unesco
Institute for Statistics

Education Data and Statistics Commission

WG/GAML/11/2.2

Benchmarking by Language Groups Framework

SDG 4, Indicator 4.1.1a for reading proficiency: Proportion of children and young people in grades 2/3 achieving at least a minimum proficiency level in reading, by sex

2025

Global Alliance to Monitor Learning

Table of contents

1. Introduction[1]

This document draws on the current evidence base on reading comprehension to propose a framework to monitor and track progress for reading proficiency in grades 2/3. The proposal is for **a Benchmarking by Language Groups Framework** that can inform global reporting on Indicator 4.1.1a. Such a framework needs to also be meaningful for country-level action based on national progress monitoring. The following summarizes the events and decisions that led to the interest in such a framework (excerpt from commissioning document; October, 2024):

> Up to late 2023, there had not been much country measurement and reporting on SDG 4.1.1a. As the result, the Inter-Agency and Expert Group on SDG Indicators (IAEG-SDGs) demoted' the indicator from Tier I to Tier II in October 2023 due to low coverage, putting its status at risk during the 2025 Comprehensive Review of the SDG Indicators Framework. The community of interest concerned with foundational learning, including many important institutional stakeholders, and thought leaders, immediately expressed deep concern in blogs and at various meetings. Those interested in measurement quickly mobilized to increase the count by laying a better technical foundation for measurement and strengthening coordination of funding for country-level measurement efforts.
>
> The UNESCO Institute for Statistics (UIS) is the UN agency with the mandate to source data on most of the education SDGs, included 4.1.1a, and make them available to the international community. In response [to the downgrading of indicator 4.1.1.a from Tier 1 to Tier 2], the UIS convened a meeting of the Global Alliance to Monitor Learning (GAML), a group of experts and stakeholders that was established by the UIS in 2016 to support its measurement and reporting efforts. It was noted by the experts at the

[1] This report was commissioned by the UNESCO Institute for Statistics (UIS) and authored by Sonali Nag, University of Oxford.

meeting that there was a divergence between established learning assessments such as PASEC, ERCE, and others, and newer assessments focusing on foundational skills such as EGRA, UNICEF's FLM, and the PAL network's "citizen led" assessments. Among the latter, while substantial measurement activities were ongoing—primarily for advocacy, program design, program tracking, and evaluation—much of this data was not being reported, was not publicly available, and technical documents were scarce and scattered. All these points highlighted the fact that existing measurement efforts, although useful as they may have been for policy advocacy, and program design and tracking, i.e. activities that have less burden for accuracy, were not good enough either for global reporting or for tracking national progress.

An issue that arose was that the newer sorts of assessments mentioned above tend to be seen as beneficial in countries with low learning levels as they assess children's skills in what one could call "precursor" skills, skills that are seen by most reading specialists as important signposts as children move towards learning to read and eventually reading to learn. The UIS position at this point, as reinforced by the Education Data and Statistics Commission (EDSC), formerly known as the Technical Cooperation Group on SDG 4 indicators - Education 2030 (TCG), that met in May 2024, is that, for purposes of counting towards a global measure of children who are proficient in reading only children's ability to answer comprehension questions be acceptable. However, the UIS also recognizes that for certain countries it is useful to measure those precursor skills.

The UIS, therefore, proposed that one could study the "benchmark" levels of these precursor skills that are associated with particular languages, or language groupings, for a reading comprehension level of 80% (of comprehension questions answered correctly) and 60%. For instance, for language X, the oral fluency rate at 80% comprehension might be 45 correct words per minute and for 60% it might be 35. Countries speaking those various languages would be given a table of benchmarks and could then measure the percent of children at those benchmarks, as a way of gauging progress towards comprehension, if children are not comprehending, or, possibly, identify causes for non-comprehension in terms of earlier skills.

It is appropriate to consider a framework for differentiated benchmarks because there is strong evidence that the first steps to gaining reading proficiency differ by languages and writing systems. Put differently, even though the monitoring of reading proficiency is global, a framework would allow the monitoring to be sensitive to variations in the rate of acquisition of, and the relative importance of, the various precursor skills.

This report includes three inter-linked tasks to develop the framework:

1.1 Review of the existing literature

A rapid review considers if the concept of 'language opacity' may be used for 'grouping languages for their learning demands at the earliest level of reading proficiency'.

1.1.1 One objective is to weigh 'the sensibility [and sense] of offering benchmarks by language group rather than for specific languages'.

1.1.2 A second objective is to develop a draft framework based on the literature.

Ideas of 'transparency' and 'opacity' usually refer to the writing system or orthography in which the language is written. Learning to read across writing systems requires **code skills** (e.g., letter/symbol knowledge, word decoding, reading fluency). But *equally important* for learning to read are **language skills** (e.g., vocabulary knowledge, sentence-level skills). Two draft frameworks will therefore be examined (described in section 5): a) a framework using the single parameter of opacity and b) a framework with multiple parameters.

1.2 Investigate data from a selection of language groups for patterns

The draft frameworks for benchmarking by language groups are applied to child performance datasets supplied by UIS.

1.2.1 First, the datasets are described.

Note that the supplied datasets are strong on code skills (e.g., letter/symbol knowledge, word decoding, reading fluency) but light on language skills (e.g., vocabulary knowledge, sentence-level skills). The pattern analysis works within this limitation.

1.2.2 Second, data patterns are examined using the single parameter of opacity. This is the commissioned framework.

1.2.3 Third, data patterns are examined using a more comprehensive multi-parameter approach to language grouping.

1.3 Propose a Benchmarking by Language Groups Framework

Based on the rapid review of the current science of reading and considerations from the supplied datasets, a Benchmarking by Language Groups Framework (BLGF) is proposed.

1.3.1 Recommended framework

1.3.2 What it might look in practice

A framework that is sensitive to precursor skills across diverse languages is needed.
*A **Benchmarking by Language Groups Framework***
looks carefully at all levels of the language in which precursors to reading proficiency are measured.

2. Literature Review[2]: Opacity, learning demands and language groups

Several characteristics of a language matter for gaining proficiency in reading comprehension. These characteristics are also influential in determining which precursor skills must be developing well (by age and grade, if tied to formal schooling) to support reading proficiency for that language. These influential language characteristics are at the level of the script in which the language is written down (described in section 2.2), the types of words in the language (section 2.3), and the nature of sentence construction (section 2.4). Some differences across languages are small but others can be substantial, with non-trivial consequences on emergent skills. The following sections list these language characteristics. These are not exhaustive lists but reflect the current evidence base[3].

2.1 Differing learning demands: Marking the differences

Many languages sit alongside each other with barely noticeable differences. For example, languages are written down in words and sentences, and sentences can be made simple for early readers. These similarities may lead to an assumption that precursor skills for reading comprehension are also similar across the world's languages. However, the cognitive task of

[2] This section is informed by the following sources:
1 Multiple narrative syntheses from a 30-year review of foundation learning, literacy and assessment in LMICs (originally a DFID funded project, new funding is from the Newton Fund and the Norwegian Research Council).
2 A multi-country programme of research on a) 3- to 6-year-olds in multilingual settings, and b) considering the oral language foundations for school readiness (UKRI-GCRF funded).
3 A multi-country analysis of datasets using Early Grades Reading Assessment (EGRA, from RTI and Save the Children), the East Asia-Pacific Early Child Development Scales (EAP-ECDS, from University of Hong Kong) and university-based research in LMICs (funded by UKRI-GCRF, the Newton Fund and Norwegian Research Council).
4 Rapid review of 3 writing systems (Arabic, Cyrillic, Indic) & language families (Bantu, Dravidian, Indo-Aryan).

[3] Examples of the current evidence base from the akshara writing system of south and south-east Asia: Nag, S. (2007). Early reading in Kannada | Nag, S. (2014). Akshara-phonology mappings | Wijaythilake, *et al.* (2018). Cognitive predictors of word reading in Sinhala. | Bhide *et al.* (2021). Spelling challenges in Hindi.

learning is somewhat different depending on script and language. The implications of differing learning demands on a benchmarking framework are illustrated below in sections titled '*Marking the differences*'.

2.2 The script in which the language is written (the orthography)

2.2.1 Orthographies differ in degree of consistency

When the link between the letter or part of a complex symbol is always to the same sound then the orthography is said to be consistent. Variations in linkages make the logic of sound-symbol connections opaque. In addition, one symbol may be used for more than one sound and this mapping may be more or less consistent. In other cases different symbols or even combinations of symbols may make the same sound (as in the *c* in *casa* in Spanish or the *qu* in *queso*, which – among other small inconsistencies -- make Spanish slightly less than the perfectly transparent language it is often assumed by non-specialists to be). For visually complex scripts, consistency is also about where each part of a symbol must be attached.

Examples:

- Letter-sound mapping is unfailingly consistent in the Sesotho and Setswana languages of Africa making reading of new words easier. Even when there is a new word, the sound-symbol links will be the same and blending these sounds will produce the word.
- Thai, Sinhala and Hindi of south Asia are easier to learn despite the symbol sets looking complicated. This is because the rules for where-to-place-which-diacritic is consistent.

2.2.2 Orthographies differ in degree of complexity

When a language has mainly a consonant-vowel pattern (as in Spanish or Italian) it is considered to have simple sounds and when there are many consonant blends it is phonologically (sound-wise) more complex (consider the Scottish surname McClellan, which, having four consonants in a row, is bewildering to someone who learnt to read in Spanish, and the double consonant ll which makes a sound in Spanish that would appear unpronounceable in English). Similarly, symbols range from simple dashes and curves to densely packed symbols with only minute differences. When there is more phonological complexity and/or visual complexity learning may need time.

Examples:

- The complex consonantal systems in the isiZulu, isiXhosa and Xitsonga languages of Africa place a greater demand on decoding of new words.
- In Asia, the simple letter forms of the Filipino, Khasi and Malay scripts are easier to learn than the visually close patterns of the Telugu, Mandarin and Sinhala orthographies.

There is, however, another line of evidence that is relevant for benchmarking precursor skills for the world's language. Children learning phonologically, morphologically and/or visually more complex systems gain in other ways. They become more sensitive to fine details in the specific area of complexity earlier than those learning simple systems.

2.2.3 A sound may have one symbol or more than one symbol

An example of one sound having one symbol is for the English letter 'p'. In contrast, the sound /z/ in English has two symbols, 'z' (as in 'zen') and 's' (as in 'his'). When a sound has more than one symbol (e.g. a lower-case and upper-case letter, such as the letter pairs 'w-W', 'a-A'), this is called allography. When a language has extensive allography there are more symbols to learn.

Examples:

- The Armenian, Cyrillic and Roman scripts have upper- and lower-case letters.
- An example of more extensive allography is Arabic. There are usually 4 possibilities for a symbol in Arabic depending on position in a word, such as |ـبـ|, |بـ|, |ـب| and |ب| for <ب>. There are also exceptions such as the symbols <و> and <ا> that have only two forms.

Marking the differences: If tests of letter knowledge (naming/sounding out symbols) are limited only to items as they look in isolation then the test is missing the essential learning target of how letters change depending on place in a word. Then an important precursor skill for reading proficiency in grade 2/3 is left unmonitored for the language with allography. Meanwhile, tests of reading accuracy and fluency are even less identical than letter knowledge tests in languages with weak and strong allography because the contributing skills needed to do well in the reading tests are not the same (e.g., English *vs* Arabic).

2.2.4 Orthographies differ in degree of completeness

Sometimes information beyond the script is needed to read.

Examples:

- In English, the word 'wind' needs to be read keeping in mind the sentence context ('it is the easterly wind', 'ask them to wind the string tight').
- In Asanti Twi, there are three nasal vowels but they are not written. For example, sentence context decides if the written word <hu> means 'to blow', but if 'to see' it then should be sounded as /hũ/)4.
- In Arabic, when symbols are written with vowel markers the information is complete. But in unvowellised Arabic information is incomplete because only the consonants of a word are written down leaving out accompanying vowels.

Marking the differences: Unvowellised Arabic is introduced around grade 4 in many countries, a grade-level that is outside the reporting range for SDG 4.1.1a. However, if comprehension of unvowellised texts is to transition well for Arabic learners, then the important precursor skills to monitor would be vocabulary and grammar knowledge[5]. This is because unvowellised words in a sentence, such as a word with a three-consonant word pattern, carry both meaning and grammatical information that must be recognised for comprehension.

2.2.5 Orthographies differ in degree of predictability

In the parameters described till now, there is the idea of a rule and the exception to the rule. When there are more exceptions to the rule, then there is more to learn. Once learnt, the rule may be generalised. It also means that such information can become redundant (automatised)

---

[4] Schroeder, L.L. & Nindow, M.O. (2023). Ghana's orthographies shape literacy curriculum design. In: Joshi, R.M., McBride, C.A., Kaani, B., & Elbeheri, G. (eds) Handbook of Literacy in Africa. Literacy Studies, vol 24. Springer, Cham. https://doi.org/10.1007/978-3-031-26250-0_14

[5] A concern is about affordable, large-scale language assessments whose results are relatively comparable across countries. A new generation of assessments may be more suitable than earlier attempts at such assessments. Example of language assessment in Arabic: Saiegh-Haddad, E & Schiff, R (2024). Diglossic and orthographic features of reading comprehension in Arabic. *Reading Research Quarterly, 60, 1,* https://doi.org/10.1002/rrq.598

releasing much-needed attentional resources for monitoring meaning for reading comprehension.

Marking the differences: Between two languages of India, Kannada and Bengali, the written consonant-consonant strings are more predictable in Kannada compared to Bengali. Tests of word reading and reading fluency that carry such strings are therefore not strictly equivalent across the two languages, and pace of learning may differ.

2.2.6 Orthographies differ in degree of asymmetry

Between the print-to-speech connections needed for reading and the speech-to-print linkages used for spelling. The asymmetry may be at all levels: consistency, complexity, completeness, allography and predictability.

Marking the differences: Although German and English belong to the West Germanic language family, German is more asymmetric than English. In English, reading and spelling can both have several inconsistencies, placing a more-or-less symmetric learning demand for both tasks (e.g., knowing not to read the silent letters [underlined] in '<u>k</u>now' and '<u>w</u>rite', and remembering to add these silent letters when spelling). However, German is more inconsistent only for spelling. Tests for reading and spelling are therefore not strictly equivalent across the two languages, and pace of learning may differ.

In conclusion, it is possible to think of language opacity narrowly as the extent of consistency and inconsistency. A broader definition would also consider the other characteristics of complexity, completeness, allography and predictability.

Based on the above, Table 1a considers implications for benchmarking by language groups.

Table 1a. Implications for a Benchmarking by Language Groups Framework based on variety in scripts

Based on the variety in the scripts in which languages are written			
Three inter-linked questions	1. Can the concept of 'language opacity' be used for 'grouping languages for their learning demands at the earliest level of reading proficiency'?	2. Is there sense in 'offering benchmarks by language group rather than for specific languages'?	3. Is there sense in 'offering benchmarks by language group using a criterion of orthographic transparency'?
Conclusion	Partial yes.	Yes.	Partial yes.
Implications	The most common parameter for grouping could be orthographic consistency. Other parameters for grouping could be complexity, completeness, allography and predictability.	Many languages are similar and common benchmarks may be provided for these. For languages with limited data there is a practical advantage because it becomes possible to offer benchmarks by identifying their language group.	Orthographic transparency (or consistency) is only one parameter on which languages differ. Keeping a tight focus only on orthographic transparency misses other parameters that also shape precursor skills for reading comprehension.

2.3 The types of words in the language (the semantics and morphology)

2.3.1 Many words are constructed with two or more meaning-units bundled together

One reason why some languages have many long words is because each word is densely packed with many smaller meaning-units. The meaning-units are called morphemes and poly-morphemic words may be constructed through different word construction processes.

Examples:

- Morphemes may be joined up to communicate specific meanings. The addition of 'mid' in English helps distinguish time between 'night' and 'midnight', similar word lengthening is seen in translation equivalents in Tagalog—'gabi' & 'hatinggabi' and Hindi—'raathri' & 'madhyaraathri'). These kinds of words are called compound words. Some languages have a lot of compound words.

- Some languages have many long words because of a phenomenon called agglutination. In agglutination, word parts called affixes are added. An example of an affix in English is the past tense -ed in a word like 'fixed' (fix + ed). Some languages attach affixes only to the ends of words (suffixes), others attach mainly to the beginnings (prefixes), within (infixes) or across the word (circumfixes). Some languages use all varieties of affixation.

Marking the differences: Albanian, Finnish, Mandarin, Tagalog and Turkish are strong in compounding. The southern Bantu languages of Africa, the Dravidian languages of South Asia and the Malayo-Polynesian languages of Southeast Asia are strong in affix use. Benchmarking separately for these language groups acknowledges their word construction processes. Word reading tests in grades 2/3 could then include common polymorphemic words because these are essential for reading comprehension in the language. In addition, children who have insights into the internal morphemic structure of words are better at word reading in these languages. Current assessments at scale do not extend to morphological awareness, leaving a gap in the tracking of an important precursor skill for some languages.

2.3.2 Words differ in degree of semantic transparency

When the parts of a word give a clue to its meaning then the word has semantic transparency (e.g., 'paint-painter'). The connections between other linked words are opaque or less clear ('deep-depth'). Words are also semantically opaque if they are incomplete (see 2.2.4). Whatever the reason for opacity, the learning becomes easier with increase in transparency.

Marking the differences: In unvowelised Arabic, words must be recognised based on knowledge about word families that share a core consonant string or 'word pattern'. For example, the consonant string <ktb> refers to 'writing' in 28 of the 31 words it appears in[6] (Seghier and Boudelaa, 2024). This means that the decoding of <ktb> will be semantically 'transparent' most of the time but there are 3 instances when it is 'opaque'. Benchmarking vocabulary knowledge

[6] Seghier, M.L., & Boudelaa, S. (2024). The view from Arabic.

becomes important to track precursor skills for reading comprehension for this language group.

2.3.3 Words differ in prosody

Word rhythm, word stress and word tone together make up word prosody. In some languages these features are used to provide the meaning of a word. One study shows that lexical tone awareness in Cantonese and lexical stress awareness in English are important precursor skills for reading comprehension[7].

Examples:

- Tonal languages such as Thai, Punjabi and Igbo communicate word meanings by tones that have names like high tone, mid tone, falling tone, and neutral tone. Children learning tonal languages show awareness for minute tonal differences that are not noticed by newcomers to the language.
- Languages with lexical stress similarly use sound variations to communicate different meanings for the same base word. In English for example, stress changes meaning such as OBject is a solid thing, but obJECT is going against someone or something.

The above list on word types is not exhaustive. By one count there are 125 different morphological processes that may be used to characterise languages and by asking the simple question 'how much of morphology is there?' scholars have ranked languages for word variety[8].

[7] Tong, *at el.* (2017). Tone matters.
[8] Donohue, M. & Gil, D. (2024). Morphology in *The Oxford Guide to the Malayo-Polynesian Languages of Southeast Asia*.

2.4 The nature of sentence construction in a language (syntax)

It appears that when a word has a predictable and regular pronunciation then symbol-sound decoding is sufficient for word reading. But when a word has an unpredictable and irregular pronunciation then clues from the sentence within which the word appears (the sentence context) helps with word reading. The type of sentence knowledge can differ by language. For example, in some languages the word order for a sentence is always fixed but in others there is considerable flexibility. Understanding how word order works is sentence-level knowledge.

There is probably a greater dependence on sentence-level skills in opaque languages than is usually appreciated. Thus, a grouping parameter of language opacity is useful but there is more than just orthographic transparency to consider.

Examples:
English, Arabic and Tamil are low in orthographic transparency. Precursor code and oral language skills are tightly interwoven in these languages, and well-developed sentence-level knowledge offer clues to support decoding and/or accurate recognition of opaque words.

The ambition of a benchmarking by language groups exercise could address one or more of the characteristics described above. Table 1b considers further implications for benchmarking based on language diversity at the word- and sentence-level.

Table 1b. Implications for a Benchmarking by Language Groups Framework based on variety in word and sentence types

Based on variety in words & sentences that make up different languages			
Three inter-linked questions	1. Can the concept of 'language opacity' be used for 'grouping languages for their learning demands at the earliest level of reading proficiency'?	2. Is there sense in 'offering benchmarks by language group rather than for specific languages'?	3. Is there sense in 'offering benchmarks by language group using a criterion of orthographic transparency'?
Conclusions	Partial yes.	Yes.	Partial yes.
Implications	A grouping parameter to consider is word length. Word length captures important language differences in word patterns.	Many languages are similar in their word patterns. Common benchmarks may be provided for these.	Keeping a tight focus only on orthography misses language-related precursor skills for reading comprehension.

	This parameter is not directly about language opacity but is a visible and intuitively easy parameter to grasp for stake holders.		A grouping parameter to particularly consider is word length.

2.5 What else to consider?

Two issues go alongside a discussion on benchmarking by language groups: Frist, is the assessments used to track precursor and reading comprehension skills. Second, is an expert-informed approach to making language groups. Several points need consideration within each.

2.5.1 Psychometric and psycholinguistic properties of assessments

The primary objective of the current exercise is to: '"benchmark" levels of these precursor skills that are associated with particular languages, or language groupings, for a reading comprehension level of 80% (of comprehension questions answered correctly) and 60%.' There are multiple assumptions inherent in this effort. Some are listed below:

2.5.1.1 The tests of precursor skills

The current evidence (introduced in brief in the previous sections) shows that code skills and oral language skills are the precursors to reading comprehension in *all* researched languages. The literature also shows that assessments of these skills carry the same name globally (e.g., letter/symbol knowledge, listening comprehension). However, *the specifics of the items that make up the test* will differ by the psycholinguistic characteristics of the script and language.

Assumptions

- The tests used to assess precursor skills will comprise questions that measure learning on key script and language characteristics important for reading comprehension.
- The differentiation of precursor levels will be particularly sensitive to children whose skills are emergent (those well below the 60% level of reading comprehension).
- The assessment of precursor levels will also be sensitive to skill differentiation among children who are not struggling but also not yet able to apply them during reading comprehension (those at the 60% and 80% level of reading comprehension).

A current strength for many languages is the advances made for the national monitoring of precursor code skills (but see *Marking the Differences* in sections 2.2 to 2.4). Assessing code skills have the benefit of simplicity when compared to assessing oral language skills, and much work remains to be done for national monitoring of language skills.

2.5.1.2 The tests of reading comprehension

Reading comprehension tests will be the reference point for the benchmarking exercise. Much has been specified for the psychometric quality assurance of these tests. Test questions will have to work well for the purposes of the benchmarking at 60% and 80% levels.

Assumptions

- The test used to assess reading comprehension will comprise questions that can pick up differences in all levels of proficiency well.
- The test questions will be particularly sensitive to children who are have developed the precursor skills essential for basic comprehension of texts (e.g., are scoring above 30% on a grade-level reading comprehension test) but who are not yet securely on their way to accomplishing the level of reading comprehension expected for their grade (e.g. scores around 60% and 80%).

Constructing tests with high-quality psychometric and psycholinguistic properties is an ongoing effort. In a later section, assumption testing will be applied to available datasets (section 3.1). This, and earlier analyses of assessments of foundation learning globally, shows that progress has been made for psychometric targets but not yet for psycholinguistic targets.

2.5.2 A system for language grouping

A framework to benchmark by language groups will require *a system for sorting languages*.

2.5.2.1 Quantifying using language corpora

It is possible to quantify characteristics such as the ones in sections 2.2 to 2.4 and let this information determine the language groups. For a small set of languages, the quantification can be based on **large language models** with automatized parsers that can split and count units within sentences and words. For many more languages, the quantification can be on the language in children's books. These books provide the written language children must read

and comprehend and are a good real-world resource to characterise the language and the script[9]. Protocols for such work have recently become available[10]. The quantification methods are based on advances in the field of **corpus linguistics**.

For many other languages, the technical infrastructure and computational resources for quantification of language characteristics are not yet available. This is because the digital resources needed (e.g., automatized parsers to split and tag language units and child language corpora) are yet to be developed. The approach for these languages will be to use expert judgement (described next) to locate their position within a language group.

2.5.2.2 Consensus building approach and expert judgement

The Delphi method draws on a knowledgeable expert panel to build consensus on a topic of interest[11]. For our purposes, the panel of experts would be psycholinguists, researchers of child language and literacy, and early grade literacy and language educators. This panel is assumed to draw upon available linguistic descriptions (the descriptive grammars) and qualitative evidence on how language and script characteristics are shaping the precursors to reading comprehension. The consensus-building would be on characterising the language and script on parameters such as the ones in sections 2.2 to 2.4. The Delphi process starts with an initial set of statements first rated independently and then everyone seeing what other panel members said, with who-said-what kept anonymised. Discussions may then lead to individual positions changing and a revision of how the language should be characterised. More rounds of ratings are elicited on revised statements about the language and script. A pre-set number

[9] For an example see Nag, (2022). How children learn to use a writing system: Mapping evidence from an Indic orthography to written language in children's books.
[10] For an example see Nag et al. (2024). NSP-SCD: A corpus construction protocol for child-directed print in understudied languages.
[11] Examples of the Delphi method informing topics linked to child language and literacy include a) Bishop et al. (2017). Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. J Child Psychol Psychiatry. 58(10):1068-1080. https://doi.org/10.1111/jcpp.12721 b) Carroll et al. (2025), Toward a consensus on dyslexia: findings from a Delphi study. J Child Psychol Psychiatr. https://doi.org/10.1111/jcpp.14123.

of rounds are conducted prior to drawing up a consensus statement. This consensus document provides the data for entering a language into a language sorting system.

2.6 Framework development: Two approaches

The second objective of the literature review (1.1.2) was to develop a framework that may be used for global monitoring of reading proficiency. Two frameworks are considered: using opacity (consistency) (section 2.6.1) and opacity *plus* other word characteristics (section 2.6.2).

2.6.1 Language grouping by language opacity

This language characteristic has received the most research attention. A classic grouping of languages by orthographic opacity is to use two categories labelled as the shallow orthographies and the deep orthographies. The prediction is that the languages that fall in the shallow cluster will be easier to learn.

> *"...because shallow orthographies have relatively simple, consistent, and complete connections between letter and phoneme, it is easier for readers to recover more of a printed word's phonology prelexically by assembling it from letter-phoneme correspondences."* *(Katz & Frost, 1992, pp. 71-72[12])*

A single parameter approach is however less efficient. A worked example is a comparison between French and English:

> *"If we were to pick two orthographies that are comparable in terms of complexity, but different in terms of predictability (e.g., French and English), we would expect that learning these correspondences would take the same amount of time. However, after the correspondences are learnt, we would expect that the accuracy in applying these*

[12] Katz, L., & Frost, L. (1992). Reading in Different Orthographies: The Orthographic Depth Hypothesis. In R. Frost, & L. Katz (Eds.), Orthography, Phonology Morphology, and Meaning (pp. 67-84). Amsterdam: Elsevier.

correspondences to new words would be higher in the more predictable orthography."

(Schmalz et al. 2015[13])

This means that if language grouping was only with one parameter then benchmarks for letter sound would be the same for both languages within the group but not the benchmarks on word reading and nonword reading. On the other hand, if both complexity and predictability are considered for the language grouping then French (and similar other European alphabetic languages) and English would fall into different groups and have their own benchmarks for word reading and nonword reading.

Table 2 lists two further approaches within the field of language opacity and weighs their usefulness for a Benchmarking by Language Group Framework.

Table 2. Approaches to ranking by language opacity with implications for a global framework

Approach	A key example	Implications
An estimation approach based on the predictability of pronunciation of the first letter in words This approach focuses on what is called onset entropy. Here, 'the number of different ways in which the initial letter of a word, on average, can be pronounced in a given orthography' is calculated. When there are more ways to pronounce, the language is more opaque.	5 European alphabetic languages Moll, et al. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learning and Instruction, 29*, 65–77.	This approach ignores other parts of the word where difficulties may arise (e.g. later vowels in a word in English and Asanti Twi, or the later complex syllables in German and Russian). The approach has limited use given the range of ways languages differ beyond the first symbol. **An approach is needed that looks at all parts of the word** and not just the start of the word (see next row).
Intuitive approach based on expert judgement This approach has typically focussed on the number and types of rules in a language focussing especially on	13 European Alphabetic languages Seymour, P., Aro, M., & Erskine, J. (2003).	May be applied to multiple languages **based on a pre-set list of language characteristics**.

[13] Schmalz X, Marinus E, Coltheart M, & Castles A. (2015). Getting to the bottom of orthographic depth. *Psychonomic Bulletin and Review.* 22(6):1614-29. https://doi.org/10.3758/s13423-015-0835-2

complex rules. When there are more rules or more complex rules then the language is more opaque. This approach does better than the onset entropy (above row) because the approach considers language characteristics in any position in a word.	Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143–174.	May be kept specific to language consistency (2.2.1) or may incorporate other parameters of orthographic transparency listed in 2.1 (complexity, completeness, positional allography and predictability).

A robust approach to quantification by opacity is available although the number of languages that have been ranked using quantitative methods is still small. Instead, the trend is to use expert judgement to identify the language group of languages that are less studied.

2.6.2 Language grouping by other word characteristics (length, morphology, agglutination)

There is a strong theoretical rationale from the science of reading for including language characteristics beyond 'Language Opacity' and 'Orthographic Transparency'. Examples from four distinct script families demonstrate the nature of the converging evidence base.

Examples[14]:

- Northern Europe: In the semi-transparent Roman-alphabet-based Norwegian, oral language skills make a two-tiered contribution to Grades 2 to 4 reading comprehension. Oral language is a direct, strong and stable precursor to reading comprehension. Oral language also works indirectly by supporting code-related precursor skills to reading comprehension. Here, oral language skills include vocabulary, listening comprehension and grammar knowledge, tapping word knowledge beyond code skills. Together, language and code skills, 'with their interaction and curvilinear effects, explain almost all (99.7%) of

[14] Hjetland, *et al.* (2019). Pathways to reading comprehension. | Crawford, *et al.* (2025). Inadequate foundational decoding skills constrain global literacy goals. | Hemelstrand, *et al.* (2023). The Impact of Character Complexity on Chinese Literacy | Drummond & Nakamura (2021). The latent data structure in Kyrgyz, Russian & Tajik. | Nag, (2025).

the variance in reading comprehension skills at 7 years of age.' (Hjetland et al., 2019, abstract)

- Multi-country: A recent 48-country study (covering 96 languages using an alphabetic script; 75% in a local or national language, and 25% in the former colonial languages of English, French, Spanish and Portuguese), signalled forcefully the critical role of code skills (section 2.2). But there was another inference implicit in this multi-language analysis (see added emphasis):

> *"Most crucially, pupils must acquire the understanding that letters represent sounds, and they must learn to retrieve letter-to-sound mappings fluently to decode printed words. This decoding process allows pupils* ***to use their spoken vocabulary*to access the meanings*** *of unfamiliar printed words, and it provides the basis for developing readers to get the reading practice vital for building proficiency through the later primary and secondary school years."* *(Crawford et al., 2024, * see next for discussion)*

* A wider approach to benchmarking of precursor skills actively monitors and tracks the multiple reasons why there is a need 'to access meanings' to decode. These reasons include word complexity, completeness, allography, word length, morphology and agglutination[15].

- East Asia: In the visually complex writing system of Chinese, word reading accuracy goes down with an increase in character complexity. Theres is a slightly curvilinear relationship over the early grades, with morphological awareness (oral language skills) reducing the challenge of learning complex characters. (Hemelstrand et al., 2024).
- Eastern Europe, Central and Northern Asia: In languages from three language families using the Cyrillic-based script (Kyrgyz, Tajik and Russian) two latent variables emerged when a comparable set of precursor skills were assessed: decoding and oral language

[15] In tandem, a language grouping framework that includes these multiple script and language dimensions clearly signals the role of meaning-focused precursor skills for high-quality reading comprehension. However, several practical constraints remain around assessment of oral language skills with conclusions such as 'benchmarking of morphological knowledge is, therefore, not feasible at this time" (pg 14). But, distillations from university-based assessment research are available. There are good examples that allow approximation to other global metrics that are "comparable (if not psychometrically equivalent) across different countries and for different languages".

comprehension. One conclusion from this multi-language study is that benchmarking code skills cannot be a 'proxy' for oral language skills. Oral language provides new information (Drummond & Nakamura, 2021).

- South Asia: In the Indic akshara writing system of Kannada, children encounter a bewildering range of inflections in story books. These words have been described as requiring 'morphological processing around polysemy, homonymy, phonologically conditioned allomorphy, and morphophonological changes' (Nag, 2025). All words follow systematic word formation rules.

The quantification of language differences by complexity, completeness, allography, the word length, morphology and agglutination characteristics is currently dependent on descriptive grammars (detailed linguistic reports) for most languages. These are often sufficient to apply a principle for language grouping using similar approaches as for Opacity (2.6.1): a) use expert judgement along with b) a child-directed print corpus to quantify characteristics.

A framework based on a multi-dimensional view of what shapes precursor skills

Language opacity (consistency) and other word characteristics (such as complexity, completeness, positional allography, predictability, word length, morphology, agglutination) together provide a more comprehensive framework for benchmarking precursor skills.

3. Patterns of child performance

In this section the two approaches to a benchmarking framework are applied to child performance datasets from a small set of languages made available by UIS. The aim is to investigate data patterns by language opacity and other characteristics. The 31 languages and the skills covered are given in Table 3. The data cover eight language families.

3.1 Description of datasets

All datasets cover multiple precursor code skills. **Word decoding and reading fluency are always assessed**. Letter/symbol knowledge and phonological skills are sometimes assessed. All datasets are light on precursor language (as opposed to just decoding) skills. There is **no specific information on morphology- and sentence-level skills** but listening comprehension is always assessed and sometimes there is a brief vocabulary test. The reading comprehension test in all but one dataset comprised 5 questions based on one short passage. Reading the passage simultaneously supplies fluency information from words correctly read per minute. Only those reading comprehension questions that apply to the text read within the pre-set time are administered.

Test characteristics were examined for psychometric (section 3.1.1) and psycholinguistic features (section 3.1.2). Based on this, decisions were made on which languages could be used to examine the two benchmarking frameworks (section 3.1.3).

Table 3. Precursor skills measured across 34 datasets for 31 languages from 8 language families

Datasets	Reading Comp. Passage Length (Number of sentences)	Precursor Code Skills					Precursor Language Skills			
		Symbol Knowledge		Decoding			Word knowledge		Sentence and Discourse level	
		Name	Sound	Syllables	Familiar Words	Invented Words	Vocabulary	Morphology	Listening Comp.	Grammar
Arabic[1]										
Region 1	42		✓	✓		✓			✓	
Region 2[2]	76		✓			✓			✓	
Bantu language family										
Chichewa	xxx[3]			✓	✓				✓	
Kinyarwanda		✓		✓	✓				✓	
Swahili				✓		✓				
Chitonga	56 (11)		✓	✓		✓			✓	
Cinyanja	48 (7)		✓	✓		✓			✓	
Icibemba	43 (8)		✓	✓		✓			✓	
Kikaonde	50 (8)		✓	✓		✓			✓	
Lunda	43 (7)		✓	✓		✓			✓	
Luvale	49 (8)		✓	✓		✓			✓	
Silozi	57 (7)		✓	✓		✓			✓	
Indo-Aryan language family[5]										
Assamese			✓[4]		✓	✓	✓		✓	
Bangla			✓[4]		✓	✓	✓		✓	
Gujarati			✓[4]		✓	✓	✓		✓	
Hindi			✓[4]		✓	✓	✓		✓	
Konkani			✓[4]		✓	✓	✓		✓	
Marathi			✓[4]		✓	✓	✓		✓	
Nepali			✓[4]		✓	✓	✓		✓	
Odia			✓[4]		✓	✓	✓		✓	
Punjabi			✓[4]		✓	✓	✓		✓	
Urdu			✓[4]		✓	✓	✓		✓	
Dravidian language family[5]										
Kannada			✓[4]		✓	✓	✓		✓	
Malayalam			✓[4]		✓	✓	✓		✓	
Tamil			✓[4]		✓	✓	✓		✓	
Telugu			✓[4]		✓	✓	✓		✓	
Sino-Tibetan[5]										
Bodo			✓		✓	✓	✓		✓	
Garo			✓		✓	✓	✓		✓	

Austroasiatic[5]										
Khasi			✓		✓	✓	✓		✓	
Tibeto-Burman[5]										
Manipuri			✓		✓	✓	✓		✓	
Mizo			✓		✓	✓	✓		✓	
West Germanic[5]										
English			✓		✓	✓	✓		✓	

Note: [1]Two datasets couldn't be tagged to a country. This was needed to establish degree of challenge on precursor skills due to diglossia. [2]One additional test available, reading comp. test has 7 items (others have 5).[3]Items not available for language description.[4]Symbol recognition in this writing system is equivalent to syllable reading. [5]A phonological test also available.

3.1.1 Psychometric details

Item- and test-level statistics are taken from supplied data tables. The sample sizes per item/test range from 2 (a final question in a reading comprehension test in a particular language) to 8087 (a syllable sound test in a particular language). Available information is item difficulty level, item-total (test score) correlation and Cronbach's alpha (test reliability). From a psychometric point of view, the items are satisfactory to excellent: item difficulty levels vary, item-total correlations are mostly above .20 with several above .90. Cronbach's alpha is consistently above .95 for one language family and between .60 and .80 for the rest.

3.1.2 Psycholinguistic details

Items do not always reflect **language characteristics relevant to track precursor code skills** for a particular language and script. These are the salient characteristics of words that would be found in books for Grades 1 to 3 (such as described in 2.2). The main reason for not finding psycholinguistically-tailored items is the use of translated items. A translation method is often adopted to ensure the same idea units are used across listening or reading comprehension tests. The method also helps provide surface level equivalence for comprehension questions (e.g. direct, simple or inferential, integrative). But the translation method does not easily allow for equivalence on most other precursor skills. For example, translated words may differ in transparency, complexity, completeness and morphological richness across the language set. Importantly, the translated word may not carry the most essential characteristics to assess.

Example:

- Tests of 19 language across 5 language families (Indo-Aryan, Dravidian, Sino-Tibetan, Austroasiatic, Tibeto-Burman) were developed by translating an English test.

The language in tests may use more common words found in early vocabularies. A later acquired vocabulary implies the words may be less familiar to children in the early grades and hence unavailable in the spoken vocabulary to support decoding and access meaning. This means, precursor language skills may play a more substantial role for the test with the later vocabularies compared to the test with the early vocabularies.

Example:

- Tests in Arabic from two regions had items with different vocabulary demand.

In contrast to precursor code skills, the datasets are limited for **precursor language skills**. Assessment is narrowly focussed on listening comprehension. For several languages, a five-item vocabulary test is available.

3.1.3 Decisions on use of the datasets

Strengths of the dataset relevant for a comparison of frameworks were identified. The first decision was to work with **high-contrast language pairs** and **script-matched language pairs.** The second decision was to work **with code skills,** in particular **reading fluency**. The rationale for the approach is given below.

Decisions:

- The datasets are excellent at the level of language pairs that are comparable on multiple psycholinguistic parameters. When they then differ on a parameter/tightly contained set of parameters this provides a high-contrast pair to examine what the precursor skills look like on a comparable test. Similarly, should script-matched languages still display the variations on comparable tests because they are across language families, this is informative.
- Pattern analysis using all languages within a language family was abandoned because many tests were not well-tailored for language and script characteristics (psycholinguistic). This is despite many tests having moderate to excellent internal consistency (psychometric).

- Test booklets were not available for some languages that may be locally written in different scripts. This made it impossible to decide what the script demand was (e.g. the decoding demands of the Latin alphabet versus the Indic akshara).
- Ideally, both sets of precursors—code and language—should have been examined. However, only code skills will be focussed on because unlike language, assessment of code skills are available per language and are psychometrically more robust. Reading fluency was picked because this skill demonstrates similarities and differences in benchmarks most vividly.

Expert judgement[16] was used to identify high-contrast examples using both language and script characteristics. The script matching was kept intuitively simple by limiting choice to a globally familiar script: the Latin alphabet. Decoding demand was hypothesised based on what is known to speed up and slow down precursor skills. Tables 4 and 5 give the pairs.

Table 4. High-contrast language pairs chosen to examine the single and multi-parameter frameworks

Pairs (language family)	What is high-contrast?	What is closely similar?
Hindi – Urdu (Indo-Aryan)	Code • The writing systems differ in transparency: Urdu has more visual complexity, substantial allography, and is more opaque.	Language • Sister languages with shared vocabulary, morphology and grammar. • The languages differ slightly in sound inventory.
Malayalam – Tamil (Dravidian)	Code • Tamil is less transparent; Malayalam has longer words. • There is more linguistic distance between spoken and written Tamil.	Language • Both are morphologically rich. Code • Both use the akshara writing system. • Both are non-linear but also have several notable linear features.
Chitonga - Silozi[1] (Bantu)	Code • Chitonga joins up morphemes (conjunctive), a few digraphs (~17). Words are long.	Language • Morphologically rich. Code • Both use the Latin alphabet. • Both are tonal languages.

[16] Including from descriptive grammars, published research, and asking language experts.

	• Silozi has many morphemes separate (disjunctive) and has many digraphs & trigraphs (~40). Words are short.	• Both are mostly transparent
Icibemba – Silozi (Bantu)	Code • Icibemba joins up morphemes (conjunctive), a few digraphs. Words are long. • Silozi has many morphemes written separate (disjunctive) and many digraphs & trigraphs (~40). Words are short.	Language • Morphologically rich Code • Both use the Latin alphabet • Both are tonal languages • Both are mostly transparent

Note: [1]Contrast chosen because reading comprehension passages had similar number of words.

Table 5. Script-matched pairs chosen to examine the single and multi-parameter frameworks

Pairs (language family)	What is closely similar?
Icibemba – Lunda (Bantu)	Language • Morphologically rich, conjunctive hence long words, tonal Code • Both use the Latin alphabet, mostly transparent
English – Khasi (West Germanic & Austroasiatic)	Code • Both use Latin alphabet, often opaque

3.2 Data patterns

Using the single parameter of language opacity works many of the times but not always. To demonstrate this, reading fluency data for language pairs are presented. The data are simply the sample means of words read correctly per minute.

Instances when the single parameter approach work are shown with high-contrast pairs that differ by opacity (section 3.2.1) and script-matched pairs that differ by opacity (3.2.2). The single parameter does not work when the language and script characteristic shaping precursor skills is another word characteristic. Here, the weakness of the single parameter approach is demonstrated using morphologically rich languages that are either conjunctive or disjunctive (with long or short words that also lead to few or many words per sentence) (section 3.2.3).

3.2.1 High-contrast pairs that differ by opacity

Using the single parameter of opacity works well when the over-riding difference between two languages is opacity, even if there are several other co-occurring differences. Table 6 provides fluency data for high-contrast pairs from 2 language families.

Table 6. High-contrast language pairs predicted just as well by both frameworks

Pairs	Language Opacity indicator	Single parameter of opacity[1]	Multi-parameter[2]
Hindi – Urdu (Indo-European language family)	H more Transparent U more Opaque	Prediction: Benchmarks for Reading Fluency will be lower for Urdu than Hindi	
		Urdu – Hindi Reading Fluency benchmarks for 80% RC: Benchmark 56 65 Lower boundary 42 55; Upper boundary 64 66 Reading Fluency benchmarks for 60% RC: Benchmark 39 61 Lower boundary 26 52; Upper boundary 52 65	
		Differences captured	Differences captured
Tamil - Malayalam ([Dravidian language family)	M more Transparent T more Opaque	Prediction: Benchmarks for Reading Fluency will be lower for Tamil than Malayalam	
		Tamil – Malayalam Reading Fluency benchmarks for 80% RC: Benchmark 40 58 Lower boundary 27 49; Upper boundary 51 61 Reading Fluency benchmarks for 60% RC: Benchmark 26 54 Lower boundary 16 43; Upper boundary 38 55	
		Differences captured	Differences captured

Note: [1]This is the commissioned framework. [2]Opacity + other word characteristics (such as complexity, completeness, positional allography, predictability, word length, morphology, agglutination).

3.2.2 Script-matched pairs that differ by opacity

The parameter of language opacity will also work well when languages use the same script, but not always. **Other differences also matter.** Table 7 provides information with Latin-based scripts. Individually, the language pairs in Panels 1 & 2 show how benchmarks are close when languages pairs are closely similar for opacity. But together, Panels 1 & 2 show how benchmarks vary substantially when languages are different for opacity *and* other word characteristics (such as complexity, word length, morphology, agglutination). The difference is of words per minute in the 20s compared to 60s.

Table 7. Script-matched pairs predicted just as well by both frameworks

Pairs	Single parameter of opacity	Multi-parameter
Prediction: Benchmarks for Reading Fluency will be similar for both languages		
Panel 1	Transparent ↕ Opaque	Isibemba – Lunda [Bantu language family, transparent] Reading Fluency benchmarks for 80% RC: Benchmark 23 21; Lower boundary 17 14; Upper boundary 29 26; SE 15.13 70.61 Reading Fluency benchmarks for 60% RC: Benchmark 12 14; Lower boundary 6 8; Upper boundary 18 20; SE 3.25 27.73
	Similarities captured	Similarities captured
Panel 2		English – Khasi [West Germanic & Austroasiatic, opaque] Reading Fluency benchmarks for 80% RC: Benchmark 67 65; Lower boundary 57 56; Upper boundary 68 68 Reading Fluency benchmarks for 60% RC: Benchmark 65 60; Lower boundary 57 50; Upper boundary 68 65
	Similarities captured	Similarities captured

3.2.3 Language pairs with precursor skills shaped by another word characteristic

A multi-parameter approach will work well when the over-riding differences between two languages is on **parameters other than language opacity**. The differences chosen for analysis is the contrast when morphologically rich languages either join up all the morphemes to create long words or keep them separate resulting in many small words. The number of words per sentence then changes. Table 8 shows the contrasting performance. All examples are from one language family. Note how the prediction changes depending on the framework (underlined).

Table 8. High-contrast pairs predicted better by the multi-parameter framework

Pairs	Single parameter of opacity[1]	Multi-parameter[2]
Chitonga - Silozi (Bantu language family)	Prediction: Benchmarks <u>should be the same</u> for Chitonga and Silozi	Prediction: Benchmarks <u>will be lower</u> for Chitonga than Silozi
	Chitonga – Silozi Reading Fluency benchmarks for 80% RC: Benchmark 26 33; Lower boundary 19 27; Upper boundary 29 33; SE 11.84 70.61 Reading Fluency benchmarks for 60% RC: Benchmark 16 27; Lower boundary 9 21; Upper boundary 23 33; SE 3.64 27.73	
	Differences ***not*** captured	Differences captured

Pairs	Single parameter of opacity[1]	Multi-parameter[2]
Isibemba – Silozi (Bantu language family)	Benchmarks should be the same for Isibemba and Silozi	Benchmarks will be lower for Isibemba than Silozi
	Isibemba - Silozi Reading Fluency benchmarks for 80% RC: Benchmark 23 33; Lower boundary 17 27; Upper boundary 29 33; SE 15.13 70.61 Reading Fluency benchmarks for 60% RC: Benchmark 12 27; Lower boundary 6 21; Upper boundary 18 33; SE 3.25 27.73	
	Differences ***not*** captured	Differences captured

Note: [1]This is the commissioned framework. [2]Opacity and other word characteristics (such as complexity, completeness, positional allography, predictability, word length, morphology, agglutination).

3.2.4 When other reasons explain the data

Quality of instruction silently shapes all benchmarking estimations. How reading is taught is outside the scope of psychometrics and psycholinguistics but most certainly explains some of the data.

Example:

- There have been multiple reports expressing concern for the stubbornly low attainments in country X. In the supplied data, benchmarks across all precursor skills for language X in country X were persistently lower than benchmarks in a sister language Y in country Y that shares closely similar morphological, complexity and transparency features. In addition, in the contrasting language Z from country Z that has more orthographic complexity and incompleteness but similar morphological features, the benchmarks were more advanced than the more transparent Language X.

Put differently, when languages are taught in poorly resourced settings, the data for benchmarking are skewed to a lower range. Therefore, it will be important to be alert to the instruction environment that the data for establishing benchmarks come from.

3.3 A sense-check of the two frameworks

Taken together, section 3.2 demonstrates the several reasons why the single parameter of language opacity may be a sub-optimal framework for the world's scripts and languages. Looking across language families, a benchmark based in a single parameter does not fit all. This is evident when looking across the languages in Tables 6 to 8. Within a language family, languages with closely similar language and script characteristics have similar benchmark estimates. But, if there is a language within a language family that is different in a key language and/or script characteristic, then the benchmark estimates may shift considerably enough to warrant separate consideration.

In other words, the differences on precursor skills do not always stem from language opacity.

Taking more parameters into account works better to explain why there are differences and provide the rationale for a benchmarking framework.

A further test would be to go beyond the datasets reviewed in section 3, to independent benchmarking efforts. One example is provided.

Example:

- Nguni Language Group: Early grade texts have multiple complex consonants (e.g., ndl, tsh, gcw, ntsw). The orthography is transparent. All languages in the group are morphologically rich but vary in word lengths due to most being conjunctive (many long words) but one disjunctive (many short words). The data from the disjunctive language (isiNdebele) differs from rest that are conjunctive (siSwati, isiZulu and isiXhosa) but is closer to estimates from the disjunctive Sesotho-Setswana language group. Comparing Nguni language group to the Sesotho-Setswana language group, the benchmarks are similar for syllable reading per minute (40 correct per minute by end of Grade 1) but different for words per minute (35 vs 60 words correct per minute by end of Grade 3)[17].

[17] South African Languages Reading Benchmarks Policy Brief 20 November 2023; NM personal communication March 2025

In summary, it would be an error to focus only on opacity to sort languages. Using multiple parameters recognises the diverse pathways to gaining reading proficiency.

Looking beyond Language Opacity for Framework Building

Differences in precursor skills do not always stem from language opacity.

Taking more parameters into account explains the differences better and provides a more nuanced benchmarking framework.

4. Recommendations for a Benchmarking by Language Groups Framework (BLGF)

The recommendation to inform a Benchmarking by Language Groups Framework (BLGF) is based on two findings: 1. there are important differences *within* language families, and 2. these differences are along multiple parameters. Each translates into specific action points.

Action point 1: Variations *within* language families should inform the framework

One immediately available language classification system comes from historical linguistics (traditional language families, genealogical grouping). In this system, languages are grouped together because they share an ancestor, called the proto-language. Daughter languages within this language family may have changed over centuries of migration and borrowings leading to changes in language and script characteristics. Thus, simply using traditional language families (such as the Romance languages, Bantu languages) for language grouping may not be the best approach because of the variations in benchmarks within these families. **Grouping languages by traditional language families will not work** because historic groupings ignore important differences in the acquisition of reading proficiency.

Action point 2: Variations along *multiple parameters* should inform the framework

The language grouping framework needs to work for early grade learners. This means several parameters must inform the grouping, such as the parameters listed in sections 2.2 (*The script in which the language is written*), 2.3 (*The types of words in the language*) and 2.4 (*The nature of sentence construction in a language.* In addition, to the rapid review provided on the current science of reading (section 2) are considerations from supplied datasets (section 3). The recommendation is for **a Benchmarking by Language Groups Framework (BLGF) that is informed by a multi-parameter approach**. If the language grouping parameters acknowledge and respond to learning demands from orthography, semantics, morphology and syntax, this can serve the primary purpose:

To publish 'the "benchmark" levels of... precursor skills that are associated with particular languages, or language groupings'. (Commissioning document, Oct. 2024)

4.1 Framework adoption

Which languages will enter the BLGF? All languages in which early grades reading instruction is offered. All other spoken languages and mother tongues will enter the BLGF when they are offered as a Language of Instruction (LoI). In multilingual and bilingual contexts, while home languages will of course shape precursor skills, it is these children's language of instruction that will enter the BLGF.

A back-of-the-envelop survey suggests the languages for the BLGF may be divided into four groups:

- Tier 1 languages: Adequate to excellent evidence base, descriptive grammars, expert community and child-directed corpus (e.g. some languages within each language family: Bantu, Nguni, Sesotho-Setswana, Indo-Aryan, Dravidian, Malayo-Polynesian)
- Tier 2 languages: Two or more of the above resources (e.g. some languages within: the World Arabics, the Cyrillic-script languages)
- Tier 3 languages: Descriptive grammars as main resource and emergent other resources (e.g. some within: the Austroasiatic languages, the Quechua family)
- Tier 4 languages: Exceptionally limited resources (to be identified)

4.2 What the BLGF may look like in practice?

Populating BLGF: Start with Tier 1 and 2 languages, with pilots for Tier 3 and 4. Follow ideal data analytic approaches such as within Item Response Theory (IRT).

Sorting into languages groups: To be developed. An example would be a three-group system:

- Group 1 languages (short words and transparent words),
- Group 2 languages (longer words and less transparent words), and

- Group 3 languages (many opaque words).

A multi-method approach to sorting the world's languages

The approach to sorting languages into groups will have to be a mix of quantification of languages on pre-set parameters plus expert decisions.
For quantification, frequency of certain characteristics in a corpus of child-directed print (children's books) may be considered.
For expert decisions, a Delphi method for decisions on which language goes into which group may be considered.

5. End note

This report provides a framework on how to structure different easy-to-refer benchmark tables.

The benchmarks may be presented as **a look-up system for all languages of instruction offered in a country.**

The benchmarks may also be presented as **a script-level benchmark table** that may be used by any country with a language of instruction using the script.

It is also possible for countries speaking various languages to be given a table of benchmarks that could 'measure the percent of children at those benchmarks, as a way of gauging progress towards comprehension, if children are not comprehending, or, possibly, identify causes for non-comprehension in terms of earlier skills'.

uis.unesco.org