SETTING READING BENCHMARKS IN SOUTH AFRICA

October 2020

PHOTO: SCHOOL LIBRARY, CREDIT: PIXABAY

Contract Number: 72067418D00001, Order Number: 72067419F00007

basic education
Department:
Basic Education
REPUBLIC OF SOUTH AFRICA

USAID
FROM THE AMERICAN PEOPLE

ISBN: 978-1-4315-3411-1

Contract Number: 72067418D00001, Order Number: 72067419F00007

SETTING READING BENCHMARKS IN SOUTH AFRICA

October 2020

AUTHORS

Matthew Jukes (Research Consultant)
Elizabeth Pretorius (Reading Consultant)
Maxine Schaefer (Reading Consultant)
Katharine Tjasink (Project Manager, Senior)
Margaret Roper (Deputy Director, Khulisa)
Jennifer Bisgard (Director, Khulisa)
Nokuthula Mabhena (Data Visualisation, Khulisa)

CONTACT DETAILS

Margaret Roper
26 7th Avenue
Parktown North
Johannesburg, 2196
Telephone: 011-447-6464

Email: mroper@khulisa.com

Web Address: www.khulisa.com

CONTRACTUAL INFORMATION

Khulisa Management Services Pty Ltd, (Khulisa) produced this Report for the United States Agency for International Development (USAID) under its Practical Education Research for Optimal Reading and Management: Analyze, Collaborate, Evaluate (PERFORMANCE) Indefinite Delivery Indefinite Quantity (IDIQ) Contract Number: 72067418D00001, Order Number: 72067419F00007.

ACKNOWLEDGEMENTS

THIS PUBLICATION WOULD NOT HAVE BEEN POSSIBLE WITHOUT THE TECHNICAL INPUT AND SUPPORT OF A NUMBER OF EXPERTS. KHULISA WOULD LIKE TO THANK EVERYONE INVOLVED IN THIS PUBLICATION.

TABLE OF CONTENTS

TABLE OF CONTENTS 4
LIST OF TABLES 6
LIST OF FIGURES 6
GLOSSARY OF TERMS 7
ACRONYM LIST 1
FOREWORD 2
EXECUTIVE SUMMARY 3
INTRODUCTION 5
NAVIGATION ICONS 6
PART ONE: THE CONTEXT OF READING IN SOUTH AFRICA 8
1.1 SOUTH AFRICA'S OFFICIAL LANGUAGES 8
1.2 LINGUISTIC FEATURES OF SOUTH AFRICA'S AFRICAN LANGUAGES 8
1.3 ORTHOGRAPHIC FEATURES OF WRITTEN LANGUAGES 9
1.4 IMPLICATIONS OF ORTHOGRAPHIES FOR EARLY READING 13
1.5 READING ASSESSMENT 15
1.5.1 ASSESSING READING COMPREHENSION 15
1.5.2 ASSESSING ORAL READING FLUENCY 17
1.5.3 ASSESSING LETTER-SOUND KNOWLEDGE 18
PART TWO: UNDERSTANDING READING BENCHMARKS 20
2.1 WHAT ARE READING BENCHMARKS? 20
2.2 WHY ARE BENCHMARKS IMPORTANT IN SOUTH AFRICA? 20
2.3 WHY SET READING BENCHMARKS? 21
2.4 WHAT READING COMPETENCIES SHOULD BE BENCHMARKED? 21
2.4.1 LETTER-SOUND BENCHMARKS 23
2.4.2 COMPLEX CONSONANT SEQUENCES BENCHMARKS 25
2.4.3 ORAL READING FLUENCY BENCHMARKS 26
2.4.4 READING COMPREHENSION BENCHMARKS 27
2.4.5 READING ACCURACY BENCHMARKS 28
2.5 CAN BENCHMARKS BE SHARED IN SIMILAR LANGUAGES? 28
2.6 IS IT NECESSARY TO SET BENCHMARKS BY GRADE? 30
2.7 WHY SET TARGETS? 30
2.8 HOW TO ESTIMATE BENCHMARKS? 31
2.8.1 NORM-REFERENCED BENCHMARKS 31
2.8.2 CRITERION-REFERENCED BENCHMARKS 31
2.8.2.1 Criterion-Referenced Data Analytical Methods 31
2.8.2.2.1 Mean, Median Logistic And Linear Regression Methods 32
2.8.2.3 Criterion-Referenced Expert-based Methods 33
2.9 ASSESSING VALIDITY OF CRITERION-REFERENCED BENCHMARKS? 35
PART THREE: ASSESSMENT, SAMPLING AND ANALYSIS 37
3.1 ASSESSMENT TOOLS FOR DEVELOPING BENCHMARKS 37
3.1.1 PILOTING AND PSYCHOMETRIC PROPERTIES OF TOOLS 37
3.1.2 LEVEL OF TEXT 38
3.1.3 PASSAGE TIMING 39
3.2 SAMPLING FOR DATA ANALYTICAL METHODS 39
3.2.1 SAMPLE SIZE 39

3.2.2 SAMPLE COMPOSITION (HIGH AND LOW PERFORMING LEARNERS). 42
3.2.3 SAMPLES AND SUBSAMPLES 42
3.3 PRE-ANALYSIS STEPS FOR BENCHMARKING ORF AGAINST COMPREHENSION 42
3.4 METHODS TO ANALYSE ORF AGAINST COMPREHENSION 43
PART FOUR: CURRENT BENCHMARKING PRACTICE IN SOUTH AFRICA 46
4.1 CURRENT BENCHMARKING INITIATIVES 46
4.2 EXISTING DATASETS (KNOWN) 46
4.3 USABILITY OF AVAILABLE DATASETS – PRELIMINARY ASSESSMENT (KNOWN) 47
4.4 EXISTING LEARNER ASSESSMENT INSTRUMENTS (KNOWN) 48
PART FIVE: PRACTICE GUIDE TO SETTING BENCHMARKS 51
5.1 THE BENCHMARKING PROCESS 51
PHASE 1: SET UP DECISION-MAKING STRUCTURES FOR BENCHMARKING 51
PHASE 2: ESTABLISH GOALS OF THE BENCHMARKING PROCESS 52
PHASE 3: DECIDE ON THE APPROPRIATE COMPETENCIES TO BENCHMARK 52
PHASE 4: SELECT ANALYSIS METHOD 54
PHASE 5: SET THE BENCHMARK 55
PHASE 6: SET TARGETS 55
PHASE 7: EVALUATE BENCHMARKS 55
5.2 COMPARING AND PRIORITISING LANGUAGE BENCHMARKS 55
PART SIX: PRACTICE NOTE ON BENCHMARKING STRATEGIES 57
STRATEGY 1: BENCHMARKS BASED ON ANALYSIS OF EXISTING DATASETS 57
Description of Strategy 1 57
Mapping datasets 58
Steps in Determining Whether the Existing Data Can Be Used for Analysis (Pre-Analysis Steps) 58
Process of benchmarking using existing datasets 58
Feasibility of the strategy 58
Test equivalence 59
STRATEGY 2: BENCHMARKS BASED ON PRIORITISED ADDITIONAL DATA COLLECTION59
Description of Strategy 2 59
Objectives for Collecting Additional Data 59
Steps in Determining Whether an Existing Dataset Can Be Topped Up 59
Process of Benchmarking Using Additional Data 59
Feasibility of the Strategy 59
Test Equivalence 59
STRATEGY 3: BENCHMARKS BASED ON PRIMARY DATA COLLECTION 60
Description of Strategy 3 60
Objectives for Benchmarking Language(s) Based on Primary Data 60
Specific Steps in Benchmarking Based on Primary Data Collection 60
Details of Data Collection Process 60
Feasibility of the Strategy 60
Test Equivalence 60
PART SEVEN: ELEMENTS TO SUPPORT BENCHMARKING 61
7.1 BENCHMARKS AND THE DBE CURRICULUM STANDARDS 61
7.2 CURRICULUM AND ASSESSMENT POLICY STATEMENTS 61
7.3 FOUNDATION PHASE 62
7.4 GAPS IN CAPS 63
7.5 EXPERTISE REQUIRED FOR SETTING BENCHMARKS 65
7.6 STAKEHOLDER ENGAGEMENT 67
7.7 COST ELEMENTS 68
7.8 FIELDWORK STANDARDS 73
CONCLUSION 76
BIBLIOGRAPHY 81

ANNEXURE 1: EXISTING DATASETS (2020) 88

ANNEXURE 2: STAKEHOLDER CONSULTATION 93
READING BENCHMARKS WORKSHOP 93
BRIDGE EARLY GRADE READING COMMUNITY OF PRACTICE 94

LIST OF TABLES

Table 1: Words per Sentence in Disjunctive or Conjunctive Orthographies 11
Table 2: Recognition of Single Consonants Versus Digraphs/Trigraphs of Grade 2 isiXhosa Readers 13
Table 3: Letter-sound Knowledge: Letters Correct per Minute 24
Table 4: Complex Consonant Sequences across South African Languages 25
Table 5: ORF Data: Mean Words Correct per Minute 26
Table 6: Four Levels of Reading Comprehension 27
Table 7: Setting Reading Benchmarks by Grade 30
Table 8: Data Analytical Methods 33
Table 9: Assumptions and Estimates Used to Calculate the Sample Size for Grade 1, 2, and 3 40
Table 10: Sample Size of Schools and Learners Based on the Assumptions from Table 9 41
Table 11: Checklist before Beginning Benchmark Analyses 42
Table 12: Summary of Benchmarking Approaches 51
Table 13: Time Allocation for Home Language and First Additional Language 62
Table 14: Stakeholders in Benchmarking 67
Table 15: Generic Cost Elements for Benchmarking Methods 70
Table 16: Summary of Available Data for Benchmarking Across Languages in Grade 1-3 78

LIST OF FIGURES

Figure 1: South African Home Languages per Population Group (2018) 8
Figure 2: Southern Bantu Language Family in South Africa 9
Figure 3: Continuum of Transparency-Opaque Letter-sound Transferring 10
Figure 4: Continuum of Word Length for Disjunctive and Conjunctive Languages 10
Figure 5: Diacritics in Tshivenda 13
Figure 6: Subcomponents of Early Reading Development 14
Figure 7: Nepal Grade 2 Child Predicted Probabilities of Scoring > 80% in ORF 44
Figure 8: EGRS 1 Learner Probability of Reading With 80% Comprehension in Setswana 48
Figure 9: Sample EGRA Training Agenda 74

GLOSSARY OF TERMS

Agglutinating language	A language in which the addition of affixes (small meaning units) to word roots (the most basic part of a word which has meaning) signal changes in meaning, rather than word order or separate words. For example, 'bangazifunda' (isiZulu for 'they may read them') includes the affixes 'ba', 'nga' and 'zi' to change the meaning of the root word, 'funda'. Agglutinating languages include Southern Bantu languages such as isiZulu and Setswana, as well as Turkish and Finnish.
Alphabetic language	In alphabetic orthographies, each symbol or letter represents spoken language at the level of the phoneme.[1]
Analytic language	A language in which word order and separate words convey grammatical information. English is mostly analytic and uses word order to signify grammatical relations. For example, in the sentence: 'The boy helped the girl' the word order indicates that the boy is the subject/agent and the girl is the object/recipient.
Angoff Method	A formal procedure for producing valid benchmarks based on both expert opinion and data.
Benchmark	Benchmarks refer to standards of proficiency in an educational competency, skill or domain. "Benchmarks are particularly useful for reading, as they establish expectations and norms for reading performance."[2]
Complex consonant sequences	Refers to Digraph/Trigraph/Quadgraph (each of which is defined below)
Concurrent validity analysis	The validity of a measure (for example, a fluency benchmark) is assessed by looking at the concurrent relationship with another measure (for example, comprehension). The term concurrent implies that both measures are evaluated at the same time.
Conjunctive orthography	How words are written. In conjunctive orthographies such as isiZulu and Siswati, the affixes to verb and noun roots are most often written next to the root, without orthographic spaces, for example, 'Ngiyamthanda' (isiXhosa: 'I love him/her').
Corpus /corpora	A large, structured set of texts used for hypothesis testing and statistical analysis in linguistics to check the frequency of occurrence of words and validate linguistic rules in a specific language.

[1] Conrad (2016)
[2] Research Triangle Institute (RTI) International's Early Grade Reading Assessment (EGRA) Toolkit (2015)

Criterion-referenced testing /assessment	Criterion-referenced testing measures individual performance against a set of: "… predetermined criteria or learning standards, for example, concise, written descriptions of what children are expected to know and be able to do at a specific stage of their education."[3] The purpose of criterion-referenced testing is to determine whether an individual has acquired a specific set of skills, or whether they have learned specific knowledge (for example, a curriculum). Scoring looks at the number of test-takers who achieve a certain proficiency level, rather than comparing results against a norm (see norm-referenced testing).
Decoding	The ability to recognize words accurately and fluently by the systematic use of letter-sound relationships, which relies on an understanding of the alphabetic principle.[4] In effect, it means to read a word by sounding the letters or letter combinations (such as digraphs and trigraphs), for example. 't-o-k-o-l-o-sh' and 'ch-ai-r'.
Diacritics	A mark on a letter indicating that it should be pronounced differently to the unmarked form; compare '**s**' and '**š**' in Setswana. The former makes the 's' sound in snake, and the latter makes the 'sh' sound in 'ship'.
Digraphs	Refers to the convention of spelling words. Digraphs are a combination of two letters that represent a single sound; for example, 'sh' in ship or 'hl' in 'hlala' (isiXhosa: 'to sit').
Diphthong	A vowel sound in which a vowel and a glide, by a change in tongue position, produce the sound of a single vowel.[5]
Disjunctive orthography	How words are written. In disjunctive orthographies, such as Setswana and Sepedi, the affixes to verb roots are most often written separately before the verb that is with orthographic spaces, for example, 'Ke a ba rata' (Sepedi: 'I like them').
Grapheme	The smallest meaningful symbol in a language, such as letters.
Infix	A small unit of meaning (affix) added to the middle of a root; for example, in Tagalog (an Austronesian language), -um- is added after the first consonant of the root to form the past tense: 'bili' 'buy'; 'bumili' 'bought'.[6]

[3] Great Schools Partnership (2014)
[4] Pretorius and Lephalala (2011)
[5] SIL International (2019)
[6] SIL International (2019)

Inflectional language (also called fusional language)	A language in which morphemes (meaning units) added to word roots simultaneously denote multiple grammatical features. For example, French is inflectional where the verb form changes to indicate both person and tense. The verb 'parler' 'to speak', conjugated depending on first, second or third person, and singular, plural as well as based on tense. 'Parle' then indicates first person singular and present tense, while 'parlons' indicates first person plural, and present tense.
Linear regression model	This method involves fitting a straight line to model the relationship. For example, between fluency and comprehension and estimating the fluency level at which the line reaches 80 per cent comprehension.
Linguistic features of a language	Factors related to the linguistic structure of a language such as syntax; grammar (how words are ordered to make sentences), phonology (the sound system), morphology (how words are formed), and semantics (how words are used).
Logistic regression model	Estimates probability. For example of reaching 80 per cent comprehension at each level of Oral Reading Fluency.
Mean method	The fluency benchmark is calculated as the mean Oral Reading Fluency of all learners (children) who have 80 per cent comprehension or higher.
Median method	The benchmark is calculated as the median Oral Reading Fluency of all children with at least 80 per cent comprehension.
Morpheme	The smallest unit of meaning in a language.[7]
Morphological structure	The internal structure of words.
Morphology	The study of the internal structure of words of a language. Depending on the language, words contain a root (the part of the word that cannot be further broken down into smaller parts), and a series of affixes, each of which adds to, or change, the meaning of the root.[8]
Morpho-syntactic	The linguistic units of a language with morphological and syntactical features.
Non-alphabetic orthography	In non-alphabetic orthographies, each symbol represents the syllable or a one-syllable unit of meaning (that is, larger than a phoneme).[9]

[7] SIL International (2019)
[8] Bauer (1983)
[9] Conrad (2016)

Norm-referenced test	Norm-referenced testing refers to the use of standardized tests designed to rank test-takers and compare them against one another. The results of norm-referenced tests demonstrate the relative performance of test-takers against the performance of a "hypothetical average student".[10] This is determined by comparing individual scores with the scores of other test-takers. Scores are usually reported as a percentage or percentile ranking.
Orthographic features of a language	Factors related to how the language is conventionally written down including the types of characters used (for example, letters or glyphs), rules for capitalisation, rules for word formation, and spelling rules.
Orthography	The rules for how a language should be written down, including conventions for spelling, capitalisation, word breaks, and punctuation.
Phoneme	A phoneme is the smallest element of speech that differentiates one word from another in a language; for example, 'm' in 'am' that differentiates this word from 'an', 'at' or 'as'.
Phonological awareness	The awareness of and ability to manipulate the sound units in one's language.[11][12]
Predictive validity analysis	The validity of a measure (for example, a fluency benchmark) is assessed by looking at how well it predicts another measure (for example, reading ability in later grades).
Prefix	A small unit of meaning (affix) added to the beginning of a word; for example, 'un-' in 'unhappy', 'u-' in 'umama' (isiZulu: 'mother').
Prosody	The patterns of stress and intonation in a language.
Quadgraphs	The convention of spelling words. Quadgraphs are four-letter combinations that represent a single sound; for example, 'ough' in 'bough', or 'ntsh' in 'ntshe' (isiZulu: 'ostrich').
Quingraphs	The convention of spelling words. Quingraphs are five-letter combinations that represent a single sound; for example, 'ntshw' in 'intshwela' (isiXhosa: 'spider').
Reading norm	A reading norm is a norm-referenced: "… judgement of what a child should be able to do,"[13] or, a norm-referenced benchmark for reading (see norm-referenced testing).

[10] Great Schools Partnership (2014)
[11] Smith et al (1998)
[12] Anthony et al (2003)
[13] Zieky & Perie, 2006, cited in RTI International (2015)

Semantics	The study of how words are used to convey meaning in a language, including vocabulary.
Suffix	A small unit of meaning (affix) added to the end of a word; for example, '-s' (plural) in 'cats', '-ng' (locative) in 'nokeng' (Setswana: 'to the river').
Syntax	The study of how languages form sentences, including word order.
Trigraphs	The convention of spelling words. Trigraphs are three-letter combinations that represent a single sound; for example, 'igh' in high, 'tch' in watch, or 'ndl' in indlu (isiZulu; 'house').
Word boundary	The written form of words. Word boundaries are the white spaces in between words, which indicate where one word ends and another starts.
Word root	The part of a word that carries the primary meaning and cannot be further broken down into smaller meaning units. For example, in the word 'impossibility', 'possible' is the root.

ACRONYM LIST

CAPS	Curriculum and Assessment Policy Statement
COP	Community of Practice
CV	Consonant-Vowel
CWPM	Correct Words Per Minute
DBE	Department of Basic Education
DFID/ESRC	United Kingdom Department for International Development / Economic and Social Research Council
EFAL	English First Additional Language
EGRA	Early Grade Reading Assessment
EGRS	Early Grade Reading Study
FAL	First Additional Language
HL	Home Language
IRT	Item-Response theory
LOE	Level of Effort
LoLT	Language of Learning and Teaching
LRL	Learner Regeneration Literacy Programme
NECT	National Education Collaboration Trust
NORC	National Opinion Research Centre, University of Chicago
ORF	Oral Reading Fluency
PanSALB	Pan South African Language Board
PIRLS	Progress in International Reading and Literacy Study
ReSEP	Research on Socio-economic Policy at the University of Stellenbosch
RC	Reading comprehension
RTI	Research Triangle Institute
SALDRU	Southern Africa Labour and Development Research Unit
SPS	Story Powered Schools
UCT	University of Cape Town
UK	United Kingdom
U.S.	United States of America
TARMII – FP	Teacher Assessment Resources for Monitoring and Improving Instruction for Foundation Phase
USAID	United States Agency for International Development
V	Vowel
ZAR	South African Rand
ZenLit	Zenex Foundation Literacy Project

FOREWORD

The National Development Plan: The Vision 2030 chapter on education emphasizes the importance of language proficiency, envisioning that: "… 90% of learners in Grades 3, 6 and 9 must achieve 50% or more in the Annual National Assessment in literacy, numeracy/ mathematics and science." Similarly, the Department of Basic Education (DBE)'s sector plan, the Action Plan to 2019: Towards the Realisation of Schooling 2030 makes its number one goal to: "Increase the number of learners in Grade 3 who, by the end of the year, have mastered the minimum language and numeracy competencies for Grade 3."

However, in the absence of established reading benchmarks, it is difficult for the education system to identify children at risk for reading failure, especially in the early years where reading skills are taught. It is also difficult to set clearly articulated, realistic milestones for teachers to monitor appropriate reading achievement at each grade.

Rigorous international research has provided robust evidence for profiling what successful reading in English as a home language entails. This evidence has been helpful, even in our context, and formed a base from which to work. The same, however, cannot be said for the remaining official languages in South Africa. The differences in structure between the language groups within African languages, Afrikaans and English, make it necessary to develop reading norms and benchmarks that are specific to each one, or at least to each language group. Such work has not been undertaken systematically in South Africa and is equally rare elsewhere on the African continent.

This report is an important contribution to addressing this gap, forming part of the sector's recent efforts to provide guidance based on research and expert advice on how best to support reading in the early grades, especially for African languages. The report provides a synthesis of progress in local benchmarking efforts, skills proposed for benchmarks as well as data-driven methods that may be used.

By consolidating what is known about approaches to developing benchmarks, identifying updated methods and acknowledging the broad collaboration required among stakeholders, this report serves as a strategic reference document for benchmarking over the next few years.

Finally, the commitment of the donor community to this important work continues to be appreciated and critical. In particular, I want to express appreciation for the partnership with the United States Agency for International Development with whose support this report was undertaken.

MR HM MWELI
DIRECTOR-GENERAL
DATE: 05/10/2020

EXECUTIVE SUMMARY

In 2019, the Department of Basic Education (DBE) and the United States Agency for International Development (USAID), South African academics and reading practitioners and international benchmarking experts began exploring early grade reading benchmarking. Benchmarking is an important educational tool that allows:

- Teachers, schools and parents to monitor that the child is **mastering components of reading** and report on them according to age, grade and milestones. For example, the South African curriculum requires that all Grade 1 learners know their letter sounds.
- The school, district, province and DBE to use **progress against benchmarks** to measure the quality of educational delivery and use this data to provide input into policy reform, to design training and to influence pre-service teacher training.

South Africa has 11 official languages that vary in terms of structure and complexity. While there are benchmarking elements in all languages embedded in the curriculum, there is no comprehensive or systematic benchmarking process. This report expands on the complexity of these languages and many of the issues that arise when trying to design benchmarks.

This document summarises more than **100 policy documents, research reports, and journal articles**.

The report lays out the choices the DBE faces in designing credible benchmarks. These include:

- Identifying a **range of competencies** that can be benchmarked, for example Oral Reading Fluency (ORF), sound-letter recognition, reading comprehension, complex consonant sequences and reading accuracy.
- Gathering **sufficient data** from learners to define the benchmarks or milestones. This report lays out the requisite **sample size** of learners reading in each language and options for data collection. In particular, some historic data can be reanalysed to elicit and inform benchmarks. Likewise, the data can be supplemented or alternatively, for some languages, new data collected.
- Deciding which **methods and approaches** will be used. There are trade-offs associated with various techniques for measuring reading (oral reading fluency, comprehension, and so on). The report also includes examples of the **statistical analyses** required and practice guidelines.
- Ensuring that **process elements** are included in the benchmarking process including:
    - Aligning with the National Curriculum Statement (NCS) Grade R-12 and the National Curriculum and Assessment Policy Statements (CAPS) for the Foundation Phase (or filling gaps in CAPS)
    - Ensuring stakeholder involvement and buy-in
    - Communicating the benchmarks in a way that makes it useful for parents, teachers, schools and policy makers.

Three recommended strategies to develop benchmarks in South Africa based on current reading and benchmarking initiatives in South Africa are provided. The three strategies are:

- Strategy 1: Analyse Existing Datasets (most inexpensive, but much of the data was not collected for the purpose of benchmarking)
- Strategy 2: Collect Prioritized Additional (Top Up) Data (enhances and fills in the gaps for Strategy 1)
- Strategy 3: Collect Primary Data (Most expensive and time consuming but fit for purpose)

Recommendations include:

- Prioritise setting benchmarks against other competencies and skills.
- Consider the extent of completeness and gaps in the relevant data for each language and across the skill sets when benchmarking.
- Rank and or select benchmarking methods based on existing knowledge and data, costs and expertise.
- Focus benchmarking on skills that improve and strengthen reading and literacy.

This document is one way to increase access to benchmarking concepts and existing progress in South Africa. It is also a further step to build a reading benchmarking community of practice made up of policy makers, academics and practitioners. Next steps include building benchmarks by language group and ultimately by language, continuing and funding data collection and analysis towards benchmarking. Ultimately, benchmarking must be used as a tool to measure education quality (through systemic evaluation processes) and translated into teaching practice to improve early grade reading.

In the ever evolving and developing field of comprehension and reading fluency, and benchmarking, the recommendations contained in this report should be contemplated and considered with any new theory in the field.

INTRODUCTION

The purpose of this document is to provide the South African Department of Basic Education (DBE), funders, and partners with information relevant to establishing reading benchmarks in South African languages. Benchmarks refer to standards of proficiency in an educational skill or domain. According to Research Triangle Institute (RTI) International's Early Grade Reading Assessment (EGRA) Toolkit (2015) (2nd Edition): "Benchmarks are particularly useful for reading, as they establish expectations and norms for reading performance." They can be used in different ways and for different purposes but, essentially, they are important for measuring reading progress. Furthermore, benchmarks in the 11 official languages of South Africa can contribute to child mastery of the components of reading and measure progress against targets.

The report objective is to present an illustrative framework and practice guidelines for the development of data-driven reading benchmarks in South Africa over the next five years, including cost, methodology, and logistical considerations.

This document summarises over 100 policy documents, research reports, and journal articles to provide a reference document to understand the context of reading norms in South Africa, the science behind reading norms and benchmarks, to identify various approaches and methods to establishing benchmarks, and provide guidelines for setting reading norms and benchmarks in South Africa.

In addition, the report summarises recent South African and international research and expert stakeholder consultation on benchmarking to agree that there is need for specificity in measuring benchmarks for the following reading competencies:

- Letter-sound knowledge
- Letter combinations referred to as complex consonant sequences[14] (or digraph, trigraph and quadgraph)
- Oral Reading Frequency (ORF)
- Reading Comprehension
- Accuracy in Reading

Methods of assessment, data collection and analysis for establishing benchmarks are discussed to inform the processes and methods to set benchmarks for the above competencies. This document provides a case for shared benchmarks in similar languages (for example. all Nguni languages), set benchmarks by grades and the importance of setting targets.

The report also sets out to draw on international and local best practice to outline three viable approaches for setting benchmarks:

- Strategy 1: Analyse Existing Datasets (most inexpensive, but much of the data was not collected for the purpose of benchmarking)
- Strategy 2: Collect Prioritized Additional (Top Up) Data (enhances and fills gaps for Strategy 1)
- Strategy 3: Collect Primary Data (most expensive and time consuming but fit for purpose)

For each strategy, we recommend using a combination of two broad methodological approaches to setting benchmarks. First, most competencies can be benchmarked by convening a group of experts to interpret performance statements (for example, from CAPS) and translate them into quantitative targets in literacy tests. A systematic, data driven approach is proposed for expert benchmark setting. Second, some competencies can be benchmarked against criterion performance in another competencies. For example,

[14] There is a lack of consensus amongst linguists as to whether these are quadgraphs or consonant blends. The term complex consonant sequences is used in this document.

we recommend benchmarking ORF at a level allowing for basic comprehension by learners. We provide a number of different analytical approaches to producing such benchmarks based on the data available.

The report has seven parts:

- **Part One**: Presents an outline of official languages, the context of reading assessment and reading in South Africa, including some linguistic characteristics of the Southern African Bantu languages
- **Part Two**: Defines the science behind reading norms and benchmarks, the difference between these concepts, and methods for estimating benchmarks. The purpose is to guide the identification of the purpose and use of reading norms and benchmarks for different competencies and stakeholders
- **Part Three**: Identifies assessment tools, recommended sampling approaches and provides technical guidance for the analysis of data to set benchmarks
- **Part Four**: Outlines current initiatives, available datasets and existing assessment instruments in South Africa and discusses, based on current initiatives, the implications for future work in this area
- **Part Five**: Based on emerging practice, a practice guide detailing seven phases of setting benchmarks is presented
- **Part Six**: Practice guides for three viable strategies for benchmarking reading in South Africa over the next five years, drawing on methods outlined in Part Three, are detailed
- **Part Seven:** Presents important elements to support the setting of benchmarks, including highlighting insights and requirements laid out in the curriculum that affect benchmarking, the level and type of expertise needed to conduct benchmarking research, and the type of stakeholder engagement required

This framework is intended for use by all interested funders, researchers and other partners as a starting point from which to engage with data-driven benchmarks, under the leadership of the DBE, across all South African languages. Various stakeholders (see Annexure 2) were involved to contribute to the quality, credibility, and transparency of the strategic framework and practical guidelines put forward.

In the ever evolving and developing field of comprehension and reading fluency, and benchmarking, the recommendations contained in this report should be contemplated and considered with any new theory in the field.

NAVIGATION ICONS

Within each of the sections, for ease of navigation, the following icons are used:

Research: This includes background information, detail on the Southern African Bantu languages, and research. Some sections provide technical details which provide deeper insights of key elements of language construction, reading competencies and benchmarking practice.

Benchmarks: Information on benchmarks including the background to benchmarking, details on how to approach benchmarking, various methods to set benchmarks, and key considerations are identified.

Approaches and methods: The approaches, methods and steps to set benchmarks and key considerations in benchmarking are identified. This will be of interest to readers engaged in setting benchmarks.

Statistics, sample and analytics: Technical details on sampling, analysis, statistical approaches, and analytics are highlighted.

Policy and implications: Implications of the research, assessment or benchmarking practice are identified. The implications may pertain to policy, curriculum, finance or management decisions. These will be of interest and importance to government departments, the donor community, stakeholders and implementation organisations working in the sector.

The reading competencies to be benchmarked are represented by the following icons:

	Reading Comprehension
	Oral Reading Fluency (ORF)
	Letter-sound Knowledge
	Complex consonant sequences or Digraph/Trigraph/Quadgraph
	Reading Accuracy

Other icons used to help navigate and identify elements of setting benchmarks include:

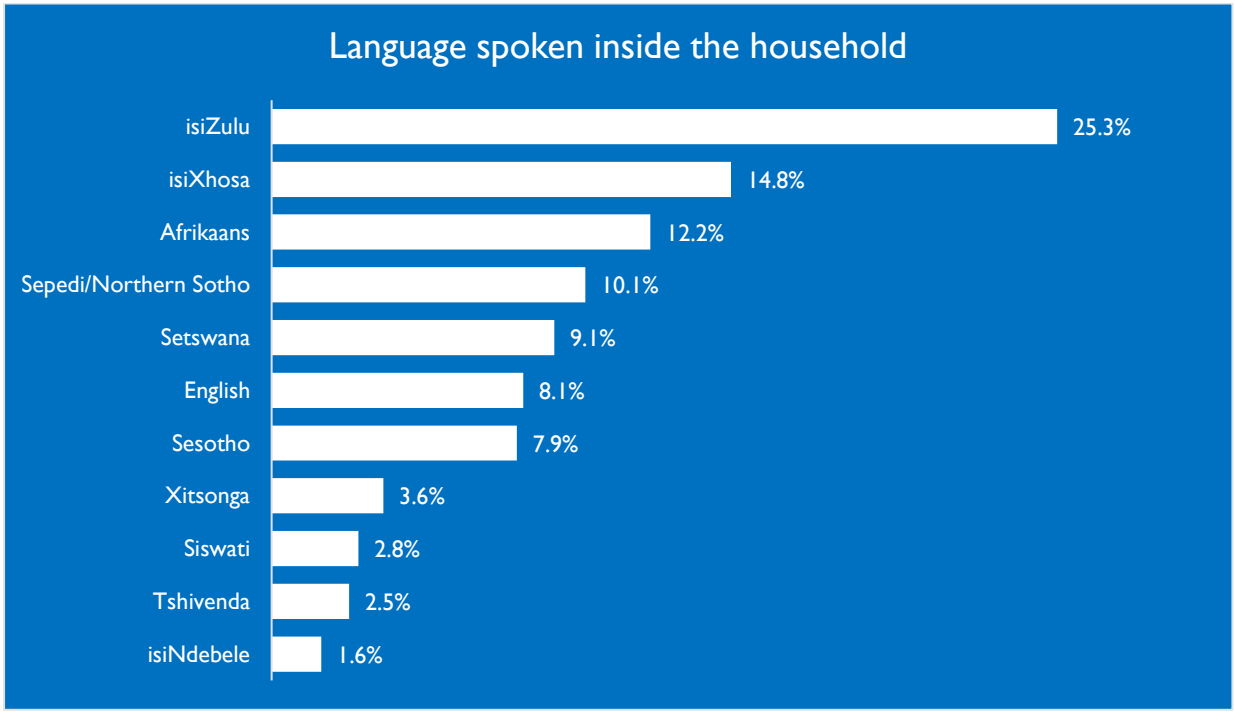
	Expert based methods
	Targets
	Goals
	Metric

PART ONE: THE CONTEXT OF READING IN SOUTH AFRICA

1.1 SOUTH AFRICA'S OFFICIAL LANGUAGES

There are 11 official languages in South Africa. English and Afrikaans are European languages belonging to the Germanic language family, and the nine African languages belong to the Southern Bantu language group. The distribution of the languages spoken by household members, by population group, are shown in Figure 1 below.[15]

Figure 1: South African Home Languages per Population Group (2018)



IsiZulu, isiXhosa and Afrikaans have the highest number of home-language users, while Siswati, Tshivenda and isiNdebele are the smallest language groups in South Africa. It should be noted that isiNdebele is also spoken in Zimbabwe (and Tshivenda to a lesser extent). Setswana is the official language of Botswana, Sesotho is the official language of Lesotho, Siswati is the official language of Eswatini, and Xitsonga (also referred to as Shangaan) is also spoken in Mozambique. Reading research may have been undertaken in these languages in neighbouring countries, and storybooks, graded readers, and other print-based resources developed in these languages. However, there may be differences in dialect.

Learning how to read depends on learning how elements of one's spoken language are represented in written language. Thus, reading development is influenced by factors related to the linguistic structure of a language (such as sound and word formation systems), as well as factors related to the way the language is written (orthographic features). It is essential to establish benchmarks that are reliable and valid per the linguistic and orthographic features of each language. These factors are addressed in Sections 7.2 and 7.3, respectively.

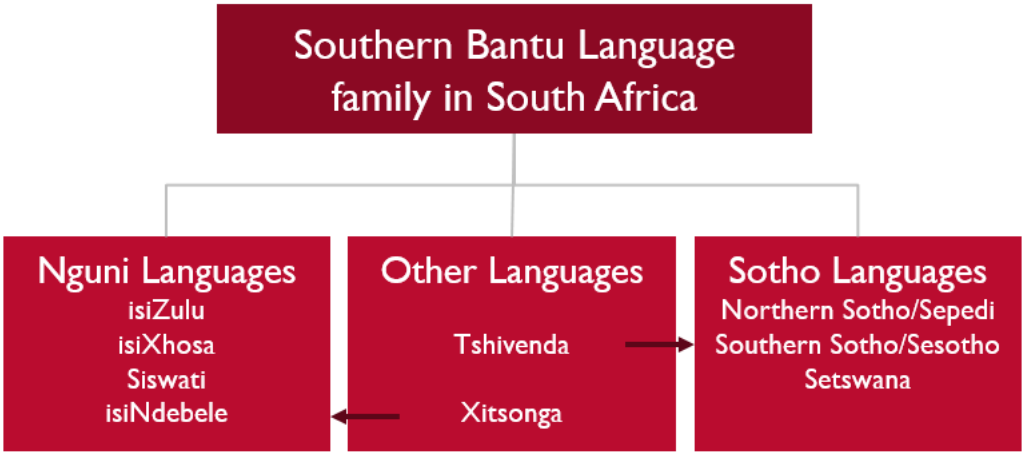
1.2 LINGUISTIC FEATURES OF SOUTH AFRICA'S AFRICAN LANGUAGES

The nine official African languages in South Africa all belong to the Southern Bantu language family. These, in turn, form two main subgroups, the Nguni (isiZulu, isiXhosa, isiNdebele and Siswati), and the Sotho language families (Setswana, Sesotho and Sepedi/Northern Sotho). Xitsonga and

[15] Statistics South Africa, General Household Survey (2018)

Tshivenda are less closely related to either the Nguni or Sotho language groups, but share features common with languages in Mozambique and Zimbabwe respectively. Xitsonga belongs to a subgroup of the Nguni language family (TswaRonga) while Venda is an isolate in the larger Sotho subgroup. Figure 2 below shows the African language groups in South Africa.

Figure 2: Southern Bantu Language Family in South Africa



Source: Adapted from Spaull et al (2020)

All Southern African Bantu languages are agglutinating languages meaning that words have complex morphological structures, comprising roots to which several prefixes, infixes, and suffixes are added to convey semantic and syntactic information. In contrast, English and Afrikaans do not have complex morphology and are classified as mildly inflectional or analytic languages.

The typical word boundaries (the spaces between words) that occur in written forms of analytic or mildly inflectional languages such as English and Afrikaans (nouns, pronouns, verbs, adjectives, adverbs, and so on) do not always occur in African languages. For example, English and Afrikaans have prepositions to indicate place ('in the sea' / 'in die see'). In contrast, African languages use a locative construction whereby a locative morpheme is added to the noun stem, either as a prefix (in IsiXhosa: the 'sea' is 'ulwandle' → 'elwandle' is 'in the sea') or a suffix (in Sesotho: 'leoatle' → 'leoatleng').

The phonology (sound system) of the Southern African Bantu languages differs from English and Afrikaans. English and Afrikaans have many vowels (20 and 15, respectively[16]). These vowels include both short and long vowels, as well as diphthongs. There are slightly more consonants than vowels in English (24) and Afrikaans (18), but all are plain consonants. On the other hand, the Southern African Bantu languages in South Africa have relatively small and symmetrical vowel inventories ranging from five to nine. However, they have complex consonant systems (ranging from 30 to almost 60 consonants, depending on the language, with consonants with various manners of articulation, including plain, click, implosive, and ejective). Note that not all the African languages in South Africa make use of all four airstream mechanisms (how airflow is created in the vocal tract).[17] Additionally, tone (pitch variation on the nucleus of a syllable) is used phonemically (that is, it changes the word meaning) in the Southern African Bantu languages. Tone is not used phonemically in English or Afrikaans.

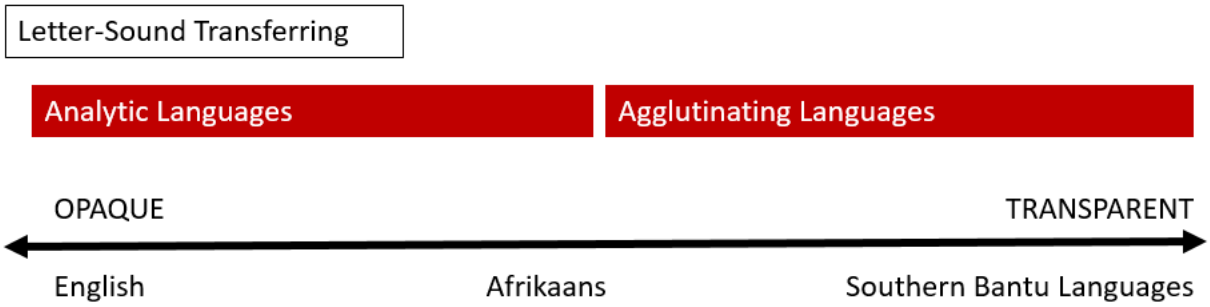
1.3 ORTHOGRAPHIC FEATURES OF WRITTEN LANGUAGES

This section briefly discusses how the orthographies of the official languages are similar to or different from, one another, and what implications this might have for early reading development.

[16] Wissing (2018)
[17] Ogden (2011)

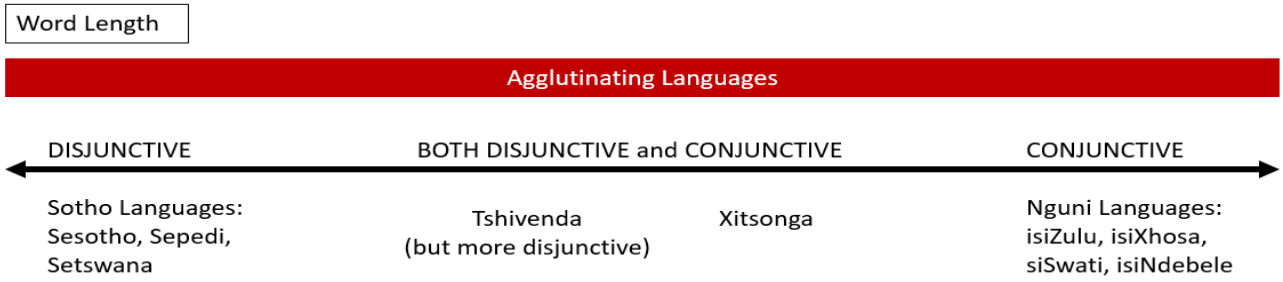
Letter-sound transparency: The orthography or writing system of a language occurs on a continuum of transparency-opacity, depending on the regularity whereby sounds are represented by letters of the alphabet (see Figure 3). English is regarded as having an **opaque** orthography since different letters can represent the same sound, for example, the letters 'f', 'ph' or 'gh' all represent the sound /f/ as in 'fog', 'phone', and 'cough'. In contrast, Afrikaans and the Southern Bantu Languages are regarded as having **transparent** orthographies, where the letter-sound relationship is regular. For example, in isiXhosa, the sound /f/ is always represented by the letter 'f', as in 'ufudo' and 'funda'. Research across diverse European languages shows that children learn to read relatively more quickly in languages with transparent orthographies such as Italian, Greek, and Finnish compared to those with opaque orthographies such as English and French.[18]

Figure 3: Continuum of Transparency-Opaque Letter-sound Transferring



Word length: For various historical and linguistic reasons, the Nguni languages have a *conjunctive* orthography (words with many morphemes are written as one word). The Sotho languages have a *disjunctive* orthography, where verbal elements with suffixes are written together with the word stem as a single word, while the prefixes are written as stand-alone morphemes. It is best to consider the concepts of conjunctivism and disjunctivism, which apply to agglutinating languages, on a continuum (Figure 4). The Sotho languages (such as Setswana) can be considered to be on the more disjunctive side. Tshivenda and Xitsonga fit somewhere in the middle of the continuum as they have features of both conjunctive and disjunctive writing. However, Tshivenda has more of a disjunctive orthography than Xitsonga. The Nguni languages (such as isiXhosa) occur more on the conjunctive side of the continuum. This has implications for word length in written texts: Nguni texts comprise many long words resulting in 'dense' text, whereas Sotho texts contain long words interspersed with many short single morpheme words comprising one or two syllables.

Figure 4: Continuum of Word Length for Disjunctive and Conjunctive Languages



[18] Ziegler and Goswami (2005)

Table 1 illustrates these orthographic differences in the first paragraph taken from the same story in four of the languages.[19] These are compared with English and Afrikaans – both considered analytic languages where the concepts of conjunctivism and disjunctivism do not apply (see glossary of terms).

Table 1: Words per Sentence in Disjunctive or Conjunctive Orthographies

Type		Language	Text (first paragraph of the same story across languages)
Agglutinating Language	Sotho	Sesotho	Ka tsatsi le leng motsamai a lapileng o ne a tle motseng. A kopa dijo feela ho ne ho se na tseo ba ka mo arolelang tsona.
		Tshivenda	Li̓nwe ḓuvha muendi we a vha e na ndala o swika mu̓ḓini. A humbela zwi̓liwa. Ho vha hu si na we a vha e na zwi̓liwa.
		Xitsonga	Siku rin'wana mufambi loyi a ri na ndlala. U fikile emugangeni. A kombela swakudya, kambe a ku nga ri na loyi.
	Nguni	isiXhosa	Kwakukho umhambi owayelambe kunene. Wahamba engena ecela amalizo. Kwakungekho kutya, kwanto tu kwaphela emizini.
Analytic Language		English	One day there was a stranger who was very hungry. He came to a village and asked for food. Nobody had any food.
		Afrikaans	Op 'n dag was daar 'n vreemdeling wat baie honger was. Hy het na 'n dorpie gekom en vir kos gevra. Niemand het kos gehad nie.

Type		Language	Words in Sentence 1	Words in Sentence 2	Words in Sentence 3	Total words	Average Words per sentence	Average Letters per word	Total single syllable words: Vowel/ Consonant-Vowel
Agglutinating Language	Sotho	Sesotho	12	15		27	13.5	3.4	16
		Tshivenda	12	3	11	26	7.5	3.8	17
		Xitsonga	8	3	10	21	7	4	9
	Nguni	isiXhosa	4	4	6	14	4.6	6.7	1
Analytic Language		English	10	9	4	23	7.7	3.9	17
		Afrikaans	11	10	5	26	8.7	3.6	18

Source: Adapted from Spaull et al (2020)

[19] Taken from the story *Stone Soup* in the Vula Bula series of graded readers produced by The Molteno Institute. These versioned stories are intended for Grades 1/2 level. They are open source texts and can be downloaded at: https://vulabula.molteno.co.za/readers_by_type/

As shown in Table 1, the paragraph has a similar number of words in the Sesotho, Venda, Afrikaans and English texts, with many single-syllable word units occurring in the sentences. In contrast, isiXhosa has the fewest and longest words, resulting in dense text, and few single-syllable words. Xitsonga is in between, with words in sentences displaying features of both conjunctive and disjunctive orthography.

Table 1 illustrates the average word length (letters per word) based on only three sentences. Corpora can be used to give a more accurate idea of average word length in different languages. Data from a corpus such as the Brown Corpus reveals that the average length of an English word is about 4.75 letters per word.[20] Data from the official word list of the Taalkommissiekorpus shows that the average length of an Afrikaans word is 10.58 characters (based on the first 100 000 words, which includes compound words; the longest of which is 34 characters).

Specialized children's literature corpora include literature aimed at children and are expected to reflect the types of words children would read or have read to them. The corpora reveal that the average word length is 8.9 letters for isiXhosa[21] and nine letters for isiZulu.[22] In comparison, the average word length for Sesotho is 7.28 letters.[23]

The texts included in a corpus can affect the word length data retrieved. Prinsloo and De Schryver[24] report somewhat different average word lengths for each language compared to the specialized children's literature corpora. Prinsloo and De Schryver examined: "… various corpora available in the Department of African languages at the University of Pretoria" before and up to 2002.[25] They report the average word lengths for various languages. The average word length for the Nguni languages ranged from 5.88 (isiXhosa) to 7.18 (isiZulu) letters per word.[26] The Sotho languages had very similar average word lengths at 3.88 for Sesotho and Sepedi and 3.89 for Setswana. Tshivenda and Xitsonga were closely aligned to the Sotho languages with words being on average 4.07 and 4.29 letters long, respectively. Collectively, the corpus data highlights that there are differences in the length of words across languages. Word length will affect reading speed measured in words per minute.

Complex Consonant Letter-sounds:

While English and Afrikaans have a complex vowel system (as discussed previously), but a set of fairly simple consonant sounds, the African languages have a simple vowel system, with five to seven vowels, but complex consonant sounds represented by single letters, digraphs (2-letter combinations that represent a single sound), trigraphs (3-letter combinations that represent a single sound), and also some complex sequences of four to five consonants ('ntsh') and ('ntshw') (collectively referred to as complex consonant sequences in this report). English only has seven consonant digraphs; for example, 'th' represents the sound '/ð/' as in 'the', 'then'; 'sh' represents '/ʃ/' as in 'ship', and one trigraph, for example, 'tch' '/ʧ/' in 'watch'. Afrikaans has 11 consonant digraphs (25 digraphs in total) and five trigraphs. In contrast, the African languages typically have several digraphs, about six to eight trigraphs, as well as many consonant sequences.

Although digraphs and trigraphs are visually more complicated to recognize than single letters for young readers and would therefore typically be introduced later in phonics programmes, some of them occur commonly (for example, 'ng', 'kh', 'th', 'nd', 'kw', 'ny', 'hl' are among the 18 most common consonants in isiXhosa).[27] So it would be difficult to delay their instruction. For example, in Table 1 above, there are

[20] Baker (nd)
[21] Randera and Rees (2019)
[22] Randera and Rees (2019)
[23] Randera and Rees (2019)
[24] Prinsloo and de Schryver (2002)
[25] Prinsloo and de Schryver (2002, p 256)
[26] Prinsloo and de Schryver (2002)
[27] Personal communication from Siân Rees at the Molteno Institute for Language and Literacy who is currently researching frequencies of graphemes (letters) in the different African languages.

11 (Sesotho), ten (Tshivenda), nine (Xitsonga) and 13 (isiXhosa) digraphs or consonant sequences that occur in the short paragraph in a graded reader intended for Grade 1 and Grade 2 readers. There is very little research on how these complex letter-sound configurations affect early reading development. The only available data pertaining to Grade 2 isiXhosa learners shows that knowledge of digraphs or trigraphs lags significantly behind that of single consonants, as shown in Table 2.[28] If children have difficulty recognizing digraphs, then they will struggle to read texts at a fairly basic level.

Table 2: Recognition of Single Consonants Versus Digraphs/Trigraphs of Grade 2 isiXhosa Readers

Grade	Configuration	Mean letters correct per minute	Percentage of learners obtaining zero
Grade 2	Single letters	28	12%
	Digraphs/trigraphs	10	52%

Diacritics:

Afrikaans, Sepedi, and Tshivenda are the only languages that make use of diacritics. Afrikaans has diacritics on different vowels such as 'ê', and 'oë' to distinguish them from 'e' and 'oe', while Sepedi distinguishes between 's' and 'š'. Tshivenda uses the most diacritics — a dot above 'n', and a turned 'v' or wedge (ˬ) below 'd', 't', 'l' in upper and lower case to signal pronunciation differences in these consonants and their plain counterparts (Figure 5).

Figure 5: Diacritics in Tshivenda

Ṅ Ḍ Ḷ Ṉ Ṱ ṅ ḍ ḷ ṉ ṱ

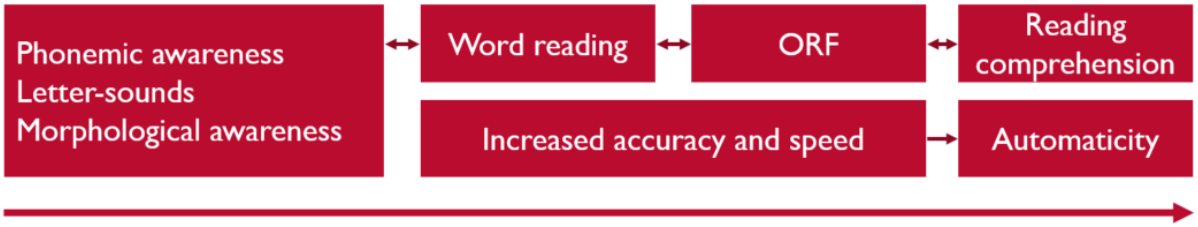
Whether the use of diacritics affects the learning of letter-sound relations and word reading among early Tshivenda readers awaits research.[29] In Afrikaans and Tshivenda, letters that have diacritics act as additional symbols to learn in the language (that is, additional phonics) but they are transparent. Children will need to learn the sound-letter correspondence for 'oe' as well as 'oë', for example. It is, therefore, fair to argue that diacritics might slow down reading acquisition because children have more phonics to learn. Additionally, finer-grained visual perception is required to notice the diacritics, so this may also take children longer to master initially.

1.4 IMPLICATIONS OF ORTHOGRAPHIES FOR EARLY READING

In languages with alphabetic writing systems, performance on tasks such as phoneme awareness, letter-naming, letter-sound knowledge, and word-reading are strongly correlated. Oral Reading Fluency (ORF) and reading comprehension are also robustly correlated[30]. In some studies, both phoneme awareness and letter-sound knowledge have been found to be strong, independent predictors of word reading.[31] Early reading research in African languages shows that morphological knowledge (knowledge of the root word) is also associated with word-reading in the agglutinating languages.[32] Word reading correlates with ORF which in turn predicts reading comprehension. Figure 6 depicts the relationship between these subcomponents.

[28] Ardington (2019)
[29] There is currently no direct research on the use of diacritics in Tshivenda and Afrikaans
[30] Fuchs et al (2001)
[31] Clayton et al (2019)
[32] Rees (2016)

Figure 6: Subcomponents of Early Reading Development



While the continuum does not imply a strictly linear sequence of development and there can be reciprocal influences between subcomponents, it is a truism that children cannot understand a text on their own unless they can read words. Reading words presupposes letter-sound knowledge that is rapidly executed. Although benchmarks can, in principle, be established for any of these subcomponents, decisions about which subcomponents to benchmark should be offset against criteria such as feasibility, reliability, validity, and utility.

Although phonemic awareness is a predictor of early word-reading ability, phonemic awareness assessments are tricky to design for African languages because of their vowels (V) or consonant-vowel (CV) syllable structure, with many words starting and ending with a vowel. Separating a word into the sounds that make up the word (phoneme awareness) is more difficult when the syllables are complex (comprise multiple consonant and V units).[33] For example, 'umama' (isiXhosa: mother) consists of three syllables: u.ma.ma. The first syllable contains only a vowel 'u', that is, it has a V structure. The second two syllables consist of a consonant and vowel and that is a CV structure. Phonemic awareness also takes a substantial amount of time to assess reliably due to the need for detailed instructions and examples.

Furthermore, distinguishing between syllables and phonemes during phonological awareness assessments can sometimes be problematic for teachers and fieldworkers who may be administering them. Due to the prevalence of the syllable unit, and perhaps due to the methods of teaching reading, it can be difficult for speakers to analyse syllables into phonemes. The most significant setback, however, is simply the time needed to administer reliable phonological awareness assessments. Setting benchmarks for phonemic awareness may thus face feasibility, reliability, and validity challenges.

Awareness of morphological (and orthographic) patterns in words is an important component of word reading and ORF. Fluency depends on children using morphological (and orthographic) information to incorporate: "… smaller and larger word chunks until full word recognition is reached."[34] However, research on morphological aspects of reading and their predictive role in word reading or ORF in agglutinating African languages is still at an early stage and benchmarking of morphological knowledge is, therefore, not feasible at this time. In comparison, assessing letter-sound knowledge and ORF is easier, less time-consuming, and more straightforward for teachers and fieldworkers. Furthermore, ORF is a stronger predictor of reading comprehension than word reading.[35] These factors make letter-sound and ORF viable candidates for reading benchmarking in all the various languages of South Africa.

[33] Goswami (2010), cited in Wilsenach (2019)
[34] Ehri (2005); Share (1995), cited in Spaull et al (2020)
[35] Fuchs et al (2001)

1.5 READING ASSESSMENT

This section provides an overview of the evolution of reading assessment internationally, in Africa, and South Africa specifically. The discussion first centres on the measurement of comprehension competencies as comprehension is the ultimate goal of reading and, therefore, has a more extended assessment history. Fluency and letter-sound recognition competencies are addressed next for their viability as measurable reading skills which can be reliably and validly benchmarked in the African languages of South Africa.

1.5.1 ASSESSING READING COMPREHENSION

The formal assessment of reading comprehension began in the 20th Century.[36] In the earliest stages, comprehension tests were included as part of intelligence batteries in the United States of America (U.S.).[37] However, it was not long before reading comprehension became a topic of study in its own right. The measurement of reading comprehension is not atheoretical or value-neutral.[38] From the outset, the assessment of comprehension has been influenced by theories of measurement, theories of reading, and contextual factors.[39] These three factors are explored in more detail below.

Two broad theories of measurement have influenced the assessment of reading comprehension, classical test theory and scaling theory.[40] Classical test theory assumes that differences in performance across participants on the same assessment are due to variations in participants' ability, or 'true' score.[41] Any other factors (such as familiarity with the topic, in the case of reading assessment) which may affect the 'true' score are assumed non-systematic due to randomisation. The non-systematic variation is noted as measurement error. When using classical test theory, test-takers answer all questions on the test and obtain a raw score, which is converted to a standardized score (age or grade norms) with a 90 and 95 per cent confidence interval. An example of a test that uses this measurement theory is the Stanford Diagnostic Reading Test, a test of English decoding, vocabulary and reading comprehension. The largest critique of tests using this measurement model is that scores on the test and the difficulty of the items can vary by sample, therefore making it difficult to compare 'true' scores across samples and time.

The scaling theory of measurement, including Item-Response Theory and Rasch measurement theory, overcomes some of the major criticisms of classical test theory.[42] Item-response theory (IRT) uses mathematical modelling to account for item difficulty, item discriminative power, and liability in terms of guessing.[43] That is to say, IRT models, such as the Rasch model, present: "… probabilistic modelling of the interaction between an individual item and an individual examinee."[44] The Progress in International Reading and Literacy Study (PIRLS),[45] which assesses reading comprehension every five years, uses this approach. In Africa, this approach has also been used to test reading comprehension and curriculum-based skills in English in the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SEACMEQ)[46] reading assessment and in French as part of the Conference of the Ministers of Education of French Speaking Countries (CONFEMEN) Programme for the Analysis of Education Systems (PASEC).[47]

---

[36] Pearson and Hamm (2005)
[37] Pearson and Hamm (2005)
[38] There are different theoretical approaches to understanding languages and comprehension. Not all linguistic experts take the same approach.
[39] Engelhard (2001)
[40] Engelhard (2001)
[41] Van der Linden and Hambleton (1998)
[42] Engelhard (2001)
[43] Van der Linden and Hambleton (1998)
[44] Van der Linden and Hambleton (1998, p 8)
[45] https://timssandpirls.bc.edu/index.html
[46] http://www.sacmeq.org/ formally abbreviated as SEAQMEC
[47] http://www.pasec.confemen.org/

Because of the more complex estimation procedures needed for IRT analysis, this approach is used predominantly for large scale assessments.

Reading theories have also influenced the development of reading comprehension assessments. There are three main approaches, the skills-emphasis approach, the constructivist approach, and the socio-cultural approach. Skills-emphasis approaches, which reached their peak in the 1980s, take the view that reading comprehension consists of various skills and strategies that readers need to master and use to achieve comprehension. Assessments based on this approach include questions on various comprehension skills and strategies, for example inferring main ideas, finding details.[48]

This skills-emphasis approach led to the use of criterion-referenced assessments. Criterion-referenced assessments stipulate a minimum benchmark that test-takers need to meet to have mastered the relevant skill. Question answering, in the form of multiple-choice assessments, proliferated using this approach and text passages were used only to assess whether a comprehension sub-skill had been mastered. Other assessment techniques included the 'cloze' task, where every n^{th} word is deleted from a passage and test-takers have to fill in the missing word based on their understanding of the passage. Passage recall is another method of assessing comprehension. After reading a passage, the reader is asked to recall what happened in the text. Of these three, question answering and the cloze procedure are faster and simpler to score.

By the 1980s, there was a move to constructivism as the paradigm used to understand reading comprehension.[49] More focus was placed on the role of the reader's prior experience, the role of text structure and story schema, as well as the reader's response to a text.[50] Consequently, reading assessments used longer, more authentic passages and prior knowledge was incorporated as part of test questions. Multiple-choice items, which had more than one correct answer, and were open-ended (also called the constructed response) questions were introduced.[51] By the late 1980s, literary theory was used to understand reading comprehension. Reader-response theories (also called socio-cultural theories) emphasize the readers' response to literature, with more open-ended questions.

> In summary, there are at least three broader theoretical understandings of reading comprehension, each with their implications for assessment. It is now known that text comprehension does not only rely on the learning and application of comprehension skills and strategies but also includes a dynamic interaction between the text and reader in a particular context.[52] Comprehension assessment should, therefore, reflect the complexity of text comprehension to the extent possible in a specific testing environment.

The development, format, and implementation of assessments are affected by the purposes of the test and the resources of the test developer. Ideally, for cost-effectiveness and efficiency, test-takers would complete individual reading comprehension assessments administered in a group setting. To capture the complexity of comprehension, again ideally, comprehension assessments would require test-takers to provide constructed responses to display their depth of reading comprehension. However, in reality, various resource constraints determine how reading comprehension is assessed. While constructed responses are often thought of as more reliable measures of reading comprehension, scoring can be subjective and time-consuming.

On the other hand, multiple choice answers can be scored automatically using a scanner. Still, test-takers may not be familiar with this method of questioning or could find the answers in the text without engaging with the content. Additionally, information resources and technology approaches reveal sample and item

[48] Pearson and Hamm (2005)
[49] Sarroub and Pearson (1998)
[50] Pearson and Hamm (2005)
[51] Sarroub and Pearson (1998)
[52] Mullis and Martin (2015); Hedgcock and Ferris (2009)

invariant scores, but samples need to be larger, thereby increasing the cost of the assessment compared to using classical test theory for measurement.

The test itself is constrained by the availability of authentic texts in the language, which is of particular concern when developing assessments for South African and other languages. At the same time, the construction of the test may be affected by the reading theory that the test developer ascribes to, which may be in opposition to the curriculum or other stakeholders' views. Finally, the purpose of the test also influences how its results will be used. A criterion-referenced test used by the government to monitor the progress of how many children have mastered a particular skill may lead to 'teaching to the test' and crowding out other curriculum activities.

1.5.2 ASSESSING ORAL READING FLUENCY

ORF, the ability to read a text accurately, with sufficient speed and appropriate expression, is a necessary skill for successful reading comprehension[53]. The assessment of ORF grew in the 1990s and early 2000s in English[54] and has gained attention in many languages since then, especially by the widespread use of the EGRA.[55]

ORF is typically assessed through a one-minute timed reading of a grade-level passage. Errors are subtracted from the total number of words read in the time limit to provide a correct words per minute (cwpm) score. Thus, ORF scores take into account both rate and accuracy. Expression is usually not assessed in large scale assessments because it is more difficult to measure objectively and reliably.

The EGRA is: "… a research-based collection of individual subtasks that measure some of the foundational skills needed for reading acquisition in alphabetic languages" such as letter-sound knowledge, ORF, and comprehension, among others.[56] Although not stated explicitly in its theoretical framework, the EGRA relies on a skills-based approach to reading comprehension. The test can be used either as a monitoring or evaluation tool. Administering the test needs to be quick (no more than 15 minutes) and easy (can be conducted by an enumerator without extensive training), and the results must be easy to interpret. These contextual factors have influenced how reading comprehension is assessed using this tool.

Reading comprehension in the original EGRA is assessed after a test taker completes the ORF task (reading a grade-level passage for one minute). A few questions (usually five) are asked after the ORF task to assess the understanding of the text. Explicit and inferential questions are asked in an open-ended format, and replies are verbal. The EGRA is scored using the classical test theory approach, and the score is either calculated as the proportion correct of the total questions, or the proportion correct of the questions asked.

The reading comprehension task of the EGRA has been criticized for different reasons. One criticism is that test-takers must give an oral response, and verbal ability can affect the possible scores.[57] Furthermore, the test is also confounded by slow reading rates over a one minute period. Children are only asked comprehension questions up to where they have read. This can lead to floor effects in the reading comprehension measure and may not test the actual reading comprehension ability of the child, had they been allowed all the time they needed to finish reading the passage. An alternative is to record ORF at one minute and at three minutes. The child therefore reads further so more questions can be asked, increasing the reliability of the reading comprehension score. Group administered written

[53] Hasbrouck and Tindal (2006)

[54] English has been given by far the most research such that norms exist for cwpm for three time points in Grades 1–8 in the United States of America. Students with a cwpm score within ten words of the 50th percentile are deemed "on track" (Hasbrouck and Tindal, 2006). These norms can be used by teachers to identify children at risk of reading failure, as well as to monitor progress over time.

[55] Dubeck and Gove (2015)

[56] Dubeck and Gove (2015, p 317)

[57] RTI International (2016)

comprehension assessments have also been used but, because of slow reading rates, and lack of writing fluency, floor effects can also occur.

An alternative method of reading comprehension assessment for the EGRA has been proposed to overcome these criticisms. This assessment, called the Sentence Choice task or test, was first piloted by Alcock and colleagues,[58] and further piloted by RTI International.[59] The task consists of ten sentence pairs, one of which is obviously true (for example, birds lay eggs), and one of which is clearly false (for example, dogs lay eggs). The sentences are jumbled in the task, so the paired sentences are separated by other items. Sentences vary from three to five words long and contain high-frequency vocabulary. The test candidate reads each sentence (of 20) and indicates whether the sentence is true or false. One point is awarded when both sentences in a sentence pair are correctly categorised as true or false, resulting in a total of ten. The test is untimed and administered in an individual or a group format.

The Sentence Choice task thus seems suitable as an alternative reading comprehension assessment task because it is independent of reading fluency, can also be created easily for new languages, and is fast and straightforward to administer. However, as an index of reading comprehension, proponents of the sociocultural reading approach may view this assessment as simplistic and reductionist.

> ORF should be considered as only one crucial part of reading achievement. For this reason, in the early stages of benchmarking or norming reading in a particular language, it is important to measure reading comprehension along with ORF. That is to say, the goal of assessing ORF should be to determine what rate of fluency is needed to enable sufficient comprehension. There is an optimal level of speed and accuracy at which comprehension can be achieved, and below which comprehension is difficult.[60] ORF, which is easier to assess than reading comprehension, acts as an index for the level of reading comprehension. Thus, as explained above, it is important to have a suitable reading comprehension assessment for the context.

1.5.3 ASSESSING LETTER-SOUND KNOWLEDGE

The assessment of letter-sound knowledge has received less attention than that of reading comprehension or ORF. Letter-sound knowledge is a foundational skill needed for decoding, word recognition, and ORF. It can be helpful to assess letter-sound knowledge in the early stages of learning to read as readers who struggle with this foundational skill are also more likely to struggle with harder skills such as word recognition and reading fluency. Letter-sound knowledge can be assessed in three ways; saying the letter sound when shown the letter, pointing to the letter when hearing the sound, or writing the letter after hearing the sound.[61] Accuracy can be determined by calculating the proportion of correct answers in a timed or untimed period. Fluency can be determined by calculating the proportion of correct answers in the amount of time given.

Commercially available letter-sound knowledge assessments include the Woodcock-Johnson Tests of Achievement and the Dynamic Indicators of Basic Early Literacy Skills. These tests are usually expensive and take a long time to administer.[62] The EGRA, used in all grades in the Foundation Phase, also includes a measure of letter-sound recognition fluency. A chart of 100 upper and lower case letters is represented randomly, and the participant has 60 seconds to say the sound of each letter. The total number of letters sounded correctly is used in further analysis. In another version of this task, participants are asked to say the name of the letter. Low scores on the letter-sound identification assessment can indicate a lack of mastery of a foundational skill of reading development, thereby highlighting that a child is at risk of reading failure. As noted in this report, it is essential to develop language appropriate letter-sound recognition

[58] Alcock et al (2000)
[59] RTI International (2016)
[60] Wang et al (2019)
[61] Dodd and Carr (2003)
[62] Piasta et al (2018)

tasks. If a language makes extensive use of complex consonant sequences and other complex graphemes, it is important to include these graphemes as part of the assessment.

> The measurement of letter-sound knowledge aligns with a skills-based approach to reading development. Proponents of the constructivist and sociocultural approaches to reading development and comprehension may take issue with the assessment of an isolated decoding skill. Nonetheless, letter-sound recognition fluency can help identify learners who are at risk of reading failure, even before they learn how to read isolated words and connected text. This is because letter-sound recognition fluency tasks assess both awareness of letter-sound correspondences, as well as the automaticity of making these correspondences. A lack of fluency or accuracy in this task can highlight that children have not started making letter-sound connections, emphasizing the need to target basic phonics skills in the classroom or remedial programme.

PART TWO: UNDERSTANDING READING BENCHMARKS

2.1 WHAT ARE READING BENCHMARKS?

Benchmarks refer to standards of proficiency in an educational skill or domain. Standards can be set relative to the performance of other children (norm-referenced) or using predetermined criteria based on what children are expected to know at their stage of education (criterion-referenced). The terms 'standards' and 'benchmarks' are used relatively interchangeably, although benchmark is used more commonly for comparing actual and expected performance.[63] A 'reading norm' is a norm-referenced benchmark for reading. It is helpful to distinguish among the following terms used in connection with benchmarking:

Goal is a long-term aspiration, maybe without a numerical value (for example, all children will be independent readers by Grade 3)

Metric is a valid, reliable unit of measurement (for example, correct words per minute (cwpm) reading connected text)

Benchmark is a numerical representation of the goal, using the metric, or a milestone on the way to achieve the goal (for example, 45 cwpm reading a passage of grade-level text)

Target is the proportion or number of learners targeted to reach the benchmark in a given time (for example, 50 per cent of learners to meet the benchmark in two years).

2.2 WHY ARE BENCHMARKS IMPORTANT IN SOUTH AFRICA?

Even though the official languages of South Africa are not distributed equally across the population (see Figure 1), it is essential to establish benchmarks that are reliable and valid per the linguistic and orthographic features of each of the languages. South Africa's National Language in Education Policy (LiEP)[64] mandates the right for all citizens to receive an education in the language of their choice (that is, one of the 11 official languages). It is, therefore, the goal of the South African Government to establish reading benchmarks in every official language.

Establishing reading benchmarks can create greater awareness of early milestones in reading development and help teachers and schools ensure that most of their learners are reaching them, thereby minimizing the chance of literacy cracks turning into gaps or chasms. The challenge in the multilingual context of South Africa lies in establishing benchmarks that are reliable and valid across the different languages used in schools.

[63] Thomas and Peng (2004)

[64] Language in Education Policy (1997)

Language proficiency (as assessed, for example, via tests of listening comprehension, vocabulary knowledge, or morpho-syntactic knowledge) is strongly associated with reading ability, in both the home language and an additional language. Children who start school with strong language skills typically find it easier to learn to read and write, while children with poor language skills are at reading risk.[65] Although these oral language proficiency skills are important for reading success, the current report addresses benchmarks of reading progress.

> The benchmarks under consideration for monitoring reading progress are not related to language proficiency more broadly construed but relate specifically to oral reading fluency and accuracy.

2.3 WHY SET READING BENCHMARKS?

There are several reasons for setting benchmarks. First, the process of setting benchmarks allows education systems to articulate their definition of reading proficiency. Second, the use of benchmarks communicates this definition of reading proficiency to others and can provide a target for teachers and learners. Third, benchmarks can be used at the school level to identify children who require additional support. Fourth, benchmarks allow a direct assessment of how many children are reading proficiently, rather than, for example, a measure of improvement in mean fluency, which may not be as easy to communicate and which may not be an indication of the general health of reading outcomes in the system. Counts (or percentages) of children are more readily interpreted, especially by non-experts. Thus, the percentage of learners reading proficiently can be included in school report cards as well as national monitoring reports. For example, Goal 1 of the DBE Action Plan to 2030: "Increase the number of learners in Grade 3 who, by the end of the year, have mastered the minimum language and numeracy competencies for Grade 3."

> As with any way of summarizing data, information is lost by converting raw fluency scores to binary indicators of whether a benchmark is achieved or not. In particular, where learner achievement is low, progress can be made in helping learners learn basic literacy skills without this improvement being evident in reported statistics related to the number of children reading at the fluency benchmark. If benchmarks are to be used for monitoring, it is vital to set them at a level that is sensitive to change. This is discussed further below.
>
> Note that the trade-off described above – that the use of benchmarks loses information to make results simpler to communicate – is only made when clear and straightforward communication of results is needed. Internal reports circulated to experts can continue to report mean reading fluency scores.

2.4 WHAT READING COMPETENCIES SHOULD BE BENCHMARKED?

When conducting benchmarking of reading competencies, it is important to determine what reading skill is to be benchmarked.

A proficient reader is one who reads with fluency (accuracy, speed/rate, and prosody)[66] and comprehension. If a benchmarking exercise were to focus only on one skill, fluency or comprehension are obvious candidates. However, comprehension can be difficult to measure. It is possible to assess a reader's ability to read connected text and fully comprehend what is read directly; however, it can be problematic to do so reliably in general[67] and with the EGRA instrument in particular.[68] Measures of comprehension vary with many factors, for example, a child's familiarity with the subject matter of the text. This makes it difficult to set reliable benchmarks for comprehension using EGRA or similar

[65] Adams (1990)

[66] Fluency assessment usually involves only accuracy and speed because prosody (referring to the patterns of stress and intonation in a language) is more difficult to assess

[67] Snow and Sweet (2003)

[68] Bartlett et al (2015)

assessments. This problem remains when comprehension is used to benchmark other skills, such as ORF. However, in such cases, additional resources and time can help to ensure a quality comprehension measure is used during the benchmarking exercise.

Instead, there are several arguments in favour of benchmarking fluency. Firstly, because fluency is a crucial skill in its own right and is part of the definition of reading proficiency. Secondly, it is relatively straightforward to measure reading fluency rates. Thirdly, fluency is a transparent measure. It is relatively easy to explain to parents and teachers how it is measured, why it is important and how fluent reading looks and sounds. Finally, there is a high level of correlation between reading fluency and comprehension. Where the relationship between fluency and comprehension is strong, fluency can be used to indicate probable levels of comprehension. Fuchs and colleagues[69] argue, based on data in English from the United States, that fluency is a reasonable indicator of overall reading proficiency. However, most of the evidence in support of this claim comes from the English language. There is a need for further work in non-English languages and African languages in particular.

How might fluency benchmarking be different in other languages? There are likely linguistic differences in at least three issues, (a) the strength of the fluency–comprehension relationship, (b) the level of fluency associated with good comprehension, and (c) the age or level of schooling at which this level of fluency is typically acquired. Each of these is addressed in turn.

First, the relationship between fluency and comprehension may be more robust in some languages than in others. In regularly spelled languages (known as having 'transparent orthographies'), consistent rules about pronouncing letters and symbols mean that text can be read fluently without comprehension, as might be the case with liturgical texts.[70] The strength of the correlation has been examined for several languages. Fuchs and colleagues[71] found that fluency in English correlated highly with an independent measure of reading comprehension (r = 0.91). The correlation was also strong for non-English speakers learning English as a second language. However, others have shown that the strength of that relationship can vary for a child's first, second or third language.[72] Research has also established a moderate to strong relationship between comprehension and fluency in a range of non-English languages, for example, Spanish (r = 0.68[73]), Turkish[74], and transparent Bantu languages such as Kiswahili and Kikuyu.[75] In Korean, Pae and Sevcik[76] found a strong correlation between reading fluency and comprehension. Even in the logographic language of Chinese, researchers found moderate to moderate-high correlations between comprehension and character-naming accuracy (r = 0.64) and character-naming speed (r = 0.55)[77]. In the Indian alpha-syllabic languages of Kannada or Telugu, however, Nakamura and De Hoop[78] did not find the same strength of relationship.

Second, there is variation across languages in the level of fluency associated with good comprehension, as discussed earlier in this report. RTI International[79] reviewed fluency benchmarks set in 35 languages across ten countries and found that most were in the range of 40–50 cwpm, except for Kenya's benchmark for English (65 cwpm). These benchmarks were set using different methods. A standardized methodology would have ensured better comparability.

[69] Fuchs et al (2001)
[70] Abadzi (2006)
[71] Fuchs et al (2001)
[72] Piper et al (2015)
[73] Jimenez et al (2014)
[74] Başaran (2013)
[75] Piper et al (2015)
[76] Pae and Sevcik (2011)
[77] Shen and Jiang (2013)
[78] Nakamura and de Hoop (2014)
[79] RTI International (2017)

Third, cross-linguistic comparisons are further complicated by the different time-scales in which fluency is acquired from one language to the next. That is, fluency is acquired more slowly in deep orthographies as compared to shallow orthographies,[80] in languages with greater consonant complexity[81] and languages with more graphemes (smallest meaningful symbols, such as letters),[82] and non-alphabetic languages.[83] Thus, there is variation among languages, both in the level of fluency benchmarks and the time taken to achieve them.

> In summary, benchmarking exercises conducted in the indigenous South African languages may find differences in the strength of the relationship between fluency and comprehension in these languages, the level of fluency associated with comprehension and the years of schooling learners need to reach this level of fluency.

The case for each type of benchmark is discussed in more detail below.

2.4.1 LETTER-SOUND BENCHMARKS

Of the lower order reading skills, letter-sound reading is a strong candidate for benchmarking. This skill is a significant competency in the process of reading development and (unlike phonological awareness) is relatively straightforward to assess and (unlike word recognition) has little overlap with higher-order reading skills.

Research on early reading development in alphabetic languages has shown that children learning to read in languages with transparent orthographies can reach ceiling levels in decoding skills faster than children learning to read in languages with opaque orthographies.[84] This also applies to reading development in agglutinating languages such as Finnish and Turkish.[85] However, research from the U.S. and the United Kingdom (UK) indicate that young learners perform well on letter-sounds from the start of schooling, even in English with its opaque orthography. In contrast, research from Africa shows that children struggle with this essential foundational reading skill. Table 3 shows data from various studies in the global North, Kenya, and South Africa, reflecting considerable variation in performance in this basic skill.

[80] Ellis et al (2004); Seymour et al (2003)
[81] Seymour et al (2003)
[82] Chang et al (2016)
[83] Chang et al (2016); Liu et al (2012); Nakamura and de Hoop (2014)
[84] Seymour et al (2003)
[85] Babayağit and Stainthorp (2007)

Table 3: Letter-sound Knowledge: Letters Correct per Minute

Letter sounds		Grade R	Grade 1	Grade 2	Grade 3
Clayton et al. 2019 (n=191)	English UK	27.9			
Good et.al 2006	English U.S.		47		
Jukes et al. 2018 (n=2,220)	Swahili Intervention Swahili Control		10.4 4.8	11.4 6.6	
Piper et al. 2015 (n= 4,385)	Swahili Intervention (PRIMR) Kenyan Intervention Home languages Swahili Control		8 4 7	17 11 13	
Taylor et al. 2018 (n = 2,600)	Setswana Coaching Intervention Setswana Control		25 (0=18%)* 22	43 (0=8%) 39	
Spaull et al. 2020 (n = 740)	Northern Sotho Xitsonga isiZulu			31 35 27	43 47 36
Menendez & Ardington 2018 (n = 8,776)	IsiXhosa baseline isiZulu baseline			19 15	31 18
Southern Africa Labour and Development Research Unit (SALDRU) / Funda Wande (n=1,180)	isiXhosa baseline		5 (0=52%)	28(0=28%)	
ZenLit. 2018	isiXhosa Control (urban) isiZulu Control (rural)		24 (0=8%) 5(0=55%)	41(0=6%) 11(0=35%)	47(0=2%) 16(0=27%)

Source: Pretorius (2019)

*Where possible, the percentage of learners obtaining zero in letter-sound knowledge is indicated in brackets

According to research from the UK, children at the end of the reception year (average age 5.2 years) can name 27 letter-sounds correctly (since preschool starts at age four in the UK, children are a year younger than Grade R children in South Africa). In the US, benchmarks for letter-naming average 47 letters per minute in Grade 1. In African languages, alphabetic code knowledge is assessed via children's ability to identify the sounds represented by letters rather than by letter naming. Acquisition of letter-sounds seems to be developing rather slowly in Africa. While performance in South Africa is marginally better than what the data reflect in Kenya, only some children start to approximate the Grade 1 U.S. mean in Grade 3 only (that is, they are further behind than their U.S. counterparts). Although letter-sound knowledge *per se* does not guarantee word reading[86] (children need to be taught how to blend letter sounds to form words), a weak letter-sound knowledge base makes it extremely difficult for children to read words.

This is a literacy 'crack' that can already be readily detected in Grade 1, so having benchmarks for letter-sounds across the languages in South Africa can help to minimize the onset of early reading problems and can also help to remediate learners who fall at this early reading hurdle. This is especially important for Grades 1-2 and reflects on the quality of learning in Grade R, where children are expected to start acquiring basics of the alphabetic code. In the Southern Africa Labour and Development Research Unit (SALDRU) / Funda Wande 2019 baseline report, the majority of children who began Grade 1 were unable to identify one letter sound, despite almost universal attendance of Grade R.

While English letter-naming benchmarks may not apply to African languages directly, knowledge of letter-sounds may be similar across alphabetic languages. In languages with a transparent orthography, this early

[86] The average word length is much shorter in English than in agglutinating languages, which makes word recognition easier in English in early literacy acquisition.

learning area is finite and should not pose a significant challenge, as indicated by the Finnish and Turkish data. However, the early reading advantage may be offset by other complexities in transparent orthographies, such as more graphemes, a high occurrence of complex consonant sequences or the use of diacritics.

> The possible impact of such features on early reading has not yet been researched in agglutinating African languages. Collecting letter-sound data across the different language families to establish letter-sound benchmarks can help to identify the interplay of frequency and complexity of different letter-sounds, and help sensitize teachers to the importance of teaching the alphabetic code to the point of mastery as early as possible.

2.4.2 COMPLEX CONSONANT SEQUENCES BENCHMARKS

Nguni and Sotho languages have a complex grapheme to phoneme mapping. In addition to individual letters sound knowledge, learners need to recognize the consonant sequences and corresponding sounds of numerous digraphs and trigraphs. These complex consonant sequences appear in early grade reading texts with high frequency, presenting an impediment to word reading for learners whose knowledge of them is weak. The Funda Wande Impact Evaluation included a digraph and trigraph sounding subtask, in addition to the standard EGRA letter-sound subtask. Results show that, at the beginning of the year, 48 per cent of Grade 1 learners could sound at least one letter correctly, but only three per cent could sound at least one digraph correctly. By the beginning of Grade 2, most learners can recognize some letter sounds, but just more than half cannot yet identify a single digraph or trigraph. By the end of Grade 2, one in six children were still unable to identify a complex grapheme.

> These data show that the recognition of complex graphemes is a skill that is more advanced than single letter recognition and sounding. Given the importance of complex graphemes in Nguni and Sotho languages, it is important to benchmark this skill.

Table 4: Complex Consonant Sequences across South African Languages

Complex Consonant Sequences		
English	**Nguni Languages**	**Sotho Languages**
Usually made up of 2-3 consonants For example, sh, ch, th, wh, ph, kn, wr, -ck, -ng, -tch	2-3 letter sounds that can be made longer and more complex (up to five consonant letters) by blending with n- and/or –w, but also always followed by a vowel For example, hl, nq, gc, tsh, ngcw, ntshw More examples in isiXhosa, isiZulu, isiNdebele, Siswati: ch, kh, ph, th, gc, dl, hl, ts, ng Tshivenda: ng, dz, vh, kh, ṱh, tsh; Xitsonga: ng, dz, ny, ch, hl	2-3 letter sounds that can be made longer and more complex (up to five consonant letters) by blending with n- and/or –w but also always followed by vowel For example, sh, tl, kg, ng, ph, tsh, tšh, tjh, ntlh

Source: Katz (2020)

2.4.3 ORAL READING FLUENCY BENCHMARKS

As discussed above, ORF has the most robust rationale as a single skill indicative of reading proficiency. At lower reading levels, reading accuracy may be a good proxy for reading proficiency,[87] but this option needs to be explored with further research.

While ORF norms in English have been researched extensively for both home language and additional language readers,[88] assessment data of ORF in African languages have only started emerging in the past five years or so. Table 5 reflects some of the data to date. Research on ORF rates in Tshivenda, Sesotho, and isiNdebele has not yet been undertaken or published.

Table 5: ORF Data: Mean Words Correct per Minute

ORF		Grade 1	Grade 2	Grade 3	Grade 4
Jukes et al. 2018	Swahili Intervention Swahili Control	7 5	20 18		
Piper et al. 2018	Swahili Intervention (PRIMR) L1 Intervention (PRIMR +mt)* Swahili Control	2 2 4	11 10 13		
Taylor et al. 2018	Setswana Coaching Intervention Setswana Control		29 24		
Taylor et al. 2019	Setswana Coaching Intervention Setswana Control	**		41 38	49-56*** 47-55
Malda et al. 2013	Setswana			37	
Spaull et al. 2018	Sepedi Xitsonga isiZulu		39 41 21	55 59 31	
Menendez & Ardington 2018	isiXhosa baseline isiZulu baseline		5 8	13 19	19 25

* Language 1 Intervention Primary Mathematics and Reading plus Mother Tongue
** Grade 1 Setswana data has been collected in 2018, but has not been analysed
*** Two passages were used and therefore the range of cwpm is presented

Overall, the evidence points to learners reading slowly and some not at all. In the baseline report by Menendez and Ardington,[89] for example, 41 and 32 percent of isiXhosa and isiZulu learners respectively could not read a single word in Grade 2. About half of the isiXhosa speaking and almost three-quarters of the isiZulu speaking children attempted at least one comprehension question. Of these, 35 and 40 percent of isiXhosa and isiZulu learners respectively scored zero on the comprehension questions in Grade 2. By Grade 3, 28 and 19 percent of isiXhosa and IsiZulu children could still not read a single word.

Because of differences in average word length in disjunctive and conjunctive written texts, it is important to examine how ORF profiles in the Sotho and Nguni language subfamilies differ or are similar and to use the data to establish reliable and valid ORF benchmarks for the African languages. These differences in word length have implications for ORF norms in that reading in the Nguni conjunctive system yields slower words per minute rates than in the disjunctive Sotho scripts because words are longer in conjunctive systems. For example, Grade 2 learners in Sepedi showed a mean of 39 cwpm versus their IsiZulu Grade 2 counterparts at 21 cwpm.[90] In contrast, Grade 2 learners in English at the fiftieth percentile average 89 cwpm.[91] It is likely, therefore, that ORF rates in the Sotho language group will be similar. Similarly, ORF

[87] RTI International (2017)
[88] Hasbrouck and Tindal (2006)
[89] Menendez and Ardington (2018)
[90] Spaull et al (2018)
[91] Hasbrouck and Tindal (2006), United States of America

rates in the Nguni language groups will be similar, but that rates between these language groups will be different.

> Given the strong relationship between ORF and reading comprehension, having benchmarks for these two aspects of reading is especially important in raising the awareness of teachers to what successful reading might look like in different languages and at various grades.

2.4.4 READING COMPREHENSION BENCHMARKS

Unlike decoding, reading comprehension is an open-ended aspect of literacy development that is fine-tuned throughout our lives. Children can differ quite considerably in their comprehension skills, and teachers need to increase the comprehension levels of their classes. The PIRLS results in all three recent assessment cycles (2006, 2011 and 2016) show how reading comprehension needs critical attention in South African schools. Although the CAPS provides some guidelines as to what kinds of comprehension questions to include, very little support is provided as to what counts as acceptable comprehension levels.

Reading comprehension is complex and draws on many knowledge bases and skills; it is far more open-ended and less reducible than the more easily computed benchmarks in foundational decoding components of reading. There is a distinction between the four levels of reading comprehension, appropriate to a specific maturation level (see Table 6).[92] These levels serve as guidelines and are not absolute.

Table 6: Four Levels of Reading Comprehension

Independent level	98% decoding accuracy 95% level of comprehension	These are highly skilled readers who can learn effectively from texts appropriate for that specific maturational level. Typically, they can respond to both literal and higher-order comprehension questions.
Instructional level	95% decoding accuracy 75% comprehension	These are readers who do not have significant reading problems but who benefit from reading instruction at their maturational level. They can usually respond to literal and inferential questions, but higher-order comprehension abilities such as integrative and evaluative questions may pose challenges.
Borderline level	90-94% accuracy in decoding 55-74% comprehension	These readers need to be given additional reading exposure and practice. Their performance on literal comprehension is better than higher-order comprehension.
Frustration level	less than 90% decoding accuracy about 50% or less comprehension	These are readers who have major reading problems and who are reading well below their maturational level. They find comprehension beyond the literal level difficult. They need intensive reading programmes to increase their reading level.

Although a score of 60 per cent for comprehension in standardized reading comprehension tests signals that a child requires additional support, for many teachers in South Africa, a score of 60 per cent would probably be regarded as a good achievement.

> Broad guidelines such as those above (Table 6) can alert teachers as to what good reading accomplishment looks like, remind them of the close relationship between decoding and reading comprehension, and call them to high standards of reading comprehension that are required for learning success.

[92] McCormick (1995)

2.4.5 READING ACCURACY BENCHMARKS

Although there is a strong argument for fluency as an indicator of reading proficiency, it may also be desirable to benchmark lower-order skills to track learners' progress towards reading proficiency, particularly those skills that benchmark earlier stages of reading development. This is useful in a context such as South Africa, where children struggle with basic reading skills.

One possibility is to measure reading accuracy – the proportion of words that are read correctly. Children first learn to read accurately before their reading speed increases. At these early stages of reading, accuracy is an essential predictor of both comprehension and fluency.[93] In other words, learners who can read words accurately are the ones who can read more fluently and with greater comprehension. Reading accuracy is an important determinant of comprehension until fluency becomes the key determinant at a later stage of reading development. Reading accuracy has the advantage that it is measured in the EGRA alongside fluency. It is expressed as the proportion of words attempted that was read correctly, from the same passage used to assess ORF.

In terms of the level required for comprehension, reading accuracy has an advantage over reading fluency in that in all languages, learners are measured on the same percentage scale and are aiming for the same target (100 per cent). Betts[94] suggested that an accuracy rate of below 90 per cent was associated with being a 'frustrated reader,' although there have been few attempts to verify that figure.[95] Johns and Magliari recorded one exception.[96] They found that fourth- to sixth-grade readers in the U.S. could achieve 70 per cent comprehension with average accuracy rates of 94 per cent. In contrast, first-grade readers had the same comprehension level at an average accuracy rate of 91 per cent. One of the few international studies of accuracy[97] found some level of consistency in reading accuracy associated with comprehension across languages. Data from seven countries in Africa, South Asia, and Southeast Asia showed that the accuracy rate among learners who could read with comprehension varied from 81 per cent in Burundi to 99 per cent in Vietnam. This may be a more manageable cross-linguistic variation than is found for reading fluency.

> Therefore, reading accuracy might serve as a more consistent foundation for benchmarks across languages at earlier stages of reading development.

2.5 CAN BENCHMARKS BE SHARED IN SIMILAR LANGUAGES?

Given the costs associated with developing benchmarks, it is worthwhile considering whether benchmarks need to be developed in each South African language, or whether ones developed in one language can be used in another.

Part one explained that reading development is affected by both the linguistic structure of the language as well as its orthography; thus, all reading benchmarks must consider these factors.

Accuracy and fluency benchmarks for letter-sound recognition and ORF could potentially be set for each language group because orthographic complexity can affect both these measures of reading. On the other hand, comprehension benchmarks can be the same across languages. If the reading comprehension assessments are somewhat equivalent across languages (for example, ask a question after each sentence), then it is reasonable to assume that children should reach similar levels of comprehension, all else being equal. Even though the total number of words read before a question may differ (due to orthographic

[93] Kim (2015); Kim et al (2014); Kim and Wagner (2015); Petscher and Kim (2011)
[94] Betts (1946)
[95] Morris et al (2011)
[96] Johns and Magliari (1989)
[97] Cardoso and Dowd (2016)

features), the content remains the same. Thus, it is reasonable to use the same comprehension benchmarks across languages.

Reading benchmarks in closely related languages, which share similar orthographic features, could potentially share the same benchmarks for accuracy and fluency in letter-sound recognition, and ORF. For example, data from isiZulu learners can be used to derive benchmarks for isiZulu as well as the other closely related Nguni languages (isiXhosa, isiNdebele, and Siswati). This may be possible because all four languages share very similar orthographic (conjunctive writing, no diacritics) and linguistic (extensive consonant inventory, phonological processes such as vowel coalescence) properties.

> It would be important to construct the letter-sound recognition task used for benchmarking carefully to include the lowest common number of complex consonant sequences so that the benchmarks are fair for languages with greater or fewer complex letter groups (that is, digraphs, trigraphs and quadgraphs).

It might also be possible to use data from one of the Sotho languages to benchmark reading in the other Sotho languages. For example, data from Sesotho could potentially be used to benchmark reading fluency for Sesotho, Setswana, and Northern Sotho. All three languages share very similar orthographic (disjunctive writing, similar average word length, zero to one diacritic) and linguistic (large consonant inventory, seven to nine vowels) properties. Again, concerning the letter-sound recognition task, it would be important to consider the prevalence of complex letter groups across languages for the letter-sound recognition assessment used.

At this stage, there is insufficient information to speculate about Xitsonga and Tshivenda. The use of diacritics in Tshivenda may warrant separate benchmarking activities in this language. Analysis of word length data may assist in deciding whether Nguni or Sotho benchmarks can be used for either language.

Before proceeding with shared benchmarks, two data sources should be consulted.

First, word length assumptions for each language must be supported by evidence. Specialized children's literature corpora were developed for all nine African languages in South Africa.[98] The average word length for each language should be determined to support the word length data reported by Prinsloo and de Schryver,[99] who used very small corpora based on texts aimed at adults. Afrikaans and English were not included in these studies and, therefore, data should be collected for these languages to produce a complete set of data for all South African languages.

Second, a pilot study could use existing data to determine the implications of using one benchmark for multiple languages. A lot of data exists for isiXhosa and isiZulu, so benchmarks can first be piloted in these languages. The pilot should determine how the fluency and accuracy benchmarks for letter-sound recognition and ORF, and the reading comprehension scores differ for each language, controlling for other factors such as socioeconomic status and gender.

> The above discussion has focused mainly on structural factors in consideration of sharing benchmarks. It will also be important to consider language identity in the discussion of whether benchmarks can be shared. Each language under consideration comes with its own culture and group of people. Speakers may object to the use of benchmarks set in one language being used for other languages.

[98] Randera and Rees (2019)

[99] Prinsloo and de Schryver (2002)

2.6 IS IT NECESSARY TO SET BENCHMARKS BY GRADE?

When considering the establishment of reading benchmarks, it is important to bear in mind that reading as a construct changes over time.[100] Different processes come into play at varying stages of development and contribute differentially to performance as reading proficiency increases. What happens in the brain of a proficient Grade 7 reader is not the same as a proficient Grade 3 reader, which, in turn, is not the same as the brain of a Grade 1 reader. This means that the importance of some processes as drivers of reading development diminish as proficiency increases and are replaced by qualitatively different processes. In other words, a benchmark that may be important for Grade 1 (for example, letter-sound reading) may not have the same importance in Grade 3, where a different benchmark may be more useful for tracking reading progress.

> Given the changes in the construct of reading over time, it is essential to consider how benchmarks should differ by grade.[101]

The suggested process for setting benchmarks by grade is as follows: First, determine the type and level of competency learners should reach for each grade. Second, determine a method for setting the benchmark. A possible plan for benchmarking at each grade could look as shown in Table 7 below.

Table 7: Setting Reading Benchmarks by Grade

Skill	Method
Proficiency in letter sound reading	Data-informed expert opinion on the number of letters named to demonstrate proficiency
Proficiency in digraph and trigraph reading	Data-informed expert opinion on the number of digraph / trigraphs read to demonstrate proficiency
Minimum fluency associated with reading a Grade 2 passage accurately	Use analytical methods to determine the level of fluency associated with comprehension
Minimum fluency to understand a Grade 3 passage	Use analytical methods to determine the level of fluency associated with comprehension
Advanced fluency	Data-informed expert opinion about fluency level Grade 4 learners should reach (above the minimum required for comprehension)

The above is an example of setting one benchmark per grade. It is possible to set benchmarks in more than one skill if desired. Note that the analytical methods described in the next section are only relevant when benchmarking the minimum fluency required for comprehension (Grades 2 and 3 in the example above). It is not relevant for benchmarking lower-order skills (letter-reading) or for advanced fluency, above the minimum level required for comprehension.

2.7 WHY SET TARGETS?

After benchmarks have been determined, it is possible to set targets for the proportion of learners reaching the benchmark. Targets should represent what is attainable, given current levels of achievement and projected improvements in instruction. Targets may increase year-on-year. For example, if 25 per cent of current Grade 3 children meet the benchmark, a target of 35 per cent could be set to reach the benchmark in one year, 50 per cent in two years and so on.

[100] Stanovich (2000)
[101] Jukes et al (2018)

> Targets must be set after benchmarks, and the benchmark level determines the targets, not the other way around. This is because benchmarks should be set at levels that are meaningful in terms of the reading development of learners in the local context. That is, the benchmark should indicate a level of proficiency in reading that is relevant to the goals of the education system.

2.8 HOW TO ESTIMATE BENCHMARKS?

Benchmarking methods can be classed as either norm-referenced or criterion-referenced.

2.8.1 NORM-REFERENCED BENCHMARKS

Norm-referenced benchmarks are those against the typical performance of a learner in a population. For example, a benchmark may be used to flag children performing below the 20th percentile in the distribution of performance of similar children. Norm-referenced benchmarks are not commonly used in resource constrained countries, where the goal for reading achievement is likely to exceed the reading level of the current population. They also rely on having reading achievement data which represents the whole population, which is more demanding than criterion-referenced benchmarks discussed below.

> The benchmarking methods under consideration in this report are criterion-referenced, where reaching the benchmark indicates that a learner has achieved a level of proficiency in a particular skill. Using ORF scores as valid predictors of reading proficiency requires that validity be established with other measures of reading achievement,[102] discussed below.

2.8.2 CRITERION-REFERENCED BENCHMARKS

There are two methods of setting criterion-referenced benchmarks: data analytical methods and expert based methods. The first set of methods is based on statistical analysis (data analytical methods and linear regression), where data are available from the same learners for both the benchmarked skill (for example, ORF) and the criterion (for example, comprehension). Other competencies cannot be benchmarked easily against performance on other subtests, and therefore the second method is expert-based benchmarking (for example based on theory, expert opinion or curriculum goals). These methods are discussed below.

2.8.2.1 CRITERION-REFERENCED DATA ANALYTICAL METHODS[103]

There are many approaches to setting benchmarks using analytic data methods. The most commonly used methods specify a comprehension benchmark and then use statistical techniques to identify the fluency levels associated with reaching that benchmark. In the examples below, we have used 80 per cent comprehension as the benchmark. This threshold definition has been prevalent in EGRA assessments, where five comprehension questions are used. However, the use of 80 per cent comprehension as a threshold is not a strong recommendation. The performance threshold for comprehension should be defined as part of the benchmarking process taking into account the competency being benchmarked and the precise definition of the criterion against which it is being benchmarked, which, in turn, depend on curriculum goals and the aims of the benchmarking process.

The first three methods described in more detail below (mean, median, logistic regression) involved treating reading comprehension as a binary variable, above or below a performance threshold. In contrast, the linear regression method treats comprehension as a continuous variable. An alternative method to

[102] Wood (2006)

[103] Jukes et al (2018b)

benchmark fluency against comprehension is to use non-parametric techniques to examine the fluency-comprehension gradient to identify fluency thresholds;

i) Below which comprehension is unlikely to develop, and

ii) Above which there are diminishing benefits to increasing fluency and comprehension proficiency becomes the limiting factor.

These non-parametric techniques methods can be particularly useful in settings where comprehension skills are weak and insufficient learners reach the specified comprehension benchmark, or the relationship between comprehension and fluency breaks down. These techniques do not make any assumptions about the fluency-comprehension relationship and are also less reliant on the difficulty level of the specific comprehension questions. The main disadvantages of non-linear methods are that they are relatively advanced and require some level of judgement and hence do not lend themselves to an entirely systematic, replicable approach across languages or studies.

2.8.2.2.1 MEAN, MEDIAN LOGISTIC AND LINEAR REGRESSION METHODS

Mean method: This method calculates the fluency benchmark as the mean ORF of all learners with 80 per cent comprehension or higher.

Median method: Similar to the mean method, the benchmark is calculated as the median ORF of all learners with a minimum of 80 per cent comprehension.

Logistic regression model: A logistic regression model is fitted to produce the benchmark estimate. The outcome variable is binary, based on whether learners achieve 80 per cent comprehension. The independent variable is ORF from the same passage. The logistic regression estimates the probability of reaching 80 per cent comprehension at each level of ORF. The value of the probability can be adjusted to the data and needs of the analysis. In previous work,[104] benchmarks were set at the predicted fluency at which learners had a 0.5 probability of reaching 80 per cent comprehension.

Linear regression model: A linear regression model is fitted to produce the benchmark estimate. The outcome variable is the percentage of comprehension questions answered correctly, and the independent variable is ORF. This method fits a straight line to model the relationship between fluency and comprehension and then estimates the fluency level where the line reaches 80 per cent comprehension. The ORF upper and lower bounds are estimated at 80 per cent comprehension with 95 per cent confidence intervals.

Table 8 below compares the mean, median, and logistic regression methods. The mean and median methods are more straightforward to use than logistic regression. However, one disadvantage of the mean and median methods is that they produce benchmark values even when the data are not reliable or when there is no strong relationship between fluency and comprehension.

Regression methods rely on a strong relationship between fluency and comprehension and good quality data. This can be a disadvantage when using poor quality datasets because the method results in unstable estimates. However, this aspect of the method should be considered a strength because it prevents setting benchmarks that are not justified by the data. Good quality data and a strong fluency-comprehension relationship are reflected in more precise regression benchmark estimates.

Another advantage of regression methods is that they can be applied consistently across languages.[105] This may be particularly important when setting benchmarks across different languages in the same country. Regional representatives may advocate different benchmarks in their region (for example, a higher

[104] Jukes et al (2018)

[105] RTI International (2017)

benchmark motivated by higher aspirations, or a lower benchmark to inflate apparent achievement levels). In such cases, a consistently applied regression method can provide objective and comparable benchmark estimates across languages on which to base regional- and language-specific decisions. This advantage applies to both regression methods, although procedures for the systematic application to benchmarking across languages have been developed in greater detail for the logistic regression method (see Section 2.6.2.2).

The linear regression method has the advantage that it treats comprehension as a continuous rather than a binary variable, and therefore makes use of more information contained in the data. This conclusion applies equally to regression using other functional forms (such as polynomial or logit functional forms).

On the other hand, the mean and median methods produce higher benchmarks with higher-performing samples. Regression methods work on the assumption that, while one sample may have a higher achievement level than another, the shape of the fluency-comprehension curve will be the same in each. Thus, there should be less variation based on specific sample characteristics in benchmarks set by the regression method compared to the mean and median methods. For this reason, regression methods are more likely to produce consistent estimates across different samples and languages.

Table 8: Data Analytical Methods

Criteria	Mean	Median	Logistic Regression	Linear Regression
Ease of use	Simple	Simple	Advanced	Advanced
Robust to outliers	×	✓	✓	✓
Models fluency-comprehension relationship	×	×	✓	✓
Produces benchmark values even with unreliable data	✓	✓	×	×
Adjustable parameters	×	×	✓	×
Estimates benchmark precision	×	×	✓	✓
Consistent estimates across languages	×	×	✓	✓
Dichotomizes comprehension	✓	✓	✓	×

2.8.2.3 CRITERION-REFERENCED EXPERT-BASED METHODS

Expert-based: Another method is to set benchmarks based on theory, expert opinion or curriculum goals. This may be most appropriate with skills for which concurrent validity analyses are not appropriate, for example, to benchmark the lower-order skill of letter-reading. In this case, a criterion-referenced method would involve setting the benchmark at the number of letters read per minute associated with competency at a higher-order skill, such as word- or passage-reading. However, there is little precedent for this type of criterion-referenced method, and the rationale is not strong because there is evidence that teachers may not be a good judge of their learners' reading proficiency.[106] Instead, curriculum experts may judge that learners should be able to read all letters by the end of Grade 1. An analysis of the distribution of scores on this test may conclude that allowing for minor errors, learners reading 90 per cent of letters have achieved competency in letter-reading.

[106] Clark et al. (2019)

Modified Angoff: The Modified Angoff method[107;108] is a formal procedure for producing valid benchmarks based on both expert opinion and data.

A typical procedure is as follows:

Step 1

Identify target grades, competencies to be benchmarked and detailed performance level descriptors for each skill. Decide how many performance categories will be used for each competency. For example, categories maybe as follows:

- Just able to read accurately
- Just able to read with basic comprehension
- Just able to read with advanced comprehension

Step 2

Identify a group of about 15 experts to make judgements about the benchmarks. The experts should include people familiar with the curriculum in South Africa and with extensive experience of teaching literacy at the relevant grades. About ten experts should be current classroom teachers in each target grade. If this is a national assessment, there should be national representation in the group as well as representation along other important lines (gender, urban/rural, and so on). Document the criteria established to recruit for the panel, the processes put in place to identify people who match the requirements and include a professional profile of each expert selected in the documentation.

Step 3

Current classroom teachers in each target grade should select three learners in their class from each performance category. The learners should be at the level of the competency but not beyond it. Thus, if there are three performance categories, nine learners are needed. For oral reading fluency, performance categories may not be easy for teachers to judge. Instead, teachers can select learners representing a range of fluency levels.

Step 4

Assessments are selected to provide benchmarking data (for example, the EGRA) and teachers administer the assessment to the learners and record their results on each question. Assessments should be designed to align with the performance categories identified in Step 1. There should be sufficient items at the level of each performance category to differentiate learners in different categories.

Step 5

Convene all experts in a benchmarking workshop. Introduce the assessments for experts who did not administer them in their classes.

Step 6

Experts should form separate groups for each grade. Ask each expert to interpret the performance level descriptor provided by the steering group. Each expert should estimate the performance of a minimally proficient learner. For timed tests, this involves estimating the number of items (for example, words or letters) a minimally proficient learner could read in a minute. Experts should mark the last word they

[107] Ricker (2006)
[108] Ferdous (2019)

would expect this child to reach in one minute and then, for all words in the passage up to that point, indicate whether the learner would read the word correctly or incorrectly. For sub-tests with individual items (for example, comprehension) experts judge whether or not a minimally proficient learner would answer the item correctly, or roughly when two out of three learners at this level would answer the question correctly. Experts should also rate their level of confidence in each estimate.

Step 7

Estimates are shared and discussed. Present overall tendencies of the group without revealing the scores of individual participants.

Step 8

Data are presented to the group. These may include a distribution of performance by grade for each sub-test and the relationship between sub-tests (for example, between ORF and comprehension). Include an item analysis, if available, showing which items learners were able to answer and which were difficult. The data should also show the proportion of learners falling in each performance category. Results are discussed.

Step 9

Each expert independently re-assesses their estimates and confidence levels. This is a repeat of Step 6. Experts also assess their confidence in the overall benchmarking process.

Step 10

The initial and final estimates of the group of experts can be used as data. Typically, there is greater convergence on the final estimates compared to the initial estimates. The final estimates can be used as part of the data on which a final benchmarking decision is made.

The advantage of the modified Angoff method is that benchmarks can be more strongly rooted in the curriculum and in real classroom experiences of teachers and experts. It is particularly useful when data-driven methods for setting criterion-reference benchmarks are not appropriate. Disadvantages of this method include the difficulty in experts conceptualising a minimally proficient student in a way that is consistent across all their judgements.[109] Benchmarks set with this method are inevitably subjective and as with all group decisions, may be affected by social influence among experts. The quality of benchmarks set using this method can be improved through training of experts at the beginning of the workshop.

2.9 ASSESSING VALIDITY OF CRITERION-REFERENCED BENCHMARKS?

After setting the criterion-referenced benchmark, it is important to assess its validity. There are two types of validity to consider when establishing benchmarks, predictive and concurrent. A best-practice approach to benchmarking is to use longitudinal data to assess whether learners reaching a benchmark in lower grades go on to develop reading proficiency in subsequent grades (that is, predictive validity). This requires collecting longitudinal data over several years, and may not be practical for new data collection to establish reading benchmarks in South Africa in the next five years. However, several existing longitudinal data sets (Early Grade Reading Study (EGRS), Story Powered Schools (SPS), Funda Wanda, and Leadership for Literacy) can be used to investigate the predictive validity of benchmarks derived from those data sets. However, it would be possible to set up data collection to allow for the possibility of this kind of analysis in the future. A common method to establish concurrent validity is to benchmark one skill against another, typically ORF against comprehension. Several data-driven methods are possible, as outlined below.

[109] Ricker (2006)

Most benchmarking exercises based on the EGRA have used a concurrent validity criterion-referenced method in which fluency is benchmarked against comprehension. Typically, this method aims to identify the level of fluency at which children read with 80 per cent comprehension (four out of five questions correct on the EGRA comprehension measure).

PART THREE: ASSESSMENT, SAMPLING AND ANALYSIS

3.1 ASSESSMENT TOOLS FOR DEVELOPING BENCHMARKS

It is critical that benchmarks are based on valid and reliable assessment tools. Errors in assessment may be incorporated into the benchmarks which then persist for the lifetime of their use.

EGRA is recommended for assessing Grades 1-4 home language tasks:[110]

1. Letter-sound recognition
2. Complex consonant sequences (digraph / trigraph) reading
3. Word-reading accuracy
4. Oral reading fluency
5. Reading comprehension as assessed after reading the passage

The task would be developed by language and literacy experts. The task would be piloted in the same research context as the main study. It should take 15 minutes or less to administer per child and should be understood by all participants (that is, dialectal variation in lexical items is minimized).

> The reading comprehension test should have ten questions to improve reliability, and should assess different levels of comprehension (literal, reorganisation, inferential, evaluation, and appreciation) as appropriate for each grade.

It is essential to consider the positioning of information required to answer comprehension questions throughout the text.

However, there are problems with using the same tool to measure two different constructs (ORF and comprehension). Specifically, this approach can artificially inflate the degree to which the different constructs are correlated. For example, if a passage relays a story about a computer, learners who are familiar with computers will probably read the passage more fluently and understand more of the passage. In such a case, familiarity with content will be part of the reason fluency and comprehension are related.

Two possible solutions to this issue are recommended:

- One solution is to administer two different passages, with different subject matter. This approach allows you to compare ORF benchmarks set against comprehension performance from the same passage and a different passage.
- The second solution involves making comparisons between fluency levels, and associated benchmarks, on the two passages. Such comparisons allow for robustness checks on benchmark estimates and allow different passages to be equated (see below). It also ensures that the idiosyncrasies of a particular passage do not overly influence benchmark estimates.

3.1.1 PILOTING AND PSYCHOMETRIC PROPERTIES OF TOOLS

The following psychometric properties of tools are recommended to be assessed during a piloting exercise to ensure high quality.

[110] If time is limited, use two reading passages and related comprehension questions instead of word-reading accuracy.

Inter-rater reliability should be determined for all assessments. Inter-rater reliability is an assessment of the quality of assessors and the assessments.

This involves one assessor administering an assessment while a second assessor observes and records scores independently. Inter-rater reliability should exceed 95 per cent.

More in-depth psychometric analyses can be conducted with the reading comprehension measure because it includes independent items which are not timed. If the reading comprehension measure contains more than one type of comprehension (for example, literal and inferential), factor analysis can be used to assess whether underlying constructs in the measure map well onto these categories. Item response theory (IRT) approaches can be used to assess the range in the difficulty of individual items. The reliability of the comprehension measure can be evaluated with Cronbach's alpha or with the Kuder Richardson Formula, if unidimensional. If more than one category of reading comprehension is assessed, reliability can be calculated overall and for each category independently.

Analysis should also determine whether sub-tasks correlate moderately to strongly with each other. For example, word reading and ORF should be highly correlated.

Finally, if parallel measures are used, such as two similar assessments with the same goal, analyses should determine the equivalence of the measures (see Section 3.1.2).

3.1.2 LEVEL OF TEXT

The difficulty level of a text can affect both reading fluency and comprehension. Text difficulty is affected by syntactic complexity, vocabulary frequency, number of unique and repeated words, average word and sentence length, and syllable complexity — whether the syllable ends in a vowel (an open syllable) or a consonant (closed syllable), and the presence of consonant clusters. When using different passages in benchmarking exercises (with the same learners, in different grades or over time), it is useful to compare the passage difficulty levels. Readability formulae can help formalize this process; however, such formulae have not been developed for African languages. Funda Wande is in the process of developing readability formulae for isiXhosa. Funda Wande is also working to verify the readability formulas by using learners' experience of the text. The principles they adopt could be applied to other African languages.

There are two possible approaches to the selection of text level. One is to select authentic grade-level texts in the target language. With this approach, the number of learners able to reach performance criteria for their grade may be low. It may be necessary to sample children from higher grades or from high-performing schools to have a sample of competent readers sufficient for benchmark estimates. There is a potential concern with this approach in that the sample may not be representative of typical learners in their grade. An alternative approach is to match the level of text to the abilities of learners in their grade or to include two or more passages at different reading levels.

If multiple passages are used, analyses should be conducted to assess the equivalence of passages. Equating is a statistical procedure used to convert scores from multiple forms of a test or assessment to the same common measurement scale. This conversion process adjusts for any existing differences in difficulty between forms so that a score on one form can be equated to its equivalent value on another form. In other words, equating makes it possible to estimate the score that a learner assessed using one form of a test would have received had they taken a different test form.

> Accordingly, if multiple forms of assessments are developed for benchmarking purposes, equating procedures are essential to ensure that any statistical differences in difficulty are well-understood and that score adjustment can be made as needed. For specific guidelines and recommendations on equating EGRA forms or subtasks, see the EGRA Toolkit Second Edition.[111]

[111] RTI International. 2015. Section 10.5 Statistical Equating and Annex M: Recommendations for Equating

3.1.3 PASSAGE TIMING

When using questions about an oral reading passage for benchmarking, a three-minute time limit is recommended to ensure most learners finish the oral reading passage. Learners who finish the passage can attempt all the reading comprehension questions. In such cases, fluency can be calculated as the number of correct words read in the first 60 seconds (pro-rated for learners who complete the passage in less than a minute) or as an average rate over three minutes. It is essential to be consistent in the choice of these methods of fluency assessment because they are likely to produce slightly different rates.

> Analysis suggests[112] that benchmarks set from passages read with a one-minute time limit can be inaccurate as many learners do not read the whole passage and are, therefore, not asked all of the comprehension questions. Children who read slowly do not have the opportunity to demonstrate their comprehension skills. If children do not read to the end of the passage, it presents difficulties in deriving a comprehension score. If students are given a score which is a proportion of questions attempted (for example, a student who correctly answers one question out of two scores 50 per cent) the comprehension score is not comparable with other students who answered all questions – particularly if questions increase in difficulty. If students are given a score as a proportion of all questions available (for example, a student who correctly answers one of five questions correctly scores 20 per cent, regardless of how many questions they attempted) then the relationship between comprehension and fluency will be artificially inflated, because slow readers will be penalized, even if they understand what they are reading. For these reasons, learners should be allowed to read to the end of the passage.

3.2 SAMPLING FOR DATA ANALYTICAL METHODS

3.2.1 SAMPLE SIZE

A separate benchmark activity should be conducted in each language of instruction. For each language, a conservative estimate is that the sample should contain 1 000 children who can read at least one word of the passage. This sample applies to each benchmark set. For example, if a different benchmark is set in each grade, using a different passage, a sample of 1 000 children is required for each grade. Additionally, the EGRA Toolkit[113] contains useful information on random *versus* cluster sampling, items that need to be included in the sample size calculation, and various formulae for calculating sample size (see the EGRA Toolkit Annexes B and C).

A formal sample size calculation supports the above estimate. The sample size calculation is provided first for the mean method and is the sample size of learners (who are able to achieve above a specified comprehension score) given a desired precision in the estimated average reading fluency of that sample. The equation and its defined components can be found in the box below.

The equation is split into three parts:

i) the sample size of learners if they were sampled using simple random sampling (first part of the equation)

ii) an adjustment for the actual sampling methodology used, sampling schools then learners within the school (centre of the equation)

[112] Jukes et al (2018)

[113] RTI International, EGRA Toolkit (2015)

iii) an adjustment to account for the proportion of learners who achieved the specified comprehension score (final part of the equation).

Where: **n** is the sampled number of learners needed

ω is the desired precision such as ± 95% confidence interval band.

α is Type I error

β is Type II error, meaning (1- β) is power

N is the population of schools in the population. If the population is greater than 5000, this becomes an irrelevant term.

ICC is the intraclass correlation within and between schools.

m is the cluster size, number of sampled learners in each school

p(rc) is the proportion of learners who are expected to achieve the specified reading comprehension score.

Table 9 provides the assumptions made for each grade. The $Z_{1-\alpha/2}$ is set to 1.96 which corresponds to a two-sided 95% confidence interval. A 95% confidence interval of ± 5.0 wpm is sufficient for determining a fluency benchmark. It should be noted that if you would like to cut the precision band in half (to ± 2.5 wpm), you would need to quadruple the sample size. The rest of the components were conservative estimates derived from data from Funda Wande, SPS, EGRS II and LFL studies.

Table 9: Assumptions and Estimates Used to Calculate the Sample Size for Grade 1, 2, and 3

Item	Grade 1	Grade 2	Grade 3
$Z_{1-\alpha/2}$ =	1.96	1.96	1.96
ω = Desired precision ± ω	**5**	**5**	**5**
N = Population Size	10,000	10,000	10,000
σ = Common Standard Deviation	18	15	14
ICC = Intraclass Correlation Coefficient	0.2	0.2	0.2
M = Cluster Size	12	12	12
p(rc80) = Proportion of Learners expected to reach comprehension benchmark	0.05	0.15	0.3

Table 10 provides the sample size of schools and learners. The top portion provides the sample size needed if all learners were to achieve the reading comprehension benchmark. As can be seen, the sample size is quite small, with only nine to 14 schools needed and 12 children in each grade at each sampled school. However, not all learners will be able to read with sufficient comprehension. If we sample learners randomly, we need to adjust the final samples to obtained 159 Grade 1, 111 Grade 2 and 97 Grade 3 learners who can read with comprehension. To do this, we divide these numbers by the proportion of

learners in the population believed to be able to read with comprehension. As can be seen from the bottom half of Table 10, the samples sizes increase significantly. This is particularly true for Grade 1 learners (159 to 3 180) where only five per cent of the population is estimated to read with comprehension. Alternatively, it may be possible to sample learners in higher-performing schools to reduce the total number of learners sampled, with the caveat that this sample would not be representative.

When calculating benchmarks using the mean method described in section 2.8.2.2.1, only learners who reach the specified comprehension benchmark are included in the analysis. If a low proportion of learners are expected to reach the comprehension benchmark, then a large initial sample of learners will be required to produce a precise fluency benchmark. For example, if an analytical sample of 100 produces fluency benchmarks of the required precision and only 25 per cent of learners reach the comprehension benchmark, then 400 learners would need to be assessed. It is suggested to calculate the required analytical sample for a one set of assumptions and then explain (1) how that would need to be inflated by inverse of the proportion of learners who were expected to reach the comprehension benchmark and (2) how changing any of the assumptions would impact the sample size.

Table 10: Sample Size of Schools and Learners Based on the Assumptions from Table 9

	Grade 1	Grade 2	Grade 3
Total learners needed if they all achieved the reading comprehension benchmark	159	111	97
Total schools needed if all learners achieved the reading comprehension benchmark	14	10	8
Final learner sample size	**3 180**	**744**	**324**
Final school sample size	**265**	**62**	**27**

To obtain the total number of learners needed if they all achieved the reading comprehension benchmark, three different standard deviations need to be used for each grade. This will hold all other components constant ($\sigma_{g1} = 18, \sigma_{g2} = 15, \sigma_{g3} = 14$) and obtain varying sample sizes of learners and schools. Grade 1, for instance, resulted in 159 learners in 14 schools. To keep the standard error estimate at 18 wpm but change the number of sampled learners in the school from 12 to 20, a sample of more Grade 1 learners (n=238) but from fewer schools (n=11.9) would be needed. If the original assumptions ($\sigma_{g1} = 18, m = 12$) for Grade 1 is kept, but the intraclass correlation changed from 0.2 to 0.3; the sample size will increase to 213 learners from 18 schools. Analysis of previous data sets[114] suggests that the logistic regression method requires a larger sample size, typically between two and five times the size of the mean method. Taking the large sample size estimates of around 200 for the mean method, a sample of 1 000 learners who can read at least one word is recommended for the logistic regression method. This is a conservative estimate designed to ensure a high probability of a precise benchmark, even with unfavourable conditions.

It is not recommended to use Grade 1 learners for benchmarking exercises where new data must be collected, as this is inefficient and expensive. It is recommended that new data collection should focus on Grade 3 learners. Given current low reading levels, Grade 3 learners sampled can be used for benchmarking on Grade 2 passages.

[114] Jukes et al, 2018, Room to Read Data Analytical Report (unpublished)

3.2.2 SAMPLE COMPOSITION (HIGH AND LOW PERFORMING LEARNERS).

The ideal sample should have a high proportion of learners who perform close to the target level of comprehension competency. Benchmark estimates are less precise if too many or too few learners perform below the target competency level. Samples between 30 and 70 per cent of learners who meet the competency level are recommended.

3.2.3 SAMPLES AND SUBSAMPLES

When benchmarking ORF against comprehension, it is not critically important for the sample to be representative of the overall population. This is because the relationship between fluency and comprehension is determined mainly by the properties of the language, rather than the characteristics of the sample. However, it would not be advisable to sample children from only one geographic or socio-economic group. Sampling from at least two provinces (where the language is spoken in more than one province) and from a lower and upper quintile of socio economic status distribution is recommended. The proportion of mother-tongue speakers in the benchmarking sample should be roughly the same (plus or minus 20 per cent) as the target population (that is, all school children being taught to read in the same language).

3.3 PRE-ANALYSIS STEPS FOR BENCHMARKING ORF AGAINST COMPREHENSION

Prior to conducting analyses of ORF data, it is essential to understand the sample and the data. Table 11 provides a list of the information to be aware of before running benchmark analyses.[115] This pre-analysis check will assist in understanding any potential limitations in the benchmark estimates and provide guidance on the most appropriate method.

Table 11: Checklist before Beginning Benchmark Analyses

✓	Step	Notes
	KNOW THE SAMPLE	
☐	Note how many children, grades, and schools are sampled and the sampling methodology used	NA
☐	Conduct initial descriptive analyses by grade and school	NA
	KNOW YOUR ASSESSMENT	
☐	Determine whether the comprehension questions are clear and relevant to the passage	Obtain an English translation of the reading passage and comprehension questions
☐	Note number of words, time limit	NA

[115] Table taken from Jukes et al (2018, p 13-14)

✓	Step	Notes
	KNOW YOUR DATA	
☐	Describe learner performance on the assessment	Plot the ORF and reading comprehension scores. Note the number and percentage of learners who - could read at least one word - could answer at least one reading comprehension question - scored at least 80 per cent on the reading comprehension
☐	Describe performance of children who could read at least one word	Note the mean ORF and the mean comprehension percentage score for children with ORF >0
☐	Describe performance on each reading passage word and for each comprehension question	For each word in the passage and the comprehension questions, run an item analysis. Take note of any word or any reading comprehension question that has a much lower (or possibly higher) percentage correct compared to the others.
☐	Describe the relationship between fluency and comprehension. Categorise the ORF scores into bins of, for example, five cwpm or ten cwpm (languages with longer words may require smaller bins). For each bin, take the average reading comprehension score. Plot ORF categories versus the mean comprehension score.	Ideally, the graph will slope upward and start to flatten out above 80 per cent comprehension. If the graph dips down or does not flatten out, check for the following: - Each ORF category has at least ten children - Problems with individual comprehension questions (see previous checklist item) - Learner(s) who read quickly without understanding

3.4 METHODS TO ANALYSE ORF AGAINST COMPREHENSION

The aim of benchmarking ORF is to determine the level of fluency associated with a criterion level of performance in the comprehension assessment. As discussed above, there is a range of analytical methods available to derive this benchmark.

The logistic regression method is one approach to setting objective benchmarks across languages (see recommendation below). The method is statistically complex relative to others. For these reasons, detail of the steps involved in a logistic regression analysis with 80 per cent comprehension as a criterion is presented.

Step 1

Consider which learners to include in the sample for analysis. As a minimum, learners with zero ORF should be excluded. Additionally learners with very low ORF can artificially inflate the strength of the ORF-comprehension relationship. Consider whether to exclude readers with ORF less than a minimum threshold. A suggested threshold is 20 cwpm, but may be lower for languages with greater word length.

Step 2

Create a binary variable by dividing the sample into two groups:

a. Those who scored at least 80 per cent on the reading comprehension (RC) (RC80 = 1)

b. Those who scored less than 80 per cent on the reading comprehension (RC) (RC80 = 0)

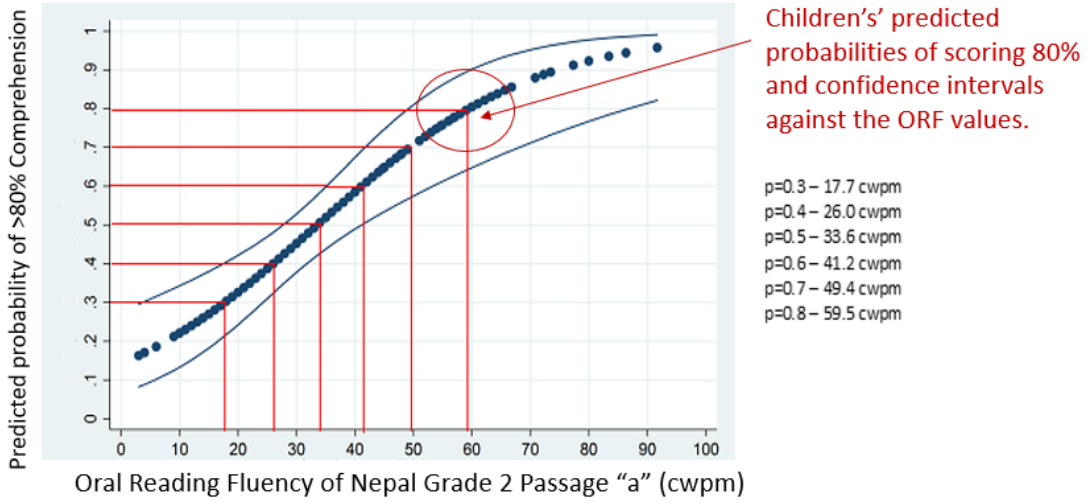
Step 3

Run a logistic regression model with RC80 as the outcome variable and ORF as the predictor. This analysis produces two fitted estimates: Beta (β) and a constant (_con). From this output, obtain each observation's predicted probability of scoring 80 per cent reading comprehension and associated 95% upper and lower bound estimates.

Step 4

Plot the learners' predicted probabilities of scoring 80 per cent (and confidence intervals) against the ORF values. An example of this type of plot can be seen in Figure 7 below (Note: This plot has been produced using data from Nepal, Grade 2 reading Passage "a". Blue data points indicate estimates and thin blue lines indicate confidence intervals).

Figure 7: Nepal Grade 2 Child Predicted Probabilities of Scoring > 80% in ORF



Source: Jukes et al, 2018, Room to Read Data Analytical Report (unpublished)

Step 5

Determine the fluency benchmark for a given predicted probability. With this method, benchmarks can be set with a choice of predicted probabilities values. The predicted probabilities that learners will have at least 80 per cent comprehension range from 0 to 1 (represented on the y-axis in Figure 7). Each predicted probability is associated with a different fluency benchmark level, indicated by the red lines. For example, in Figure 7, a fluency benchmark of 33.6 cwpm corresponds to a value of p=0.5. At this level, learners are predicted to have a 50 per cent chance of scoring at least 80% on comprehension. In other words, about half the learners should have 80 per cent comprehension at this level of fluency.

Step 6

The 95 per cent confidence intervals can be determined from the graph by observing the points on the thin blue lines (indicating confidence intervals) that map onto the chosen predicted probability. For example, a predicted probability of 0.5 has an associated upper bound of 42 cwpm and a lower bound of 27 cwpm.

Step 7

Finalize the predicted probability to use based on the following considerations:

a. What is the definition of reading proficiency? Do the estimated benchmarks align with the definition of reading proficiency for this sample?

b. How will the benchmark be used? Is the intended use appropriate to a benchmark set at the upper or lower end of the range of possible values, or something in the middle?

c. What is the precision of the benchmark estimate? In Figure 7, confidence intervals are wide (> +/- 10 cwpm) for predicted probabilities of 0.3 and less and predicted probabilities of 0.7 or more. This suggests that predicted probabilities of between 0.4 and 0.6 would be appropriate.

d. What is the proportion of learners achieving 80 per cent comprehension in the sample? The most precise estimates will involve predicted probabilities close to this proportion.

e. When setting comparable benchmarks (for example, for two regions in a country or with two different instruments), estimates will be most easily compared if using the same predicted probability for both estimates.

PART FOUR: CURRENT BENCHMARKING PRACTICE IN SOUTH AFRICA

4.1 CURRENT BENCHMARKING INITIATIVES

Room to Read has been assessing reading outcomes in Sepedi for several years using the EGRA. These data have been used to determine the impact of their literacy programme and to work out fluency benchmarks. At present, they are piloting new versions of the EGRA which, in addition to the traditional passage and question set, will include a Sentence Choice task as an alternate measure of comprehension. The Sentence Choice task presents a pair of sentences and asks the learner to determine which is correct and which is not. Based on the results of the 2019 pilots, they will decide whether to incorporate the Sentence Choice comprehension measure into their evaluation measures. Room to Read also piloted the EGRA in IsiZulu in 2019.

ReSEP, at the University of Stellenbosch, has funding from the Allan Gray Orbis Foundation Endowment and the United Kingdom Department for International Development /Economic and Social Research Council (DFID/ESRC) to conduct secondary analyses of existing early grade reading datasets in 2020. The DBE is supporting and participating in this effort through oversight forums and as part of the research team. They will attempt to establish ORF benchmarks, where possible. They will also track reading trajectories and intend to analyse other skills and variables as part of their task. The languages targeted in these analyses will include isiXhosa, isiZulu and some Siswati data. ReSEP is conducting this benchmarking work in close collaboration with Prof. Cally Ardington (University of Cape Town) and Prof. Alicia Menendez (University of Chicago).

The Southern Africa Labour and Development Research Unit (SALDRU) / Funda Wande conducted a baseline report in 2019 and is undertaking an impact evaluation to assess causal impact on learners' ability to read with meaning in the Foundation Phase. Funda Wande is developing readability formulae for isiXhosa.

Furthermore, data from EGRS I and EGRS II are large scale evaluations led by the DBE in collaboration with academics and international donor organisations. The study is contributing to building evidence on what works to improve learning and teaching of early grade reading in schools. Formal impact evaluations using randomised experiments and mixed methods (such as classroom observations) provide quantitative estimates of the interventions on home language and first additional language. The EGRS data could be used to develop reading benchmarks in African languages.

4.2 EXISTING DATASETS (KNOWN)

Stakeholder consultation revealed that there are currently eight known datasets, which already include data for letter-sound knowledge, ORF, and reading comprehension in the African languages of South Africa (see Annexure 1).[116]

[116] Adapted from Ardington (2018)

These include the following:

- Early Grade Reading Study I (Multiple donors/DBE)
- Early Grade Reading Study II (USAID/DBE)
- Emergent Literacy Intervention (USAID/Western Cape Department of Education)
- Zenex Literacy Project (Zenex Foundation/Evaluation Research Agency)
- Story Powered Schools Impact Evaluation (USAID/NORC)
- Literacy intervention and assessment data (Zenex Foundation/NECT)
- Systematic Method for Reading Success (USAID/RTI International)
- Economic and Social Research Council (DFID/ESRC)
- Funda Wande (Multiple donors/DBE)

These datasets could be explored for suitability for use in generating reading benchmarks. Access to these datasets may or may not be problematic, depending on who owns the data. Some datasets can be requested by a formal research request, and others will be made available to the public in the future.

The data is presented in Annexure 1, and a summary is provided in narrative form. Half of these datasets report on experiments (randomized control trials), three report on quasi-experiments, and one dataset reports non-experimental data. The majority of studies (seven) report on languages from the Nguni language group (isiXhosa, isiZulu and Siswati), four report on languages from the Sotho group (Setswana and Sepedi), and only one study reports data for Xitsonga. The available data sets do not have data on Sesotho, isiNdebele or Tshivenda. Concerning grade levels, the start of Grade 1, end of Grades 2 and 3 have the most data points. All the studies used EGRA. The majority of studies included measures of letter-sound recognition, word reading, ORF, and reading comprehension as assessed after the ORF task.

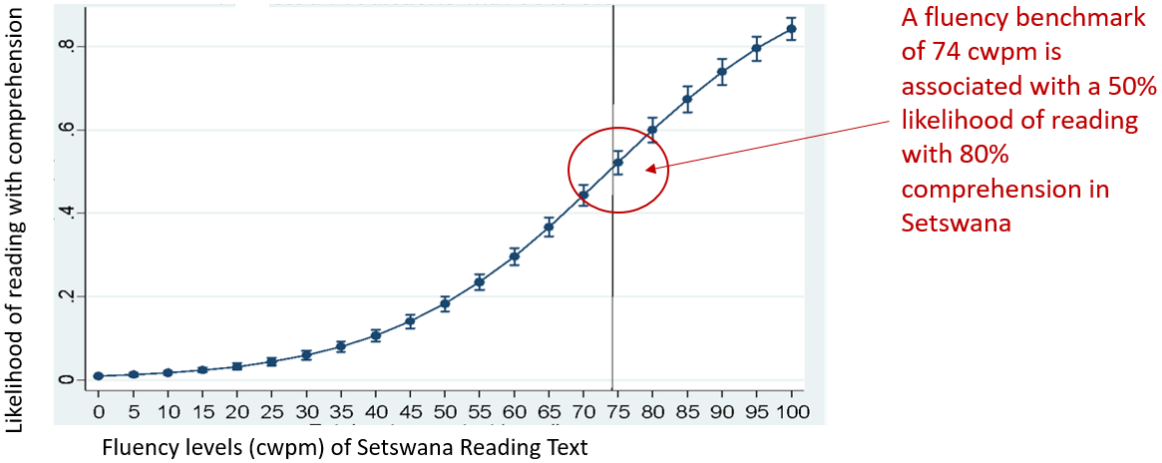
The PIRLS 2006, prePIRLS 2011, and PIRLS Literacy (2016) were included as available datasets, which present information on written comprehension. These datasets are publicly available for download. Unfortunately, because ORF is not assessed as part of these assessments, the data cannot be used to plan the development of ORF benchmarks.

4.3 USABILITY OF AVAILABLE DATASETS – PRELIMINARY ASSESSMENT (KNOWN)

The EGRS I sample in Setswana included 3 302 learners, of whom at least 75 per cent were able to read one word of the passage. This sample, therefore, exceeds the minimum sample requirement for 1 000 learners who can read one word of the passage.

Estimates of the probability of reading with 80 per cent comprehension have relatively narrow confidence intervals (see Figure 8). Based on this graph, a fluency benchmark of 74 cwpm (+/- roughly five cwpm) is associated with a 50 per cent likelihood of reading with 80 per cent comprehension in Setswana for a one minute reading passage. Estimates may change if children read to the end of the passage.

Figure 8: EGRS I Learner Probability of Reading With 80% Comprehension in Setswana



Source: Adapted from EGRS I and Taylor presentation (2019) (see Annexure 2)

Indications from the data alone suggest that this dataset is a good candidate for secondary analysis for benchmarking purposes. Further information is needed on the methods of data collection and passages used, particularly the difficulty of the passages and the comprehension questions.

The latest round of EGRS II included 1 922 Siswati and 762 isiZulu Grade 3 learners. Learners were allowed up to three-minutes to read the passage making this another highly relevant dataset for secondary analysis.

The data collected by the National Opinion Research Centre University of Chicago (NORC) for the impact evaluation of the USAID-funded Story Powered Schools (SPS) project is a promising dataset for benchmark development in isiXhosa and isiZulu with around 1 600 learners in each of Grade 2, 3 and 4 at baseline. The sample cannot be combined across the three grades because different instruments were used in each grade. For the second cohort, three-minute timings were used for the oral reading fluency and associated comprehension subtasks.

4.4 EXISTING LEARNER ASSESSMENT INSTRUMENTS (KNOWN)

The above review of existing datasets highlights the dominant use of the EGRA to assess African language reading skills in South Africa. There are versions of the EGRA available in each of the 11 official South African languages, but they warrant further research. Alternative reading assessments in South Africa in the African languages tend to be targeted at the Early Childhood Development phase (Grade R and below). For example, the South African Early Learning Outcomes Measure (ELOM)[117] measures gross motor development, fine motor coordination and visual-motor integration, emergent numeracy and mathematics, cognition and executive functioning, emergent literacy and language for children aged 50-69 months. Emergent literacy is measured in isiZulu, Setswana, and isiXhosa as part of the norming sample. The only developing literacy skills assessed included phonological awareness; otherwise, the rest of the tasks assess language skills. This test is not suitable for use in developing formal literacy benchmarks as the children have not commenced formal literacy instruction, so reading skills are not assessed.

Given its dominance in South Africa and other early grade reading interventions on the continent, the suitability of the EGRA as a tool to gather data for reading benchmarks must be considered. As mentioned earlier in this report, EGRA is a tool used in many African countries to assess a variety of reading skills. The benefit of using this tool is it is easy and quick to administer and score. Enumerators need up to a

[117] Snelling et al (2019)

week of training, and the test takes 10-15 minutes to administer per child. Additionally, data capturing can be facilitated using software such as Tangerine®, which was developed specifically by RTI International for EGRA and is being used increasingly in EGRA projects. Tangerine® is not the only way to capture EGRA data, but the platform enhances data quality and reduces assessment time spent with learners. Although word-reading and ORF cwpm scores cannot be compared directly across all languages due to structural and orthographic differences, the proportion of children scoring zero can be.

In terms of the construct validity on the tasks, the sub-tasks correlate with each other to an expected extent. For example, word and non-word reading has been found to have robust correlations (r = 0.91)[118] to one another as well as to ORF (rs = 0.89)[119] in Setswana at the end of Grade 1 (Wave 2, EGRS I)[120], in Grade 2 (rs = 0.91-0.94) (Wave 3, EGRS 1) and for word and text reading in Grade 4 (r = 0.91) (Wave 4, EGRS 2). Similar magnitudes of bivariate correlations are reported for Sepedi, Setswana and isiZulu in the Systematic Method for Reading Success (SMRS) study.[121] EGRA assessments of sub-skills in three sub-Saharan African languages (Kiswahili, Lubukusu, and Kikamba) show that results match those in other dominant languages.[122] Although path analysis has not been used to examine these structural relations in the African languages of South Africa, it can be expected that there will be similar results. Additionally, EGRA has predictive validity. Data from EGRA can predict reading scores at the next time point (for example, EGRS 1).[123] In conclusion, EGRA measures letter-sound fluency, word reading, text reading, and oral reading comprehension validly.

Little research has been done on the reliability of the EGRA tool in the South African context. Piper[124] reports on the use of EGRA with isiZulu, Sepedi and Setswana speaking children. He reports Cronbach's alpha of 0.95 for the whole test (letter-sound recognition, word reading, ORF, and reading comprehension). Additionally, a principal component analysis revealed that the test loaded highly on one factor, that of literacy skills. The uniqueness factors for each task are quite low; indicating that, to some extent, the tasks measure the same abilities. This finding is not unexpected, given that letter-sound recognition and word reading are precursors for ORF, and all three are precursors for comprehension. Generally, assessments at the start of Grade 1 are less reliable than those at later points.

Data from EGRS I shows that the start of Grade 1 assessment composite score has low bivariate correlations to other waves of data collection (rs = 0.18 - 0.23). A similar finding is reported for EGRS II where the Wave 1 composite correlates weakly with later waves (all rs < 0.35) as compared to the correlation among later waves (rs > 0.5). The relatively weaker reliability of test results collected at the start of Grade 1 is expected because such young children do not test well. For example, children may just not be test ready, could be shy of unknown adults (enumerators), and not have a long enough attention span for the assessment. In summary, the EGRA reliably assesses letter-sound recognition, word reading, and ORF, although reliability statistics should be reported more often. As mentioned previously, the way reading comprehension is measured by the EGRA typically is not very reliable. Use of the Sentence Choice test (with a reported alpha of 0.8) may be a worthwhile addition to existing EGR assessments.[125]

Dubeck and Gove[126] provide a guideline on how to develop various EGRA tasks for different languages. For the most part, these guidelines appear to be followed by test developers in South Africa. EGRA assessments should be developed by language and literacy experts familiar with the structural and orthographic features of the language. Thus, it is important for the tasks to reflect the typical structural and orthographic features of the language at a particular grade level (for example, the letter-sound

[118] r= denotes Pearson's correlation coefficient for continuous variables
[119] rs= Spearman's correlation coefficient for variables that have been converted into ranked scores
[120] EGRS I and II were set up as longitudinal studies with several waves of data collection over the years
[121] Piper (No Date)
[122] Kim and Piper (2019)
[123] Schaefer and Kotze (2019)
[124] Piper (No Date)
[125] Jukes et al (2018)
[126] Dubeck and Gove (2015)

recognition tasks should include complex consonant sequences as these are common in the African languages of South Africa). It is essential to pilot the assessment in the same context as the final study because contextual factors (such as dialect and school functionality) can affect the suitability of the assessment and the available time at the school. The pilot should examine whether the correlations between tasks are as expected, the reliability of the comprehension task is adequate, the test takes the intended amount of time, and that there is sufficient time available during a typical school day.

PART FIVE: PRACTICE GUIDE TO SETTING BENCHMARKS

The table below summarises the previous discussions about different approaches to benchmarking, highlighting the advantages and disadvantages of each.

Table 12: Summary of Benchmarking Approaches

Approach	Advantages	Disadvantages
1. Norms-based Set benchmark based on the performance of population. For example, what level of fluency is achieved by 80 per cent of Grade 2 learners	The analysis is simple and transparent	Norms of the population are too low. Hard to find a sample of competent readers. Needs to be a representative sample.
2. Criterion-based Benchmark represents the achievement of proficiency in a particular skill	The number of participants achieving the criterion level of performance can be calculated	
a) Predictive validity based on longitudinal data. For example, what level of fluency in Grade 2 is associated with being able to read proficiently in Grade 4?	Highly valid. Describes the trajectory of learners, not just where they are now. Sample need not be representative Objective	It takes time – ideally two years. Longitudinal studies require tracking of learners over time, which can be expensive. The analysis is difficult to communicate to non-statistical audiences
b) Concurrent validity. For example, what level of fluency is associated with comprehension?	Sample need not be representative Objective	The analysis is difficult to communicate to non-statistical audiences Requires a valid and appropriate level of comprehension assessment
c) Expert-based. For example, what level of fluency do local experts think learners should achieve by Grade 2?	Ensures relevance to curriculum and language Can incorporate data into the process	Subjective Potentially influenced by the workshop facilitator

This section of the report provides a guiding practice note to setting benchmarking in South Africa. A description of the process of benchmarking, based on best practice, is presented.

5.1 THE BENCHMARKING PROCESS

PHASE 1: SET UP DECISION-MAKING STRUCTURES FOR BENCHMARKING

To ensure buy-in and input from a wide range of stakeholders, setting up a steering group to oversee the work is recommended. This entails inviting participants from a range of departments and institutions, with a variety of responsibilities and oversight. At the very least, a benchmarking workshop should include (as appropriate) decision-makers from DBE responsible for the curriculum, teacher training, school supervision or inspection, policy and planning, examinations and assessment board, as well as national and regional directors, language experts (in the ministry and academic institutions), and representation from teachers and school leaders. Additionally, a small group of technical experts from DBE is suggested to investigate the results of data analysis and subsequent recommendations.

PHASE 2: ESTABLISH GOALS OF THE BENCHMARKING PROCESS

The steering group should determine the goals of the benchmarking process. The aim should be to review the proposals presented here and to identify benchmarks to be developed. The main questions to be addressed are:

1. What are the competencies to be included in the benchmarking exercise?

   Competencies should be a verbal description rather than referencing a metric. Ideally the competencies should already be described in the curriculum. If not, this is an opportunity to identify gaps in the curriculum. Initial discussions with DBE suggest that benchmarks should be set for letter reading, digraph and trigraph reading, oral reading fluency and comprehension.

2. What level of competency should be achieved at each grade?

   This should also be a performance level description for each competency, derived from the curriculum. For example, "… by Grade 2, children should be able to read a grade-appropriate passage fluently and with comprehension." For each grade, a description of how the competency develops should be provided.

3. How will benchmarks be used?

   An important decision is whether the benchmarks will be used only as a system diagnostic – a metric against which to measure progress – or will be used by teachers in formative assessments to guide instruction. Initial discussions suggest that benchmarks will be used primarily as a system diagnostic with plans in development for their use by teachers.

4. How many benchmarks should be used for each proficiency level?

   We recommend using either one or two benchmarks for each proficiency. If only one benchmark is used, learners either meet the benchmark or they do not. If two benchmarks are used, three categories are possible, such as 'not proficient', 'proficient' and 'highly-proficient'.

PHASE 3: DECIDE ON THE APPROPRIATE COMPETENCIES TO BENCHMARK

The following competencies are based on the assumption that benchmarks are requested for letter reading (letter-sound knowledge and complex consonant sequences), ORF and comprehension in Grades 1-6. One or more of the following competencies can be benchmarked:

BENCHMARKING LETTER-SOUND KNOWLEDGE

Letter-sounding is a good indicator of early reading skills and is recommended to chart progress among learners who are not yet able to read passages with minimal fluency. There is no strong theoretical justification or precedent for using statistical methods to benchmark letter reading against another skill, such as ORF. Instead, letter reading benchmarking using expert judgement, following the modified Angoff method, described above in Section 2.6.2.3, is proposed.

BENCHMARKING COMPLEX CONSONANT SEQUENCES READING

Reading of complex graphemes is a more advanced skill than letter-sounding and is important in the reading of Nguni languages. Benchmarking complex consonant sequences reading using expert judgement, following the modified Angoff method, is recommended.

BENCHMARKING ORAL READING FLUENCY

Benchmarking ORF is recommended because this skill has the most substantial claim to being a proxy for overall reading proficiency. Two methods are recommended for determining the ORF benchmark, depending on the level of the benchmark (minimum or higher level) required. These methods are described below.

Method 1: ORF for minimum comprehension proficiency and reading accuracy

When learners read on their own (as opposed to when teachers read them a story), they begin to understand text when they read with a minimum level of fluency and accuracy. The relationship between fluency, accuracy and comprehension allows for empirical methods to be used to identify the level of fluency at which minimum comprehension is achieved. The recommended approach is to assess learners on their reading fluency, accuracy, and comprehension of a passage and to examine the relations among them. It is recommended to use only literal questions in the comprehension measure, if possible, to measure minimum proficiency. For existing datasets with a mix of literal and higher-level, consider using the data to define a basic level of comprehension for benchmarking purposes

Two broad approaches to estimate this benchmark are recommended:

i. Understand the relationships between ORF, accuracy and comprehension.

   Produce plots of mean ORF *vs* mean accuracy and mean ORF *vs* mean comprehension. Additionally, describe the median and 25 and 75 percentile values of fluency at each level of accuracy and comprehension (for example, using box plots). Based on these analyses, it should be possible to identify a range of values of ORF at which minimum proficiency in comprehension is achieved. Support for the use of the range of ORF values is more reliable if learners achieve both good comprehension and >90 per cent accuracy in this range. These exploratory analyses should also allow for the identification of fluency thresholds above which there are no benefits, in terms of improved comprehension, to increasing fluency. This can be particularly useful in settings where learners do not perform well on the comprehension questions.

ii. Derive formal empirical estimates of the fluency benchmark.

   This is particularly important if objective empirical evidence is needed to justify different benchmark levels in different languages. One option is to use the logistic regression method described above to determine the level of fluency at which there is an X per cent probability of reading with Y per cent comprehension or Z per cent reading accuracy. The values of X and Y can be determined based on analysis and guidance by DBE technical advisors. Estimates are most precise when X is around 50 per cent. The value of Y depends on the nature of the comprehension test. It should be set at a value that is indicative of a minimum level of proficiency in comprehension (that is, there is little probability of achieving this score by chance). The level of required reading accuracy is typically 90 or 95 per cent.

Method 2: Higher levels of ORF

For higher levels of fluency, above the minimum required for basic comprehension, it is less justified to base the benchmark estimate solely on the relationship between fluency and comprehension. It is recommended to use expert judgement for higher levels of fluency. The expert judgement (for example, with the modified Angoff method) could be informed by data, such as the distribution of fluency and the relationship between fluency and different levels of comprehension (for example, literal, inferential).

BENCHMARKING COMPREHENSION

Benchmarking comprehension using the modified Angoff method is recommended. The first step is for the steering group to produce comprehension performance level descriptors for each grade, perhaps in terms of comprehension levels identified in the CAPS, literal, reorganisation,

inferential, evaluation and appreciation. The procedure follows that described in Section 2.6.2.3. Teachers are selected from all relevant grades and identify children in their classes who have borderline proficiency concerning the performance level. A comprehension assessment is administered to these children before the benchmarking workshop takes place. In the workshop, teachers make judgements on a question-by-question basis as to whether a minimally proficient child in each grade would be able to answer the question correctly. The modified Angoff benchmark-setting process can take place at the same time, with the same group of experts, for all three key skills, letter reading, ORF and comprehension.

PHASE 4: SELECT ANALYSIS METHOD

The analysis method chosen depends on the characteristics of the language being benchmarked. As described in this report, little is known about many of the indigenous South African languages. Thus, any benchmarking process would begin with basic data analysis in the chosen language. An essential consideration is the strength of the relationship between fluency and comprehension. As discussed above, it is possible that the fluency-comprehension relationship is less strong in transparent orthographies and all indigenous languages of South Africa are more transparent than English or Afrikaans. If the fluency-comprehension relationship is weak, it will lead to less precise estimates of the benchmark when using comprehension as a criterion. Typically, the benchmarking process will produce a range of values and will require experts and stakeholders to set a value within that range. Where the fluency-comprehension relationship is weaker, the range will be wider, and there will be more reliance on expert opinion and less on the data than in other languages. There is some evidence to show that languages with transparent orthographies have strong text reading fluency, and reading comprehension relationships that allow for precise benchmark-setting, as do languages with varying depths of orthography.[127] On the other hand, there is some evidence pointing to poor comprehension despite high fluency, particularly in languages with very transparent orthographies.[128]

There are several other unknowns in the indigenous South African languages that this work has the potential to investigate with little or no additional data collection. For example, there are many implications for reading instruction, assessment and benchmarking in the contrast between conjunctive orthographies (where several morphemes are combined in longer words) and disjunctive orthographies (where each morpheme or two constitutes a separate word and thus words are shorter and more numerous). Languages with conjunctive orthographies include isiXhosa, isiZulu, Siswati, and isiNdebele. Languages with disjunctive orthographies include Sepedi, Sesotho, Setswana, Tshivenda, and Xitsonga. Several questions are posed by the contrast between these two types of orthography. For example, ORF is typically assessed by the number of words read per minute. Presumably, learners reading in a language with conjunctive orthography (longer words) read fewer words per minute.

> Given that syllables are of a similar length in the indigenous African languages, perhaps syllables-per-minute is a better metric to assess reading skills across languages.

This has implications for readability formulae to equate passages in benchmarking exercises. Is the difficulty of reading simply due to the number of syllables per sentence and in the passage overall? Or is a three-syllable word in a conjunctive orthography more difficult to read than three similar one-syllable words in a disjunctive orthography? There are similar implications for reading accuracy. Do learners make the same number of mistakes per syllable or words across all languages or do longer words lead to more errors? These are questions that could be addressed simply by EGRA data collection in which learners' reading accuracy is assessed at the syllable level, a relatively simple change to the procedure. These analyses would have important implications for the benchmarking process and for understanding the development of reading in South African languages more generally.

---

[127] Jenkins et al (2003); Jenkins and Jewell (1993); Kim et al (2010, 2014); Kim and Wagner (2015); Roehrig et al (2008), cited in Kim and Piper (2019)

[128] Graham and van Ginkel (2014)

PHASE 5: SET THE BENCHMARK

The outcome of the above processes will be a recommended benchmark (or more than one if the approach of multiple benchmarks is being adopted) for each performance level descriptor. The recommendations may include a possible range of values and should be presented with supporting data and rationales. The steering group can then adopt and finalize benchmarks.

PHASE 6: SET TARGETS

Targets can be set for the proportion of learners to achieve each benchmark at each grade level over time, depending on government cycles of assessment and planning.

PHASE 7: EVALUATE BENCHMARKS

After benchmarks are set, their quality can be assessed. Several metrics of quality are available for ORF benchmarks set against comprehension. These are discussed below.

The width of the 95 per cent confidence interval of the benchmark estimate is an indication of its precision. As discussed above, if benchmarks are not set precisely, more weight is placed on expert and government opinion to determine the value from within the 95 per cent confidence interval range.

The benchmark can also be evaluated for its ability to classify learners as above and below the 80 per cent comprehension threshold. If the fluency benchmark is set correctly, learners performing above it will be those who can read with comprehension, and those below will be learners who struggle.

Three statistics summarise how well a fluency benchmark classifies children as reading with good or poor comprehension, namely, sensitivity, specificity, and correct classification. Sensitivity is a measure of how accurately the benchmark identifies children who can read with good comprehension, and specificity is a measure of how well it classifies children with poor comprehension. Correct classification is an overall statistic that combines specificity and sensitivity.

For benchmarks determined by expert judgement, additional data can be analysed to assess quality. For example, the concurrent validity of the letter reading benchmark could be assessed by analysing the performance on higher-level tests, such as digraph reading, for learners above and below the letter reading benchmark. If longitudinal datasets are available, analysis can consider two groups of learners, those above and below the letter reading benchmark in Grade 1, and assess their progress on other assessments in Grade 2 and above.

5.2 COMPARING AND PRIORITISING LANGUAGE BENCHMARKS

It is difficult to make direct comparisons in achievement between different languages for reasons previously discussed in this report. However, benchmarks can be set so that they have the same meaning in each language. That is, learners who reach the benchmark in any given language will have achieved the same level of reading proficiency. It is recommended that assessments are at a similar level of difficulty across languages, based on considerations outlined in Section 3.1.1 to achieve this. For reading fluency benchmarks set against comprehension, cross-linguistic comparisons are facilitated by using the same analytical method in each language. Using the logistic regression method for such cross-linguistic comparisons is recommended.

Due to budgetary constraints, it may be necessary to prioritize the development of benchmarks for certain languages before others. The budget can inform the prioritisation of languages for benchmarking of speakers of each language, the availability of viable existing data, the urgency of the problem (focusing on languages with lower performance first), and the capacity and availability of the main stakeholders.

These are discussed below. Furthermore, prioritisation can be affected by a decision to share benchmarks across languages of the same group.

Number of speakers:

Figure 1 indicates the distribution of the 11 official South African languages spoken by household members, by population group. This should be used as a starting point for understanding which languages should be prioritized in the benchmarking process.

Availability of existing data:

There is relatively more data available for the Nguni languages (specifically isiZulu and isiXhosa) than the Sotho languages, Xitsonga, or Tshivenda.

The urgency of the problem:

Performance in the PIRLS Literacy assessment in 2016 was poor for all African language readers, with 80 per cent or more of children unable to reach the low international benchmark. Sepedi readers performed the worst with 93.3 per cent of children being unable to read for meaning, followed closely by Setswana with 89.9 per cent of children being unable to read for meaning.

Capacity and availability of key stakeholders:

Benchmarking will require the collaboration of multiple stakeholders. Any decision on which languages to prioritize must take into account the available capacity.

PART SIX: PRACTICE NOTE ON BENCHMARKING STRATEGIES

Three strategies are presented in this section to set benchmarks in South Africa.

STRATEGY 1: BENCHMARKS BASED ON ANALYSIS OF EXISTING DATASETS

The focus of this strategy is to identify or generate datasets that can be used primarily for estimating ORF benchmarks against comprehension measures.

DESCRIPTION OF STRATEGY 1

This strategy involves finding existing datasets which include measures of fluency and comprehension for early grades in the required languages. An analysis of the relationship between fluency and comprehension can be used to set benchmarks as described above, providing these datasets meet specific criteria.

For existing data (see Annexure 1), it is vital to run analyses to check the precision of benchmark estimates (confidence intervals less than approximately five cwpm) before collecting more data. Strategy 1 will benefit from collaboration between funders, researchers and other stakeholders with similar interests in doing secondary analysis of existing data for benchmarking purposes.

Apart from EGRS II and SPS datasets, there are three other formerly USAID-funded literacy projects with datasets that could potentially be mined for reading benchmarks (see Annexure 1). These include the Ukusiza Project, the kaMhinga Literacy Project and the Teacher Assessment Resources for Monitoring and Improving Instruction for Foundation Phase (TARMII – FP).

The Ukusiza project was implemented as part of USAID's School Capacity and Innovation Programme (SCIP) to: "… improve reading skills among learners in Foundation Phase and Intermediate Phase grades."[129] The programme targeted Grade 1 – 3 home language and Grade 1 – 6 English as First Additional Language (EFAL) teachers, to improve their practices in the classroom. The project was evaluated using a Randomized Control Trial (RCT) design using repeated cross-sectional studies to determine the quantitative impact of the intervention on literacy skills. The baseline and midterm studies tested 2 560 learners in Grades 2 and 3 on letter naming fluency, familiar word fluency, ORF, and reading comprehension across three home languages (isiZulu, Sesotho, Setswana) and EFAL groups.

The kaMhinga Literacy Project (siyaJabula siyaKhula Learner Regeneration Literacy (LRL) programme) was implemented in Limpopo province in Grade 1, 4 and 7 classrooms. The project was assessed using a non-randomized intervention and comparison group design. Altogether 5 351 learners (2 192 Grade 1, 1 639 Grade 4 and 1 520 Grade 7) were assessed at baseline for EFAL using the standard international EGRA, Schonell's and PIRLS assessments. Three subtasks (Letter Sound, Word Recognition, and Non-Word Decoding) of the Grade 2 version of the EGRA was applied to Grade 1 and Grade 4 learners.[130]

The TARMII-FP project was implemented in four provinces in South Africa, the Free State, Limpopo, Mpumalanga and North West provinces. The research design involved selecting 20 experimental and 20 control schools within a district in each province. Data were collected on literacy achievement levels (oral and written) at the beginning and end of the 2014 academic year from a random sample of 20 learners from Grades 1, 2, and 3 in the selected schools. A total of 3 200 Grade 1, 3 200 Grade 2, and 3 200 Grade 3 children were targeted in the intervention. In a pre-test post-test of Grade 3 learners in 2014, the researchers obtained an 83 per cent response rate for the Grade 3 post-test (2 671 learners). The full datasets have not been cleared for public release and still require de-identification. The data are therefore not currently available for download. Once available, these datasets could potentially be explored against the criteria for secondary analysis.

[129] Chimere-Dan et al (2015, p 10)
[130] Murray (2013)

MAPPING DATASETS

Existing datasets should be mapped to understand whether the data is of sufficient quality to produce precise benchmark estimates and whether the process of data collection is described in enough detail to allow interpretation of the estimates.

STEPS IN DETERMINING WHETHER THE EXISTING DATA CAN BE USED FOR ANALYSIS (PRE-ANALYSIS STEPS)

Different data sets may lend themselves to various methods of analysis. The following aspects of the data collection and instruments should be considered when assessing the use of existing data sets:

- Passage reading difficulty: the passage read by learners should be available for analysis. The passage should be at an appropriate level of reading difficulty for the target learners.
- Comprehension question reliability and difficulty: responses to comprehension questions should be entered individually, which allows internal reliability to be assessed. Refer to Section 3.1.1 on piloting and psychometrics.
- Passage timing: learners should be allowed to read the passage for at least three minutes. If learners are stopped after one minute but have not yet completed the passage, the benchmark estimates will be less reliable.
- Sample size: ideally, there should 1 000 learners who can read at least one word of the passage. However, the adequacy of the sample size depends on the precision of the benchmark estimates derived. Datasets with less than 1 000 learners can be analysed to see whether benchmark estimates are reliable.
- Passage numbers: ideally, learners should be assessed on two different passages.
- Sample achievements: benchmarks will be most reliable if based on a reasonable number of learners (for example, between 30 and 60 percent of them have at least minimal comprehension proficiency).

It is recommended that these criteria are reviewed over time based on the results of benchmarking, and comprehension thresholds, cut-offs and once the reliability of samples is confirmed.

PROCESS OF BENCHMARKING USING EXISTING DATASETS

The process of benchmarking is relatively straightforward once pre-analysis checks have been done. Although pre-analysis checks are useful, the benchmarking analysis process itself will indicate whether the data is of sufficient quality by producing a precise benchmark estimate. If a reliable estimate were produced, the additional primary consideration would be whether the passage and comprehension questions were set at the right level.

FEASIBILITY OF THE STRATEGY

Once identified, it is straightforward to investigate and analyse existing data sets. It is recommended that this process be undertaken regardless of whether or not additional data are collected. The main disadvantage of secondary data analysis is that there is no control over the design of the instruments or the data collection procedures at this stage.

This may lead to a benchmark estimate that is not precise or one that is set too low or too high due to an inappropriate choice of passage (for example, one that is too difficult, too easy, or relies too much on knowledge of the subject).

Another disadvantage of this strategy is that data analysts are beholden to the methods of others. Even if these methods are considered adequate, they are unlikely to comprise a standardized strategy, which would have allowed performance against benchmarks to be comparable over time and across languages.

TEST EQUIVALENCE

Depending on the language that is prioritized, it will be essential to develop a strategy to ensure test equivalence for each language and across languages. Test equivalence ensures that differences in benchmarks across languages are due to language characteristics and not an artefact of the assessment.

STRATEGY 2: BENCHMARKS BASED ON PRIORITISED ADDITIONAL DATA COLLECTION

DESCRIPTION OF STRATEGY 2

This strategy involves assessing additional learners to increase the size of a sample. There are few existing opportunities where this 'top-up' data collection could be feasible. At the time of writing this report, for the SPS project, the evaluation team planned for data collection in August 2019 for the end of Grade 3; Room to Read also intended to pilot the Early Grade Reading Assessment in isiZulu in 2019 and will have baseline data in early 2021 to inform isiZulu benchmarks. The EGRS I and II are examples of a dataset that could potentially be topped up with data collection scheduled for 2020.

OBJECTIVES FOR COLLECTING ADDITIONAL DATA

This strategy would be valuable in the case that there were insufficient learners in an existing sample or because there was an inadequate balance of learners above and below the 80 percent comprehension threshold.

STEPS IN DETERMINING WHETHER AN EXISTING DATASET CAN BE TOPPED UP

The criteria in Strategy 1 of passage reading difficulty, comprehension question reliability, and passage timing also apply to Strategy 2. An additional consideration is whether a topped up sample would include 1 000 learners who could read at least one word of a passage with between 30 and 60 per cent reading at 80 per cent comprehension.

PROCESS OF BENCHMARKING USING ADDITIONAL DATA

A necessary precursory step to benchmarking using existing data involves analysing the existing dataset to determine if the above criteria are met and finding details of the data collection procedures used so that they can be replicated in the top-up data collection. Over time, the criteria may need to be amended.

FEASIBILITY OF THE STRATEGY

The advantage of Strategy 2 compared to Strategy 3 (collecting new data) is that data collection will already be a budgeted cost. However, it cannot be assumed that this strategy will involve significant cost savings. In some cases, the cost may be minimal (for example, where cost-sharing for training and data collection is included). In other cases, this strategy may entail substantial additional costs (for example, if data collection expands geographically or if more days are required for fieldwork). The cost advantages should be balanced against the disadvantages that the data collection will necessitate the use of existing instruments, regardless of their quality. Similar to Strategy 1, with Strategy 2, the potential to develop and apply a standardized strategy to benchmarking over time and across languages is lost.

TEST EQUIVALENCE

Depending on the languages prioritized, it will be important to develop methods to ensure test equivalence for each language and across languages. Test equivalence ensures that differences in benchmarks across languages are due to language characteristics and are not an artefact of the assessment.

A systematic process for translating the assessment into different languages should be determined. The texts used for ORF should have, as far as is possible, equivalent meaning with similar syntactic difficulty, and vocabulary frequency in different languages. Current knowledge about word frequency and its influence in reading is limited. It is recognized, however, that this could be difficult, given the different lengths of sentences with the same meaning across the South African languages. Translated texts should be back-translated to check that the meaning of the original text has been preserved.

STRATEGY 3: BENCHMARKS BASED ON PRIMARY DATA COLLECTION

DESCRIPTION OF STRATEGY 3

This strategy involves collecting primary data to estimate benchmarks. Stakeholders who were interviewed almost unanimously highlighted the need to develop reading benchmarks in South African languages that currently have little to no available reading data.[131] As noted earlier in this report, Tshivenda is one of these languages. It has the most diacritics, so this might affect letter-sound relations and word reading, and its orthography is partially disjunctive. This means that ORF in Tshivenda could be somewhere between the Sotho and Nguni languages.

OBJECTIVES FOR BENCHMARKING LANGUAGE(S) BASED ON PRIMARY DATA

To develop a standardized strategy to benchmarking that is consistent over time and across languages.

To collect primary data on fluency and accuracy in indigenous South African languages to inform our understanding of the process of reading in these languages.

SPECIFIC STEPS IN BENCHMARKING BASED ON PRIMARY DATA COLLECTION

The steps include:

- Identify the goals of benchmarking
- Identify skills to be benchmarked (for example, ORF, letter reading)
- Identify criteria for readability of passages and difficulty of comprehension questions
- Engage experts in the design of passages, comprehension questions and an independent measure of comprehension (for example, the sentence choice test)
- Conduct piloting, revision and re-piloting of instruments
- Train data collectors
- Collect data
- Analyse data
- Recommend benchmark

DETAILS OF DATA COLLECTION PROCESS

Data should be collected with standard EGRA procedures. Enumerators should also be trained to assess reading accuracy on a syllable as well as a word basis so benchmarks in the various languages can be compared more easily. However, evaluating reading accuracy on syllables is not a meaningful unit on which teachers can assess their learners' progress. The syllables correct per minute metric would only be useful if there is a heuristic to convert this to words per minute for each language so that teachers can understand and check the fluency of their readers. The analysis team would need to determine with the language and linguistic experts what the best unit of analysis for benchmarking would be well in advance of developing tools and training fieldworkers

FEASIBILITY OF THE STRATEGY

This process is more expensive than the others but allows more control over data, instrument quality and calibration and, with it, the potential to develop a standardized method for use across languages and time.

TEST EQUIVALENCE

A systematic process for translating the assessment into different languages should be determined. The texts used for ORF should have equivalent meaning with similar syntactic difficulty and vocabulary frequency. Translated texts should be back-translated to check that the meaning of the original text has been preserved.

[131] A more cost-efficient approach might be to build on existing datasets for other South African languages.

PART SEVEN: ELEMENTS TO SUPPORT BENCHMARKING

7.1 BENCHMARKS AND THE DBE CURRICULUM STANDARDS

The creation of reading benchmarks needs to be carried out in the broader context of the South African curriculum. The National Curriculum Statement (NCS) Grades R-12 stipulates the curriculum, rules for promotion, and assessment procedures teachers should use. CAPS documents were developed for each subject and specify what should be taught and how it should be taught at quite a fine level of detail.

> Reading benchmarks should provide clarity on the standards expected at each grade, as well as the lines of progression in literacy, in line with the curriculum. They should clarify what learners should know and do to reach these standards and what teachers need to do to support them to reach this goal.

7.2 CURRICULUM AND ASSESSMENT POLICY STATEMENTS

The normative frameworks for time allocation across subjects, for language skills, reading activities, and so on are specified in the CAPS as follows:

Time allocation for languages, home language (HL) and first additional language (FAL), Mathematics, and Life Skills are set out.

Different **language skills** are identified for the Foundation Phase, namely,

Foundation Phase			
Listening and speaking	Reading and Phonics	Writing	Handwriting
Intermediate Phase			
Listening	Reading	Writing	Language structure and use

Thinking and Reasoning, and *Language Structure and Use,* are integrated into the four language skills in the Foundation Phase and Intermediate Phase. The time allocations and requirements for each are specified per grade, per week, for each of the four terms.

CAPS refers to (but does not elaborate on) the **five main components of reading** identified by the National Reading Panel, namely, phonological awareness, word recognition (sight words and phonics), comprehension, vocabulary, and fluency.

Different **types of comprehension questions** are identified (namely, literal, reorganisation, inferential, evaluation, and appreciation) and teachers are encouraged to ask questions across the range of question types.

CAPS emphasizes four different **reading methods** to be adopted in the Foundation Phase, namely, Shared Reading, Group Guided Reading, Paired Reading, and Independent Reading.

CAPS specifies **core and support materials,** and states what materials teachers should have.

CAPS specifies the different types and **genres of text** that should be dealt with across the grades.

The recommended **length of texts** for English First Additional Language (EFAL) across the grades is given, but not for home languages.

Assessment codes ranging from one (not achieved) to seven (outstanding) and percentages are given for reading and reporting in the Intermediate Phase. However, these are not conventional reading measures used to assess reading.

In the EFAL CAPS, **vocabulary ranges** are suggested for Grade 1 (700-1,000 words), Grade 2 (1,000-2,000 words), and Grade 3 (1,500-2,500 words).[132] A list of the most common 300 words in English is given towards the end of the document.[133] This is not provided for home language.

A **vocabulary chart** is provided in CAPS Intermediate Phase.[134] This schema provides a developmental perspective on how many words EFAL learners in different grades should know per term. Two types of words are identified namely, common spoken words and new words in reading vocabulary, but the difference between them is not explained.

7.3 FOUNDATION PHASE

At the Foundation Phase level (Grades R-3), there are four subjects: HL, FAL, Mathematics, and Life Skills. Language and literacy teaching takes place during the HL and FAL lessons. Most commonly, in the Foundation Phase, the language taught at HL level is the same as the Language of Learning and Teaching (LoLT) also called the 'medium of instruction' at the school. The medium of instruction, and therefore HL, is typically an African language. In these schools, the FAL generally is English. In many schools, English becomes the medium of instruction from Grade 4, but the HL and FAL subjects remain as an African language and English, respectively.

There are differences between language and literacy teaching in HL and FAL lessons, including the amount of time dedicated to the subject, as well as the curriculum expectations. The curriculum stipulates minimum and maximum times that should be devoted to each language subject. If the maximum time is dedicated to HL, then the minimum time is dedicated to FAL, and *vice versa*. The time allocation for Grades 1-3 is displayed in Table 13 below.

Table 13: Time Allocation for Home Language and First Additional Language

Grade	HL maximum time (hours)		FAL maximum time (hours)		Total time (hours)
	HL time	FAL time	HL time	FAL time	
1	8	2	7	3	10
2	8	2	7	3	10
3	8	3	7	4	11

The second difference between languages offered at HL and FAL level is the curriculum focus and expectations. For example, in Grade 1 the EFAL curriculum mostly focuses on developing oral language proficiency and learners are expected to develop only emergent reading (for example, recognizing a few high-frequency words). On the other hand, the HL curriculum requires that children learn to read in the language from Grade 1. The home language curriculum has four areas of development. These include:

- Listening and speaking (oral language proficiency),
- Phonics and reading,
- Writing, and
- Language use.

[132] Department of Education (2011, p.22)
[133] Department of Education (2011, p. 87-89).
[134] Department of Basic Education (2011b, p.27)

The final HL outcomes for each grade are summarised below from the CAPS.[135]

> **Grade 1**. Children start developing their formal reading and writing skills in Grade 1. By the end of Grade 1, they should be able to listen with attention to what others are saying, express their feelings about stories, and tell stories with a defined beginning, middle, and end. They should recognize all the single letters and common digraphs, and be able to spell words learned in phonics lessons. Children should be able to read with: "… increasing fluency and expression"[136] and monitor their comprehension. They should be able to answer open-ended questions about texts. Children should be able to write correctly (for example, hold a pencil, use spaces, capital letters, and punctuation).

> **Grade 2**. Children continue to develop their reading and writing abilities, although they may receive some support using shared/paired reading and writing. By the end of Grade 2, children should be able to read independently at instructional level (word recognition between 90 and 95 per cent accuracy) during group guided reading and start to read available books for pleasure or information. They should develop responses to literature, explaining how events in stories make them feel, or justify their opinions of a text. Children should develop their awareness of conversation norms (listening to the other speaker before replying, asking relevant questions), and register (adapting speech/writing for different contexts). Children should develop their use of cursive script and should be able to write a range of genres (instructional texts, narratives, poems) using the process approach.

> **Grade 3**. By the end of Grade 3, children should be prepared for 'reading to learn' in their home language or the Foundation Phase LoLT of the school. That is, they should be proficient readers and writers so that they are prepared for subject learning from Grade 4. Ironically, the LoLT in the majority of schools switch to English in Grade 4, and these learning outcomes are then expected of children in English, their first additional language. Children should be able to read independently (orally and silently) at the instructional level (word recognition between 90 and 95 per cent accuracy) during group guided reading. They should also develop reading strategies that they can use to read and understand unknown words as well as employ self-correction strategies as part of monitoring their reading. They should be able to work with a range of genres (advertisements, poems, fiction) to identify the intended audience and purpose. They should be able to answer higher-order comprehension questions ("should her grandmother have told her that…?") and display their appreciation of a text with substantiation. Children should be able to read stories already covered independently, and they should be able to write words and sentences in both print and cursive scripts. They should be able to write in different genres independently (for example, narratives, newspaper articles, recipes).

7.4 GAPS IN CAPS

Except for the vocabulary norms, the normative reading framework in CAPS is focused far more on the teaching, or pedagogical aspects of reading instruction and little is specified in terms of learner accomplishment. Teachers are told in detail **what** to teach for reading and **how** to teach it; far less attention is given to how to nurture a culture of reading and motivate children to enjoy reading, and very little guidance is given as to how a good reader in the different grades looks.

Questions to consider in developing teacher's ability to guide and nurture learners reading abilities could include the following:

[135] Department of Basic Education (2011c)
[136] Department of Basic Education (2011c, p 75)

- If children are expected to read fluently, accurately, with good intonation, and to comprehend and enjoy what they read, then how do fluency, accuracy, good intonation, comprehension and enjoyment look?
- How should teachers assess the children and, having done the assessments, how should the scores be interpreted?
- If a Grade 1 isiZulu reader reads 19 words correctly in one minute (but has made three mistakes) and reads the words syllable-by-syllable in a sing-song voice, and can only answer one literal question on the text correctly, is the reader a good or struggling Grade 1 isiZulu reader?
- If a Grade 5 child achieves 60 per cent on a reading comprehension passage in EFAL, can they be regarded as a good comprehender?
- If the only questions that were answered correctly are literal, does the child have a comprehension problem?

For schools and teachers to have a common understanding of what good (and weak) reading accomplishment looks like, teachers need more specific guidelines on how to assess the different components of reading accurately, and how to interpret those assessments appropriately in terms of developmental reading norms. Reading benchmarks could go some way to providing this information to teachers.

In principle, one could establish reading benchmarks for every element of reading and every component in those elements. However, normative 'overkill' in the everyday classroom should be avoided. Teachers should be provided with a limited and manageable set of benchmarks that target core reading skills at different stages of reading proficiency and provide them with useful information that enables them to identify problem areas easily and early and take corrective action.

There are currently three obvious gaps in CAPS concerning providing teachers with clear guidelines as to how reading success looks. Two of them relate directly to the development of the more constrained skills in decoding (namely, letter-sound knowledge and ORF), and the third relates to the more open-ended, unconstrained knowledge-based aspect of reading comprehension.

Constrained skills are those which include a finite set of abilities, and are therefore skills which can be mastered over time with explicit and systematic instruction. In contrast, unconstrained skills or abilities include those which develop over time, and do not necessarily have an endpoint.[137] For example, phonics is a constrained skill; there are a finite number of letters and letter groups in a language, which can be learned and automated in a short time. On the other hand, vocabulary and comprehension are unconstrained skills as the reader develops their word knowledge, background knowledge, story schema and so on over time, and a maximum level for each of these abilities is not necessarily easy to define at any given time.

It is vital to make the distinction between constrained and unconstrained abilities. Constrained skills form the basis of early literacy acquisition, but are not necessarily sufficient for comprehension. They can be automated in a short time. Therefore, as children progress through school, less time should be devoted to constrained skills as they reach ceiling level (that is, children can achieve mastery of them) and more time should be devoted to the unconstrained skills, such as vocabulary and reading comprehension.

[137] Stahl and Dougherty (2011)

Letter-sound knowledge:

Although phonics is put firmly on the curriculum agenda in CAPS from an instructional point of view, it does not give teachers clear guidelines of how to assess letter-sound knowledge and how 'good performance' in this skill looks. This is a constrained skill, so high levels of performance are expected. The more accurately and automatically learners can identify sounds associated with letters, the stronger their decoding skills will be. It is important for Grade 1 teachers, in particular, to ensure that their learners attain high levels of mastery of letter-sound knowledge in each term. By Grade 3, such knowledge is already assumed, so assessment of this foundational skill may not be as critical as it is in Grade 1 (unless there are still struggling readers in a Grade 3 class).

ORF:

No guidelines or normative frameworks are provided in CAPS to help teachers determine whether their learners are on track with their ORF. This is understandable, given the current lack of research on early reading in the African languages and on ORF for English FAL learners. However, because ORF is a bridge between decoding and comprehension and because ORF scores generally correlate strongly with comprehension, ORF is arguably one of the most critical reading tools in a teacher's assessment toolkit. In some standardized reading tests in the U.S., for example,[138] strong readers decode with at least 98 per cent accuracy, while 90 per cent accuracy or less, signals weak readers and identifies early cracks in reading development. Unless teachers are made aware of the importance of accuracy and automaticity in alphabetic decoding, they may regard a score of 90 per cent in decoding as very good. A score of 90 per cent for comprehension is very good, but for a decoding skill, where fast accuracy is essential, 90 per cent is not good enough. Children who read slowly and haltingly are at risk of reading failure and ORF benchmarks can help teachers be alert to such cracks in literacy development. This area calls for local research to be undertaken urgently so that development of fluency across grades in the African languages' conjunctive and disjunctive orthographies can be established.

Reading comprehension:

Although CAPS provides examples of different types of comprehension questions that can be asked, very little support is provided as to what counts as acceptable comprehension levels. In South Africa, many teachers would regard 60 per cent as a good reading comprehension score, while reading research suggests that 60 per cent for comprehension signals a reader in need of additional support.[139] As noted earlier in this report, the 2016 PIRLS results show that 78 per cent of Grade 4 learners did not meet minimum comprehension standards (answering literal questions), and higher-order questions were beyond the reach of most of them. Results such as these indicate that South African teachers would benefit from guidelines that show how good performance in comprehension looks. While specific benchmarks in reading comprehension are not easily quantified, some framework that sets out guiding principles for distinguishing different levels of comprehension success would be beneficial.

7.5 EXPERTISE REQUIRED FOR SETTING BENCHMARKS

Establishing reading benchmarks will require collaboration among experts with different specialisations. These specialisations and their purposes are presented here. These are gleaned from Key Informant Interviews and are far from comprehensive. Specific expertise should be sought, depending on the language and type of benchmark selected.

[138] McCormick (1995)
[139] McCormick (1995)

Given the shortage of skills in African language benchmarking in South Africa, it is recommended to build the capacity of Masters and PhD students to conduct this kind of work. There is no readily available repository of specialists in South Africa, and some work should be done to identify students with interest and potential.

The core team conducting the work should be inclusive of African languages specialists, aside from their inclusion as part of the expertise needed for conducting benchmarking. This would enrich the work as it is primarily focused on African languages.

Early grade reading experts: these experts are subject matter experts. They can evaluate the rigor of the research and assure the quality of the assessment tools. For the benchmarking approaches set out in this report, these experts must agree with scientific evidence about the importance of automaticity.

Translation experts: if benchmarking is to be conducted in multiple languages and the research team anticipates equivalent tests, it will be up to the translation experts to ensure rigorous translations. There may be a need for two groups, those who can work from English into an African language and translators that can work from an African language into English (back-translation). Currently, available information suggests that there may be a capacity constraint in this area of expertise.

African language linguistics experts: these experts are involved in the test development phase. They ensure that the reading assessment is fair for all participants (taking into account dialect). They also will check the translation quality.

African language writers: if the research team decides to use original texts in each language, African language writers will be needed.

EGRA test development experts: a team with experience in EGRA development will be required to convert the early grade reading, translation, and African language experts' ideas into an assessment. These experts may be experts in other areas too.

Quantitative data analysis experts: these experts are responsible for guiding the research so that the data will be reliable and relevant to develop reading benchmarks. They will assist in providing feedback to the test development team after the pilot studies so that the assessments can be revised. They will also conduct the final analysis to determine the benchmarks.

Early Grade Reading data collection experts: these experts have proven experience in collecting data from children using Tangerine® or other data collection software, can organize the logistics of data collection and have several quality assurance measures in place. This group of experts will guide the data collection process to ensure the data is collected according to the research team's specifications. This group also has access to experienced fieldworkers.

Master teachers: these experts would evaluate the proposed benchmarks for their utility in the classroom if this were one of the aims of creating reading benchmarks.

Benchmarking experts: it is essential to tap into national, regional and international expertise in setting reading benchmarks across countries and languages. International experts have the experience and skills necessary to develop meaningful benchmarks. For example, one international organisation that has done extensive benchmarking of the EGRA is RTI International. From 2014 to 2016, for example, RTI benchmarked multiple languages across 12 countries in the USAID Education Data for Decision Making (EdData II) project. RTI International is by no means the only international expert working in this area. However, they have good experience across a range of contexts. Currently there are few national experts. There is no repository known to the authors that houses comprehensive information on international organisations working in this area.

7.6 STAKEHOLDER ENGAGEMENT

Partnerships to develop benchmarking in South African languages is critical to prevent duplication, share learning from the various benchmarking approaches and methods undertaken, and build our expertise across the languages.

It is essential to base an approach to stakeholder engagement on the intended use of the benchmarking data. The primary purpose of engaging stakeholders should be to increase understanding of the concepts of benchmarks and how they should be used. Advocacy will be valuable in terms of how to 'sell' the idea of benchmarks, and to avoid misunderstanding of its purpose and usefulness.

Although the roles of key stakeholders will differ depending on the benchmarking approaches, engagement should include, amongst others, stakeholders listed in Table 14.

Table 14: Stakeholders in Benchmarking

Communities of Practice • BRIDGE – Early Grade Reading Community of Practice • Bua-lit
DBE directorates – Curriculum, Teacher Development, Research, Assessment
Donors working in this area, such as USAID, United Nations Children's Fund (UNICEF), Zenex Foundation and others
Literacy Association of South Africa (LITASA)
Non-governmental organisations working in this space, for example, Room to Read, Nal'ibali, Read Educational Trust, Wordworks, and others
Pan South African Language Board (PanSALB)
Parents and communities
Professional development bodies (for example, the South African Council for Educators)
Teacher Unions
Teachers and Principals
Universities involved in reading research, African languages, and/or linguistics, including but not limited to: • North-West University (Bachelor of Arts in Language Technology, English, Linguistics) • Rhodes University (Department of English Language and Linguistics) • Sol Plaatje University

- University of Cape Town (Department of Linguistics and Southern African Languages)
- University of Johannesburg (Department of Childhood Education)
- University of KwaZulu-Natal (Linguistics Programme)
- University of Pretoria (Department of African Languages)
- University of South Africa (Department of Linguistics and Modern Languages, Department of Early Childhood Development)
- University of Stellenbosch (Department of African Languages, Department of Afrikaans and Dutch, Department of General Linguistics)
- University of the Free State (Department of African Languages, Department of Afrikaans, Dutch and Modern European Languages)
- University of the Witwatersrand (Department of Linguistics)
- Vista University

PanSALB, in particular, has a constitutional mandate to promote the development and use of the 11 official languages of South Africa, as well as the Khoe and San languages, and South African sign language. The board was established by an Act in the Constitution (Act 106 of 1996), which highlights the role of the board in initiating studies and research in this area. The Act also stipulates that the board is responsible for promoting the use of, and respect for, other common languages spoken in South Africa.

According to its latest Annual Report (2017-2018), PanSALB collaborated with certain South African universities to classify their research, showcase the languages targeted in research reports, and identify gaps in language research. In its next annual cycle (2018 – 2019), PanSALB intends to collaborate with the National Research Foundation (NRF) and Sol Plaatje University on language research.

It is recommended that the Language Executive Office of PanSALB, Language Research Division, be engaged as part of the stakeholder group informing reading benchmarking in South Africa.

7.7 COST ELEMENTS

The authors could find little publicly available budget information for the development of reading benchmarks in South Africa and other countries. The PIRLS, which utilizes similar training and data collection processes to the EGRA (in terms of fieldworker recruitment, training, data collection and so on) cost approximately South African Rand (ZAR) 20 million in 2011 and was estimated to cost approximately ZAR 45 million in 2016.[140] This is for a nationally representative sample of 350 schools, and two intact classrooms per school. The location of schools within provinces is an important factor concerning cost.

Creating benchmarks from scratch would be dependent on the goals of benchmarking and several other factors. Assuming one can assess 15 learners per school, a sample size of 1 000 would require visits to 67 schools. This relates to one language and one grade. A further assumption is made that a team of two fieldworkers are able, between them, to assess these 15 children in a single school day. Assuming that one can field 15 teams (30 fieldworkers) at one time (dependent on the number and availability of fieldworkers with the language skill required), and that data collection can take place back-to-back (with no breaks), then theoretically, one could collect this data in one week.

[140] Interview with PIRLS implementers in South Africa

The length of time to complete each method would have to be calculated based on:

(1) The availability of instruments for the selected language/grade

(2) The number of languages to be tested

(3) The need for additional instruments

Developing the tools, piloting, re-piloting, and so on, is what would probably take the most time, and could be several months, particularly if there is no existing corpus, assessment instruments, or data. Linguists would have first to work out whether assessing ORF using an EGRA-type assessment were feasible or fair, given the structure of the language being evaluated for benchmarking purposes.

As it is impossible to calculate the costs of creating reading benchmarks upfront, below is a table with generic cost elements for the illustrative benchmarking approaches and methods presented above.

Table 15: Generic Cost Elements for Benchmarking Methods

Cost Element	Considerations	Method 1 (existing data analysis)	Method 2 (top up data collection)	Method 3 (benchmarking from scratch)
Relevant experts in early grade reading, linguistics, African languages, and so on (Refer to Section 2.6.2.3)	Expert consulting time (Level of Effort (LOE))[141] would be required from a variety of experts for several purposes. At a minimum, LOE would need to be provided for a language-specific linguistic specialist (for example, a Tshivenda linguist specialist when benchmarking in Tshivenda). ○ These specialist would be responsible for the development, translation, and revision of texts, or for critically assessing tests developed previously. ○ Where tools exist, these experts would need to conduct quantitative and/or qualitative analysis to determine whether the tools used were valid, reliable, and fair. Should existing tools be deemed insufficient, these specialists would be responsible for tool revision or redevelopment ○ Where tools do not exist, these specialists would be responsible for: ▪ Tool design (for example, in Tangerine®, Open Data Kit (ODK), on paper) ▪ Item construction ▪ Tool validation	X	X	X
	Where primary data is being collected, fieldworkers, or researchers fluent in the language being benchmarked. Note that there may not be any specialists or experienced fieldworkers readily available in a particular language and, if this were the case, it would be necessary to identify or build skills first.		X	X
Stakeholder engagement	A benchmarking team would need to collaborate closely with implementers and funders. This could include: – Workshops (venue hire, catering, transportation, LOE) – Meetings (transportation, LOE) – Telephone calls (cost of calls, LOE)	X	X	X
Tool development	If there are no reliable, valid, and fair tools available, this could take substantial time and cost to develop.	Potentially	Potentially	X

[141] Level of effort is dependent on complexity of language, availability of existing or new data requirements, scope of the study amongst other elements.

Cost Element	Considerations	Method 1 (existing data analysis)	Method 2 (top up data collection)	Method 3 (benchmarking from scratch)
Tool translation	LOE may potentially be required for: – Translation – Tool back-translation – Translation of fieldworker instructions and manuals (if relevant)		Potentially	X
Tool pilot testing	This would be relevant in the case that existing tools are insufficient and new tools are added to future rounds of data collection. This cost element can be broken down further into: – Fieldworker recruitment – Fieldworker screening – Fieldworker training (a minimum of one week) – Fieldworker testing – Transportation (training) – Accommodation (training) – Meals and incidentals (training) – Printing /lamination /binding (training) – Stationery (training) – Time for item analysis and reliability testing.		Potentially	X
Training	It is possible that USAID would share the costs of training with the existing training plans of the relevant data collection project, including the following elements: – Fieldworker recruitment – Fieldworker screening – Fieldworker training – Fieldworker testing – Transportation (training) – Accommodation (training) – Meals and incidentals (training) – Printing /lamination /binding (training) – Stationery (training)		X	X
Data collection	This would only be relevant if collecting data for benchmarking. It is possible that donors would share the costs of data collection with the existing data collection plans of the research project, including the following elements: – Fieldwork supervision		X	X

Cost Element	Considerations	Method 1 (existing data analysis)	Method 2 (top up data collection)	Method 3 (benchmarking from scratch)
	– Quality control (in the field – spot checks; statistical quality assurance checks for the first two weeks of data collection) – Transportation (fieldwork) – Accommodation (fieldwork) – Meals and incidentals (fieldwork) – Tablets (for example, for Tangerine® assessments) – Printing /lamination /binding (for materials to display, for any paper and pencil testing) – Stationery (for paper and pencil testing) – Airtime (for fieldwork teams to communicate with supervisors)			
Data cleaning	This could include (depending on the tool used) LOE for a statistician to check: – Data entry – Data cross checking – Database structuring	Potentially	X	X
Data pre-check	This would include LOE for a statistician to conduct all pre-analysis data checks	X	X	X
Data analysis	This would include LOE for a statistician to (1) clean data (2) conduct data analysis on the existing datasets. LOE is dependent on the quality of the data, the number of variables to be assessed, analysis crosschecks, and so on. This would include LOE for a statistician to conduct data analysis. LOE is dependent on the quality of the data, the number of variables to be assessed, analysis crosschecks, and so on.	X	X	X
Reporting	The same level of effort will be required for reporting across methods	X	X	X
Project coordination and management	LOE for project oversight, planning, reporting	X	X	X
Dissemination to the public and stakeholders	This could include: – Workshops (venue hire, catering, transportation, LOE) – Meetings (transportation, LOE) If dissemination is taking place in more than one location, this could also include transportation, accommodation, meals, incidentals, and additional LOE.	X	X	X

7.8 FIELDWORK STANDARDS

The EGRA toolkit[142] defines the minimum standards for fieldwork and the selection and training of fieldworkers.[143]

The EGRA experience shows that the EGRA takes approximately 15 to 20 minutes per child to administer. If there are three assessors in a school, they will be able to complete approximately nine or ten assessments in an hour or evaluate about 30 children in three hours.[144]

Emerging practice suggests the following:

- An essential criterion for recruiting and selecting a fieldworker is the candidates' fluency in reading and speaking the languages required for administering the EGRA and for training.
- Candidates must have previous experience with the administration of assessments or data collection.
- The fieldworker must have experience of, and be comfortable with, working with children in primary schools.
- Candidates must be willing, able, and available to collect data in the relevant study areas.
- Candidates have to be proficient and have experience using information and communication technology (ICT), including a computer and hand-held electronic device (for example, a tablet or smartphone).
- The selection of fieldworkers must be based on the fieldworkers' ability to:
    - Administer the EGRA efficiently and accurately. Fieldworkers must be aware of the standards for the assessment and must also be mindful that, if they do not meet them, they may not be selected for fieldwork;
    - Show skill in administering the EGRA, including the rules and procedures for administration, rules for recording responses, and use of electronic devices;
    - Listen to the child while scoring results and entering them into an electronic device, all at the same time;
    - Establish a positive rapport with the learner. Fieldworkers must be able to assess young learners in a non-threatening way. This skill can be learnt, but not necessarily mastered, by training;
    - Work well with other team members to ensure all activities are completed in a school. They must also be able to interact well with school management and teachers;
    - Be available for the entire period of data collection and be flexible enough to sometimes work under difficult conditions.

Fieldwork needs supervision from skilled supervisors who, in the EGRA experience, oversee teams of three assessors at a school. These supervisors should have the experience and skills to:

- Lead the fieldworker teams
- Conduct quality assurance, paying attention to detail
- Understand the EGRA rules of administration, supervise fieldworkers in the administration of the EGRA, and determine where fieldworkers are making mistakes

[142] RTI International (2015)

[143] The information in this section is taken from the, RTI International EGRA Toolkit (2015). For further information the document is available from: www.globalreadingnetwork.net

[144] RTI International, EGRA Toolkit (2015)

- Have the skills and knowledge to use electronic devices effectively, so that they can assist fieldworkers with any problems that arise
- Work well with children and school management and teachers.

Furthermore, the EGRA sample agenda for training (see below) includes several crucial elements that must be included in the training of fieldworkers. The training facilitators need to:

- Engage high-level officials in training to demonstrate the importance of EGRA and gain their approval and interest in the results
- Explain the structure, purpose and rationale for assessing reading and how to administer the EGRA tools
- Explicate the rationale behind monitoring the performance of fieldworkers during training, and outline the criteria against which they will be measured
- Provide an overview of the EGRA subtasks and how they should be administered
- Describe and present any other materials and instruments that must be used along with the EGRA
- Allow fieldwork trainees to rehearse in pairs and groups while providing oversight and support. Simulation in a school after a few days of training is essential
- Observe fieldworker training, assist fieldworkers who are struggling, and retrain as and when required, both in terms of the administration of the EGRA and the use of electronic devices for administration
- Assess the accuracy and reliability of fieldworker assessments using observation (and formal reliability testing, where possible). The results should be used for retraining purposes and should ultimately be used as criteria for fieldworker selection.

It is feasible to train fieldworkers for different grades together as the tasks are similar.

Below is a sample agenda based on a typical Early Grade Reading programme taken from the EGRA toolkit. A specific agenda would need to be developed based on the specifications of the benchmarking study. For example, more time may be required for fieldworkers to learn how to administer different tools to several grades *versus* one grade.

Figure 9: Sample EGRA Training Agenda

Day & Time	Day 1	Day 2	Day 3	Day 4	Day 5
9:00-9:30 a.m.	Welcome and introduction	Review of Day 1	Review of Day 2	Review of Day 3	**Visit schools to field test instruments and questionnaires**
9:30-10:30 a.m.	**Project overview** and EGRA context	**Review draft EGRA instrument** (e.g., non-words)	**Development of Listening Comprehension Passages**	**Modify/develop additional subtasks and questionnaires, as applicable**	
10:30-11:00 a.m.	***Break***				
11:00-12:30 p.m.	**Overview of EGRA:** purpose, instrument content, results use	**Development of Oral Reading Fluency Passages**	Continue listening comprehension stories and develop questions	Modify/develop additional subtasks and questionnaires, as applicable	School visit debrief
12:30-1:30 p.m.	***Lunch***				
1:30-3:00 p.m.	**Presentation on language:** orthography and issues to consider vis-à-vis EGRA development	Continue ORF stories and develop questions	**Review and Update Pupil Questionnaire**	Review and practice EGRA administration for field test	Finalization of instruments
3:00-3:45 p.m.	***Break***				
3:45-5:00 p.m.	**Review draft EGRA instrument:** (e.g., phonemic awareness and letter sounds)	Finalize stories and questions	Finalize stories, questions, pupil questionnaire as needed	Review and practice EGRA administration for field test	Workshop Closure
Daily Objectives:	***Understanding of EGRA purpose and content***	***Oral reading passages and questions developed***	***Listening comprehension passages and stories developed; Pupil Questionnaire Developed***	***Additional subtasks/questionnaires developed***	***Instruments finalized***

This sample agenda only pertains to training for the main fieldwork. Pilot testing would require separate training. However, school-based simulation should be done as part of both the pilot and main training.

Fieldworkers should receive, at a minimum, one week of training. The PIRLS fieldworker team is trained for two weeks while some countries (for example, Tanzania) provide their fieldworkers with three or more weeks of training.

The fieldwork management team needs to ensure that data quality is checked as and when data comes in from the field. There should be at least one supervisor for every team of fieldworkers.

For the PIRLS, the quality assurance process includes:

- Monitoring 10 per cent of the sample and conducting unannounced visits at sample schools during fieldwork
- Once written assessments are scored, some are marked as reliability booklets and scored again by someone with a degree in education
- At that point, the results are captured, and reliability is scored
- In addition, the PIRLS team conducts extensive data-cleaning to ensure the quality of the data.

For the SPS project, funded by USAID, the evaluation team included the following additional criteria and checks:

- They only selected 80 per cent of the fieldworker pool, and therefore deliberately over recruited by 20 per cent
- They observed fieldworkers as they interacted with learners in the field
- At the training, they conducted inter-rater reliability assessments using mock scenarios, with the whole group assessing the scenario at one time
- They also paired two fieldworkers, and both scored a single learner
- The team used statisticians to assess the quality of the data as it came in via electronic devices

The DBE has collected data in several waves in EGRS I and at least three waves in EGRS II. There are protocols, adverts, and manuals available.

CONCLUSION

This document gives a point in time consolidation and understanding of available research, approaches and methods to set reading benchmarks in South Africa. The document provides an accessible resource which:

- Presents the context of the 11 official languages in South Africa and outlines the similarities and differences of language structure. This includes understanding the implications of language structure on setting benchmarks for individual languages as well as across languages in the Southern African Bantu group.
- Explores and unpacks reading benchmarks pertaining to a range of competencies, for example ORF, sound-letter recognition, reading comprehension, complex consonant sequences and reading accuracy.
- Details practice guidelines on the ideal process to guide future setting of benchmarks (including competency areas, methods, existing datasets and tools).
- Offers three recommended strategies to develop benchmarks in South Africa based on current reading and benchmarking initiatives in South Africa. The three strategies are:
    - Strategy 1: Analyse Existing Datasets (most inexpensive, but much of the data was not collected for the purpose of benchmarking)
    - Strategy 2: Collect Prioritized Additional (Top Up) Data (enhances and fills gaps for Strategy 1)
    - Strategy 3: Collect Primary Data (most expensive and time consuming but fit for purpose)
- Identifies and discusses key elements for setting benchmarks, including costs, training of fieldworkers, working collaboratively, identifying stakeholders, and so on.
- Guides future prioritisation of setting benchmarks in the South African languages.

The research, approaches and practice guidelines are presented for decision-makers and practitioners to prioritize and advance setting of benchmarks across all South African languages. Although recommended practices are proposed for approaches and methods, the process of setting benchmarks for specific languages and grades requires answering a number of questions and following steps to identify what is being benchmarks and why.

Table 16 summarises the availability of datasets for benchmarking reading competencies. The extent to which data is available is presented in Annexure 1 Existing Datasets (2020). The table further illustrates current gaps in available datasets.

Recommendations to improve the setting of benchmarks and improving literacy in the different languages include:

- Prioritise setting benchmarks against other competencies and skills.
- Consider the extent of completeness and gaps in the relevant data for each language and across the skill sets when benchmarking.
- Rank and or select benchmarking methods based on existing knowledge and data, costs and expertise.
- Focus benchmarking on skills that improve and strengthen reading and literacy.

The aim of setting benchmarks is to increase the mastery of learners in language and reading competencies. Therefore,

- Benchmarks should be used by teachers and the education system to identify children at risk for reading failure in the early years, and guide activities at child and classroom level to address skill

failures. Teachers, therefore, need to be able to apply the benchmarks in their teaching practice, identify children at risk of reading failure, and implement specific reading and comprehension activities to improve learner performance.

- School management teams can use benchmarks to identify if targets are being achieved, provide support to teachers and learners, and build an enabling environment for mastery of language in the school.
- District departments can furthermore articulate realistic milestones to monitor appropriate reading achievement at each grade and school, and provide supportive interventions where necessary particularly to school management and teachers.
- At a national and provincial level, knowing if benchmarks are being achieved provides essential data to inform language and literacy policy, take informed decisions and make adaptive management decisions.
- Finally, benchmarks can be used for systemic evaluation purposes, as currently South Africa lacks data on which the PIRLS results can be compared or validated.

This compilation aims to increase the understanding and collaboration between policy makers, academics and practitioners. There is an ongoing need to build the community of practice pertaining to benchmarking across languages, and strengthen education research, monitoring, evaluation and learning for language and literacy. Ultimately, benchmarking must be translated into teaching practice to improve early grade reading.

This document should remain a work in progress, whereby both progress and practice is reflected upon collaboratively and incorporated into the collective knowledge base. In the ever evolving and developing field of comprehension and reading fluency, and benchmarking, the recommendations contained in this report should be contemplated and considered with any new theory in the field.

Table 16: Summary of Available Data for Benchmarking Across Languages in Grade 1-3

	Oral Comprehension			Oral Communication			Word Reading			ORF			Letter Sound Knowledge			Written Comprehension			
Grade	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	4*
Afrikaans																PIRLS			PIRLS
English[145]																PIRLS			PIRLS
English FAL	EGRS II		EGRS II	EGRS II		EGRS II	LRL	SCIP	SCIP	EGRS II	SCIP	EGRS II SCIP	LRL	SCIP	SCIP	PIRLS	SCIP	SCIP	EGRS I PIRLS
isiNdebele																PIRLS			PIRLS
isiXhosa	ZenLit	ZenLit	E-LIT ZenLit	ZenLit	ZenLit	E-LIT ZenLit	ZenLit	ZenLit	E-LIT ZenLit	ZenLit	ZenLit SPS	E-LIT SPS	ZenLit SPS	ZenLit Funda Wande	ZenLit E-LIT Funda Wanda	PIRLS	SPS	SPS	PIRLS

[145] Data available for the United States and United Kingdom

	Oral Comprehension			Oral Communication			Word Reading			ORF			Letter Sound Knowledge			Written Comprehension			
Grade	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	4*
isiZulu	ZenLit	ZenLit ESRC	EGRS II ZenLit	ZenLit	ZenLit		EGRS II ZenLit	EGRS II SCIP	SCIP	ZenLit	ESRC EGRS II SCIP	EGRS II SCIP	EGRS II ZenLit SPS	EGRS I ZenLit SCIP	EGRS I EGRS II ZenLit SCIP	PIRLS	SPS	SPS	PIRLS
Sepedi/N Sotho		ESRC								Sis	ESRC			EGRS I	EGRS I	PIRLS			PIRLS
Sesotho								SCIP	SCIP		SCIP	SCIP		SCIP	SCIP	PIRLS			PIRLS
Setswana		EGRS I	EGRS I		EGRS I	EGRS I	EGRS I	EGRS I SCIP	EGRS I SCIP	EGRS I	EGRS I SCIP	EGRS I Malda et al SCIP	EGRS I	EGRS I SCIP	EGRS I SCIP	EGRS I PIRLS			PIRLS EGRS I
Siswati			EGRS II				EGRS II	EGRS II				EGRS II	EGRS II		EGRS II	PIRLS			PIRLS
Tshivenda																PIRLS			PIRLS

	Oral Comprehension			Oral Communication			Word Reading			ORF			Letter Sound Knowledge			Written Comprehension			
Grade	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	4*
Xitsonga		ESRC									ESRC			EGRS I	EGRS I	PIRLS			PIRLS

* Grade 4 data for Written Comprehension is included as PIRLS data is available for all languages.

BIBLIOGRAPHY

Abadzi, H. (2006). Literacy acquisition: Some general findings from recent research. Adapted from: Abadzi, H. (2006). *Literacy acquisition and the biology of reading*. In: Abadzi, H. 2006. *Efficient learning for the poor: insights from the frontier of cognitive neuroscience*. Washington, DC: The World Bank, pp. 36-49.

Adams M. (1990). *Beginning to read: Thinking and learning about print*. MIT Press

Alcock, K. J., Nokes, K., Ngowi, F., MU.S.bi, C., Mbise, A., Mandali, R., Bundy & Baddeley, A. (2000). The development of reading tests for use in a regularly spelled language. *Applied Psycholinguistics*, *21*(4), 525–555. http://dx.doi.org/10.1017/S0142716400004069

Anthony, J. l., Lonigan, C. J., Driscoll, K., Phillips, B. M., & Burgess, S. R. (2003). Phonological sensitivity: A quasi-parallel progression of word structure units and cognitive operations. *Reading Research Quarterly*, *38*(4), 470-487.

Ardington, C. (2018). Reading data summary. Distributed at Quantitative Literacy Workshop, November 2018 in Kalk Bay, South Africa.

Ardington, C. (2019). Impact Evaluation of Funda Wande Coaching Intervention: Baseline findings. May 2019. UCT: SALDRU Report.

Babayağit, S., & Stainthorp, R. (2007). Preliterate phonological awareness and early literacy skills in Turkish. *Journal of Research in Reading*, *30*(4), 394-413. https://doi.org/10.1111/j.1467-9817.2007.00350.x

Baker, P. (2016) Recent change in American and British English: A corpus-driven approach. Retrieved July 24, 2019, from http://ucrel.lancs.ac.uk/crs/attachments/UCRELCRS-2016-02-04-Baker-Paul_Baker_Presentation.pdf Accessed 24 July 2019

Bartlett, L., Dowd, A. J., & Jonason, C. (2015). Problematizing early grade reading: Should the post-2015 agenda treasure what is measured? *International Journal of Educational Development*, 40, 308–314. https://doi.org/10.1016/j.ijedudev.2014.10.002

Basaran, M. (2013). Reading fluency as an indicator of reading comprehension. *Educational Sciences: Theory and Practice*, *13*(4), 2287-2290.

Bauer, L. (1983). English Word-formation. Cambridge University Press.

Betts, E. A. (1946). Foundations of reading instruction. American Book.

Cardoso, M., & Dowd, A. J. (2016). Using literacy boost to inform a global, household-based measure of Children's Reading Skills. In *UNESCO Institute of Statistics, Understanding what works in oral reading assessments* (pp. 106-117). Montreal, Canada: UNESCO.

Chang, L.-Y., Plaut, D. C., & Perfetti, C. A. (2016). Visual complexity in orthographic learning: Modelling learning across writing system variations. *Scientific Studies of Reading, 20*(1), 64–85.

Chimere-Dan, O. D., Ennachan, L. M., Mpye, P., Nkondo, G., & Chimere-Dan, G. C. (2015). Mindset Teach Ukusiza in South Africa, 2013-2015: External Evaluation Report. Africa Strategic Research Corporation. Johannesburg, South Africa.

Clark, A., Naidoo, K., & Lilenstein, A. (2019), Adapting a screening tool for dyslexia in isiXhosa. *Reading & Writing, 10*(1), a235. https://doi.org/10.4102/rw.v10i1.235

Clayton, F. J., West, G., Sears, C., Hulme, C., & Lervåg, A. (2019). A longitudinal study of early reading development: Letter-sound knowledge, phoneme awareness and RAM, but not letter-sound integration, predict variations in reading development. *Scientific Studies of Reading, 24*(2), 91–107. DOI:10.1080/10888438.2019.1622546

Conrad, N. J. (2016). Does the brain read Chinese or Spanish the same way it reads English? *Frontiers for Young Minds*. *4*(26) doi: 10.3389/frym.2016.00026

Department of Basic Education. (2011). National *Curriculum Statement (NCS) Curriculum and Assessment Policy Statement Foundation Phase Grades 1-3. English First Additional Language.* Pretoria: Government Printing Works.

Department of Basic Education. (2011b). National *Curriculum Statement (NCS) Curriculum and Assessment Policy Statement Intermediate Phase Grades 4-6. English First Additional Language.* Pretoria: Government Printing Works.

Department of Basic Education. (2011c). National *Curriculum Statement (NCS) Curriculum and Assessment Policy Statement Foundation Phase Grades 1-3. Home Language.* Pretoria: Government Printing Works.

Department of Basic Education. (1997). Language in education policy. Announcement by the Minister of Education. https://www.education.gov.za/Portals/0/Documents/Policies/GET/LanguageEducationPolicy1997.pdf?ver=2007-08-22-083918-000

Dodd, B., & Carr, A. (2003). Young children's letter-sound knowledge. *Language, Speech and Hearing Services in Schools, 34*(2), 128 – 137.

Du Plessis, E. (2013). Introduction to CAPS. Curriculum & Instructional Studies, UNISA. Retrieved September 11, 2019, from http://www.unisa.ac.za/contents/colleges/col_education/docs/CAPS%20INTRODUCTION%20TO%20CAPS%202013.pdf Accessed 11 September 2019

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development,* 40, 315-322.

Ehri, L. C. (2005). *Development of sight word reading: Phases and findings.* In M. J. Snowling, & C. Hulme (Eds.), *Blackwell handbooks of developmental psychology. The science of reading: A handbook* (p. 135–154). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch8

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., Daal, V. H. P., & Polyzoe, N. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly, 39*(4), 438–468.

Engelhard, G. E. (2001). Historical view of the influences of measurement and reading theories on the assessment of reading. *Journal of Applied Measurement*, *2*(1), 1 – 26.

Ferdous, A. (2019). *Setting multiple performance standards for a timed reading fluency and comprehension assessment*. Draft Doctoral study: PhD Management Systems International (unpublished)

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239–256.

Graham, B., & van Ginkel, A. (2014). Assessing early grade reading: The value and limits of 'words per minute', *Language Culture and Curriculum,* 27(3), 244 – 259, https://doi.org/10.1080/07908318.2014.946043

Great Schools Partnership. (2014). The glossary of education reform: Criterion-Referenced Test. Retrieved September 11, 2019, from https://www.edglossary.org/criterion-referenced-test/

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: a valuable assessment tool for reading teachers, *The Reading Teacher*, *59*(7), 636-644.

Hedgcock, J. S., & Ferris, D. R. (2009). Teaching readers of English. Routledge.

Horowitz-Kraus, T. (2016). The role of executive functions in the reading process. In A., Khateb, & I., Bar-Kochva (Eds), *Reading fluency: Current insights from neurocognitive research and intervention studies*. Springer.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal,* 2, 127-160.

International Study Centre TIMSS & PIRLS. Lynch School of Education, Boston College. https://timssandpirls.bc.edu/about.html

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, 59, 421–432

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95(4), 719–729.

Jiménez, J. E., Gove, A., Crouch, L., & Rodríguez, C. (2014). Internal structure and standardized scores of the Spanish adaptation of the EGRA (Early Grade Reading Assessment) for early reading assessment. *Psicothema*, *26*(4), 531-537.

Johns, J., & Magliari, A. (1989). Informal reading inventories: Are the Betts criteria the best criteria? *Reading Improvement, 26*(23), 124-132.

Johnson, E. R. (2012). *Academic language & academic vocabulary: A K-12 Guide to content learning and response to intervention (RTI).* Achievement for All Publishers.

Jukes, M., Cummiskey, C., Gargano, M., & Dubeck, P. (2018). Guidance note: Setting data-driven oral reading fluency benchmarks. RTI International, Research Triangle Office. Accessed from https://www.roomtoread.org/media/984470/room-to-read_fluency-benchmarking-guidance-note_published-may-2018.pdf

Jukes, M., Cummiskey, C., Gargano, M., & Dubeck, P. (2018b). Data-Driven methods for setting reading proficiency benchmarks. RTI International, Research Triangle Office. Accessed from https://www.roomtoread.org/media/984466/room-to-read_fluency-benchmarking-analysis_report_may-2018.pdf

Katz, J. (2020). Back to Basics. PrimTEd Literacy Working Group Seminar Materials for literacy teacher programs, 7th February 2020. Available from: file:///C:/Users/mroper/Downloads/Jenny%20Katz%20PrimTEd%20Language%20comparisons%2007-02-2020.pdf. Accessed: 9 August 2020

Kim, Y. S., & Wagner, R. K. (2015). Text (Oral) reading fluency as a construct in reading development: An investigation of its mediating role for children from grades 1 to 4. *Sci.Stud. Read. 19*(3), 224–242. https://doi.org/10.1080/10888438.2015.1007375

Kim, Y. S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102*(3), 652–667. doi:10.1037/a0019643

Kim, Y. S. (2015) Language and cognitive predictors of text comprehension: Evidence from multivariate analysis. *Child Development,* 86,128–144.

Kim, Y. S., Park, C., & Wagner, R. K. (2014). Is oral/text reading fluency a "bridge" to reading comprehension? *Reading and Writing: An Interdisciplinary Journal*, 27, 79–99.

Kim, Y. S. G., & Piper, B. (2019). Component skills of reading and their structural relations: evidence from three sub-Saharan African languages with transparent orthographies. *Journal of Research in Reading, 42*(2), 326-348.

Liu, T., Chen, W., Liu, C. H., & Fu, X. (2012). Benefits and costs of uniqueness in multiple object tracking: The role of object complexity. *Vision Research,* 66, 31–38.

Malda, M., Nel, C., & Van de Vijver, F. (2013). The road to reading for South African learners: The role of orthographic depth. *Learning and Individual Differences*. 30, 34 -45, https://doi.org/10.1016/j.lindif.2013.11.008

McCormick, S. (1995*). Instructing students who have literacy problems.* Merrill.

Menendez, A., & Ardington, C. (2018). Impact Evaluation of USAID/South Africa Story Powered School Program – Baseline. Available at https://nalibali.org/sites/default/files/media/rsa_sps_baseline_report_final_public.pdf

Morris, D., Bloodgood, J. W., Penney, J., Frye, E. M., Kuban, L., Trathen, W., & Schlagal, R. (2011). Validating craft knowledge: An empirical examination of elementary- grade students' performance on an informal reading assessment. *The Elementary School Journal, 112*(2), 205 – 233.

Mullis, I. V. S., & Martin, M. O. (2015). PIRLS 2016 Assessment Framework (2nd ed.). TIMMS & PIRLS International Study Center and IEA, United States.

Murray, A. (2013). First Interim Impact Assessment Community Based Model for Learner and Teacher English Literacy Development, kaMhinga Villages, Limpopo Province, South Africa.

Nakamura, P., & Hoop, T. (2014). Facilitating Reading Acquisition in Multilingual Environments in India (FRAME-India).

Ogden, R. (2011). An introduction to English phonetics. *Phonetica*. 68. 111-2. 10.1159/000328775.

Pae, H., & Sevcik, R. (2011). The role of verbal working memory in second language reading fluency and comprehension: A comparison of English and Korean. *International Electronic Journal of Elementary Education*. 4, 47-66.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices – Past, present and future. In S. G Paris, & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13 – 69). Lawrence Erlbaum Associates, Inc.

Petscher, Y., & Kim, Y. S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology,* 49, 107-29.

Piasta, S. B., Farley, K. S., Phillips, B. M., Anthony, J. L., & Bowles, R. P. (2018). Assessment of young children's letter-sound knowledge: Initial validity evidence for letter-sound short forms. *Assessment for Effective Intervention, 43*(4), 249-255.

Piper, B. (2009). Integrated Education Program Impact Study of SMRS Using Early Grade Reading Assessment in Three Provinces in South Africa. RTI International: Research Triangle Park.

Piper, B., Destefano, J., Kinyanjui, E., & Ong'ele, S. (2018). Scaling up successfully: Lessons from Kenya's Tusome national literacy program. *Journal of Educational Change.* 19, 293 – 321.

Piper, B., Schroeder, L., & Trudell, B. (2015). Oral reading fluency and comprehension in Kenya: Reading acquisition in a multilingual environment: Fluency and Comprehension in Kenya. *Journal of Research in Reading, 39*(2), 133–152

Piper, B., Zuilkowskib, S. S., Kwayumbac, D., & Oyanga, A. (2018). Examining the secondary effects of mother-tongue literacy instruction in Kenya: Impacts on student learning in English, Kiswahili, and mathematics. *International Journal of Educational Development, 59,* 110 – 127

Polselli, S, A., & Snow, C. E. (2003). *Rethinking reading comprehension: Solving problems in the teaching of literacy*. Guilford Press

Pretorius, E. J., & Spaull, N. (2016). Exploring relationships between oral reading fluency and reading comprehension amongst English second language readers in South Africa. *Reading and Writing: An Interdisciplinary Journal, 29*(7), 1449–1471

Pretorius, E. J. & Lephalala, M. (2011). Reading comprehension in high poverty schools: How should it be taught and how well does it work? *Per Linguam, 27*(2), 1-24.

Pretorius, E. J. (2019). Small and big problem spaces in reading. ReSep seminar on reading research, Stellenbosch University, 13 September 2019.

Prinsloo, D. J., & de Schryver, G. M. (2002). Towards an 11 x 11 array for the degree of conjunctivism/disjunctivism of the South African languages. *Nordic Journal of African Studies, 11*(2), 249-265.

Probert, T. (2016). A comparative study of syllables and morphemes as literacy processing units in word recognition: IsiXhosa and Setswana. Unpublished MA dissertation. Grahamstown: Rhodes University

Programme d'anlsye des systemes educatifs de la confemen. http://www.pasec.confemen.org/

Randera, A., & Rees, S. (2019). Specialized children's literature corpus: isiXhosa. Molteno Institute for Language and Literacy: Johannesburg.

Rees, S. (2016). Morphological awareness in readers of isiXhosa. Unpublished MA dissertation, Rhodes, University, Grahamstown;

Ricker, K. L. (2006). Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *Alberta Journal of Educational Research*, *52*(1), 53–64.

Roehrig, A., Bohn-Gettler, C., Turner, J., & Pressley, M. (2008). Mentoring beginning primary teachers for exemplary teaching practices. *Teaching and Teacher Education, 24*(3). 684-702. https://doi.org/10.1016/j.tate.2007.02.008

RTI International. (2015). Early Grade Reading Assessment (EGRA) Toolkit, Second Edition. Washington, DC: United States Agency for International Development.

RTI International. (2016). Measurement and Research Support to Education Strategy Goal 1: Development and Pilot Testing of Additional Subtasks for the Early Grade Reading Assessment: EGRA 2.0. RTI International: Research Triangle Park.

RTI International. (2017). EGRA benchmarks and standards research report. A report for All Children Reading – Asia (ACR – Asia). Washington, DC: United States Agency for International Development.

RTI International. 2016. Early Grade Reading Assessment (EGRA) Toolkit. Second Edition. Washington, DC: United States Agency for International Development. https://shared.rti.org/content/early-grade-reading-assessment-egra-toolkit-second-edition

Sarroub, L. K., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *Faculty Publications: Department of Teaching, Learning and Teacher Education*. 39. http://digitalcommons.unl.edu/teachlearnfacpub/39

Schaefer, M., & Kotzé, J., (2019), 'Early reading skills related to Grade 1 English Second Language literacy in rural South African schools', *South African Journal of Childhood Education, 9*(1), a644. https://doi.org/10.4102/sajce.v9i1.644.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology,* 94(2), 143–174

Shen, H. H., & Jiang, X. (2013). Character reading fluency, word segmentation accuracy, and reading comprehension in L2 Chinese. *Reading in a Foreign Language*, 25(1), 1-25.

SIL International (2019). Accessed 24 July (2019) from https://glossary.sil.org/

Smith S. B., Simmons, D. C., & Kame'enui, E. J. (1998). Phonological awareness: Instructional and curricular basics and implications. In D. C. Simmons & E. J. Kame'enui (Eds.). *What reading research tells us about children with diverse learning needs: Bases and basics.* Lawrence Erlbaum Associates.

Snelling, M., Dawes, A., Biersteker, L., Girdwood, E., & Tredoux, C. (2019). The development of a South African early learning outcomes measure: A South African instrument for measuring early learning programme outcomes. *Child: Care, Health and Development, 45*(2), 257-270.

Snow, C. E., & Sweet, A. P. (2003). Reading for Comprehension. In A. P. Sweet, & C. E. Snow (Eds.), *Rethinking Reading Comprehension*. The Guilford Press.

South African Language in Education Policy. (1997). Available at: https://www.education.gov.za/Portals/0/Documents/Policies/GET/LanguageEducationPolicy1997.pdf?ver=2007-08-22-083918-000

Spaull, N., Pretorius, E. J., & Mohohlwane, N. (2020). Investigating the comprehension iceberg: Developing empirical benchmarks for early grade reading in agglutinating African languages. *South African Journal for Early Childhood Education, 10*(1), a773. https://doi.org/ 10.4102/sajce.v10i1.773

Stahl, K., & Dougherty, A. (2011). Applying new visions of reading development in today's classrooms. *The Reading Teacher, 65*(1), 52-56.

Stanovich, K.E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: The Guilford Press.

Statistics South Africa. 2018. Community survey. http://cs2016.statssa.gov.za/

Statistics South Africa. 2018. General Household Survey. Accessed 19 August 2020 from http://www.statssa.gov.za/publications/P0318/P03182018.pdf

Tavakol, M., & Dennick, R. (2011). Post-examination analysis of objective tests. *Med Teach*. 33, 447-58.

Taylor, R., Oberle, E., Durlak, J., & Weissberg, R. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development,* 88. 1156-1171.

Taylor, S., Cilliers, J., Prinsloo, C., Fleisch, B., & Reddy, V. (2018). The Early Grade Reading Study: Impact evaluation after two years of interventions. Available from: https://www.education.gov.za/Portals/0/Documents/Reports/EGRS/EGRS%20I%20Wave%203%20Report%202018.pdf?ver=2019-05-31-111222-217

Taylor, S., Kotze, J., Wills, G., Burger, C., & Cilliers, J. (2019). The Early Grade Reading Study sustainability evaluation. Available from: https://www.education.gov.za/Portals/0/Documents/Reports/EGRS/EGRS%20I%20Wave%204%20Report%202019.pdf?ver=2019-05-31-111638-587

The Southern and Eastern Africa Consortium for Monitoring Educational Quality. http://www.sacmeq.org/

Thomas, S., & Peng, W. J. (2004). The use of educational benchmarks in indicator publications. In Scheens J. & Hendricks, M. (eds.) *Benchmarking the quality of education,* Faculty of Behavioural Sciences. University of Twente: Enschede.

UNESCO. (2017). More than one-half of children and adolescents are not learning worldwide. Fact Sheet No. 46, UIS/FS/2017/ED/46. Montreal, QC: UNESCO Institute for Statistics. ttp://uis.unesco.org/sites/default/files/documents/fs46-more-than-half-childrennot-learning-en-2017.pdf

Van der Linden, W. J., & Hambleton, R. K. (1998). Item response theory: Brief history, Common Models and Extensions. In Van der Linden, WJ and Hambleton, RK. (Eds.). *Handbook of Modern Item Response Theory*. Springer

Vula Bula. Stone Soup. A5 Reader, Setwana. Available from: https://vulabula.molteno.co.za/system/tdf/newreaders/Molteno%20Grade%201-3%20Combined%20Web%20pdfs/Grade%201-2%20Combined%20Web%20pdfs/Grade1-2Setswana/5star%20Setswana%20Reader%20Sopo%20ya%20leje%20%28LR%29.pdf?file=1&type=commerce_product&id=1230. Accessed: 19 August 2020

Wang, Z., Sabatini, J., O'Reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: A Test of the Decoding Threshold Hypothesis. *Journal of Educational Psychology*, *111*(3), 387-401.
Wilsenach, C. (2019), 'Phonological awareness and reading in Northern Sotho – Understanding the contribution of phonemes and syllables in Grade 3 reading attainment', *South African Journal of Childhood Education, 9*(1), 1-10. https://doi.org/10.4102/sajce.v9i1.647

Wissing, D. (2018). "Overview of Afrikaans Vowels." Accessed 24 July 2019 from http://www.taalportaal.org/taalportaal/topic/pid/topic-14613241254929634

Wolf, S., Turner, E, I., Jukes, M. C. H., & Dubeck, M. M. (2017). Changing literacy instruction in Kenyan classrooms: Assessing pathways of influence to improved early literacy outcomes in the HALI intervention. *International Journal of Educational Development,* 62, 27-34

Wood, D. E. (2006). Modelling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11*(2), 85-104.

Ziegler, J., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131, 3-29.

Ziegler, J. C., & Goswami, U. (2006). Becoming literate in different languages: Similar problems, different solutions. *Developmental Science, 9*(5), 429-36.

Zieky, M., & Perie, M. (2006). A primer on setting cut scores on tests of educational achievement. Princeton, NJ: Educational Testing Service.

Websites referenced:

https://timssandpirls.bc.edu/index.html

http://www.pasec.confemen.org/

http://www.sacmeq.org/

Personal communication:

Siân Rees at the Molteno Institute for Language and Literacy who is currently researching frequencies of graphemes (letters) in the different African languages.

ANNEXURE 1: EXISTING DATASETS (2020)

WAVES																						
Language	Start year	Schools	Learners	Nature	start gr 1	mid gr 1	end gr 1	start gr2	end gr 2	start gr 3	Written Comp	Written Comp	Written Comp	Written Comp	Written Comp	Grade (baseline)	Letter sounds[146]	word reading	ORF	Oral Comp	Written Comp	Study	Holder of data	Available
All	2006	429	16073	NON-EXP							1	1	1	1	1	4					1	PIRLS	IEA	public
All	2011	341	15744	NON-EXP							1	1	1	1	1	4					1	prePIRLS	IEA	public
All	2016	293	12810	NON-EXP							1	1	1	1	1	4					1	PIRLS Literacy	IEA	public
isiZulu, English, Afrikaans	2016	125	5282	NON-EXP							1	1	1	1	1	5					1	PIRLS	IEA	public
Setswana	2015	230	4538	EXP	1		2		3		4	4	4	4	4	1	all	all	2,3,4	2,3,4	4	EGRS I	DBE	on request
isiZulu	2017	22	969	EXP	1		2		3							1	1,3	1,2	3	3		EGRS II	DBE	on request
Siswati	2017	58	2358	EXP	1		2		3							1	1,3	1,2	3	3		EGRS II	DBE	on request
isiXhosa	2016	50	940	QUASI-EXP			3									R	3	3	3	3		E-LIT	WCED/USAID	Public

[146] Numbers in this section refer to the waves in which the data was collected.

					WAVES																		
Language	Start year	Schools	Learners	Nature	start gr 1	mid gr 1	end gr 1	start gr2	end gr 2	start gr 3	Written Comp	Written Comp	Written Comp	Written Comp	Written Comp	Grade (baseline)	Letter sounds[146]	word reading	ORF	Oral Comp	Written Comp	Study	Holder of data	Available
isiZulu	2017		?	QUASI-EXP			1									1	1	1	1	1		ZenLit	Zenex Foundation	On request
isiZulu	2017		?	QUASI-EXP					1							2	1	1	1	1		ZenLit	Zenex Foundation	On request
isiZulu	2017		?	QUASI-EXP						end of year						3	1	1	1	1		ZenLit	Zenex Foundation	On request
isiXhosa	2017		?	QUASI-EXP			1									1	1	1	1	1		ZenLit	Zenex Foundation	On request
isiXhosa	2017		?	QUASI-EXP					1							2	1	1	1	1		ZenLit	Zenex Foundation	On request
isiXhosa	2017		?	QUASI-EXP												3	1	1	1	1		ZenLit	Zenex Foundation	On request
Sepedi	2017	8	135	NON-EXP						1						3	1,2	1,2	1,2	1,2		ESRC	DFID/ESRC	?
Xitsonga	2017	10	111	NON-EXP						1						3	1,2	1,2	1,2	1,2		ESRC	DFID/ESRC	?

					WAVES																		
Language	Start year	Schools	Learners	Nature	start gr 1	mid gr 1	end gr 1	start gr2	end gr 2	start gr 3	Written Comp	Written Comp	Written Comp	Written Comp	Written Comp	Grade (baseline)	Letter sounds[146]	word reading	ORF	Oral Comp	Written Comp	Study	Holder of data	Available
isiZulu	2017	42	509	NON-EXP						1						3	1,2	1,2	1,2	1,2		ESRC	DFID/ESRC	?
Sepedi	2017	8	93	NON-EXP												6			2	2		ESRC	DFID/ESRC	?
Xitsonga	2017	10	110	NON-EXP												6			2	2		ESRC	DFID/ESRC	?
isiZulu	2017	42	386	NON-EXP												6			2	2		ESRC	DFID/ESRC	?
isiXhosa	Current		1180	BASELINE												1,2						Funda Wande	DFID/ESRC	?
isiXhosa	2017	172	1780	EXP				1	2							2	1,2	1,2	1,2	1,2		SPS	USAID	avail 2020
isiZulu	2017	188	1676	EXP				1	2							2	1,2	1,2	1,2	1,2		SPS	USAID	avail 2020
isiXhosa	2017	172	1784	EXP						1						3	1	1,2	1,2,3,4	1,2		SPS	USAID	avail 2020
isiZulu	2017	188	1671	EXP						1						3	1	1,2	1,2,3,4	1,2		SPS	USAID	avail 2020
isiXhosa	2017	172	1685	EXP							2	2	2	2	2	4	1	1,2	1,2,3,4	1,2	2, 3	SPS	USAID	avail 2020
isiZulu	2017	188	1808	EXP							2	2	2	2	2	4	1	1,2	1,2,3,4	1,2	2, 3	SPS	USAID	avail 2020
Setswana	2017	40	880	QUASI-EXP	1				2							1	1,2	1,2	1,2	1,2		NECT	NECT/Zenex Foundation	On request

					WAVES																		
Language	**Start year**	**Schools**	**Learners**	**Nature**	**start gr 1**	**mid gr 1**	**end gr 1**	**start gr2**	**end gr 2**	**start gr 3**	**Written Comp**	**Written Comp**	**Written Comp**	**Written Comp**	**Written Comp**	**Grade (baseline)**	**Letter sounds[146]**	**word reading**	**ORF**	**Oral Comp**	**Written Comp**	**Study**	**Holder of data**	**Available**
isiXhosa	2017	40	893	QUASI-EXP	1				2							1	1,2	1,2	1,2	1,2		NECT	NECT/Zenex Foundation	On request
isiZulu	2017	70	1616	QUASI-EXP	1				2							1	1,2	1,2	1,2	1,2		NECT	NECT/Zenex Foundation	On request
Sepedi	2009	15	210	EXP	1	2										1	1,2	1,2	1,2	1,2		SMRS	USAID?	USAID/ DBE
isiZulu	2009	14	190	EXP	1	2										1	1,2	1,2	1,2	1,2		SMRS	USAID	USAID/ DBE
Setswana	2009	15	250	EXP	1	2										1	1,2	1,2	1,2	1,2		SMRS	USAID	USAID/ DBE

ANNEXURE 2: STAKEHOLDER CONSULTATION

Khulisa began a process of stakeholder consultation in 2019 to address some of the questions that this report sets out to answer. The team held Key Informant Interviews (KIIs) and discussions (see list of stakeholders interviewed below) with USAID (two), the DBE research team (three respondents), the University of Pretoria PIRLS team (one), Room to Read (two), Research on Socio-economic Policy at the University of Stellenbosch (ReSEP) (one), Funda Wande (two), UCT (one), SEED Educational Trust (two) and the Catholic Institute for Education (CIE) (one). The team consulted with stakeholders on what they are doing in the reading and reading benchmarking space in South Africa and their interests. The purpose of these interviews was to position new initiatives in the context of other interventions and look for potential areas of collaboration. The team also consulted with African language and benchmarking experts to determine methods for establishing reading benchmarks, potential target population(s), and tools available with rationale for use or not, sampling and analysis methods. This feedback is incorporated into this report.

READING BENCHMARKS WORKSHOP

On 9 May 2019, the DBE hosted a workshop on the development of reading benchmarks in South African languages. Dr Matthew Jukes from RTI international co-facilitated the workshop with the DBE. The workshop was attended by (see list of participants below) Khulisa staff (four), RTI (one), the DBE research directorate (five), the DBE languages curriculum team (two), the DBE teacher development team (two), Room to Read (one), BRIDGE (one), USAID (four), Funda Wande (two), the UCT (one), the University of Pretoria (UP) (one), the Zenex Foundation (two), ReSEP (two), and a linguistics expert affiliated with the University of Johannesburg (UJ).

The objectives of the workshop were to:

- Outline the current situation of reading in South Africa
- Summarise the work that the DBE Curriculum branch is doing in the development of reading norms
- Describe the science behind reading norms and benchmarks, the difference between these concepts, and related statistical data collection processes and methods
- Determine the need, purpose, and use of reading norms and benchmarks for different stakeholders
- Outline the reading data that is currently available and discuss the implications for future work in this area

There were presentations and question and answer sessions in the workshop. Part of the workshop involved group work, whereby the participants were split into three groups which were tasked with addressing one of three questions each:

1. What is the purpose or goal of reading benchmarks in the South African context?
2. How, and by whom, will reading benchmarks be used in South Africa?
3. What are the language-specific issues that need to be taken into consideration in developing reading benchmarks for South African languages?

In terms of the purpose and goals of reading benchmarks in South Africa, the first group felt that there may be different purposes and uses based on what is feasible in the next five years. It was clear from the discussion that reading benchmarks should assist in understanding the development of language and components across different languages, in diagnosing the status of reading development and in monitoring the progress of language development in the country. Stakeholders felt it would be essential to document an approach to developing reading benchmarks that could be used by different parties for benchmarking languages systematically and comparably.

In terms of the use of reading benchmarks in South Africa for different stakeholder groups, the second group highlighted five potential applications for different situations. This includes:

- Teachers using benchmarks as a diagnostic tool to understand what is happening in the classroom and to inform their teaching practice;
- Parents using benchmarks to understand their child and the school's performance and progress
- DBE, funders, and partners monitor national performance and progress against benchmarks
- District officials monitor reading in the classroom; and
- Resource developers ensure that resources are relevant and useful for different languages.

Stakeholders felt that a first step would be to develop a corpus of the 500 most common words per language and build reading assessment passages around that.

Regarding the language-specific issues that need to be considered, the third group felt that decoding should receive attention; that different dialects within a language should be considered (for example, Setswana and Sotho have significant differences in dialect, depending on the area in which the language is spoken). The fact that judging words read correctly is different for subjunctive *versus* conjunctive languages was highlighted as an issue for consideration, as well as the need to assess the reliability of the texts used in reading assessments. The final issue raised was that, if teachers were to use benchmarks, they would require support to understand how benchmarks work and which ones are important for determining whether a child is reading for meaning. Room to Read noted in their work, that once teachers used ORF benchmarks, children started reading faster and with expression.

BRIDGE EARLY GRADE READING COMMUNITY OF PRACTICE

Following the workshop, the Khulisa team approached the BRIDGE Early Grade Reading Community of Practice (COP) sponsored by the Zenex Foundation. The BRIDGE COP is affiliated to the National Education Collaboration Trust (NECT). With more and more organisations coming on board to help solve the reading crisis, BRIDGE provides a forum for the dissemination of information about early grade reading and reading norms. The purpose was to share information on the proceedings of the workshop and to find out what others are doing in this space. The COP meeting allowed the team to access multiple stakeholders in the reading community at one time.

STAKEHOLDERS INTERVIEWED

The following stakeholders were interviewed during the development of this report in 2019:

STAKEHOLDER	DESCRIPTION
Centre for Evaluation and Assessment (CEA) at the University of Pretoria (UP)	The CEA is a research unit within the faculty of education at the University of Pretoria. The CEA has led the implementation of the international Progress in International Reading Literacy Study (PIRLS) in South African since 2005. (https://www.up.ac.za/centre-for-evaluation-and-assessment).
Department of Basic Education Research Directorate	The South African Department of Basic Education is a department of the South African government, which oversees primary and secondary education in South Africa (https://www.education.gov.za/).
Funda Wande	Funda Wande is a not-for-profit organisation that provides Creative Commons licensed video and print materials to train teachers how to teach reading (for meaning) in Grades 1 – 3 (https://fundawande.org/).
Research on Socio Economic Policy Unit at the University of Stellenbosch (ReSEP)	ReSEP comprises a group of researchers situated in the department of economics at the University of Stellenbosch. It focuses its research on issues of poverty, income distribution, social mobility, economic development and social policy, and is a leading institution involved in reading research in South Africa. (https://resep.sun.ac.za/).
Room to Read	Room to Read is an international non-profit organisation aimed at improving literacy and gender equality in education. It has branches multiple countries, including one in South Africa (https://www.roomtoread.org/).
SEED Educational Trust	The SEED Educational Trust is a charitable Trust that has been working in schools and Districts since 2006 and has run leadership programmes for more than 800 leaders from eight districts across five provinces (http://seedtrust.org.za/).
South African Labour and Development Research Unit (SALDRU), University of Cape Town (UCT)	SALDRU delivers applied empirical research and capacity building. It is situated within the school of economics at the University of Cape Town. It focuses its research on issues of poverty and inequality, labour markets, human capital and social policy. (https://www.saldru.uct.ac.za/).

Stakeholders Engaged in Workshops and BRIDGE COP

Stakeholder	Description
BRIDGE Communities of Practice	A South African non-profit organisation that convenes communities of practice among stakeholders in the education system to improve their collective impact on the education system (https://www.bridge.org.za/).
Department of Basic Education Languages Curriculum Team	The South African Department of Basic Education is a department of the South African government, which oversees primary and secondary education in South Africa (https://www.education.gov.za/).
Department of Basic Education Teacher Development Team	
Gauteng Department of Education (GDE)	The Gauteng Department of Education is a department of the South African government, which oversees primary and secondary education in Gauteng province (https://www.gauteng.gov.za/Departments/DepartmentDetails?departmentId=CPM-001004).
PRAESA	The Project for the Study of Alternative Education in South Africa (PRAESA) is a multilingual, early literacy research and development organisation, affiliated with the University of Cape Town (http://www.praesa.org.za/).
University of Johannesburg (UJ) Department of Linguistics	The department of linguistics is situated within the faculty of humanities at the University of Johannesburg (https://www.uj.ac.za/faculties/humanities/department-of-linguistics).
University of South Africa (UNISA) Department of Linguistics and Modern Languages	The department of linguistics and modern languages is situated within the University of South Africa (https://www.unisa.ac.za/sites/corporate/default/Colleges/Human-Sciences/Schools,-departments,-centres,-institutes-&-units/School-of-Arts/Department-of-Linguistics-and-Modern-Languages)
Wordworks	A South African non-profit organisation focusing on early language and literacy development in the first eight years of life (http://www.wordworks.org.za/#).
Zenex Foundation	Established in 1995, the Zenex Foundation is an independent trust that funds research, programmes and projects in Mathematics, Science and Language education in South Africa (https://www.zenexfoundation.org.za/about).

Contract Number: 72067418D00001, Order Number: 72067419F00007

Published by the Department of Basic Education

Sol Plaatje Building, 222 Struben Street

Private Bag X895, Pretoria, 0001

Telephone: 012 357 3000 Fax: 012 323 0389 Hotline: 0800 202 933

ISBN Number: 978-1-4315-3411-1

Website: www.education.gov.za

Facebook: www.facebook.com/BasicEd

Twitter: www.twitter.com/dbe_sa