B., Louden, W. and Rohl, M. (2005). 'Education for equality' and 'education for life': Examining reading literacy and reading interest in Papua New Guinea primary schools. *Asia-Pacific Journal of Teacher Education*, 33(1), 77–96.

*Hoxhallari, L., Van Daal, V. and Ellis, E. (2004). Learning to read words in Albanian: A skill easily acquired. *Scientific Studies of Reading*, 8(2), 153–166.

*Hungi, N. (2008). Examining differences in mathematics and reading achievement among grade 5 pupils in Vietnam. *Studies in Educational Evaluation*, 34(3), 155–164.

*Jere-Folotiya, J., Chansa-Kabali, T., Munachaka, J. C., Sampa, F., Yalukanda, C., Westerholm, J., . . . Lyytinen, H. (2014). The effect of using a mobile literacy game to improve literacy levels of grade one students in Zambian schools. *Education Tech Research Dev*, 62, 417–436.

*Johnson, D. (2003). Activity theory, mediated action and literacy: Assessing how children make meaning in multiple modes. *Assessment in Education: Principles, Policy and Practice*, 10(1), 103–129.

*Johnson, D., Hayter, J. and 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.

*Jukes, M., Vagh, S. and Kim, Y. (2006). *Development of assessments of reading ability and classroom behaviour*. Washington, DC: World Bank.

*Jukes, M. C. H. and Grigorenko, E. L. (2010). Assessment of cognitive abilities in multiethnic countries: The case of the Wolof and Mandinka in the Gambia. *British Journal of Educational Psychology*, 80(1), 77–97.

*Kalia, V. (2007). Assessing the role of book reading practices in Indian bilingual children's English language and literacy development. *Early Childhood Education Journal*, 35(2), 149–153.

*Kalia, V. (2009). *English oral language, narrative, and literacy development in Indian bilingual pre-school children: Exploring the role of home literacy environment*. Dissertation Abstracts International Section A: Humanities and Social Sciences, 70(11-A), 4172.

*Kalia, V. and Reese, E. (2009). Relations between Indian children's home literacy environment and their English oral language and literacy skills. *Scientific Studies of Reading*, 13(2), 122–145.

*Kormi-Nouri, R., Moradi, A.-R., Moradi, S., Akbari-Zardkhaneh, S. and Zahedian, H. (2012). The effect of bilingualism on letter and category fluency tasks in primary school children: Advantage or disadvantage? *Bilingualism: Language and Cognition*, 15(2), 351–364.

*Lakshminarayana, R., Eble, A., Bhakta, P., Frost, C., Boone, P., Elbourne, D. and Mann, V. (2013). The support to rural India's public education system (STRIPES) trial: A cluster randomised controlled trial of supplementary teaching, learning material and material support. *PLoS ONE*, 8(7), e65775. doi:10.1371/journal.pone.0065775

*Ledesma, H. M. L. (2002). *Language factors influencing early reading development in bilingual (Filipino–English) boys*. Dissertation Abstracts International Section A: Humanities and Social Sciences, 63(6-A), 2096.

*Lee, L. and Wheldall, K. (2011). Acquisition of Malay word recognition skills: Lessons from low-progress early readers. *Dyslexia*, 17(1), 19–37.

*LeVine, R., LeVine, S., Schnell-Anzola, B., Rowe, M. L. and Dexter, E. (2012). *Literacy and mothering: How women's schooling changes the lives of the world's children*. Oxford, UK: Oxford University.

*Mahapatra, S., Das, J., Stack-Cutler, H. and Parrila, R. (2010). Remediating reading comprehension difficulties: A cognitive processing approach. *Reading Psychology*, 31(5), 428–453.

*Mishra, R. and Stainthorp, R. (2007). The relationship between phonological awareness and word reading accuracy in Oriya and English: A study of Oriya-Speaking fifth-graders. *Journal of Research in Reading*, 30(1), 23–37.

*Mohamed, W., Elbert, T. and Landerl, K. (2011). The development of reading and spelling abilities in the first 3 years of learning Arabic. *Reading and Writing: An Interdisciplinary Journal*, 24(9), 1043–1060.

*Mohsin, M., Nath, S. R. and Chowdhury, A. M. R. (1996). Influence of socioeconomic factors on basic competencies of children in Bangladesh. *Journal of Biosocial Science*, 28(1), 15–24.

*Moore, A. C., Akhter, S. and Aboud, F. E. (2008). Evaluating an improved quality preschool program in rural Bangladesh. *International Journal of Educational Development*, 28(2), 118–131.

*Mwaura, P., Silva, K. and Malmberg, L.-E. (2008). Evaluating the madrasa preschool programme in East Africa: A quasi-experimental study. *International Journal of Early Years Education*, 16(3), 237–255.

*Nag, S. (2007). Early reading in Kannada: The pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading*, 30(1), 7–22.

*Nag, S. (2013). Low literacy attainments in school and approaches to diagnosis: An exploratory study. *Contemporary Education Dialogue*, 10(2), 197–221.

Nag, S., Chiat, S., Torgerson, C. and Snowling, M. J. (2014). *Literacy, foundation learning and assessment in developing countries: Final report*. London, UK: EPPI-Centre, Social Science Research Unit, University of London.

*Nag, S. and Snowling, M. J. (2011). Cognitive profiles of poor readers of Kannada. *Reading and Writing: An Interdisciplinary Journal*, 24(6), 657–676.

*Nag, S. and Snowling, M. J. (2012). Reading in an alphasyllabary: Implications for a language-universal theory of learning to read. *Scientific Studies of Reading*, 16(5), 404–423.

Nag, S., Snowling, M.J. and Asfaha, Y. (2016). Classroom literacy practices in low- and middle-income countries: an interpretative synthesis of ethnographic studies. *Oxford Education Review*, 42(1), 36–54. doi: 10.1080/03054985.2015.1135115

*Nag, S., Treiman, R. and Snowling, M. (2010). Learning to spell in an alphasyllabary: The case of Kannada. *Writing Systems Research*, 2(1), 1–12.

*Nag-Arulmani, S., Reddy, V. and Buckley, S. (2003). Targeting phonological representations can help in the early stages of reading in a non-dominant language. *Journal of Research in Reading*, 26(1), 49–68.

*Nakamura, P. (2014). *Facilitating Reading Acquisition in Multilingual Environments in India (FRAME-India): Final report*. American Institutes for Research.

*Nonoyama-Tarumi, Y. and Bredenberg, K. (2009). Impact of school readiness program interventions on children's learning in Cambodia. *International Journal of Educational Development*, 29(1), 39–45.

*Ocampo, D. J. (1996). Development of an early reading program for daycare centers in urban poor communities in the Philippines. Retrieved from the 'Literacy Online' website: www.literacyonlineorg

*Oktay, A. and Aktan, E. (2002). A cross-linguistic comparison of phonological awareness and word recognition in Turkish and English. *International Journal of Early Years Education*, 10(1), 37–48.

*Opel, A., Ameer, S. S. and Aboud, F. E. (2009). The effect of preschool dialogic reading on vocabulary among rural Bangladeshi children. *International Journal of Educational Research*, 48(1), 12–20.

*Paxson, C. and Schady, N. (2007). Cognitive development among young children in Ecuador: The roles of wealth, health, and parenting. *The Journal of Human Resources*, 42(1), 49–84.

Perfetti, C. and Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18, 22–37.

*Pinto, C. (2010). *Literacy boost Kailali Nepal (Year 1 report)*. Kailali, Nepal: Save the Children.

*Piper, B. (2010). *Ethiopia early grade reading assessment (Data analysis report)*. Research Triangle Park, NC: RTI.

*Piper, B. and Korda, M. (2011). *EGRA Plus: Liberia (Program evaluation report)*. Research Triangle Park, NC: RTI.

*Piper, B., Zuilkowski, S. S. and Mugenda, A. (2014). Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. *International Journal of Educational Development*, 37, 11–21.

Pisani, L., Borisova, I. and Dowd, A., (2010). *International Development and Early Learning Assessment Technical Working Paper*. Save the Children.

*Pretorius, E. and Currin, S. (2010). Do the rich get richer and the poor poorer? The effects of an intervention programme on reading in the home and school language in a high poverty multilingual context. *International Journal of Educational Development*, 30(1), 67–76.

*Ramchandra, V. and Karanth, P. (2007). The role of literacy in the conceptualization of words: Data from Kannada-speaking children and non-literate adults. *Reading and Writing: An Interdisciplinary Journal*, 20(3), 173–199.

*Rao, N., Sun, J., Zhou, J. and Zhang, L. (2012). Early achievement in rural China: The role of preschool experience. *Early Childhood Research Quarterly*, 27(1), 66–76.

*Rochdi, A. (2010). *Developing pre-literacy skills via shared book reading: The effect of linguistic distance in a diglossic context*. Dissertation Abstracts International: Section B: The Sciences and Engineering, 70(8-B), 4801.

*Rolla San Francisco, A., Arias, M., Renata, V. and Snow, C. (2006). Evaluating the impact of different early literacy interventions on low-income Costa Rican kindergarteners. *International Journal of Educational Research*, 45(3), 188–201.

*Rolleston, C. and Krutikova, S. (2014). Equalising opportunity? School quality and home disadvantage in Vietnam. *Oxford Review of Education*, 40(1), 112–131.

*Rout, E. L. (2001). Learning how to read. *Psychological Studies*, 46(1-2), 34–39.

*Schagen, I. and Shamsan, Y. (2007). *Analysis of Hyderabad data from 'Jolly Phonics' Initiative to Investigate its Impact on Pupil Progress in reading and Spelling – India*. Slough: National Foundation for Educational Research.

*Sen, R. and Blatchford, P. (2001). Reading in a second language: Factors associated with progress in young children. *Educational Psychology*, 21(2), 189–202.

*Shah-Wundenberg, M., Wyse, D. and Chaplain, R. (2012). Parents helping their children to read: The effectiveness of paired reading and hearing reading in a developing country context. *Journal of Early Childhood Literacy*, 13(4), 471–500.

*Sharma, R. (1997). Dynamics of learning three R's in Madhya Pradesh. *Economic and Political Weekly*, 32 (17), 891–901.

*Sharma, U. (2014, July). *Can computers increase human capital in developing countries? An evaluation of Nepal's one laptop per child program*. Paper presented at the Annual Meeting of Agricultural and Applied Economics Association, Minneapolis, Minnesota.

* Singh, A. (2014). Test score gaps between private and government sector students at school entry age in India. *Oxford Review of Education*, 40(1), 30–49.

*Spratt, J., Seckinger, B. and Wagner, D. (1991). Functional literacy in Moroccan school children. *Reading Research Quarterly*, 26(2), 178–195.

*Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., . . . Bundy, D. A. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence*, 30(2), 141–162.

*Strasser, K. and Lissi, M. R. (2009). Home and instruction effects on emergent literacy in a sample of Chilean kindergarten children. *Scientific Studies of Reading*, 13(2), 175–204.

*Tahan, S., Cline, T. and Messaoud-Galusi, S. (2011). The relationship between language dominance and pre-reading skills in young bilingual children in Egypt. *Reading and Writing: An Interdisciplinary Journal*, 24(9), 1061–1087.

*Tambulkani, G. and Bus, A. G. (2011). Linguistic diversity: A contributory factor to reading problems in Zambian schools. *Applied Linguistics*, 33(2), 141–160.

Treiman, R. and Kessler, B. (2014). *How children learn to write words*. New York, NY: Oxford University Press.

UIS (2016). *Understanding what works in oral reading assessments: Recommendations from donors, implementers and practitioners*. Montreal, Canada: UNESCO Institute for Statistics.

*Vagh, S. B. (2009). *Learning at home and at school: A longitudinal study of Hindi language and emergent literacy skills of young children from low-income families in India*. Dissertation Abstracts International Section A: Humanities and Social Sciences, 70(11-A), 4183.

*Veii, K. and Everatt, J. (2005). Predictors of reading among Herero–English bilingual Namibian school children. *Bilingualism: Language and Cognition*, 8(3), 239–254.

*Wagner, D. A. (1993). *Literacy, culture, and development: Becoming literate in Morocco*. New York, NY: Cambridge University.

Westbrook, J., Durrani, N., Brown. R., Orr, D., Pryor, J., Boddy, J. and Salvi, F. (2013). *Pedagogy, curriculum, teaching practices and teacher education in developing countries: final report* (ISBN: 978-1-907345-64-7). Education Rigorous Literature Review. London, UK: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

*Williams, E. (1993). First and second language reading proficiency of year 3, 4, and 6 children in Malawi and Zambia. *Reading in a Foreign Language*, 10(1), 915–929.

*Williams, E. (1998). *Investigating bilingual literacy: Evidence from Malawi and Zambia* (Education Research Paper, p. 110). London, UK: Department for International Development (DFID).

*Williams, E. (2007). Extensive reading in Malawi: Inadequate implementation or inappropriate innovation? *Journal of Research in Reading*, 30(1), 59–79.

*Winch, C. and Gingell, J. (1994). Dialect interference and difficulties with writing: An investigation in St. Lucian primary schools. *Language and Education* (pp. 157–182).

*Winskel, H. and Widjaja, V. (2007). Phonological awareness, letter knowledge and literacy development in Indonesian beginner readers and spellers. *Applied Psycholinguistics*, 28(1), 23–45.

Annex A List of measures by area of assessment

A.1 Emergent literacy

S no.	Author(s) (Year). *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Chinyama *et al*. (2012) *Zimbabwe, Shona* Low SES, L1	CAP (Grade 3)
2	Davidson, M. and Hobbs, J. (2013) Piper, B. and Korda, M. (2011)* *Liberia, English* Low SES, L2	Orientation to print (EGRA) (reported in *) (Grades 2–3)
3	Davidson, M. and Hobbs, J. (2013) Piper, B. and Korda, M. (2011) *Liberia, English* Low SES, L2	Unfamiliar words: nonsense or pseudo-words (EGRA) (Grades 2–3)
4	Friedlander *et al*. (2014) *Zambia, Bemba* (inferred) Low SES, L1	CAP (Grade 3)
5	Jere-Folotiya *et al*. (2014) *Zambia, CiNyanja* Mixed SES, Mixed	CiNyanja orthography test (Grade 1)
6	Kalia, V. (2007) *India, English* Middle/Upper SES, L2	Awareness of print (print concepts and book handling skills, Clay's CAP test). (Preschool)
7	Kalia, V. (2009) Kalia, V. and Reese, E. (2009) *India, English* Middle/Upper SES, L2	Emergent literacy (Preschool)
8	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Concept of print (Grades 1–5)
9	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Concept of print (Grades 1–5)
10	Ocampo, D. J. (1996) *Philippines, Filipino* Low SES, L1	The reading readiness test (Preschool)
11	Pinto, C. (2010) *Nepal, Nepali* Low SES, Mixed	CAP (Grade 2)
12	Rao *et al*. (2012) *China, Mandarin Chinese* Low SES, Not clear	School Readiness Composite (with literacy component) (Grade 1)
13	Rao *et al*. (2012) *China, Mandarin Chinese*	School readiness composite (without literacy component – Chinese characters subtest)

	Low SES, Not clear	(Grade 1)
14	Rochdi, A. (2010) *Morocco, Modern Standard Arabic* Low SES, L2	The moving word problem (Out-of-school children)
15	Rochdi, A. (2010) *Morocco, Modern Standard Arabic* Low SES, L2	Book-related concepts (Out-of-school children)
16	Rolla San Francisco *et al*. (2006) *Costa Rica, Spanish* Low SES, L1	CAP (Preschool)
17	Sen, R. and Blatchford, P. (2001) *India, English* Mixed SES, L2	CAP (Preschool and Grade 1)
18	Strasser, K. and Lissi, M. (2009) *Chile, Spanish* Mixed SES, L1	Emergent writing (Preschool and Grade 1)
19	Vagh, S .B. (2009) *India, Hindi* Low SES, Mixed	The grapheme concept task (Preschool)
20	Vagh, S. B. (2009) *India, Hindi* Low SES, Mixed	The word concept task (Preschool)
21	Vagh, S .B. (2009) *India, Hindi* Low SES, Mixed	The book task (Preschool)
22	Vagh, S. B. (2009) *India, Hindi* Low SES, Mixed	*Akshara*/grapheme writing task (Preschool)

A.2 Symbol knowledge

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (grades targeted)
1	Alcock *et al*. (2000) *Tanzania, Kiswahili* Low SES, L1	Kiswahili letter reading (Grades 2–5)
2	Alcock *et al*. (2000) *Tanzania, Kiswahili* Low SES, L1	Kiswahili Wide Range Achievement Test (WRAT) (Grades 1–5)
3	Alcock *et al*. (2010) *Tanzania, Kiswahili* Low SES, L1	Kiswahili reading task (letter recognition) (Grades 1–2)
4	Asfaha *et al*. (2009) *Eritrea, Saho* Mixed SES, L1	Saho letter knowledge (Grades 1 and 4)
5	Asfaha *et al*. (2009) *Eritrea, Kunama* Mixed SES, L1	Kumana letter knowledge (Grades 1 and 4)

6	Asfaha *et al*. (2009) *Eritrea, Tigrinya* Mixed SES, L1	Tigrinya letter knowledge (Grades 1 and 4)
7	Asfaha *et al*. (2009) *Eritrea, Tigre* Mixed SES, L1	Tigre letter knowledge (Grades 1 and 4)
8	Chinyama *et al*. (2012) *Zimbabwe, Shona* Low SES, L1	Shona letter knowledge (inferred) (Grade 3)
9	Chinyama *et al*. (2012) *Zimbabwe, English* Low SES, L2	English letter knowledge (inferred) (Grade 3)
10	Davidson, M. and Hobbs, J. (2013) Piper, B. and Korda, M. (2011) *Liberia, English* Low SES, L2	English letter-naming fluency (EGRA) (Grades 2–3)
11	Elbeheri, G. and Everrett, J. (2007) *Egypt, Arabic* Low SES, L1	Arabic grapheme discrimination (Grades 4–5)
12	Friedlander *et al*. (2014) *Zambia, not clear* Low SES, L1	Letter knowledge (Grade 3)
13	Jere-Folotiya *et al*. (2014) *Zambia, CiNyanja* Mixed SES, Mixed	CiNyanja orthography test (Grade 1)
14	Jukes et. al. (2006) *Kenya, Swahili* Mixed SES, L1	Swahili letter reading fluency (Grades 2–3)
15	Jukes *et al*. (2006) *Kenya, English* Mixed SES, L2	English letter reading fluency (Grades 2–3)
16	Kalia, V. (2009) *India, English* Middle/Upper SES, L2	English letter identification (Preschool)
17	Kormi-Nouri *et al*. (2012) *Iran, Persian* Mixed SES, Mixed	Persian letter fluency task (Grades 1–5)
18	Ledesma, H. M. (2002) *Philippines, English* Mixed SES, Mixed	English/Filipino letter naming (Preschool and Grade 1)
19	Lee, L. and Wheldall, K. (2011) *Malaysia, Malay* Middle/Upper SES, L1	Malay letter knowledge test (Grade 1)
20	Nag, S. (2007) *India, Kannada* Mixed SES, L1	Kannada *akshara* knowledge (Grades 1–4)
21	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada *akshara* knowledge (Grades 4–6)

22	Nag, S. and Snowling, M. (2012) *India, Kannada* Mixed SES, L1	Kannada *akshara* knowledge (Grades 4–6)
23	Nag-Arulmani *et al*. (2003) *India, English* Middle/Upper SES, L2	English letter–sound correspondence (Grade 3)
24	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada letter (*akshara*) naming test (Grades 1–5)
25	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu letter (*akshara*) naming test (Grades 1–5)
26	Nakamura, P. (2014) *India, English* Mixed SES, L2	English letter-naming test (Grades 1–5)
27	Oktay, A. and Aktan, E. (2002) *Turkey, Turkish* Middle/Upper SES, L1	Turkish letter identification task (Preschool and Grade 1)
28	Oktay, A. and Aktan, E. (2002) *Turkey, English* Middle/Upper SES, L2	English letter identification task (Preschool and Grade 1)
29	Oktay, A. and Aktan, E. (2002) *Turkey, Turkish* Middle/Upper SES, L1	Turkish letter usage task (Preschool and Grade 1)
30	Oktay, A. and Aktan, E. (2002) *Turkey, English* Middle/Upper SES, L2	English letter usage task (Preschool and Grade 1)
31	Pinto, C. (2010) *Nepal, Nepali* *Low SES*	Nepali letter identification (Grade 2)
32	Piper, B. (2010) *Ethiopia, Tigrinya* Low SES, L1	Tigrinya *fidel* identification fluency (Sabean script) (from EGRA battery) (Grades 2–3)
33	Piper, B. (2010) *Ethiopia, Amharic* Low SES, L1	Amharic *fidel* identification fluency (Sabean script) (from EGRA battery) (Grades 2–3)
34	Piper, B. (2010) *Ethiopia, Hararigna* Low SES, L1	Hararigna *fidel* identification fluency (Sabean script) (from EGRA battery) (Grades 2–3)
35	Piper, B. (2010) *Ethiopia, Afan Oromo* Low SES, L1	Afan Oromo letter-naming fluency (from EGRA battery) (Grades 2–3)
36	Piper, B. (2010) *Ethiopia, Somaligna* Low SES, L1	Somaligna/Somali letter-naming fluency (from EGRA battery) (Grades 2–3)
37	Piper, B. (2010) Ethiopia, Sidaamu Afoo Low SES, L1	Sidaamu Afoo letter-naming fluency (from EGRA battery) (Grades 2–3)
38	Rolla San Francisco *et al*. (2006)	Spanish letter identification

	Costa Rica, Spanish Low SES, L1	(Preschool)
39	Schagen, I. and Shamsan, Y. (2007) Dixon *et al*. (2011) *India, English* Low SES, L2	NFER – A (Grade 1)
40	Schagen, I. and Shamsan, Y. (2007) Dixon *et al*. (2011) *India, English* Low SES, L2	NFER – B (Grade 1)
41	Schagen, I. and Shamsan, Y. (2007) Dixon *et al*. (2011) *India, English* Low SES, L2	NFER – C (Grade 1)
42	Sen, R. and Blatchford, P. (2001) *India, English* Mixed SES, L2	English letter naming (Preschool and Grade 1)
43	Sen, R. and Blatchford, P. (2001) *India, English* Mixed SES, L2	English letter and word association (Preschool and Grade 1)
44	Sousa *et al*. (2010) *South Africa, English* Mixed SES, L2	English letter name (Grade 2)
45	Sousa *et al*. (2010) *South Africa, English* Mixed SES, L2	English letter sound (Grade 2)
46	Sousa *et al*. (2010) *South Africa, Zulu* Mixed SES, L1	Zulu letter name (Grade 2)
47	Sousa *et al*. (2010) *South Africa, Zulu* Mixed SES, L1	Zulu letter sound (Grade 2)
48	Sousa *et al*. (2010) *Chile, Spanish* (Mixed SES, L1)	The Woodcock–Muñoz letter identification subtest (Preschool and Grade 1)
49	Tahan *et al*. (2011) *Egypt, Arabic* Middle/Upper SES, L1	Arabic orthographic recognition task (Preschool)
50	Vagh, S. B. (2009) *India, Hindi* Low SES, L1	Knowledge of the Hindi alphasyllabary (ALPHA) – identification (Preschool)
51	Vagh, S. B. (2009) *India, Hindi* Low SES, L1	Knowledge of the Hindi alphasyllabary (ALPHA) – writing (Preschool)
52	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* *Low SES, L1*	Moroccan Arabic letter concept task (Preschool, Grade 1, Grade 5)
53	Wagner, D. A. (1993) *Morocco, Moroccan Arabic*	Moroccan Arabic letter boundary task

	Low SES, L1	(Preschool, Grade 1, Grade 5)
54	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* *Low SES, L1*	Moroccan Arabic letter recognition for positional variants (Preschool, Grade 1, Grade 5)
55	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* *Low SES, L1*	Moroccan Arabic letter form recognition (Preschool, Grade 1, Grade 5)
56	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* *Low SES, L1*	Moroccan Arabic letter identification (name or sound) (Preschool, Grade 1, Grade 5)
57	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* *Low SES, L1*	Moroccan Arabic letter-vowel (CCV) pronunciation (Preschool, Grade 1, Grade 5)
58	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia* Mixed SES, L1	Bahasa Indonesia letter knowledge (Grades 1–2)

A.3 Reading accuracy

S No.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (grades targeted)
1	Alcock *et al*. (2000) *Tanzania, Kiswahili* Low SES, L1	Kiswahili translation of WRAT (Grades 1–5)
2	Alcock *et al*. (2000) *Tanzania, Kiswahili* Low SES, L1	Kiswahili school-based reading tests (includes words, non-words) (Grades 1–5)
3	Alcock *et al*. (2000) *Tanzania, Kiswahili* Low SES, L1	Kiswahili word reading tests (word–pseudo-word discrimination) (Grades 2–5)
4	Asfaha *et al.* (2009) *Eritrea, English* Mixed SES, L2	English word reading test (Grade 4)
5	Asfaha *et al.* (2009) *Eritrea, Arabic* Mixed SES, Mixed	Arabic word reading test (L1) (Grade 4)
6	Asfaha *et al.* (2009) *Eritrea, Kunama* Mixed SES, L1	Kunama word reading test (Grade 4)
7	Asfaha *et al.* (2009). *Eritrea, Saho* Mixed SES, L1	Saho word reading test (Grade 4)
8	Asfaha *et al.* (2009) *Eritrea, Tigre* Mixed SES, Mixed	Tigre word reading test (Grade 4)
9	Asfaha *et al.* (2009). Eritrea, Tigrinya Mixed SES, Mixed	Tigrinya word reading test (Grade 4)

10	Asfaha *et al*. (2009) *Eritrea, Kunama* Mixed SES, Mixed	Kunama word reading (Grades 1 and 4)
11	Asfaha *et al*. (2009) *Eritrea, Saho* Mixed SES, Mixed	Saho word reading (Grades 1 and 4)
12	Asfaha *et al*. (2009) *Eritrea, Tigre* Mixed SES, Mixed	Tigre word reading (Grades 1 and 4)
13	Asfaha *et al*. (2009) *Eritrea, Tigrinya* Mixed SES, Mixed	Tigrinya word reading (Grades 1 and 4)
14	Babayigit, S. and Stainthorp, R. (2010). *Northern Cyprus, Turkish* Mixed SES, L1	Text reading accuracy (Grades 1–2)
15	Chinyama *et al*. (2012) *Zimbabwe, Shona* Low SES, L1	Shona single word reading of MUW (Grade 3)
16	Chinyama *et al*. (2012) *Zimbabwe, English* Low SES, L2	English single word reading of MUW (Grade 3)
17	Chinyama *et al*. (2012) *Zimbabwe, Shona* Low SES, L1	Shona accuracy (Grade 3)
18	Chinyama *et al*. (2012) *Zimbabwe, English* Low SES, L2	English accuracy (Grade 3)
19	Chowdhury *et al*. (1994) *Bangladesh, Bengali* Mixed SES, L1	Reading Bengali words (Grade 5)
20	Chowdhury *et al*. (1994) *Bangladesh, Bengali* Mixed SES, L1	Reading Bengali sentences (Grade 5)
21	Davidson, M. and Hobbs, J. (2013) Piper, B. and Korda, M. (2011) *Liberia, English* Low SES, L	Familiar words test (high-frequency sight words subtest in EGRA) (Grades 2 and 3)
22	Elbeheri, G. and Everrett, J. (2007) *Egypt, Arabic* Low SES, L1	Arabic word chain (Grades 4–5)
23	Elbeheri, G. and Everrett, J. (2007) *Egypt, Arabic* Low SES, L1	Arabic pseudo-word reading (Grades 4–5)
24	Farukh, A. and Vulchanova, M. (2014) *Pakistan, Urdu* Mixed SES, L2	Urdu reading accuracy (Grade 3)
25	Friedlander *et al*. (2014) *Zambia, Bemba* Low SES, L1	Bemba word recognition (Grade 3)

26	Friedlander *et al*. (2014) *Zambia, English* Low SES, L2	English word recognition (Grade 3)
27	Friedlander *et al*. (2014) *Zambia, Bemba* Low SES, L1	Bemba accuracy (Grade 3)
28	Friedlander *et al*. (2014) *Zambia, English* Low SES, L2	English accuracy (Grade 3)
29	Hopkins *et al*. (2005) *Papua New Guinea, English* Mixed SES, Mixed	The Martin and Pratt non-word reading test (Grades 3,5 and 7)
30	Hopkins *et al*. (2005) *Papua New Guinea, English* Mixed SES, L2	The Burt word reading test (Grades 3,5 and 7)
31	Hoxhallari *et al*. (2004) *Albania, Albanian* Mixed SES, L1	Albanian reading test (Grade 1)
32	Johnson, D. (2003) *Bangladesh, Bengali* Mixed SES, L1	Bengali running record – Reading accuracy (Grade 4)
33	Johnson *et al*. (2000) *Malawi, English* Mixed SES, L2	English reading accuracy (component of a running record) (Grades 1–8)
34	Johnson *et al*. (2000) *Malawi, Chichewa* Mixed SES, L1	Chichewa reading accuracy (component of a running record) (Grades 1–8)
35	Johnson *et al*. (2000) *Sri Lanka, English* Mixed SES, L2	English reading accuracy (component of a running record) (Grades 1–8)
36	Johnson *et al*. (2000) *Sri Lanka,* Sinhalese Mixed SES, L1	Sinhala reading accuracy (component of a running record) (Grades 1–8)
37	Johnson *et al*. (2000) *Sri Lanka,* Tamil Mixed SES, L1	Tamil reading accuracy (component of a running record) (Grades 1–8)
38	Jukes *et al*. (2006) *Kenya, Swahili* Mixed SES, L1	Swahili word reading (similar to a written lexical judgement task) (Grades 2–3)
39	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	Filipino word identification test (Preschool and Grade 1)
40	Ledesma, H. M. (2002) *Philippines, English* Mixed SES, Mixed	English word identification (subtest of the Woodcock Reading Mastery test, Revised: Woodcock, 1987) (Preschool and Grade 1)
41	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	Filipino word attack (non-words and very low-frequency words) (Preschool and Grade 1)

42	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	English word attack (non-words and very low-frequency words) (Preschool and Grade 1)
43	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	English passage accuracy (Preschool and Grade 1)
44	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	English word efficiency (TOWRE, Torgesen, Wagner and Rashotte, 1999) (Preschool and Grade 1)
45	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	English non-word efficiency (TOWRE, Torgesen, Wagner and Rashotte, 1999) (Preschool and Grade 1)
46	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	Filipino word efficiency (similar to TOWRE) (Preschool and Grade 1)
47	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	Filipino non-word efficiency (similar to TOWRE) (Preschool and Grade 1)
48	Lee, L. and Wheldall, K. (2011) *Malaysia, Malay* Middle/Upper SES, L1	Malay word recognition test (Grade 1)
49	Mahapatra *et al*. (2010) *India, English* Middle/Upper SES	English word identification (Grade 4)
50	Mishra, R. and Stainthorp, R. (2007) *India, Oriya* Mixed SES, L1	Das Oriya word reading test (Das and Kendrick, 1997) (Grade 5)
51	Mishra, R. and Stainthorp, R. (2007) *India, Oriya* Mixed SES, L1	Oriya experimental pseudo-word reading test (Grade 5)
52	Mishra, R. and Stainthorp, R. (2007) *India, English* Mixed SES, L2	English word reading test (British Ability Scales, Elliot *et al.*, 1996) (Grade 5)
53	Mishra, R. and Stainthorp, R. (2007) *India, English* Mixed SES, L2	English non-word reading test (the Phonological Assessment Battery, Frederickson *et al*. 1997) (Grade 5)
54	Nag, S. (2007) *India, Kannada* Mixed SES, Mixed	Reading (word and non-word); (subtest from the Literacy Acquisition Battery (LAB), 2004) (Grades 1–4)
55	Nag, S. and Snowling, M. (2012). *India, Kannada* Mixed SES, L1	Kannada reading accuracy (Grades 4–6)
56	Nag, S. and Snowling, M. (2011). *India, Kannada* Mixed SES, L1	Kannada reading accuracy (Grades 4–6)
57	Nag-Arulmani *et al*. (2003) *India, English*	Non-word reading (Grade 3)

	Middle/Upper SES	
58	Nag-Arulmani *et al*. (2003) *India, English* Middle/Upper SES	WORD Single word reading (Grade 3)
59	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada word and non-word decoding test (Grades 1–5)
60	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu word and non-word decoding test (Grades 1–5)
61	Nakamura, P. (2014) *India, English* Mixed SES, L2	English word and non-word decoding test (Grades 1–5)
62	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada Slasher test (Grades 1–5)
63	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu Slasher test (Grades 1–5)
64	Nakamura, P. (2014) *India, English* Mixed SES, L2	English Slasher test (Grades 1–5)
65	Oktay, A. and Aktan, E. (2002) *Turkey, English* Middle/Upper SES, Mixed	English decoding task (Woodcock letter–word identification test) (Preschool and Grade 1)
66	Oktay, A. and Aktan, E. (2002) *Turkey, Turkish* Middle/Upper SES, Mixed	Turkish decoding task (Preschool and Grade 1)
67	Pinto, C. (2010) *Nepal, Nepali* Low SES, Mixed	Reading accuracy (Grade 2)
68	Ramchandra, V. and Karanth, P. (2007) *India, Kannada* Mixed SES, L1	Kannada reading (of self-written material) (Preschool and Grade 1)
69	Ramchandra, V. and Karanth, P. (2007) *India, Kannada* Mixed SES, L1	Kannada word cover (Preschool and Grade 1)
70	Ramchandra, V. and Karanth, P. (2007) *India, Kannada* Mixed SES, L1	Kannada word circle (Preschool and Grade 1)
71	Rolla San Francisco *et al*. (2006) *Costa Rica, Spanish* Low SES, L1	Spanish reading (Spanish version of the Woodcock letter–word identification subtest) (Preschool)
72	Rout, E. L. (2001) *India, Oriya* Middle/Upper SES, L1	Odia graded oral reading (Grades 3, 5, 7)
73	Schagen, I. and Shamsan, Y. (2007) Dixon *et al*. (2011)	The Burt reading test (1974) (Grade 1)

	India, English Low SES, L2	
74	Sen, R. and Blatchford, P. (2001) *India, English* Mixed SES, L2	English word matching (Preschool and Grade 1)
75	Sen, R. and Blatchford, P. (2001) *India, English* Mixed SES, L2	English word reading (Preschool and Grade 1)
76	Sen, R. and Blatchford, P. (2001) *India, English* Mixed SES, L2	Neale Analysis of Reading Ability (Neale, 1989) (Preschool and Grade 1)
77	Sen, R. and Blatchford, P. (2001) *India, English* Mixed SES, L2	Word reading subtest (British Ability Scales; Elliot *et al*., 1979) (Preschool and Grade 1)
78	Shah-Wundenberg *et al*. (2012) *India, English* Low SES, L2	Gates-MacGinitie Reading Test – Level BR (beginning reading skills) (MacGinitie *et al*., 2002) (Grade 1)
79	Shah-Wundenberg *et al*. (2012) *India, English* Low SES, L2	Individual running record (Clay, 1979) on grade-level text (Grade 1)
80	Sternberg *et al*. (2002) *Tanzania, Kiswahili* Low SES, L1	Kiswahili word reading (Grades 2–5)
81	Veii, K. and Everatt, J. (2005) *Namibia, English* Mixed SES, L2	English word reading (Grades 2–5)
82	Veii, K. and Everatt, J. (2005) *Namibia, Herero* Mixed SES, L2	Herero word reading (Grades 2–5)
83	Veii, K. and Everatt, J. (2005) *Namibia, English* Mixed SES, L2	English non-word reading (Grades 2–5)
84	Veii, K. and Everatt, J. (2005) *Namibia, Herero* Mixed SES, L2	Herero non-word reading (Grades 2–5)
85	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* Low SES, L1	Moroccan Arabic word concept task (Preschool, Grade 1, Grade 5)
86	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* Low SES, L1	Moroccan Arabic word boundary task (Preschool, Grade 1, Grade 5)
87	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* Low SES, L1	Moroccan Arabic word decoding test (Grade 1)
88	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* Low SES, L1	Moroccan Arabic word–picture matching test (Grade 1)
89	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia*	Bahasa Indonesia word reading (Grades 1–2)

	Mixed SES, L1	
90	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia* Mixed SES, L1	Bahasa Indonesia non-word reading (Grades 1–2)

A.4 Spelling

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Alcock, K. and Ngorosho, D. (2007) *Kenya, Kiswahili* Low SES, L1	Kiswahili spelling (based on WRAT Spelling: Jastak and Wilkinson, 1984) Grade 5
2	Alcock, K. and Ngorosho, D. (2003) *Tanzania, Kiswahili* Low SES, L1	Spelling task – Study1 (Grades 2–5)
3	Alcock, K. and Ngorosho, D. (2003) *Tanzania, Kiswahili* Low SES, L1	Spelling task – Study 2 (Grades 2–5)
4	Alcock, K. and Ngorosho, D. (2003) *Tanzania, Kiswahili* Low SES, L1	Writing task (Grades 2–5)
5	Asfaha *et al*. (2009) *Eritrea, Saho* Mixed SES, L1	Saho spelling (Grades 1 and 4)
6	Asfaha *et al*. (2009) *Eritrea, Kunama* Mixed SES, L1	Kunama spelling (Grades 1 and 4)
7	Asfaha *et al*. (2009) *Eritrea, Tigrinya* Mixed SES, L1	Tigrinya spelling (Grades 1 and 4)
8	Asfaha *et al*. (2009) *Eritrea, Tigre* Mixed SES, L1	Tigre spelling (Grades 1 and 4)
9	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish single word spelling (Grades 1–2)
10	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish sentence spelling (Grades 1–2)
11	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish composition writing – spelling error rate (Grades 1–2)
12	Chowdhury *et al*. (1994) *Bangladesh, Bangla* Mixed SES, L1	Bangla writing words (Grade 5)
13	Chowdhury *et al*. (1994) *Bangladesh, Bangla*	Bangla writing one's own name (Grade 5)

	Mixed SES, L1	
14	Chowdhury *et al*. (1994) *Bangladesh, Bangla* Mixed SES, L1	Bangla – writing a sentence (Grade 5)
15	Elbeheri, G. and Everett, J. (2007) *Egypt, Arabic* Low SES, L1	Arabic spelling (Grades 4–5)
16	Jere-Folotiya *et al*. (2014) *Zambia, CiNyanja* Mixed SES, Mixed	CiNyanja spelling test (multiple-choice format: Ojanen *et al*., 2013). (Grade 1)
17	Ledesma, H. M. (2002) *Philippines, English* Mixed SES, L1	Filipino spelling (Preschool and Grade 1)
18	Ledesma, H. M. (2002) *Philippines, English* Mixed SES, L2	English spelling (Preschool and Grade 1)
19	Mohamed *et al.* (2011). *Egypt, Arabic* Middle/Upper SES, L1	Arabic spelling (Grades 1–3)
20	Nag-Arulmani *et al*. (2003) *India, English* Middle/Upper SES, L2	WORD spelling (Wechsler Objective Reading Dimensions: Rust *et al.*, 1993) (Grade 3)
21	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada spelling (Grades 4–6)
22	Nag *et al*. (2010) *India, Kannada* Mixed SES, L1	Kannada spelling (Grades 4–5)
23	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada spelling test (Grades 1–5)
24	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu spelling test (Grades 1–5)
25	Nakamura, P. (2014) *India, English* Mixed SES, L2	English spelling test (Grades 1–5)
26	Nonoyama-Tarumi, Y. and Bredenberg, K. (2009) *Cambodia, Khmer* Low SES, L1	Khmer writing (Grade 1)
27	Schagen, I. and Shamsan, Y. (2007) Dixon *et al*. (2011) *India, English* Low SES, L2	English spelling test (Schonell and Schonell, 1952) (Grade 1)
28	Schagen, I. and Shamsan, Y. (2007) Dixon *et al*. (2011) *India, English* Low SES, L2	English dictation test (Grade 1)

29	Sousa *et al*. (2010) *South Africa, Zulu* Mixed SES, L1	Zulu real-word spelling (Grade 2)
30	Sousa *et al*. (2010) *South Africa, English* Mixed SES, L2	English real-word spelling (Grade 2)
31	Sousa *et al*. (2010) *South Africa, Zulu* Mixed SES, L1	Zulu non-word spelling (Grade 2)
32	Sousa *et al*. (2010) *South Africa, English* Mixed SES, L2	English non-word spelling (Grade 2)
33	Sternberg *et al*. (2002) *Tanzania, Kiswahili* Low SES, L1	Kiswahili spelling (Grades 2–5)
34	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia* Mixed SES, L1	Spelling stem words (Grades 1–2)
35	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia* Mixed SES, L1	Spelling affixed words (Grades 1–2)

A.5 Reading fluency

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Alcock *et al*. (2000) *Tanzania, Kiswahili* Low SES, L1	Sentence reading test (Grades 2–5)
2	Asfaha *et al*. (2009) *Eritrea, Saho* Mixed SES, Mixed	Saho number of words read in 3 minutes (Grades 1 and 4)
3	Asfaha *et al*. (2009) *Eritrea, Kunama* Mixed SES, Mixed	Kunama number of words read in 3 minutes (Grades 1 and 4)
4	Asfaha *et al*. (2009) *Eitrea, Tigrinya* *Mixed SES, Mixed*	Tigrinya number of words read in 3 minutes (Grades 1 and 4)
5	Asfaha *et al*. (2009) *Eritrea, Tigre* *Mixed SES, Mixed*	Tigre number of words read in 3 minutes (Grades 1 and 4)
6	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish text reading speed (Grades 1–2)
7	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish word reading (one-minute word reading) (Grades 1–2)
8	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish*	Turkish non-word reading (one-minute word reading)

	Mixed SES, L1	(Grades 1–2)
9	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish agglutinating word reading (one-minute word reading) (Grades 1–2)
10	Chinyama *et al*. (2012) *Zimbabwe, Shona* Low SES, L1	Shona fluency (Grade 3)
11	Chinyama *et al*. (2012) *Zimbabwe, English* Low SES, L2	English fluency (Grade 3)
12	Davidson, M. and Hobbs, J. (2013) Piper, B. and Korda, M. (2011). *Liberia, English* Low SES, L2	Passage/oral reading fluency (connected texts) (EGRA) (Grades 2–3)
13	Friedlander *et al*. (2014) *Zambia, Bemba* Low SES, L1	Bemba fluency (Grade 3)
14	Friedlander *et al*. (2014) *Zambia, English* Low SES, L2	English fluency (Grade 3)
15	Johnson, D. (2003) *Bangladesh, Bangla* Mixed SES, L1	Reading fluency (Grade 4)
16	Jukes *et al*. (2006) *Kenya, English* Mixed SES, L2	English oral reading fluency (Grades 2–3)
17	Jukes *et al*. (2006) *Kenya, Swahili* Mixed SES, L1	Swahili oral reading fluency (Grades 2–3)
18	Jukes *et al*. (2006) *Kenya, English* Mixed SES, L2	English non-word reading efficiency (Grades 2–3)
19	Jukes *et al*. (2006) *Kenya, Swahili* Mixed SES, L1	Swahili non-word reading efficiency (Grades 2–3)
20	Jukes *et al*. (2006) *Kenya, English* Mixed SES, L2	English sentence reading (timed) (Grades 2–3)
21	Jukes *et al*. (2006) *Kenya, Swahili* Mixed SES, L1	Swahili sentence Reading (timed) (Grades 2–3)
22	Mohamed *et al*. (2011) *Egypt, Arabic* Middle/Upper SES, L1	One-minute Arabic word reading test (Grades 1–3)
23	Mohamed *et al*. (2011) *Egypt, Arabic* Middle/Upper SES, L1	One-minute Arabic non-word reading test (Grades 1–3)
24	Nag, S. and Snowling, M. (2012). *India, Kannada*	Kannada reading speed (Grades 4–6)

	Mixed SES, L1	
25	Nag, S. and Snowling, M. (2011). *India, Kannada* Mixed SES, L1	Kannada reading rate (Grades 4–6)
26	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada oral reading fluency test (Grades 1–5)
27	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu oral reading fluency test (Grades 1–5)
28	Nakamura, P. (2014) *India, English* Mixed SES, L2	English oral reading fluency test (Grades 1–5)
29	Pinto, C. (2010) *Nepal, Nepali* Low SES, Mixed	Nepali reading fluency (Grade 2)
30	Piper, B. (2010) *Ethiopia, Tigrinya* Mixed SES, L1	Tigrinya familiar word fluency (EGRA) (Grades 2 and 3)
31	Piper, B. (2010) *Ethiopia, Amharic* Mixed SES, L1	Amharic familiar word fluency (EGRA) (Grades 2 and 3)
32	Piper, B. (2010) *Ethiopia, Hararigna* Mixed SES, L1	Hararigna familiar word fluency (EGRA) (Grades 2 and 3)
33	Piper, B. (2010) *Ethiopia, Afan Oromo* Mixed SES, L1	Afan Oromo familiar word fluency (EGRA) (Grades 2 and 3)
34	Piper, B. (2010) *Ethiopia, Somaligna* Mixed SES, L1	Somaligna familiar word fluency (EGRA) (Grades 2 and 3)
35	Piper, B. (2010) *Ethiopia, Sidaamu Afoo* Mixed SES, L1	Sidaamu Afoo familiar word fluency (EGRA) (Grades 2 and 3)
36	Piper, B. (2010) *Ethiopia, Tigrinya* Mixed SES, L1	Tigrinya nonsense word fluency (decoding fluency, EGRA) (Grades 2 and 3)
37	Piper, B. (2010) *Ethiopia, Amharic* Mixed SES, L1	Amharic nonsense word fluency (decoding fluency, EGRA) (Grades 2 and 3)
38	Piper, B. (2010) *Ethiopia, Hararigna* Mixed SES, L1	Hararigna nonsense word fluency (decoding fluency, EGRA) (Grades 2 and 3)
39	Piper, B. (2010) *Ethiopia, Afan Oromo* Mixed SES, L1	Afan Oromo nonsense word fluency (decoding fluency, EGRA) (Grades 2 and 3)
40	Piper, B. (2010). *Ethiopia, Somaligna* Mixed SES, L1	Somaligna nonsense word fluency (decoding fluency, EGRA) (Grades 2 and 3)

41	Piper, B. (2010) *Ethiopia, Sidaamu Afoo* Mixed SES, L1	Sidaamu Afoo nonsense word fluency (decoding fluency, EGRA) (Grades 2 and 3)
42	Piper, B. (2010) *Ethiopia, Tigrinya* Mixed SES, L1	Tigrinya oral reading fluency (reading connected text, EGRA) (Grades 2 and 3)
43	Piper, B. (2010) *Ethiopia, Amharic* Mixed SES, L1	Amharic oral reading fluency (reading connected text, EGRA) (Grades 2 and 3)
44	Piper, B. (2010) *Ethiopia, Hararigna* Mixed SES, L1	Hararigna oral reading fluency (reading connected text, EGRA) (Grades 2 and 3)
45	Piper, B. (2010) *Ethiopia, Afan Oromo* Mixed SES, L1	Afan Oromo oral reading fluency (reading connected text, EGRA) (Grades 2 and 3)
46	Piper, B. (2010) *Ethiopia, Somaligna* Mixed SES, L1	Somaligna oral reading fluency (reading connected text, EGRA) (Grades 2 and 3)
47	Piper, B. (2010) *Ethiopia, Sidaamu Afoo* Mixed SES, L1	Sidaamu Afoo oral reading fluency (reading connected text, EGRA) (Grades 2 and 3)
48	Piper *et al*. (2014) *Kenya, Kiswahili* Mixed SES, L1	Kiswahili oral language fluency (Grades 1 and 2)
49	Piper *et al*. (2014) *Kenya, English* Mixed SES, L2	English oral language fluency (Grades 1 and 2)
50	Tambulkani, G. and Bus, A. G. (2011) *Zambia, English* Low SES, L2	One minute of reading English words (Grade 1)
51	Tambulkani, G. and Bus, A. G. (2011) *Indonesia, Lozi* Low SES, L1	One minute of reading Lozi words (Grades 1 and 2)
52	Tambulkani, G. and Bus, A. G. (2011) *Zambia, Mbunda* Low SES, L1	One minute of reading Mbunda words (Grade 1)

A.6 Reading comprehension

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Asfaha *et al*. (2009) *Eritrea, Tigre* Mixed SES, Mixed	Tigre reading comprehension (Grade 4)
2	Asfaha *et al*. (2009) *Eritrea, Kunama* Mixed SES, Mixed	Kunama reading comprehension (Grade 4)
3	Asfaha *et al*. (2009) *Eritrea, Saho*	Saho reading comprehension (Grade 4)

	Mixed SES, Mixed	
4	Asfaha *et al*. (2009) *Eritrea, Arabic* Mixed SES, Mixed	Arabic reading comprehension (Grade 4)
5	Asfaha *et al*. (2009) *Eritrea, Tigrinya* Mixed SES, Mixed	Tigrinya reading comprehension (Grade 4)
6	Asfaha *et al*. (2009) *Eritrea, English* Mixed SES, L2	L2 (English) reading comprehension (Grade 4)
7	Berry, C. (2001). *The Turks and Caicos Islands, English* Low SES, L2	The McLeod gap test (Grades 3–5)
8	Chinyama *et al*. (2012) *Zimbabwe, Shona* Low SES, L1	Shona reading comprehension (Grade 3)
9	Chinyama *et al*. (2012) *Zimbabwe, English* Low SES, L2	English reading comprehension (Grade 3)
10	Chowdhury *et al*. (1994) *Bangladesh, Bangla* Mixed SES, L1	Bangla comprehension passage (Grade 5)
11	Clarkson, P. C. (1993) *Papua New Guinea, Pidgin/ Tok Pisin* Mixed SES, Mixed	Pidgin/Tok Pisin Cloze test (Grade 5)
12	Clarkson, P. C. (1993) *Papua New Guinea, English* Mixed SES, L2	English Cloze test (Grade 5)
13	Clarkson, P. and Galbraith, P. (1992) *Papua New Guinea, Pidgin/ Tok Pisin* Mixed SES, Mixed	Pidgin/Tok Pisin Cloze test (Grade 6)
14	Clarkson, P. and Galbraith, P. (1992) *Papua New Guinea, English* Mixed SES, L2	English Cloze test (Grade 6)
15	Davidson, M. and Hobbs, J. (2013) Piper, B. and Korda, M. (2011) *Liberia, English* Low SES, L2	Passage comprehension (EGRA) (Grades 2–3)
16	Friedlander *et al*. (2014) *Zambia, Bemba* Low SES, L1	Bemba reading comprehension (Grade 3)
17	Friedlander *et al*. (2014) *Zambia, English* Low SES, L2	English reading comprehension (Grade 3)
18	Guild, D. E. (2000) *Solomon Islands, English* Mixed SES, L1	English reading comprehension (Grade 2)
19	Hungi, N. (2008). *Vietnam, Vietnamese*	Vietnamese reading test (Grade 5)

	Mixed SES, L1	
20	Johnson, D. (2003) *Bangladesh, Bangla* Mixed SES, L1	Bangla reading comprehension (Grade 4)
21	Jukes *et al*. (2006) *Kenya, English* Mixed SES, L2	English passage comprehension (Grades 2–3)
22	Jukes *et al*. (2006) *Kenya, Swahili* Mixed SES, L1	Swahili passage comprehension (Grades 2–3)
23	Jukes *et al*. (2006) *Kenya, English* Mixed SES, L2	English Maze test (Grades 2–3)
24	Jukes *et al*. (2006) *Kenya, Swahili* Mixed SES, L1	Swahili Maze test (Grades 2–3)
25	Mahapatra *et al*. (2010) *India, English* Middle/Upper SES, L2	English passage comprehension (Grade 4)
26	Nag-Arulmani *et al*. (2003) *India, English* Middle/Upper SES, L2	English reading comprehension (Wechsler Objective Reading Dimensions: Rust *et al.*, 1993) (Grade 3)
27	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada reading comprehension (Grades 4,5 and 6)
28	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada comprehension test (Grades 1–5)
29	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu comprehension test (Grades 1–5)
30	Nakamura, P. (2014) *India, English* Mixed SES, L2	English comprehension test (Grades 1–5)
31	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada passage reading comprehension test (Grades 1–5)
32	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu passage reading comprehension test (Grades 1–5)
33	Nakamura, P. (2014) *India, English* Mixed SES, L2	English passage reading comprehension test (Grades 1–5)
34	Pinto, C. (2010) *Nepal, Nepali* Low SES, Mixed	Nepali reading comprehension (Grade 2)

35	Piper, B. (2010) *Ethiopia, Tigrinya* Mixed SES, L2	Tigrinya reading comprehension (EGRA) (Grades 2–3)
36	Piper, B. (2010) *Ethiopia, Amharic* Mixed SES, L2	Amharic reading comprehension (EGRA) (Grades 2–3)
37	Piper, B. (2010) *Ethiopia, Hararigna* Mixed SES, L2	Hararigna reading comprehension (EGRA) (Grades 2–3)
38	Piper, B. (2010) *Ethiopia, Afan Oromo* Mixed SES, L2	Afan Oromo reading comprehension (EGRA) (Grades 2–3)
39	Piper, B. (2010) *Ethiopia, Somaligna* Mixed SES, L2	Somaligna reading comprehension (EGRA) (Grades 2–3)
40	Piper, B. (2010) *Ethiopia, Sidaamu Afoo* Mixed SES, L2	Sidaamu Afoo reading comprehension (EGRA) (Grades 2–3)
41	Piper *et al*. (2014) *Kenya, Kiswahili* Mixed SES, L1	Kiswahili reading comprehension (Grades 1–2)
42	Piper *et al*. (2014) *Kenya, English* Mixed SES, L2	English reading comprehension (Grades 1–2)
43	Pretorius, E. and Currin, S. (2010) *South Africa, English* Low SES, L2	English reading comprehension test (Grade 7)
44	Pretorius, E. and Currin, S. (2010) *South Africa, Northern Sotho* Low SES, L1	Northern Sotho reading comprehension test (Grade 7)
45	Rout, E. L. (2001) *India, Oriya* Middle/Upper SES, L1	Oriya graded reading comprehension (Preschool, Grades 1, 3, 5)
46	Sharma, R. (1997) *India, Hindi* Mixed SES, L1	Hindi reading comprehension (Grades 4–5)
47	Spratt *et al*. (1991) *Morocco, Arabic* Low SES, L1	School-based reading assessment (Grades 3–6)
48	Spratt *et al*. (1991) *Morocco, Arabic* Low SES, L1	Household literacy assessment (Grades 3–6)
49	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* Low SES, L1	Moroccan Arabic sentence maze (Grade 1)
50	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* Low SES, L1	Moroccan Arabic paragraph comprehension (Grade 1)
51	Williams, E. (1998)	English Modified Cloze

	Malawi, English Mixed SES, L2	(Grades 3,4,6)
52	Williams, E. (1998) *Zambia, English* Mixed SES, L2	English Modified Cloze (Grades 3,4,6)
53	Williams, E. (1998) *Malawi, Chichewa* Mixed SES, Mixed	Chichewa Modified Cloze (Grades 3, 4, 6)
54	Williams, E. (1998) *Zambia, Nyanja* Mixed SES, Mixed	Nyanja Modified Cloze (Grades 3,4,6)
55	Williams, E. (1998) *Malawi, English* Mixed SES, L2	English word find (Modified Cloze) (Grades 5)
56	Williams, E. (1998) *Zambia, English* Mixed SES, L2	English word find (Modified Cloze) (Grade 5)
57	Williams, E. (1998) *Malawi, Chichewa* Mixed SES, Mixed	Chichewa word find (Modified Cloze) (Grade 5)
58	Williams, E. (1998) *Zambia, Nyanja* Mixed SES, Mixed	Nyanja word find (Modified Cloze) (Grade 5)
59	Williams, E. (1998) *Zambia, Nyanja* Mixed SES, Mixed	Nyanja reading proficiency (Grade 5)
60	Williams, E. (1993) *Zambia, Nyanja* Mixed SES, Mixed	Nyanja Modified Cloze (Grades 3, 4, 6)
61	Williams, E. (1993) *Malawi, English* Mixed SES, L2	English Modified Cloze (Grades 3, 4, 6)
62	Williams, E. (1993) *Zambia, English* Mixed SES, L2	English Modified Cloze (Grades 3, 4, 6)
63	Williams, E. (1993) *Malawi, Chichewa* Mixed SES, Mixed	Chichewa Modified Cloze (Grades 3, 4, 6)
64	Williams, E.(1993) *Zambia, English* Mixed SES, L2	English Modified Cloze (Grades 3,4,6)
65	Williams, E. (2007) *Malawi, English* Mixed SES, L2	English Modified Cloze (Grades 4–5)
66	Williams, E. (2007) *Malawi, English* Mixed SES, L2	Informal assessment of structured reading (Grades 4–5)

A.7 Narrative writing

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish composition writing – fluency (Grade 2)
2	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish composition writing – content (Grade 2)
3	Chowdhury *et al.* (1994). *Bangladesh, Bangla* Mixed SES, L1	Bangla writing a letter (Grade 5)
4	Griffin, P. and Anh, P. N. (2005) *Vietnam, Vietnamese* Mixed SES, L1	Vietnamese literacy (Grade 5)
5	Johnson, D. (2003) *Bangladesh, Bangla* Mixed SES, L1	Bangla writing scale (Grade 4)
6	Johnson *et al*. (2000) *Malawi, English* Mixed SES, L2	English language: writing (Grades 1–8)
7	Johnson *et al*. (2000) *Malawi, Chichewa* Mixed SES, L1	Chichewa language: writing (Grades 1–8)
8	Johnson *et al*. (2000) *Sri Lanka, English* Mixed SES, L2	English language: writing (Grades 1–8)
9	Johnson *et al*. (2000) *Sri Lanka, Sinhala* Mixed SES, L1	Sinhala language: writing (Grades 1–8)
10	Johnson *et al*. (2000) *Sri Lanka, Tamil* Mixed SES, L1	Tamil language: writing (Grades 1–8)
11	Mohsin *et al*. (1996) *Bangladesh, Bangla* Mixed SES, L1	ABC survey instrument: Bangla writing (Grade 5)
12	Nag, S. (2013) *India, Kannada* Mixed SES, L1	Kannada narrative writing (Grades 3–4)
13	Sen, R. and Blatchford, P. (2001) *India, English* Mixed SES, L2	English copying a sentence (Preschool and Grade 1)
14	Winch. C. and Gingell. J. (1994) *Southern Caribbean, Creole* Mixed SES, L1	Creole narrative completion task (Grades 2–3)
15	Winch. C. and Gingell. J. (1994) *Southern Caribbean, Creole* Mixed SES, L1	Creole letter of complaint to shopkeeper (Grades 2–3)

A.8 Grade-level tests

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Abeberese *et al*. (2011) Abeberese *et al*. (2014) *Philippines, Filipino* Mixed SES, Unclear	Reading skills (national reading exams) (Grade 4)
2	Asfaha *et al*. (2009) *Eritrea, English* Mixed SES, L2	English grade-level test (Grade 4)
3	Asfaha *et al*. (2009) *Eritrea, Tigre* Mixed SES, L1	Tigre grade-level test (Grade 4)
4	Asfaha *et al*. (2009) *Eritrea, Tigrinya* Mixed SES, L1	Tigrinya grade-level test (Grade 4)
5	Asfaha *et al*. (2009) *Eritrea, Saho* Mixed SES, L1	Saho grade-level test (Grade 4)
6	Asfaha *et al*. (2009) *Eritrea, Kunama* Mixed SES, L1	Kunama grade-level test (Grade 4)
7	Asfaha *et al*. (2009) *Eritrea, Arabic* Mixed SES, L1	Arabic grade-level test (Grade 4)
8	Aturupane *et al*. (2014) *Sri Lanka, Sinhala* Mixed SES, L1	Sinhala academic test (Grades 4,8)
9	Aturupane *et al*. (2014) *Sri Lanka, Tamil* Mixed SES, L1	Tamil academic test (Grades 4,8)
10	Aturupane *et al*. (2014) *Sri Lanka, English* Mixed SES, L2	English academic test (Grade 4)
11	Griffin, P. and Thanh, M. T. (2006) *Vietnam, Vietnamese* Mixed SES, L1	Vietnamese reading test (Grade 5)
12	Lakshminarayana *et al*. (2013) *India, Telugu (inferred)* Low SES, Unclear	Language test (Grades 2–4)
13	Mohsin *et al*. (1996) *Bangladesh, Bangla* Mixed SES, L1	ABC survey instrument: reading (Grades 5–6)
14	Rolleston, C. and Krutikova, S. (2014) *Vietnam, Vietnamese* Mixed SES, L1	Vietnamese (Grade 5)
15	Sharma, U. (2014) *Nepal, Nepali*	Nepali (Grades 2, 3, 4 and 6)

	Mixed SES, L1	

A.9 Vocabulary

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Alcock *et al*. (2010) *Tanzania, Kiswahili* Low SES, L1	Kiswahili counting words (Grades 1–2)
2	Alcock *et al*. (2010) *Tanzania, Kiswahili* Low SES, L1	Kiswahili vocabulary (Grades 1–2)
3	Ardila *et al*. (2005) *Colombia, Spanish* Mixed SES, L1	Spanish semantic verbal fluency (Grades 1–8)
4	Ardila *et al*. (2005) *Colombia, Spanish* Mixed SES, L1	Spanish phonemic verbal fluency (Grades 1–8)
5	Ardila *et al*. (2005) *Mexico, Spanish* Mixed SES, L1	Spanish semantic verbal fluency (Grades 1–8)
6	Ardila *et al*. (2005) *Mexico, Spanish* Mixed SES, L1	Spanish phonemic verbal fluency (Grades 1–8)
7	Baydar *et al*. (2013) *Turkey, Turkish* Mixed SES, L1	Turkish receptive language test (TRLT) *(*Preschool and out of school)
8	Bekman *et al*. (2011) *Turkey, Turkish* Low SES, L2	Turkish PPVT (Preschool)
9	Brouwers *et al*. (2006) *India, Hindi* Low SES, L1	Hindi (inferred) picture vocabulary (Grades 1–4)
10	Brouwers *et al*. (2006) *India, Hindi (inferred)* Low SES, L1	Hindi (inferred) fluency (Grades 1–4)
11	Castilla, A. (2008) *Colombia, Spanish* Middle/Upper SES, L1	Spanish Test de Vocabulario en Imágenes Peabody (Preschool)
12	Crookston *et al*. (2014) *Ethiopia, Amaringa* Mixed SES, L1	Amaringa PPVT (Grades 4–5)
13	Crookston *et al*. (2014) *Ethiopia, Oromifa* Mixed SES, L1	Oromifa PPVT (Grades 4–5)
14	Crookston *et al*. (2014) *Ethiopia, Tigrigna* Mixed SES, L1	Tigrigna PPVT (Grades 4–5)

15	Crookston *et al*. (2014) *India, Telugu* Mixed SES, L1	Telugu PPVT (Grades 4–5)
16	Crookston *et al*. (2014) *India, Kannada* Mixed SES, L1	Kannada PPVT (Grades 4–5)
17	Crookston *et al*. (2014) *Peru, Spanish* Mixed SES, L1	Spanish PPVT (Grades 4–5)
18	Crookston *et al*. (2014) *Peru, Spanish* Mixed SES, L1	Quechua PPVT (Grades 4–5)
19	Crookston *et al*. (2014) *Vietnam, Tieng Viet Nam* Mixed SES, L1	Tieng Viet Nam PPVT (Grades 4–5)
20	Crookston *et al*. (2014) *Vietnam, H'mong* Mixed SES, L1	H'mong PPVT (Grades 4–5)
21	Crookston *et al*. (2014) *Multiple countries, multiple minor languages* Mixed SES, L1	Minor (other) language PPVT (Grades 4–5)
22	Fedda, O. D. and Oweini, A. (2012) *Lebanon, English* Middle/Upper SES, L2	English WJ-III Tests of Achievement: Picture vocabulary subtest (Preschool, Grades 1–5)
23	Fedda, O. D. and Oweini, A. (2012) *Lebanon, Arabic* Middle/Upper SES, L1	Arabic WJ-III Tests of Achievement: Picture vocabulary subtest (Preschool, Grades 1–5)
24	Fernald *et al*. (2011) *Madagascar, Malagasy* Low SES, L1	Malagasy receptive language (Preschool)
25	Jere-Folotiya *et al*.(2014) *Zambia, CiNyanja* Mixed SES, Mixed	CiNyanja picture vocabulary test (PVT) (Grade 1)
26	Jukes, M. C.H. and Grigorenko, E. L. (2010) *Gambia, Wolof* Low SES, L1	Wolof categorical fluency (Grades 6–8)
27	Jukes, M. C.H. and Grigorenko, E. L. (2010) *Gambia, Mandinka* Low SES, L1	Mandinka categorical fluency (Grades 6–8)
28	Jukes, M. C.H. and Grigorenko, E. L. (2010) *Gambia, Wolof* Low SES, L1	Wolof vocabulary test (Grades 6–8)
29	Jukes, M. C.H. and Grigorenko, E. L. (2010) *Gambia, Mandinka* Low SES, L1	Mandinka vocabulary test (Grades 6–8)
30	Kalia, V. (2007) *India, English* Middle/Upper SES, L2	English PPVT III-B. (Preschool)
31	Kalia, V. (2009)	English PPVT III-B

	India, English Middle/Upper SES, L2	(Preschool)
32	Kalia, V. and Reese, E. (2009) *India, English* Middle/Upper SES, L2	English PPVT III-B (Preschool)
33	Kormi-Nouri *et al*. (2012) *Iran, Persian* Mixed SES, Mixed	Persian category fluency task (Grades 1–5)
34	Kormi-Nouri *et al*. (2012) *Iran, Persian* Mixed SES, Mixed	Persian letter fluency task (words starting with target letter) (Grades 1–5)
35	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	Filipino verbal fluency (Preschool and Grade 1)
36	Ledesma, H. M. (2002) *Philippines, English* Mixed SES, Mixed	English verbal fluency (Preschool and Grade 1)
37	Lee, L. and Wheldall, K. (2011) *Malaysia, Malay* Unclear SES, L1	Malay vocabulary test (Grade 1)
38	Moore *et al*. (2008) *Bangladesh, Bangla* Mixed SES, L1	Bangla vocabulary (Preschool)
39	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada vocabulary (Grades 4–6)
40	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada oral vocabulary knowledge (Grades 1–5)
41	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu oral vocabulary knowledge (Grades 1–5)
42	Nakamura, P. (2014) *India, English* Mixed SES, L2	English oral vocabulary knowledge (Grades 1–5)
43	Opel *et al*. (2009) *Bangladesh, Bangla* Low SES, L1	Bangla vocabulary test (Preschool)
44	Mwaura *et al*. (2008) *Kenya, Kiswahili* Mixed SES, Mixed	Kiswahili verbal meaning (Preschool)
45	Mwaura *et al*. (2008) *Uganda, Luganda* Mixed SES, Mixed	Luganda verbal meaning (Preschool)
46	Mwaura *et al*. (2008) *Tanzania/Zanzibar, Kiswahili* Mixed SES, Mixed	Kiswahili verbal meaning (Preschool)
47	Paxson, C. and Schady, N. (2007) *Ecuador, Spanish* Mixed SES, L1	Test de Vocabularioen Imageries Peabody (TVIP) (Preschool)

48	Rochdi, A. (2010) *Morocco, Arabic* Low SES, L2	Arabic word learning (Out of school)
49	Rochdi, A. (2010) *Morocco, Arabic* Low SES, L2	Arabic fast mapping (Out of school)
50	Rolla San Francisco *et al*. (2006) *Costa Rica, Spanish* Low SES, L1	Spanish vocabulary (Preschool)
51	Strasser, K. and Lissi, M. (2009) *Chile, Spanish* Mixed SES, L1	Spanish receptive vocabulary (Preschool and Grade 1)
52	Singh, A. (2014) Cueto *et al*. (2009) *India, Telugu* Mixed SES, Mixed	Telugu PPVT (Preschool and Grade 1)
53	Tahan *et al*. (2011) *Egypt, English* Middle/Upper SES, L2	English PPVT (Preschool)
54	Tahan *et al*. (2011) *Egypt, Arabic* Middle/Upper SES, L1	Arabic PPVT (Preschool)
55	Tambulukani, G. and Bus, H. (2012) *Zambia, Lozi* Low SES, L1	Lozi familiar language test (Grade 2)
56	Tambulukani, G. and Bus, H. (2012) *Zambia, Mbunda* Low SES, L1	Mbunda familiar language test (Grade 2)
57	Tambulukani, G. and Bus, H. (2012) *Zambia, Nyanja* Low SES, L1	Nyanja familiar language test (Grade 2)
58	Tambulukani, G. and Bus, H. (2012) *Zambia, Town Nyanja* Low SES, L1	Town Nyanja familiar language test (Grade 2)
59	Tambulukani, G. and Bus, H. (2012) *Zambia, English* Low SES, L2	English familiar language test (Grade 2)
60	Vagh, S. B. (2009) *India, Hindi* Low SES, Mixed	Hindi picture identification task (Preschool)
61	Vagh, S. B. (2009) *India, Hindi* Low SES, Mixed	Hindi picture naming task (Preschool)
62	Wagner, D. A. (1993) *Morocco, Moroccan Arabic* Low SES, L1	Moroccan Arabic picture vocabulary (MP) (Preschool, Grade 1, Grade 5)
63	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia* Mixed SES, L1	Bahasa Indonesia morpheme deletion (Grade 2)

A.10 Other language measures

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish grammatical awareness (judgement) (Grades 1 and 2)
2	Babayigit, S. and Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish grammatical awareness (production) (Grades 1 and 2)
3	Bekman *et al*. (2011) *Turkey, Turkish* Low SES, L2	Turkish listening comprehension (Preschool and not yet enrolled children)
4	Bekman *et al*. (2011) *Turkey, Turkish* Low SES, L2	Turkish listening to and following verbal instructions (Preschool and not yet enrolled children)
5	Bekman *et al*. (2011) *Turkey, Turkish* Low SES, L2	Turkish story comprehension (Preschool and not yet enrolled children)
6	Bekman *et al*. (2011) *Turkey, Turkish* Low SES, L2	Turkish elicited imitation (Preschool and not yet enrolled children)
7	Castilla, A. (2008) *Colombia, Spanish* Middle/Upper SES, L1	Spanish language sample (Preschool)
8	Castilla, A. (2008) *Colombia, Spanish* Middle/Upper SES, L1	Spanish elicitation task (Preschool)
9	Chinyama *et al*. (2012) *Zimbabwe, Shona* Low SES, L1	Shona listening comprehension (Grade 3)
10	Chinyama *et al*. (2012) *Zimbabwe, English* Low SES, L2	English listening comprehension (only for non-readers) (Grade 3)
11	Davidson, M. and Hobbs, J. (2013) Piper, B. and Korda, M. (2011) *Liberia, English* Low SES, L1	English listening comprehension (EGRA) (Grades 2–3)
12	Fernald *et al*. (2011) *Madagascar, Malagasy* Low SES, L1	Malagasy memory of phrases (Preschool)
13	Friedlander *et al*. (2014) *Zambia, Bemba* Low SES, L1	Bemba listening comprehension (Grade 3)
14	Friedlander *et al*. (2014) *Zambia, English* Low SES, L2	English listening comprehension (Grade 3)

15	Jukes, M. C.H. and Grigorenko, E. L. (2010) *Gambia, Wolof* Low SES, L1	Wolof proverb understanding (Grades 6–8)
16	Jukes, M. C.H. and Grigorenko, E. L. (2010) *Gambia, Mandinka* Low SES, L1	Mandinka proverb understanding (Grades 6–8)
17	Kalia, V. (2009) *India, English* Mixed SES, L2	English story comprehension (Preschool)
18	Kalia, V. (2009) *India, English* Mixed SES, L2	English story quality (Preschool)
19	Nag, S. (2007) *India, Kannada* Mixed SES, Mixed	Kannada language proficiency (Grades 1–4)
20	Nag-Arulmani *et al*. (2003) *India, Kannada* Middle/Upper SES, L1	Kannada language comprehension (Grade 3)
21	Nag-Arulmani *et al*. (2003) *India, English* Middle/Upper SES, L2	English Test for the Reception of Grammar (Bishop, 1989) (Grade 3)
22	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada sentence repetition (Grades 4–6)
23	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada listening comprehension (Grades 1–5)
24	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu listening comprehension test (Grades 1–5)
25	Nakamura, P. (2014) *India, English* Mixed SES, L2	English listening comprehension test (Grades 1–5)
26	Mwaura *et al*. (2008) *Kenya, Kiswahili* Mixed SES, L1	Kiswahili verbal comprehension (Preschool)
27	Mwaura *et al*. (2008) *Uganda, Luganda* Mixed SES, L1	Luganda verbal comprehension (Preschool)
28	Mwaura *et al*. (2008) *Tanzania/Zanzibar, Kiswahili* Mixed SES, L1	Kiswahili verbal comprehension (Preschool)
29	Piper, B. (2010) *Ethiopia, Tigrinya* Mixed SES, L1	Tigrinya listening comprehension (EGRA) (Grades 2–3)
30	Piper, B. (2010) *Ethiopia, Amharic* Mixed SES, L1	Amharic listening comprehension (EGRA) (Grades 2–3)
31	Piper, B. (2010)	Hararigna listening comprehension (EGRA)

	Ethiopia, Hararigna Mixed SES, L1	(Grades 2–3)
32	Piper, B. (2010) *Ethiopia, Afan Oromo* Mixed SES, L1	Afan Oromo listening comprehension (EGRA) (Grades 2–3)
33	Piper, B. (2010) *Ethiopia, Somaligna* Mixed SES, L1	Somaligna listening comprehension (EGRA) (Grades 2–3)
34	Piper, B. (2010) *Ethiopia, Sidaamu Afoo* Mixed SES, L1	Sidaamu Afoo listening comprehension (EGRA) (Grades 2–3)
35	Veii, K. and Everett, J. (2005) *Namibia, English* Mixed SES, L2	English listening comprehension (Grades 2–5)
36	Veii, K. and Everatt, J. (2005) *Namibia, Herero* *Mixed SES, L1*	Herero listening comprehension (Grades 2–5)

A.11 Phonological awareness

S. no.	Author(s) (Year) *Country, Language of assessment* SES, Assessment is in L1/L2/Mixed	Name of measure (Grades targeted)
1	Alcock *et al*. (2010) *Tanzania, Kiswahili* Low SES, L1	Kiswahili blending (Grades 1–2)
2	Alcock *et al*. (2010) *Tanzania, Kiswahili* Low SES, L1	Kiswahili segmenting (Grades 1–2)
3	Alcock *et al*. (2010) *Tanzania, Kiswahili* Low SES, L1	Kiswahili counting sounds (Grades 1–2)
4	Alcock *et al*. (2010) *Tanzania, Kiswahili* Low SES, L1	Kiswahili 'odd one out' (Grades 1–2)
5	Babayigit, S., Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish word analysis test (syllable deletion) (Grades 1–2)
6	Babayigit, S., Stainthorp, R. (2010) *Northern Cyprus, Turkish* Mixed SES, L1	Turkish word analysis test (phoneme deletion) (Grades 1–2)
7	Bekman *et al*. (2011) *Turkey, Turkish* Mixed SES, L2	Turkish discriminating first and last sounds (Preschool and out of school)
8	Davidson, M. and Hobbs, J. (2013) Piper, B. and Korda, M. (2011) *Liberia, English* Low SES, L2	English phoneme awareness (initial sound) (EGRA) (Grades 2–3)
9	Elbeheri, G. and Everett, J. (2007) *Egypt, Arabic*	English phoneme deletion (Grades 4–5)

	Low SES, L1	
10	Elbeheri, G. and Everett, J. (2007) *Egypt, Arabic* Low SES, L1	Arabic rhyme detection (Grades 4–5)
11	Elmonayer, R. (2012) *Egypt, Arabic* Middle/Upper SES, L1	Arabic kindergarten inventory of phonological awareness (Preschool)
12	Farukh, A. and Vulchanova, M. (2014) *Pakistan, Urdu* Mixed SES, L2	Urdu non-word repetition (Grade 3)
13	Kalia, V. (2007) *India, English* Middle/Upper SES, L2	English blending (Preschool Comprehensive Test of Phonological and Print Processing: Lonigan *et al*., 2002) (Preschool)
14	Kalia, V. (2007) *India, English* Middle/Upper SES, L2	English elision (Preschool Comprehensive Test of Phonological and Print Processing: Lonigan *et al*., 2002) (Preschool)
15	Kalia, V. (2009) *India, English* Middle/Upper SES, L2	English Preschool Comprehensive Test of Phonological and Print Processing (Lonigan *et al*., 1998) (Preschool)
16	Kalia, V. and Reese, E. (2009) *India, English* Middle/Upper SES, L2	English Preschool Comprehensive Test of Phonological and Print Processing (Lonigan, et. al., 1998) (Preschool)
17	Ledesma, H. M. (2002) *Philippines, English* Mixed SES, L2	English elision (Preschool and Grade 1)
18	Ledesma, H. M. (2002) *Philippines, English* Mixed SES, L2	English blending words (Preschool and Grade 1)
19	Ledesma, H. M. (2002) *Philippines, Filipino* Mixed SES, Mixed	Filipino verbal learning (Preschool and Grade 1)
20	Lee, L. and Wheldall, K. (2011) *Malaysia, Malay* SES unclear, L1	Malay phonological blending (Grade 1)
21	Lee, L. and Wheldall, K. (2011) *Malaysia, Malay* SES unclear, L1	Malay phonological segmentation (Grade 1)
22	Mishra, R. and Stainthorp, R. (2007) *India, English* Middle/Upper SES, L2	English Test of Phonological Awareness (Hatcher *et al.*, 1994). (Grade 5)
23	Mishra, R. and Stainthorp, R. (2007) *India, Oriya* Middle/Upper SES, L2	Oriya adaptation of the Test of Phonological Awareness (Hatcher, *et al*., 1994). (Grade 5)

24	Nag, S. (2007) *India, Kannada* Mixed SES, L1	Kannada short phonological processing battery (Grades 1–4)
25	Nag-Arulmani *et al*. (2003) *India, English* Middle/Upper SES, L2	English–Kannada phonological skills battery (Grade 3)
26	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada phoneme substitution (Grades 4–6)
27	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada syllable substitution (Grades 4–6)
28	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada syllable deletion (Grades 4–6)
29	Nag, S. and Snowling, M. (2011) *India, Kannada* Mixed SES, L1	Kannada phoneme deletion (Grades 4–6)
30	Nag, S. and Snowling, M. (2012) *India, Kannada* Mixed SES, L1	Kannada phoneme deletion (Grades 4–6)
31	Nag, S. and Snowling, M. (2012) *India, Kannada* Mixed SES, L1	Kannada phoneme substitution (Grades 4–6)
32	Nag, S. and Snowling, M. (2012) *India, Kannada* Mixed SES, L1	Kannada syllable deletion (Grades 4–6)
33	Nag, S. and Snowling, M. (2012) *India, Kannada* Mixed SES, L1	Kannada syllable substitution (Grades 4–6)
34	Nakamura, P. (2014) *India, Kannada* *Mixed SES, L1*	Kannada syllable blending (Grades 1–5)
35	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu syllable blending (Grades 1–5)
36	Nakamura, P. (2014) *India, English* Mixed SES, L2	English syllable blending (Grades 1–5)
37	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Syllable deletion test – Kannada (Grades 1–5)
38	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Syllable deletion test – Telugu (Grades 1–5)
39	Nakamura, P. (2014) *India, English* Mixed SES, L2	Syllable deletion test – English (Grades 1–5)
40	Nakamura, P. (2014)	Kannada phoneme blending test

	India, Kannada Mixed SES, L1	(Grades 1–5)
41	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu phoneme blending test (Grades 1–5)
42	Nakamura, P. (2014) *India, English* Mixed SES, L2	English phoneme blending test (Grades 1–5)
43	Nakamura, P. (2014) *India, Kannada* Mixed SES, L1	Kannada phoneme deletion (Grades 1–5)
44	Nakamura, P. (2014) *India, Telugu* Mixed SES, L1	Telugu phoneme deletion (Grades 1–5)
45	Nakamura, P. (2014) *India, English* Mixed SES, L2	English phoneme deletion (Grades 1–5)
46	Oktay, A. and Aktan, E. (2002) *Turkey, English* Middle/Upper SES, L2	English syllable segmentation (Preschool and Grade 1)
47	Oktay, A. and Aktan, E. (2002) *Turkey, Turkish* Middle/Upper SES, L1	Turkish syllable segmentation (Preschool and Grade 1)
48	Oktay, A. and Aktan, E. (2002) *Turkey, English* Middle/Upper SES, L2	English phoneme segmentation (Preschool and Grade 1)
49	Oktay, A. and Aktan, E. (2002) *Turkey, Turkish* Middle/Upper SES, L1	Turkish phoneme segmentation (Preschool and Grade 1)
50	Oktay, A. and Aktan, E. (2002) *Turkey, English* Middle/Upper SES, L2	English phoneme deletion (Initial sound) (Preschool and Grade 1)
51	Oktay, A. and Aktan, E. (2002) *Turkey, Turkish* Middle/Upper SES, L1	Turkish phoneme deletion (initial sound) (Preschool and Grade 1)
52	Oktay, A. and Aktan, E. (2002) *Turkey, English* Middle/Upper SES, L2	English phoneme deletion (final sound) (Preschool and Grade 1)
53	Oktay, A. and Aktan, E. (2002) *Turkey, Turkish* Middle/Upper SES, L1	Turkish phoneme deletion (final sound) (Preschool and Grade 1)
54	Piper, B. (2010) *Ethiopia, Tigrinya* Mixed SES, L1	Tigrinya phonological awareness (EGRA) (Grades 2 and 3)
55	Piper, B. (2010) *Ethiopia, Amharic* Mixed SES, L1	Amharic phonological awareness (EGRA) (Grades 2 and 3)
56	Piper, B. (2010) *Ethiopia, Hararigna* Mixed SES, L1	Hararigna phonological awareness (EGRA) (Grades 2 and 3)

57	Piper, B. (2010) *Ethiopia, Afan Oromo* Mixed SES, L1	Afan Oromo phonological awareness (EGRA) (Grades 2 and 3)
58	Piper, B. (2010) *Ethiopia, Somaligna* Mixed SES, L1	Somaligna phonological awareness (EGRA) (Grades 2 and 3)
59	Piper, B. (2010) *Ethiopia, Sidaamu Afoo* Mixed SES, L1	Sidaamu Afoo phonological awareness (EGRA) (Grades 2 and 3)
60	Rochdi, A. (2010) *Morocco, Arabic* Low SES, L2	Arabic phonological awareness (Preschool)
61	Rolla San Francisco *et al*. (2006) *Costa Rica, Spanish* Low SES, L1	Spanish phonological awareness (Preschool)
62	Sousa *et al*. (2010) *South Africa, Zulu* Mixed SES, L1	Zulu syllable segmentation (Grade 2)
63	Sousa *et al*. (2010) *South Africa, English* Mixed SES, L2	English syllable segmentation (Grade 2)
64	Sousa *et al*. (2010) *South Africa, Zulu* Mixed SES, L1	Zulu onset–rime detection (Grade 2)
65	Sousa *et al*. (2010) *South Africa, English* Mixed SES, L2	English onset–rime detection (Grade 2)
66	Sousa *et al*. (2010) *South Africa, Zulu* Mixed SES, L1	Zulu phoneme deletion (Grade 2)
67	Sousa *et al*. (2010) *South Africa, English* Mixed SES, L2	English phoneme deletion (Grade 2)
68	Strasser, K. and Lissi, M. (2009) *Chile, Spanish* Middle SES, L1	Spanish phonemic awareness (Preschool and Grade 1)
69	Tahan *et al*. (2011) *Egypt, Arabic* Middle/Upper SES, L1	Arabic elision (syllable deletion) (Preschool)
70	Tahan *et al*. (2011) *Egypt, English* Middle/Upper SES, L2	English elision (syllable deletion) (Preschool)
71	Tahan *et al*. (2011) *Egypt, Arabic* Middle/Upper SES, L1	Arabic blending task (Preschool)
72	Tahan *et al*. (2011) *Egypt, English* Middle/Upper SES, L2	English blending task (Preschool)
73	Veii, K. and Everett, J. (2005)	Herero phoneme recognition

	Namibia, Herero Mixed SES, L1	(Grades 2–5)
74	Veii, K. and Everett, J. (2005) *Namibia, English* Mixed SES, L2	English phoneme recognition (Grades 2–5)
75	Veii, K. and Everett, J. (2005) *Namibia, Herero* Mixed SES, L1	Herero sound discrimination (Grades 2–5)
76	Veii, K. and Everett, J. (2005) *Namibia, English* Mixed SES, L2	English sound discrimination (Grades 2–5)
77	Veii, K. and Everett, J. (2005) *Namibia, Herero* Mixed SES, L1	Herero non-word sequence repetition (Grades 2–5)
78	Veii, K. and Everett, J. (2005) *Namibia, English* Mixed SES, L2	English non-word sequence repetition (Grades 2–5)
79	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia* Mixed SES, L1	Bahasa Indonesia syllable segmentation (Grades 1–2)
80	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia* Mixed SES, L1	Bahasa Indonesia rhyme detection (Grades 1–2)
81	Winskel, H. and Widjaja, V. (2007) *Indonesia, Bahasa Indonesia* Mixed SES, L1	Bahasa Indonesia onset detection (Grades 1–2)
82	*Winskel, H.* and *Widjaja, V. (2007)* *Indonesia, Bahasa Indonesia* *Mixed SES, L1*	Bahasa Indonesia phoneme deletion (Grades 1–2)
83	*Winskel, H.* and *Widjaja, V. (2007)* *Indonesia, Bahasa Indonesia* *Mixed SES, L1*	Bahasa Indonesia syllable deletion (Grades 1–2)

Annex B Summary of psychometric, administrative and contextualisation data

B.1 Literacy measures (1)

Measures→ Area of assessment ↓	Emergent literacy (22 measures from 15 studies)	Symbol knowledge (58 measures from 30 studies)	Reading accuracy (90 measures from 37 studies)	Spelling (35 measures from 17 studies)
Psychometric characteristics				
Reliability indices	Reliability of >.8 = 7 m Between .6 and .79 = 1 m <.59 = 1 m No information = 13 m	Reliability of >.8 = 13 m Between .6 and .79 = 4 m <.59 = 2 m No information = 39 m	Reliability of >.8 = 21 m Between .6 and .79 = 4 m <.59 = 1 m No information = 64 m	Reliability of >.8 = 11 m Between .6 and .79 = 6 m <.59 = 0 m No information = 18 m
Administrative details				
Procurement	Free-ware = 5 m Commercial-ware = 2 m Researcher developed = 13 m Not clear = 2 m	Free-ware = 11 m Commercial-ware = 2 m Researcher developed = 43 m Not clear = 2 m	Free-ware = 16 m Commercial-ware = 16 m Researcher developed = 58 m	Free-ware = 0 m Commercial-ware = 3 m Researcher developed = 32 m
Mode of data gathering	Individual testing = 19 m Group testing = 0 m Not clear = 3 m	Individual testing = 43 m Group testing = 3 m Not clear = 12 m	Individual testing = 88 m Group testing = 2 m	Individual testing = 28 m Group testing = 6 m Not clear = 1 m
	Observation data = 0 m Reported information = 0 m Performance data = 20 m Not clear = 2 m	Observation data = 0 m Reported information = 0 m Performance data = 54 m Not clear = 4 m	Observation data = 1 m Reported information = 0 Performance data = 89 m	Observation data = 0 m Reported information = 0 m Performance data = 35 m
Contextualisation				
Pilot before use	Pilot reported = 12 m Unclear = 10 m	Pilot reported = 25 m Unclear = 33 m	Pilot reported = 28 m Unclear = 62 m	Pilot reported = 11 m Unclear = 24 m

Localising	Original (no change) = 1 m Adapted = 13 m Bespoke = 8 m	Original (no change) = 4 m Adapted = 19 m Bespoke = 35 m	Original (no change) = 15 m Adapted = 21 m Bespoke = 50 m Not clear = 4 m	Original (no change) = 3 m Adapted = 5 m Bespoke = 27 m

Note: m = measures.

B.2 Literacy measures (2)

Measures→ Area of assessment ↓	Reading fluency (52 measures from 16 studies)	Reading comprehension (66 measures from 27 studies)	Narrative writing (17 measures from eight studies)
Psychometric characteristics			
Reliability indices	Reliability of >.8 = 13 m Between .6 and .79 = 8 m <.59 = 2 m No information = 29 m	Reliability of >.8 = 13 m Between .6 and .79 = 14 m <.59 = 0 m No information = 39 m	Reliability of >.8 = 4 m Between .6 and .79 = 1 m <.59 = 0 m No information = 12 m
Administrative details			
Procurement	Free-ware = 28 m Commercial-ware = 0 m Researcher developed = 21 Not clear = 3 m	Free-ware = 22 m Commercial-ware = 3 m Researcher developed = 39 Not clear = 2 m	Free-ware = 0 m Commercial-ware = 0 m Researcher developed = 16 Not clear = 1 m
Mode of data gathering	Individual testing = 43 m Group testing = 2 m Not clear = 7 m	Individual testing = 38 m Group testing = 23 m Not clear = 5 m	Individual testing = 15 m Group testing = 2 m
	Observation data = 0 m Reported information = 0 m Performance data = 46 m Not clear = 6 m	Observation data = 1 m Reported information = 0 Performance data = 65 m	Observation data = 0 m Reported information = 0 Performance data = 17 m
Contextualisation			
Pilot before use	Pilot reported = 36 m Unclear = 16 m	Pilot reported = 42 m Unclear = 24 m	Pilot reported = 4 m Unclear = 13 m
Localising	Original (no change) = 2 m Adapted = 26 m Bespoke = 21 m Not clear = 3 m	Original (no change) = 7 Adapted = 25 m Bespoke = 33 m Not clear = 1 m	Original (no change) = 0 Adapted = 0 m Bespoke = 17 m

B.3 Language measures

Measures→ Area of assessment ↓	Other language measures (36 measures from 17 studies)	Vocabulary (63 measures from 22 studies)	Phonological skills (83 measures from 27 studies)
Psychometric characteristics			
Reliability indices	Reliability of >.8 = 5 m Between .6 and .79 = 4 m <.59 = 0 m No information = 27 m	Reliability of >.8 = 31 m Between .6 and .79 = 7 m <.59 = 2 m No information = 23 m	Reliability of >.8 = 11 m Between .6 and .79 = 15 m <.59 = 1 m No information = 56 m
Administrative details			
Procurement	Free-ware = 10 m Commercial-ware = 6 m Researcher developed = 17 m Not Clear = 3 m	Free-ware = 0 m Commercial-ware = 28 m Researcher developed = 35 m	Free-ware = 8 m Commercial-ware = 13 m Researcher developed = 62 m
Mode of data gathering	Individual testing = 35 m Group testing = 0 m Not clear= 1m	Individual testing = 63 m Group testing = 0 m	Individual testing = 83 m Group testing = 0 m
	Observation data = 0 m Reported information = 0 m Performance data = 35 m Not clear = 1 m	Observation data = 0 m Reported information = 0 m Performance data = 63 m	Observation data = 0 m Reported information = 0 m Performance data = 83 m
Contextualisation			
Pilot before use	Pilot reported = 19 m Unclear = 17 m	Pilot reported = 23 m Unclear = 40 m	Pilot reported = 30 m Unclear = 53 m
Localising	Original (no change) = 1 m Adapted = 17 m Bespoke = 18 m	Original (no change) = 9 m Adapted = 29 m Bespoke = 25 m	Original (no change) = 12 Adapted = 27 m Bespoke = 44 m

Note: m = measures.

Annex C Search strategy employed in Nag *et al.* (2014)

Excerpted from

Technical Report No. 1, Review Methodology: Search and Selection Process, Carole Torgerson, Sonali Nag, Shula Chiat and Margaret J. Snowling (2014).[25]

1.1 Methodology

1.1.1 Search methods used in draft review

Electronic searches were undertaken to identify studies primarily about child literacy, reading, writing, numeracy and mathematics in developing countries. The searches were also designed to identify studies about basic education, educational achievement and school attendance of children in the developing world.

The searches were limited by date range (1990 to the present). 1990 was chosen as the cut-off year because this is the year of the *Education for All Jomtien Summit* and all our target countries are signatories to this UN Declaration. There has been a steady (and sometimes, rapid) rise in school coverage and attention to literacy in all countries following the Jomtien Summit.

The searches were not limited by language of publication.

1.1.1.1 Concepts used in the search strategy

The search strategies were devised using a combination of indexed keyword terms and free text search terms appearing in the title and/or abstracts of database records. Search terms were identified through discussion between the research team, by scanning background literature and 'key articles' already known to the project team, and by browsing database thesauri.

Initially, the project team identified a group of 13 'key articles' to use as a test set in the development of the search strategy. Five databases (ERIC, PsycINFO, Social Science Citation Index (SSCI), EconLit and ASSIA) were searched to check if each of the 13 'key articles' were present and what indexing terms had been assigned to the database record. A draft search strategy was then created and run in the ERIC and PsycINFO databases and the results scanned to see how many of the 'key articles' were retrieved. Of the 13 'key articles', nine were present in the ERIC database and three were in PsycINFO. The draft search strategy initially retrieved only four of the nine 'key articles' in ERIC, and two of the three in PsycINFO.

When a 'key article' was not identified by the search strategy (or did not use relevant search terms), the record was checked for potential search terms, which were then added to the search strategy. This procedure was followed after amendments had been made to the second and third drafts of the search strategy. After each draft, the search strategy was sent to the research team for comments, and further iterations were made, until a fourth and final search strategy was agreed upon.

An additional test of the search strategy involved sending random sample sets of 100 records identified in ERIC and PsycINFO using the second draft search strategy to members of the team (SN and CT) to check the relevance of records retrieved (and to help confirm inclusion criteria).

[25] Search strategy team: Steven Duffy (Centre for Reviews and Dissemination, University of York); Prerna Menon (PM), Kamila Polisenska (KP) and Gurpreet Reen (GP) (University of Oxford) and Angshuman Phukan (AP) (The Promise Foundation) with Sonali Nag (SN), Carole Torgerson (CT), Shula Chiat (SC) and Maggie Snowling (MS).

Both tests ensured that the final search strategy identified the 'key articles' and also, more importantly, that it identified other similar studies.

During development of the search strategy it was found that a very large literature about 'adult literacy' in developing countries was being retrieved. It was therefore necessary to introduce in the search terms a concept for 'children', with additional search terms for school type (primary, elementary, kindergarten, etc.) and school grade (Grade 1 to Grade 8). The use of age-related terms in the title and abstract of database records may have been restrictive but was unavoidable.

Similarly, it is not ideal to limit searches geographically but without including the concept of 'developing countries' in the search strategy an extensive literature about child literacy in North America, Western Europe and Australia was accessed. The research team agreed that this concept should be included in the search strategy to prevent retrieval of a large and irrelevant literature. Early search strategy development suggested that generic search terms for 'developing countries' were not identifying studies relevant to the review, including 'key articles'. The team decided to include named countries to help capture this literature. Countries with poor literacy and low income rates were identified from sources such as the World Bank, DFID and UNESCO; including named countries in the search strategy improved the identification of relevant studies. For example, a number of studies did not include terms for 'developing countries' or a named country in the subject indexing, title or abstract of database records, but did include reference to the child's language (e.g. 'Kannada', 'Arabic' or 'Swahili'). Therefore, the main languages spoken in developing countries were included in our search.

The final search strategy was developed by Information Specialist, Steven Duffy of the Centre for Reviews and Dissemination (CRD), and peer reviewed for accuracy by another Information Specialist based at CRD (Lisa Stirk).

The literature searches involved searching a wide range of databases covering education, mental health, economics and social care. The following databases and resources were searched: ERIC, PsycINFO, SSCI, Conference Proceedings Citation Index-Social Science and Humanities (CPCI-SSH), EconLit, British Education Index, Australian Education Index (AEI), ASSIA, Dissertation Abstracts, Index to Theses, BLDS, Eldis, OAISTER, Zetoc, RePEc, ScienceDirect and JSTOR.

Details of the ERIC search strategy and the results of all searches are listed in Appendix 3.

As with the research itself, the claims made by any review rest on the methodology used in collecting the sources reviewed. Although our methods for searching and screening were rigorous, time and resources did not allow for full systematic methods to be used here, such as are used in many systematic reviews, for example comprehensive combinations of hand searches, electronic database searches, 'snowballing' of references, citation tracking, personal knowledge and serendipitous discovery of sources. A more comprehensive review of the evidence would require the use of a consistent and more thorough combination of search approaches for all the research questions proposed. Capturing grey literature and dissertations from universities in the developing countries was a particular challenge.

Given that a number of databases were searched, some degree of duplication resulted. In order to manage this issue, the titles and abstracts of bibliographic records were downloaded and imported into EndNote bibliographic management software and duplicate records removed.

Further material was added to the material collected from the electronic searches through:

- A call for suggestions from key academics.

- ‘Snowballing’ of references and citation tracking.

For an update of this review in the future we recommend systematic search of the following:

- theses from key African and South Asian universities;
- reports from key international aid agencies, NGOs and civic bodies; and
- call for suggestions from teachers, field workers and NGO workers.

1.1.2 Search terms

The final search strategy was structured using the following concepts:

(literacy OR reading OR writing OR numeracy OR mathematics OR school attendance OR school achievement)

AND

(children OR primary education OR school grade)

AND

(developing countries OR named countries OR named languages)

OR

named literacy programmes

AND

Date limit 1990–2013

1.1.3 Inclusion criteria and guidance for independent pre-screening of titles and abstracts

A guidance note was developed by members of the team for the first stage of pre-screening. In this stage rapid checks were undertaken to exclude records that do not meet criteria for review.

1.1.3.1 Steps for pre-screening

Scan title and abstract and decide whether this fulfils 1 and 2 and 3 and 4 (see above).

If it does, decision is ‘include’

[using ‘Y’] means record ‘included’. Leaving blank means ‘excluded’

If in doubt be inclusive

1.1.3.2 Inclusion criteria

1. Topic: Literacy and/or numeracy

AND

2. Country: the 143 countries of low, lower-middle and upper-middle countries (*World Bank and DAC List of ODA Recipients*, OECD, 2012)

AND

3. Age of children: 3–13 (literacy studies) 3–8 (numeracy studies)
OR

4. Grade of children: Up to Grade 8 (literacy studies) Grade 2 (numeracy studies)
AND

5. Publication dates: between 1990 and January 2013

1.1.3.3 Guidance for moderation for pre-screening

Decisions between pairs of reviewers were displayed in EndNote. If there was agreement to include the study it was included at the first stage; similarly if there was agreement to exclude. If one reviewer included and the other excluded a study, a third person arbitrated.

1.1.4 Inclusion criteria and guidance for independent screening of titles and abstracts

A guidance note was developed by members of the team for the second stage of screening. In this stage each abstract was checked for inclusion criteria, thematic focus and research design.

1.1.4.1 Steps for screening

Check title and abstract and decide whether this fulfils 1 and 2 and 3 and 4 and 5 and 6a OR 6b and 7 (see below)

If it does, decision is to 'include'

If it does not, decision is to 'exclude'

If in doubt, be inclusive

1.1.4.2 Inclusion criteria

1. Topic: Literacy and/or numeracy and/or teacher training and/or assessment
   AND
2. Country: developing country (based on World Bank and OECD list of low-, lower-middle- and upper-middle-income countries)
   AND
3. Age of children: 3–13 (literacy studies) 3–8 (numeracy studies)
   OR
   Grade of children: Up to Grade 8 (literacy studies), Grade 2 (numeracy studies)
   AND
4. Publication dates: between 1990 and present
   AND
5. Language: (based on a background note of languages for each target country in the review)
   AND

6a. Study types for narrative synthesis: Cross-linguistic, cross-orthography and cross-cultural studies examining individual differences, group differences, cross-sectional and longitudinal predictors. These may be either single or multi-factorial studies, and deploy various assessment tools and products.

OR

6b. Study types for systematic synthesis/meta-analysis: Quasi-, true and natural experimental designs evaluating literacy and numeracy interventions undertaken in developing countries to improve literacy and numeracy.

1.1.4.3 Guidance for moderation for independent screening

Decisions between pairs of reviewers (AP and KP) were displayed in EndNote. If there was agreement to include the study, it was included at the second stage; similarly if there was agreement to exclude. If one included and the other excluded a study, a third person arbitrated (SN for excludes for first 8,000 records and SC for the rest).

1.1.5 Guidance for preliminary data extraction from titles and abstracts

In this stage prior to the first wave of data extraction, reviewers read each title and abstract and coded for what the record appeared to contain. This preliminary mapping of the records was by theme, research design and descriptors of interest to the interdisciplinary team. The guidance note was developed by the core team based on two meetings attended by all team members.

1.1.5.1 Steps for preliminary data extraction

Check title and abstract and decide whether the records can be described by a theme, a research design, and descriptors from Psychology, Linguistics, Education, Sociology and Economics (see codes below 1.1.5.2)

Check title and abstract and decide whether the record can answer the broad research questions that were the focus of the review (see broad research questions 1.1.5.1 above)

If it does, decision is ‘include’. If it does not, decision is ‘exclude’

If in doubt be inclusive and include.

1.1.5.2 Data extraction codes

The following codes were used for data extraction.

A. Theme

Literacy* Numeracy* Teacher training Assessment* Intervention*

B. Research design*

Includes case study, survey, correlational, longitudinal, experimental/group comparison, quasi-experimental, randomised controlled trials, design unclear

[Codes relevant for the Assessment of Literacy and Foundation Learning Review are marked with an asterisk.]

1.1.5.3 Guidance for quality checks for preliminary data coding

One reviewer coded the disciplinary descriptors and a second reviewer coded for themes and research design. All codes were displayed in EndNote. If there was a query, SN was consulted as

arbiter. Quality assurance for the preliminary data coding was done by an independent reviewer to check if records matched the code.

An identical search strategy was repeated for the timeframe from January 2013 to December 2014. This search was funded by a Grant from the British Academy and Royal Society to SN. The CRD once again conducted the searches and members from the original team (SN, GR, PM) followed the same guidance notes for the pre-screening, screening and quality appraisal.

Annex D Guidance note for developing strength of evidence

Steps for identification of assessment measures

Check the methods section of paper[26] and decide whether a reported measure belongs to the literacy, numeracy and affective-motivational domains. If it does, the decision is to 'include'; if it does not, the decision is to 'exclude'; if in doubt, be inclusive.

Inclusion criteria

1. Assessment measure: Literacy and/or numeracy and/or affective-motivational

AND

2. Assessment of children

Age of children: 3–13 (literacy studies) 3–8 (numeracy studies)

OR

Grade of children: Up to Grade 8 (literacy studies), Grade 2 (numeracy studies)

Guidance for moderation for inclusion of a measure

Decisions of pairs of reviewers are displayed in a shared folder. If there is agreement to include the tool, include for the second stage of data extraction; similarly if there is agreement to exclude. If one reviewer includes and the other excludes a tool, a third person arbitrates.

Guidance for data extraction about a measure

Read the full paper for a detailed extraction of information using a bespoke extraction template. Check introduction, methods, results and discussion sections for stated information. If there is relevant information, the decision is to 'include' in the relevant field in the extraction template. If there is not, the decision is to 'leave blank'. If information is inferred then state this along with a note on what led to the inference. If specific information (e.g. about pilots, psychometric properties, etc.) is available elsewhere (e.g. a technical report, an earlier publication, etc.), procure this and extract.

The specific guidance note for each is as follows:

Psychometric robustness: Examine a measure for its distributional characteristics. Note presence of floor effects (test items are overly difficult), ceiling effects (test items are too easy) and clustering of data (lack of variation or adequate sensitivity). Note report of reliability. Note correlation with theoretically associated measures to assess validity. If information is not available 'leave blank'.

Contextualisation: Examine procurement details (free-ware, commercial, or researcher developed), mode of use (observed, reported or performance data; individual or group formats), and source (original, adapted, or bespoke). Note training cost (for administrators, scorers and quality assurance staff). Note time taken (for test preparation, administration, scoring and analysis, and communication of findings in a manner meaningful for teachers and lay users). If information is not available 'leave blank'.

[26] Here, 'paper' refers to an article in a peer-reviewed journal, a book chapter or a technical report.

Guidance for quality checks for data extraction

Details about each measure are to be double extracted, that is, two reviewers are to read the study and fill out the extraction template. For 20% of the studies, two extractions are done independently and then collated to check for consistency. If there are discrepancies due to omissions, note this. If there are discrepancies of interpretation, SN or GA arbitrates. For the rest of the studies, one reviewer is to read, and to confirm or edit extractions completed by another reviewer with the aim of ensuring completeness of extraction. Thus in 20% the two reviewers are blind to the other's extraction and in 80% they are not.

Guidance for evaluating strength of evidence

Determine strength of evidence for each area of assessment based on how many criteria of robustness and usability are met by how many measures within each area of assessment.