elom

early learning measurement tools

JUNE 2025

ELOM-R (v1)
TECHNICAL MANUAL 2
Language Assessment

1ST EDITION 2025

Developed on behalf of DataDrive2030 by Matthew Kleineibst (Psychology Department, University of Cape Town and AX consultgroup), Jürgen Becker (Industrial Psychology Department, University of the Western Cape and AX consultgroup) and Andrew Dawes (Psychology Department, University of Cape Town and DataDrive2030).

With contributions from Sonja Giese, Linda Biersteker, Elizabeth Girdwood, and Caylee Cook (all DataDrive2030).

This Technical Manual accompanies the tablet-based ELOM-R (v1) Direct Assessment Manual, as well as the ELOM-R (v1) Technical Manual 1: The Development of ELOM-R, and ELOM-R (v1) Technical Manual 3: Mathematics).

TO CITE THIS MANUAL:
Kleineibst, M., Becker, J. & Dawes, A. (2025). ELOM-R (v1) Technical Manual 2: Language Assessment. DataDrive2030, Westlake Cape Town.
For further information, please refer to **https://DataDrive2030.co.za.**

ISBN

First Edition 2025

CONTENTS

ACKNOWLEDGMENTS

Linda Biersteker Emeritus Professor Andrew Dawes Sonja Giese Elizabeth Girdwood Matthew Snelling Dr Temitope Ogunyoku	ELOM-R (v1) Development and Pilot Project Team
Professor Jürgen Becker Matthew Kleineibst Emeritus Professor Andrew Dawes	Psychometry
Emeritus Professor Elizabeth Pretorius, Department of Linguistics and Modern Languages, University of South Africa Dr Shelley O'Carroll, Director and Early Literacy Specialist, Wordworks	Specialist consultants on Foundation Phase Literacy and Language
Ms Dikeledi Mathebe and Department of Basic Education colleagues	Translation and Advice on African languages
Roots and Shoots JET Education Services Kelello Consulting	Research projects and Early Learning Programmes: For access to ELOM-R (v1) data
Genesis Analytics	Psychometric Study Field Work and Data Collection 2023
Social Surveys Africa	Pilot Study Field Work and Data Collection 2018

PROJECT FUNDERS

We have benefited over many years from the financial support of Innovation Edge, and the Maitri Trust. We also acknowledge the contribution of the Zenex Foundation, Anglo American and Mr Price Foundation in supporting research studies that contributed to this Manual.

ACRONYMS

CFA	Confirmatory Factor Analysis
1PL	One Parameter Logistic Model
CTT	Classical Test Theory
CFI	Comparative Fit Index: used to assess model fit in Confirmatory Factor Analysis
CI	Confidence Interval
DIF	Differential Item Functioning
EFA	Exploratory Factor Analysis
EF	Executive Functioning
IRT	Item Response Theory
ITC	International Test Commission
LOGIT	Log-Odds Unit. Logits are linear measures on the same equal interval scale and can be summed.
MGCFA	Multi-Group Confirmatory Factor Analysis
MHX2	Mantel-Haenszel chi-square test for DIF
PC	Percent Correct
PCM	Partial Credit Model
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RMSEA	Root Mean Square Error of Approximation: used to assess model fit in Confirmatory Factor Analysis.
TIMSS	International Mathematics and Science Study
TLI	Tucker-Lewis Index: used to assess model fit in Confirmatory Factor Analysis.

CHAPTER 1. INTRODUCTION: USING THE ELOM-R LANGUAGE (v1) ASSESSMENT

THE ELOM-R (v1) TECHNICAL MANUALS ARE IN THREE PARTS:

1. **ELOM-R (v1) Technical Manual 1: Development of the ELOM-R Language and Mathematics Assessments** *(the first phase for both tools)*
2. **ELOM-R (v1) Technical Manual 2: Language Assessment** *(this Manual)*
3. **ELOM-R (v1) Technical Manual 3: Mathematics Assessment**

All are available on the DataDrive2030 website. Prior to consulting Technical Manuals 2 and 3, we strongly recommend readers familiarise themselves with Technical Manual 1, as we do not cover the same ground in this Manual. That Manual outlines the background to the development of the ELOM-R Mathematics (v1) and ELOM-R Language (v1) measures, including translation procedures and the importance of establishing their cross-language equivalence and measurement invariance. It also summarises the ELOM-R (v1) Pilot study designed to test and adjust items prior to finalisation for the analyses.

In this chapter, we provide a brief outline of the purpose, content and use of the ELOM-R Language (v1) Assessment. Chapter 2 presents the psychometric analyses undertaken to assess scale reliability, measurement equivalence and bias in the eight languages in which the tool has been developed thus far. In Chapter 3 we present final psychometric analyses undertaken on the combined sample of all eight languages to establish the construct validity, reliability and Differential Item Functioning (Test DIF) to establish whether the ELOM-R Language (v1) Assessment shows test bias in any of the languages. Here we also report on the standardisation and norms of the measure.

WHAT THE ELOM-R LANGUAGE (v1) ASSESSMENT MEASURES:

PURPOSE

The ELOM-R Language (v1) Assessment is primarily intended for use in research studies, surveys, and evaluations of literacy and language interventions designed to enable readiness for Grade 1. It is, therefore, appropriate for the assessment and descriptions of groups of children and is not a diagnostic test of individual child school readiness. The Language assessment items (revised since the pilot described in Manual 1) are closely aligned with the skills and knowledge expected of children who have completed the Grade R curriculum. It, therefore, permits users to identify the levels of knowledge and skill at which groups of children are functioning by the end of the Grade R year. The tool may, therefore, be regarded as a summative assessment of children's literacy and language, and unless there is a good reason such as addressing a specific research question, the test should be administered close to the end of the Grade R year or early in Grade 1.

When used at a population level (e.g. a random sample of Grade R classes in an Education District) this tool enables users to a) look back at the Grade R year and make recommendations for attention to areas of weakness in children's literacy and language abilities that show up in the findings that may benefit subsequent cohorts, and b) look forward to Grade 1 by drawing attention to areas in which populations of children require particular support in the early phases of that Grade. Findings can then be used to inform strategies for enhancing preschool, Grade R and Grade 1 curricula, quality and training in the CAPS language area.

This test can, therefore, be used in population surveys to estimate the proportion of children who are on Track for Grade 1 in language knowledge and skills, similar to the assessment of pre-Grade R children in the **Thrive By Five** Index Survey series (see **https://thrivebyfive.co.za**).

Like the ELOM 4&5 Years Assessment tool, the ELOM-R Language (v1) Assessment is a direct individual assessment of children's abilities designed for administration by trained assessors using standard test kits. Test performance is captured on tablets and records are uploaded to a server for analysis. This practice standardises administration for each language group and minimises measurement error.

The tools may also be used in research studies and to assess the performance of groups of children following their participation in interventions designed to enhance inputs to numeracy or literacy education programmes.

ELOM-R LANGUAGE (v1) ASSESSMENT ITEMS

The Pilot measure (see ELOM-R (v1) Technical Manual 1) included two items to assess Short Term and Auditory Memory (Pilot Item 1: Digits Forward) and Working and Auditory Memory (Pilot Item 2: Non-Word Repetition). Both items assess cognitive skills related to language and numeracy abilities. As these Executive Function (EF) items will be included in a separate EF measure in development, they were removed from the final Language measure used for psychometry. The item set for psychometric analysis and norming is presented in Table 1.

*Table 1. ELOM-R Language (v1) Items**

GRADE R CAPS AREA	ITEM	NUMBER OF TRIALS
LISTENING & SPEAKING Vocabulary and oral language	1. Productive Vocabulary (3)	36
	7. Listening Comprehension (9)	10
READING & PHONICS Phonemic awareness and the underpinning auditory, visual and spatial perception required for reading. Letter, word and initial consonant recognition.	2. Beginning Sounds (4)	8
	3. Letter Sounds (5)	8
WRITING & HANDWRITING Drawing and emergent writing skills; underpinning perceptual & motor skills; spatial and visual awareness	4. Copy Shapes (6)	4
	5. Write Name (7)	1
	6. Writing with encouragement (8)	1
Understanding of print: Understanding the orthographic system and written language	8. Book concept, orientation, and word concept (10)	9

**Appendix 1 provides item scoring.*

Given varying numbers of trials, raw scores on each ELOM-R (v1) item are on different scales. For example, item 1 (Productive Vocabulary) has 36 trials, and a child can obtain a score from 0-36; Beginning Sounds (item 2) has eight trials and a child can score from 0-8. When a test is standardised, all scores must be converted to the same scale. For this reason, all ELOM-R (v1) item scores are converted to percentage correct total scores on the test, ranging from 0-100.

ASSESSING EQUIVALENCE AND BIAS IN MEASURES FOR A DIVERSE SOCIETY

The psychometric methods used in standardising the ELOM-R Language (v1) Assessment follow ITC *Confirmation Guidelines* C-1(9), C-2(10), C-3 (11) and C-4 (12) as described in ELOM-R (v1) Technical Manual 1. These Guidelines have informed the psychometric procedures followed in the cross-national and South African adaptations of both the International Mathematics and Science Study (TIMSS) (assesses Grade 9s[1]), and the Programme for International Student Assessment (PISA) (assesses literacy in Grade 4[2]).

[1] https://www.timss-sa.org/publication/the-south-african-timss-2019-grade-9-results

[2] https://www.up.ac.za/media/shared/164/ZP_Files/2023/piirls-2021_highlights-report.zp235559.pdf

When a test is intended for more than one cultural or linguistic group as is the case with the ELOM-R (v1) Language Assessment, it is necessary to undertake procedures to establish whether the psychometric properties of the test are the same when adapted and translated into other languages. In recommending procedures for test adaptation for use in different ethnolinguistic groups, Hambleton (2001[3]) states that: *"Evidence is needed to support the use of an adapted test in each language where it is used"* (p. 168). We follow him in assessing whether the various languages of testing have the same factor structure, they each measure the same underlying trait. Furthermore, we follow ITC Guideline C-2 (10) which states that test developers should "provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations" (ITC, 2017, p. 114[4]).
As van de Vijver and Tanzer, (2004)[5] put it:

> **Both bias and equivalence are pivotal concepts in cross-cultural assessment. Equivalence of measures (or lack of bias) is a prerequisite for valid comparisons across cultural populations** *(p. 120)*

Van de Vijver and Rothmann (2004[6]) remarked at that time that psychometric research work on this issue was in its infancy in South Africa. We are not aware of significant advances in measures designed to assess the skills in the ELOM-R (v1) assessments since then. However, one example is the work conducted on the ELOM 4&5 Years Assessment tool to assess the cross-language equivalence of that instrument (Dawes et al., 2020[7]; Snelling et al., 2019[8]).

A taxonomy of bias and equivalence issues relevant to the ELOM-R (v1) assessments drawn from Van de Vijver and Rothmann (2004[9]) pp. 2-3} and Poortinga (1998[10]) is presented in Table 2. Note that in their papers, the above authors refer to cross-cultural bias and equivalence. Our primary concern in developing the ELOM-R (v1) is to reduce bias as far as possible because of language differences between groups. In South Africa, language is, of course, a key component of culture. However, it would be a grave mistake to see each South African language group as embodying a distinct isolated culture. Multiple cultural commonalities will be evident across linguistic groups, particularly in modern urban communities and among children who have received a Grade R education (the target group for this measure).

[3]Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. European Journal of Psychological Assessment, 17(3), 164-172. Doi 10.1027//1015-5759.17.3.164
[4]International Test Commission. (2017). ITC Guidelines for Translating and Adapting Tests (Second edition). [**www.InTestCom.org**].
[5]Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology,* 54(2), 119-135.
[6]Van de Vijver, and Rothmann (2004). Assessment in multicultural groups: The South African case. *South African Journal of Industrial Psychology,* 30(4), 1-7
[7]Dawes, A., Snelling, M.J.T.L., Henning, T. & Moonsamy, J. (2020). ELOM Teacher Assessment. In Dawes, A., Biersteker, L., Girdwood, E., Snelling, M.J.T.L., Tredoux, C.G. et al. Early Learning Outcomes Measure. Technical Manual (pp. 40-44). Claremont, Cape Town: The Innovation Edge **https://datadrive2030.co.za/wp-content/uploads/2022/09/ELOM-Technical-Manual_2020-1.pdf**
[8]Snelling, M.J.T.L., Tredoux, C.G., Dawes, A., Anderson, K., Henning, T. Moonsamy, J. & Scott, M. (2020). Psychometry and statistical analyses. In Dawes, A., Biersteker, L., Girdwood, E., Snelling, M.J.T.L., Tredoux, C.G. et al. Early Learning Outcomes Measure. Technical Manual (pp.14-25). Claremont, Cape Town: The Innovation Edge. **https://datadrive2030.co.za/wp-content/uploads/2022/09/ELOM-Technical-Manual_2020-1.pdf**
[9]Van de Vijver, and Rothmann (2004). Assessment in multicultural groups: *The South African case. South African Journal of Industrial Psychology,* 30(4), 1-7
[10]Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. International Journal of Psychology, 24, 737-756.

TYPE OF BIAS	DEFINITION	SOURCE / EXAMPLE
BIAS	*"Nuisance factors that threaten the comparability of scores across groups"* (Van de Vijver & Rothmann, p.3).	Construct is not understood in the same or similar way across groups.
CONSTRUCT BIAS	The *"construct measured is not identical across groups"* (Van de Vijver & Rothmann, p.3).	Skills measured may not be familiar to one or another group.
METHOD BIAS	*"Factors, resulting from sample incomparability (sample bias), instrument characteristics (instrument bias), tester effects and communication problems administration bias)"* (Van de Vijver & Rothmann, p.3).	Incomparability of samples; test instructions understood differently (functional inequivalence); instructions to assessors unclear.
ITEM BIAS	*"Nuisance factors at the item level"* (Van de Vijver and Rothmann, p.3).	"Nuisance factors" influence test performance that introduces measurement error. They need to be accounted for or explained. For example: poor translation; item unfamiliar to the culture.
EQUIVALENCE	*"Comparability of test scores across cultures"* (Van de Vijver & Rothmann, p.3).	Items are similar in difficulty across groups. Children of similar ability perform similarly across items.
STRUCTURAL EQUIVALENCE	*"Instrument measures the same construct in the groups studied"* (Van de Vijver & Rothmann, p.3).	Test Factor structure is the same across language groups.
SCALAR OR FULL SCORE EQUIVALENCE	*"Scores are fully comparable"* across language groups (Van de Vijver & Rothmann, p.3).	The same item and measurement unit is used to assess all groups.

INVESTIGATING BIAS IN THE ELOM-R LANGUAGE (v1)

We begin with a brief overview of approaches to establishing reliability, equivalence and bias between measures adapted from a source (in this case English) to other languages.

The factor structure of each language version of the ELOM-R Language (v1) was compared using Multi-Group Confirmatory Factor Analysis (MGCFA). The procedure is also used to establish whether the relationship between the items and the total test score is the same or similar in each of the languages. Where this is established (known as cross-validation) in the languages of adaptation, one can assume that the test is measuring the same properties in all languages, and, therefore, a child's test scores have the same meaning regardless of their language or cultural background. Where this is not so, adjustments to test items may be necessary. For further detail on these topics, readers are referred to Fischer & Karl, (2019[11]); van de Vijver and Tanzer, (2004[12]) and Geisinger (1994[13]).

Test reliability (in these investigations internal consistency), item difficulty and item discrimination (between more and less able children) were also assessed to establish whether these are comparable across the languages. Item-level Differential Item Functioning (DIF) and Test DIF were investigated using Item Response Theory (IRT) Rasch analyses which compare individuals' performance on each item in each language to assess whether children in a particular language group perform the same (uniform bias), better (benign DIF) or worse (adverse DIF) than other groups on an item despite their similar overall ability. Test-Level (cumulative) DIF analysis provides the same information for the entire test. The metric equivalence of a test adapted and translated from a base language (in this case, English), is established when an item difficulty does not vary significantly between English and the languages of translation (Milfont & Fischer, 2015[14]; 2007[15]). None of these investigations could be undertaken on Pilot data as a) the samples were too small, and b) some adjustments were made after the Pilot (see ELOM-R (v1) Technical Manual 1: Development of the ELOM-R Language and Mathematics Assessments.

[11]Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in psychology*, 10, 1507- . doi: 10.3389/fpsyg.2019.01507
[12]Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology, 54(2), 119-135.
[13]Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment,* 6(4), 304-312.
[14]Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research,* 3(1), 111-130.
[15]Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development,* 78(4), 1255-1264.

CHAPTER 2. PSYCHOMETRY AND STATISTICAL ANALYSES

In this chapter, we summarise preliminary psychometric analyses undertaken using Classical Test Theory (CTT) and Item Response Theory (IRT) modelling procedures to investigate the factor structure and internal consistency of the ELOM-R Language (v1) in isiZulu, isiXhosa, Sepedi, Sesotho, Setswana, Tshivenda, English and Afrikaans. Full psychometric reports for each language are available.

CONSIDERATIONS FOR SAMPLE SIZE

Recommendations for sample size for these analyses vary (e.g. Kline, 1979[16]; Kyriazos, 2018[17]; Mundfrom et al, 2005[18]). Kline, among others, recommends at least n=100. However, Mundfrom et al. (p. 159) note that: "*Suggested minimums for sample size include from 3 to 20 times the number of variables and absolute ranges from 100 to over 1,000. For the most part, there is little empirical evidence to support these recommendations*".

In their paper, Mundfrom et al. (2005) report that an empirically informed guide to sample size for factor analysis is the variables to factors ratio (or test items to factors ratio). In the ELOM-R Language (v1) Assessment, we have eight items and tested a single-factor solution (8 items and 1 factor i.e. a ratio of 8:1).

As in the case of Factor Analysis, the research literature provides various guidelines on sample size for IRT Rasch analyses, making it challenging for the researcher to choose which to follow. Some have recommended at least n=1,000 / group – an unfeasible and unaffordable prospect for ELOM-R (v1) IRT analyses. Linacre (1994[19]) provides support for reliable findings in one-parameter logistic models (1PL) analyses (as used here), with samples as small as 50. However, Chen et al. (2014[20]) caution against samples of less than 100 and show that parameter estimates in Rasch analyses are more reliable when samples exceed 250.

Based on these considerations, we decided to realise minimum sample sizes of at least 275 children in each language to cover requirements for both IRT Rasch and CTT Factor analysis and reliability.

As noted in the ELOM-R (v1) Technical Manual 1, while it is best practice to include representative numbers of children from all socio-economic strata in each language, this was not feasible in a study of this scope. Furthermore, as we shall observe, language and socio-economic status are often confounded in South Africa. As a long-term consequence of apartheid policy which prior to 1994 discriminated both structurally and personally against people of colour, far greater proportions of African language speakers than English and Afrikaans reside in households in the lower three quintiles. Inequality remains and affects language comparisons on psychometrics and must be borne in mind throughout this report.

[16]Kline, P. (1979). *Psychometrics and psychology.* London: Academic Press.
[17]Kyriazos, T. A. (2018). Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology,* 9(08), 2207. DOI: 10.4236/psych.2018.98126
[18]Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159-168.
[19]Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch measurement transactions,* 7, 328.
[20]Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of life research,* 23, 485-493.

SAMPLE

The sample for the analyses that follow was drawn from two sources:

1. **Sample 1:** Studies using the measure in research and evaluation studies (see below): n = 1,713 randomly selected children in 225 schools and Grade R classrooms)
2. **Sample 2:** Data collected in public school Grade 1 classes to make up the required sample sizes for psychometric analyses: n = 890 randomly selected children in 77 schools and Grade 1 classrooms.

Note that even though they are included in this number, isiNdebele, Siswati, and Xitsonga language samples were not included in analyses that follow as sample sizes were not sufficient to establish baseline psychometric properties. Data on these groups will be collected for analysis at a later point.

VARIATIONS IN SAMPLE SIZES FOR ANALYSES

It is important to note that sample sizes will vary for the psychometric analyses undertaken due to missing values and the outlier cases removed.

Table 3. Descriptive Statistics for Child Age (Months)

N	MEAN	SD	MEDIAN	SD	MINIMUM	MAXIMUM
2564	77.4	3.88	77.3	3.88	70.0	89.0

Figure 1. Distribution of Child Age (Months)

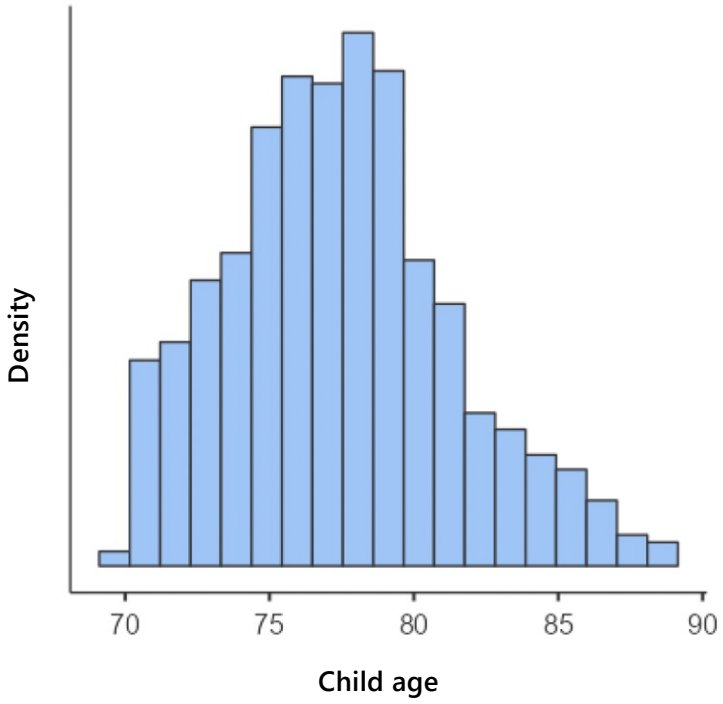
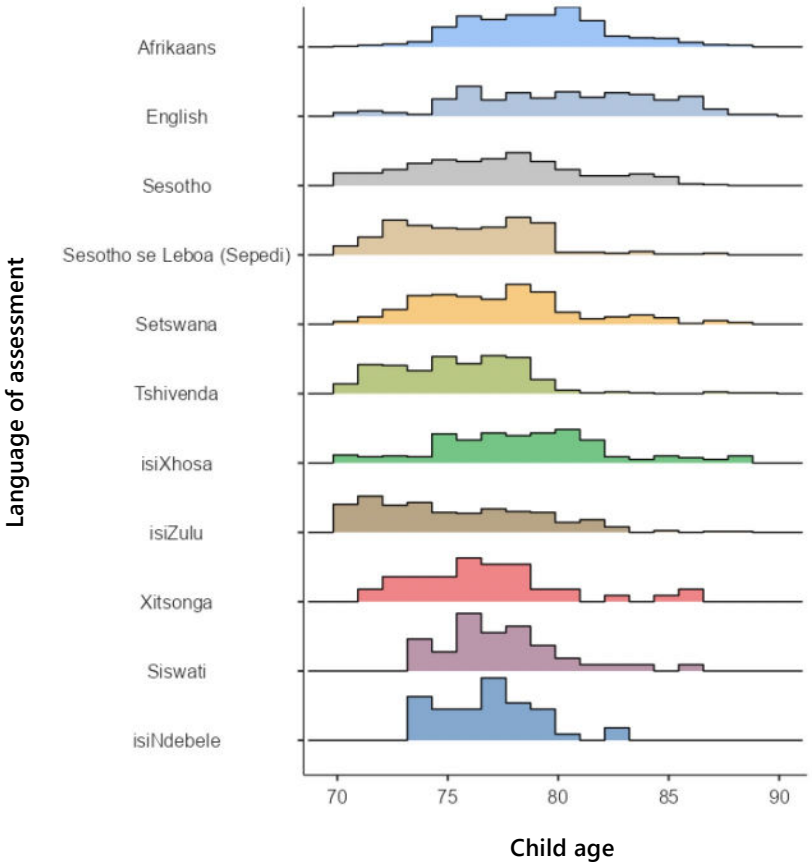


Table 4. Descriptive Statistics for Child Age (Months) Per Language Group

	language_assessment	N	Mean	Median	SD	Minimum	Maximum
child_age	Afrikaans	448	79.1	79.0	3.15	70.6	88.6
	English	282	80.1	80.1	4.17	70.2	88.9
	Sesotho	287	77.2	77.0	3.85	70.0	86.6
	Sesotho se Leboa (Sepedi)	286	75.9	75.8	3.21	70.0	87.1
	Setswana	276	77.4	77.2	3.58	70.3	88.0
	Tshivenda	290	75.5	75.4	3.07	70.0	89.0
	isiXhosa	290	78.4	78.4	3.89	70.3	88.6
	isiZulu	283	75.3	74.8	3.68	70.0	87.8
	Xitsonga	41	76.9	76.5	3.62	71.7	86.5
	Siswati	40	77.4	77.3	2.74	73.5	86.2
	isiNdebele	41	76.9	77.0	2.32	73.3	82.9

Figure 2. Distribution of Child Age (Months) Per Language Group



ETHICS PROCEDURES

1. **Sample 1** Approval for the research and evaluation studies followed different channels: The Roots and Shoots study received ethical approval from the Faculty of Commerce at the University of Cape Town. Kellelo Consulting received approval from the Gauteng Department of Basic Education. JET Education Services did not go through an IRB process. However, their caregiver consent forms asked for consent to use the data for research purposes beyond the Anglo American programme.

2. **Sample 2** was approved by the Provincial Departments of Education of the schools where the data was collected and by an Ethics Committee of the University of Cape Town Humanities Faculty on 7 November 2023 (reference No. PSY2023-031).

Children's caregivers were requested to provide informed consent for their children's participation. Forms explaining the study were sent to them by the child's school. Caregivers were requested to sign and return the form should they consent. If the form was not returned, and as the study constituted a minimal risk to participants, opt-out / passive consent was approved by the Committee. Children were asked to assent to testing; if they refused, another child was recruited. They were able to discontinue the test at any time.

ASSESSOR TRAINING

Assessors of children in both samples 1 and 2 attended four days of ELOM-R (v1) training and only proceeded to the field if judged competent in administering the tests. Inter-scorer reliability was established as part of training and accreditation. Only assessors who scored a minimum of 85%, scoring concordance with a standardised scoring of a demonstration video, were accredited to use the ELOM-R (v1) tools.

DATA COLLECTION

Data for sample 1 was provided by the various research and evaluation study teams. Fieldwork for Sample 2 was undertaken by Genesis Analytics. Their field report notes: *"Data was collected to make up sufficient numbers for the analyses and was drawn from children enrolled in Grade 1 classes in primary schools in KwaZulu-Natal, the Free State, Limpopo, the Eastern Cape and Mpumalanga. Schools were purposively selected to enrol children from the range of school quintiles (Q) in each language. However, matching the home language with the language of instruction in the Grade 1 class (essential for this study) proved challenging in higher quintiles (Q4 and Q5), where English predominates. To address this, fieldwork staff identified schools in these quintiles with a significant number of students speaking the target language at home despite being taught in another language and noted these instances in the final dataset. Achieving the target in the upper quintiles was challenging due to the insufficient number of learners in the schools to fully meet the target. This also necessitated adjusting to include more schools and learners from Q3."*

ASSESSMENT OF CHILDREN

Children were tested in a quiet space on both ELOM-R Mathematics and Language (v1) Assessments in their home languages on the same day (with a break between tests). While the order of assessments was not predetermined, often assessors started with the ELOM-R Mathematics (v1) and proceeded to the ELOM-R Language (v1) with a short break in between. Children were returned to their classrooms post the assessment.

Scale reliability, factor structure, item difficulty and bias

Methods commonly used to assess test internal consistency (reliability) and factor structure fall within the Classical Test Theory (CTT) approach to psychometrics, a longstanding approach to assessing the integrity and performance of psychometric tests. In this approach, the variance between individuals in their responses to test items is attributed to their standing on a latent (unobservable but inferable) ability or trait such as IQ (Furr, 2021[21]). In CTT, only one measurement term is specified – the (latent) ability represented by the Total score on the measure.

RELIABILITY (INTERNAL CONSISTENCY)

To assess whether the ELOM-R (v1) items are consistent in their measurement of the underlying construct, reliability was tested using McDonald's omega (ω), a version of Cronbach's alpha statistic that does not assume equal variances for all items. Generally, a value of ω **= 0.70** and higher indicates scale reliability (Kline, 2000[22]). To assess reliability on the item level, ω is calculated with each item excluded sequentially. If the reliability of the scale improves when an item is excluded, that item is detracting from the internal consistency of the scale.

While **0.70** is regarded as acceptable for many purposes, Nunnally (1978[23]) notes that in applied settings where important high-stakes decisions are made about individuals based on their test scores, a reliability of **.90** is the standard to realise. We do not regard ELOM-R (v1) as a "high stakes" test in Nunnally's terms as it is not intended to inform high-stakes decisions made on individual children as would be the case, for example, where a child would be kept back a year from the Grade 1 year. Rather, the ELOM-R (v1) tests are intended to provide descriptions of populations or smaller groups to inform curriculum and programme inputs to the Grade R and Grade 1 year and to assess the performance of groups of children following their participation in interventions designed to enhance inputs to numeracy or literacy education programmes. For these purposes the reliability standard recommended by Nunally is regarded as too stringent and not applied here.

Item-rest correlations indicate the strength of each item's correspondence to the rest of its scale. Item-rest correlations are generally considered adequate above r **= 0.3.** Test-retest reliability is not considered here as it has not yet been investigated.

CONFIRMATORY FACTOR ANALYSIS (CFA)

CFA is a statistical modelling method for the probabilistic testing of specified factor models within the covariance structure of the test items. The analysis tests whether or not the hypothesised factor structure is confirmed. For example, does the ELOM-R Language (v1) Assessment measure one underlying construct or not? CFA, therefore, provides an assessment of how well a set of items reflect the theoretical structure of the constructs they are purported to measure - in this case CAPS Language skills following exposure to Grade R.

As mentioned previously, when a test has been translated (in this case from English) and adapted for use in other languages, CFA is conducted on all the languages so that the factor structure can be compared, an approach known as Multi-Group Confirmatory Factor Analysis (MGCFA). If the resulting factor structure is the same in all the languages, then we can be reassured that the test measures the same construct in all. Translation procedures are described in ELOM-R (v1) Manual 1.

[21]Furr, R. M. (2021). Psychometrics: An Introduction. Sage Publications. ISBN: 9781071824108
[22]Kline, P. (2000). Handbook of Psychological Testing. London, United Kingdom: Routledge.
[23]Nunnally, J. C. (1978). An overview of psychological measurement. Clinical diagnosis of mental disorders: A handbook. Springer.

A unidimensional (single-factor) model was tested for all the languages for which the sample size is adequate. Fit statistics are used to assess the fit of the model to the observed data (is the hypothesised factor structure evident). Factor loadings of individual items to the single-factor model are evaluated to assess potential misfit at the item level. The goal is to have a good-fitting model. Table 5 describes the main statistics used in this section of the report as well as rough guidelines to their interpretation (Barrett, 2007[24]; Hu & Bentler, 1999[25]; Tavakol & Wetzel, 2020[26]).

Table 5: CFA Statistics and their Interpretation

STATISTIC	INTERPRETATION
CHI-SQUARE (χ^2)	An overall test of the fit of observed variance within and between items to a specified statistical model. Smaller values with non-significant p-values are considered indicative of model fit. However, this test is considered highly sensitive and often shows misfit for generally well-fitting models tested in larger samples or with complex factor structures. For this reason, fit indices such as RMSEA, CFI, and TLI are usually considered more important for assessing CFA model fit.
FACTOR LOADINGS	A correlation coefficient between an item score and its latent factor. **Factor loadings > 0.3** indicate a sufficiently strong relationship between the item and the underlying factor.
STANDARDISED LOADINGS	As the unstandardised factor loading is calculated on the same scale as item scores, it does not allow for meaningful interpretation of the strength of factor loadings. Standardised factor loadings are calculated on a universally comparable scale, in which factor **loadings >0.3 are acceptable.**
RMSEA	An Absolute Fit Index where a value of 0 indicates a perfect model. Values closer to 0 indicate a better model fit. Values **<0.08 indicate good fit.**
CFI & TLI	The Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) are both fit statistics which compare the fit of a factor model to a baseline model. Values both range from 0 to 1 and are considered **acceptable > 0.9 and > 0.95.**

As will be evident below, a single (unidimensional) factor structure was not clearly established for the ELOM-R Language (v1) Assessment in CFA. Given that more than one factor was plausible within the CAPS domains (see Table 1), Exploratory Factor Analyses (EFA) were also undertaken in each language to explore any subfactor structure evident in the data.

[24]Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815–824. **https://doi.org/10.1016/j.paid.2006.09.018**

[25]Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modelling, 6(1), 1–55. **https://doi.org/10.1080/10705519909540118**

[26]Tavakol M, Wetzel A. (2020). Factor Analysis: a means for theory and instrument development in support of construct validity. Int J Med Educ. 2020 Nov 6;11:245-247. doi: 10.5116/ijme.5f96.0f4a. PMID: 33170146; PMCID: PMC7883798

RASCH ANALYSIS

The Rasch model is a popular implementation of Item Response Theory (IRT), which can be used in conjunction with the CFA described above. IRT Rasch specifically models responses on any test item as a product of both the child's ability and the difficulty of the item, which are not taken into consideration in CCT methods. When item difficulty is estimated, scoring within the IRT paradigm offers more rigorously modelled – and therefore, more accurate –estimates of respondents' true level of ability (Baker, 2001[27]; Bond & Fox, 2015[28]; Fan, 1998[29]).

Based on the item scoring in the ELOM-R Language (v1) Assessment and the presumption of a unidimensional factor structure (necessary for Rasch analysis), a dichotomous one-parameter logistic model (1PL) Rasch model was initially used for analyses. Percent Correct (PC) scores for each item were first dichotomised using WINSTEPS® software. This common approach requires that a score of 100% (correct) on the item is transformed to 1 and all other percentages are converted to 0. This is unproblematic when the item can only be correct or incorrect. But when the item has gradations of correctness (e.g. 50% or 60% correct) as is the case in multi-trial ELOM-R (v1) items, these are lost.

While this method was selected as the most suitable for this purpose, the results of the Rasch portion of these analyses should be interpreted with caution. The dichotomisation of item responses may misrepresent ELOM item response variances, and item difficulty estimates should be interpreted as the difficulty of attaining a perfect response rather than the overall difficulty of the original polytomous scale[30]. Other modelling methods to take into account polytomous items were explored and are discussed at a later point.

Scores on the ELOM-R (v1) Language were subjected to Rasch modelling to determine item difficulty and a more accurate assessment of the validity and reliability of the test. However, as will be evident in the analyses, model fit was poor, reinforcing the findings from CFA. Important metrics to consider in Rasch analysis are described in Table 6 below, along with guidelines for their interpretation (Bond & Fox, 2015; Linacre, 2002[31]; Yen, 1993[32]).

[27]Baker, F. (2001). The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.
[28]Bond, T., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences. New York, NY: Routledge
[29]Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person parameters. Educational and Psychological Measurement, 58, 357–381.
[30]Polytomous scales have more than two possible scores for an item. This is the case in ELOM-R (v1) Language where item trials are individually scored and summed to derive the item score.
[31]Linacre, J. (2002). What Do Infit and Outfit, Mean-Square and Standardized mean? Rasch Measurement Transactions, 16. Retrieved from **https://www.rasch.org/rmt/contents.html.**
[32]Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30(3), 187-213.

Table 6: CFA Statistics and their Interpretation

STATISTIC	INTERPRETATION
MEASURE (ITEM INTERCEPT)	Indicates the probabilistic Rasch model estimate (in logits) for item difficulty and person ability. **An item estimate of 0 indicates that it is of average difficulty, with negative and positive numbers indicating lower and higher difficulty respectively.** Difficulty estimates **typically range between -3 and +3.** As a foundational principle of the Rasch model, it is expected that for an item with a logit of 0, respondents with an ability estimate of 0 have an equal chance of responding correctly or incorrectly.
MEAN SQUARE INFIT	Fit statistic indicating the accuracy of the Rasch model in predicting responses. The Infit statistic is sensitive to model misfit weighted towards inliers, or those who score close to the item difficulty estimate. **An infit statistic of 1 is ideal, with lower values (<0.6) indicating overfit, and higher values (>1.4) indicating misfit.** Typically, the Infit statistic is given greater consideration than the outfit, as it is less of a threat to accurate measurement.
MEAN SQUARE OUTFIT	Fit statistics indicate the accuracy of the Rasch model in predicting responses. The Outfit statistic is sensitive to model misfit caused by outliers. **An outfit statistic of 1 is ideal, with lower values (<0.6) indicating overfit, and higher values (>1.4) indicating misfit.**
PERSON RELIABILITY	An overall measure of the consistency of response scoring, interpreted similarly to Cronbach's alpha. Values of 1 are ideal, with person **reliabilities above 0.5 considered acceptable.**
POINT-MEASURE CORRELATION	Correlation between raw item or scale score and Rasch ability estimates. **Considered acceptable above 0.2.**
MADaQ3	MADaQ3 offers an overall estimate of model fit and is an adjusted aggregate of Q3 coefficients (residual correlation coefficients) across items. It is reported on the logit scale. Smaller MADaQ3 values are preferred, and model fit is indicated when the associated p-value exceeds 0.05. However, it should be noted that the MADaQ3 statistic tests perfectly fit the Rasch model and are sensitive to sample size, so is prone to type II error. High Q3 correlations are indicative of local dependence, which violates the statistical integrity of Rasch modelling.

SUMMARY PSYCHOMETRIC PROPERTIES ELOM-R (v1) LANGUAGE ASSESSMENT

Language group sample sizes and findings are provided in Table 7. Detailed psychometric reports are available for each language from DataDrive2030. While reliability is sound in all languages, using RMSEA, model misfit is evident in CFA for all, including English (the language in which the test was designed). However, Afrikaans displays model fit using CFI and CFA in this language can be regarded as unidimensional. Note that the Afrikaans finding may be attributable to the much greater sample size in this language. Despite the findings for CFA above, scree plots in all languages indicated that the Language Assessment has a unidimensional scale.

In Rasch analyses, the only language with an acceptable fit is Afrikaans (again, perhaps due to much greater sample size). Rasch's low person reliability indicates that the variance explained by ability estimates is not large enough relative to that explained by their standard error. However, the point-measure correlations are all quite strong, indicating a stable relationship between ability measures and test scores. This is the **key consideration, as a child's ability level is related to their performance on more or less difficult items.**

Table 7. ELOM-R (v1) Language: Comparison of CFA and Rasch findings by Language

LANGUAGE	SAMPLE	RELIABILITY[33]	CFA[34]	RASCH[35]
ENGLISH	n=282 Q1,2 &3 = 63% Q4&5 =37%	ω =0.746	Model Misfit (RMSEA = 0.103) Scree plot** indicates Unidimensional	Model Misfit: point-measure correlation (r = 0.813); person reliability (0.420) below threshold
AFRIKAANS	n=448 Q1,2 &3 = 46% Q4&5 =54%	ω =0.834	Model misfit RMSEA (0.108) and TLI (0.875) Model Fit on CFI (0.910). Unidimensional	Model misfit RMSEA (0.108) and TLI (0.875) Model Fit on CFI (0.910). Unidimensional
ISIXHOSA	n=291 Q1,2 &3 = 68% Q4&5 =32%	ω =0.747	Model misfit (RMSEA = 0.105) Scree plot** indicates Unidimensional	Model misfit (RMSEA = 0.105) Scree plot** indicates Unidimensional
ISIZULU	n=280 Q1,2 &3 = 57% Q4&5 =43%	ω =0.769	Model misfit (RMSEA = 0.123) Scree plot** indicates Unidimensional	Model misfit (RMSEA = 0.123) Scree plot** indicates Unidimensional
SETSWANA*	n=277 Q1,2 &3 = 95% Q4&5 =5%	ω =0.765	Model misfit (RMSEA = 0.121) Scree plot** indicates Unidimensional	Model misfit (RMSEA = 0.121) Scree plot** indicates Unidimensional
SEPEDI *	n=282 Q1,2 &3 = 93% Q4&5 =7%	ω =0.774	Model Misfit: RMSEA = 0.112 Scree plot** indicates Unidimensional	Model Misfit. point-measure correlation (r = 0.769); person reliability (0.152); below threshold.
TSHIVENDA*	n=292 Q1,2 &3 = 93% Q4&5 =7%	ω =0.718	A degree of Misfit (RMSEA = 0.095) Scree plot** indicates Unidimensional	Model Misfit. point-measure correlation (r = 0.769); person reliability (0.152); below threshold.
SESOTHO*	n = 282 Q1,2 &3 = 72% Q4&5 =28%	ω =0.785	Model Misfit: RMSEA = 0.120. Scree plot** indicates Unidimensional	Some Model Misfit. point-measure correlation (r = 0.796); person reliability (0.268). below threshold.

(*Note that in four languages (highlighted in red), a very high proportion of children are in the lower school quintiles. Language and quintiles are clearly confounded, and this is likely to affect all results for that group.

It is important to note that the score binning method employed in the Rasch analyses reported above, which aims to achieve dichotomous item values, does not take polytomous scoring (see below), which will have distorted these results. The results of Rasch modelling presented above therefore cannot be considered reliable. Alternative approaches were undertaken and are reported below.

[33]ω should => 0.7.
[34]Confirmatory Factor Analysis tests a model of the number of factors / item clusters / domains expected for the test. A single factor model was tested as this is what is required for standardisation. RMSEA should be < 0.08. CFA does not control for item difficulty.
[35]Dichotomous Rasch modelling was used here and takes into account both item difficulty and person ability.; Point measure correlation should >0.2

TESTING RASCH MODELS TO TAKE ACCOUNT OF ELOM-R LANGUAGE (v1) POLYTOMOUS ITEM DESIGN

As noted, dichotomous Rasch modelling was used for analyses presented in Table 7. This is appropriate for tests where test items are scored correct/incorrect. However, dichotomised scores represent an oversimplification of the ELOM-R Language (v1) responses as the items have several trials in which scores contribute to the total and are not dichotomised.

To address this, we undertook an investigation of an IRT model based on its original, polytomous (multi-trial) response structure. Several were considered, most notably the Partial Credit Model (PCM). However, this was not successful. Polytomous IRT methods such as PCM require data to be available for all possible scores on an item, and the analysis involves estimating the difficulty of not just the item itself, but also the difficulty of achieving each of its possible (trial) responses. For PCM, the dataset must include sufficient data (child scores) for all these levels of performance.

This was not the case for all ELOM-R Language (v1) items, as some had empty or sparse response levels for certain items due to the child choosing not to proceed to answer a trial or item, or because of stop rules (the item is discontinued if the child fails a certain number of trials). If an insufficient number of children achieve a particular response to an item trial, the difficulty of doing so cannot be accurately modelled, and the item-level difficulty estimate is undermined (Linacre, 2000[36]). For example, on item 1 (productive vocabulary), a child may score between 0 and 36 depending on the number of trials passed. However, there may be too few or no records of children scoring trials 12, 30, or 36 correct.

The PCM results reflected this issue with model fit approximating that of a dichotomous Rasch model, but with severe misfit at the item level.

Continuous and Poisson Count Rasch models were also considered. However, all IRT modelling efforts on the ELOM-R Language (v1) Assessment were challenged by the varying item response scales. No function describing the relationship between item responses and difficulty/ability parameters will perfectly suit all items within the test, as their response scales differ

In the next version of ELOM-R Language (v1) Assessment additional items will be added, and factor structure will be re-assessed. Once a clear structure is established, Rasch analyses using a hybrid approach wherein items with similar scale properties are treated as separate testlets will be considered. Each testlet will need to contain sufficient items for parameter estimation.

[36]Linacre, J. (2000). Comparing and Choosing between “Partial Credit Models” (PCM) and “Rating Scale Models” (RSM). RMT. https://www.rasch.org/rmt/rmt143k.htm

SUMMARY EXPLORATORY FACTOR ANALYSIS (EFA) FINDINGS

Given CFA and Rasch misfit findings, Exploratory Factor Analysis (EFA) was conducted on all languages to establish possible additional factors and cross-loadings (items are associated with more than one construct). The summary findings are as follows:

- **English**: EFA produced two factors using parallel analysis, but a significant drop in eigenvalue (factor 2 eigenvalue = 0.435). The two factors produced through EFA modelling do not align with the CAPS: Listening and Speaking' and CAPS: 'Emergent writing and handwriting skills' subdomains. Another two-factor CFA model tested the viability of a CAPS 'Listening and Speaking' factor and a CAPS 'Developing writing and handwriting skills' factor, but again, the model was not supported.
- **Afrikaans:** EFA produced three factors using parallel analysis, although eigenvalues for the second (0.319) and third (0.121) factors were small, and multiple cross-loadings were indicated. A two-factor CFA model tested the viability of CAPS Listening and Speaking' and 'CAPS 'Developing writing and handwriting skills' factors, but a two-factor model was not supported.
- **isiXhosa:** EFA produced three factors on parallel analysis, although the second two factors produced several cross-loadings. Eigenvalues for the second ($\lambda = 0.299$) and third ($\lambda = 0.269$) factors explained very little variance; a scree supported a single-factor solution. A two-factor CFA model tested the viability of a CAPS 'Listening and Speaking' and 'CAPS 'Developing writing and handwriting skills' factor, but the model was not supported.
- **isiZulu:** EFA produced two factors on parallel analysis but a significant drop in eigenvalue (factor 2 eigenvalue = 0.490). A two-factor CFA model tested the viability of a CAPS 'Listening and Speaking' and 'CAPS 'Developing writing and handwriting skills' factor, but the model was not supported.
- **Setswana:** EFA produced two factors on parallel analysis, but a drop in eigenvalue (factor 2 eigenvalue = 0.507). The factor solution was not clear. Only one item produced a clear loading on the first factor. The second factor was comprised a mix of CAPS domain items. A two-factor CFA model tested the viability of a CAPS 'Listening and Speaking' and 'CAPS 'Developing writing and handwriting skills' factor, but the model was not supported.
- **Tshivenda:** EFA produced two factors on parallel analysis, but a drop in eigenvalue (factor 2 eigenvalue = 0.415). CAPS domains were not clearly supported by factors. A two-factor CFA model tested the viability of a CAPS 'Listening and Speaking' and 'CAPS 'Developing writing and handwriting skills' factor, but the model was not supported.
- **Sepedi:** Produced three relatively sound factors on parallel analysis, although eigenvalues for the second (0.406) and third (0.207) factors were small. A two-factor CFA model was constructed, but the sub-threshold fit statistics do not justify the lower parsimony of a two-factor model. A unidimensional model is better supported for the Sepedi Language.

MULTIPLE GROUP FACTOR STRUCTURE CONCLUSION

1. From a construct validity perspective, these results indicate that the ELOM-R Language (v1) Assessment items do not clearly describe the CAPS areas. Analyses indicate that a clear factor structure for the current eight item version of ELOM-R Language (v1) is not apparent in any language. EFA did not support the presence of clear underlying factors consistent with CAPS Literacy and Language domains.

2. Three items are the minimum required for single-factor CFA. These analyses suggest that the current number of items (8) is likely too small to provide a reliable model for both CFA and Rasch in each language. The next step to improve the ELOM-R Language (v1) Assessment will be to source additional items to be tested on samples of 100 children in each language.

3. This assumption was tested using two-factor CFA models for each language group as an addendum to the MGCFA process. As foreshadowed by low eigenvalues for secondary factors on EFA, added factors were relatively weak, and model fit was not improved for any language.

4. As unidimensionality is required for Rasch analyses, the failure to detect a clear single factor meant that Rasch misfit was inevitable.

ASSESSMENT OF BIAS: DIFFERENTIAL ITEM FUNCTIONING IN THE ELOM-R LANGUAGE (v1) ASSESSMENT

Following IRT Guideline TD-4 (7) which requires test developers to provide evidence that items are suitable for all intended populations, we assessed the extent to which the ELOM-R Language (v1) Assessment items assess children's abilities fairly in each language group.

Differential Item Functioning (DIF) is an IRT-based method for detecting bias at the item level and works on the assumption that people who have the same level of ability on an underlying trait should have a similar probability of responding correctly (Magis et al., 2010[37]). In this case, DIF is used to assess whether latent ability scoring on the ELOM-R Language (v1) Assessment differs across gender and language groups. DIF detection is performed using the Mantel-Haenszel chi-square test in addition to the Rasch-Welch t-test. Both provide estimates of DIF as well as their statistical significance, and are described in more detail in Table 8 below (Holland & Thayer, 1985[38]; Linacre, 2016[39]; Magis et al, 2010).

[37]Magis, D., Beland, S., Tuerlincks, F., & De Boeck, P. (2010). difR: A general framework and an R package for the detection of dichotomous differential item functioning. (Version 5.1.0) [R package]. Retrieved from **https://CRAN.R-project.org/package=difR.**
[38]Holland, P.W. and Thayer, D.T. (1985). An alternate definition of the ets delta scale of item difficulty. ETS Research Report Series, 1985: i-10. **https://doi.org/10.1002/j.2330-8516.1985.tb00128.x**
[39]Linacre, J. M. (2016). Index. Retrieved from **http://www.winsteps.com/index.htm**

Table 8. DIF Statistics and Their Interpretation

STATISTIC	INTERPRETATION
MHχ^2	The Mantel-Haenszel is a chi-square test for DIF. For each item and at each ability level, it compares the probability of a correct response between the "reference group" (**English** in this analysis) and a "focal group" **(one of the other languages)**. It then aggregates the odds of a correct response across the sample ability levels to produce an overall item DIF estimate. Values are positive with a lower limit of 0. **Higher values indicate larger differences between the groups compared**. Significance is set to **($p < 0.05$). When significant, DIF is observed.**
RASCH-WELCH t	The Rasch-Welch t-test involves the application of the student's t-test to compare Rasch model difficulty estimates between groups. The t statistic is distributed around 0. **Higher negative numbers indicate potential bias in favour of the focal group, and higher positive numbers indicate potential bias in favour of the reference group.**
DIF CONTRAST	DIF contrasts are effect size measures for DIF representing the overall difference in the probability of a correct response between a reference and focal group on the logit scale. A value of 0 indicates no difference between groups in terms of their probability of responding correctly, with **higher positive and negative values indicating DIF in favour of the reference and focal groups, respectively.** The ETS Delta scale is commonly used for interpreting the magnitude of DIF; contrasts > **0.43 logits** are considered slight to moderate; contrasts >0.64 logits are considered moderate to large.

SUMMARY OF DIF FINDINGS FOR ELOM-R LANGUAGE (v1)

Full reports of DIF analyses are available from DataDrive2030 on request. A summary of the findings is presented below.

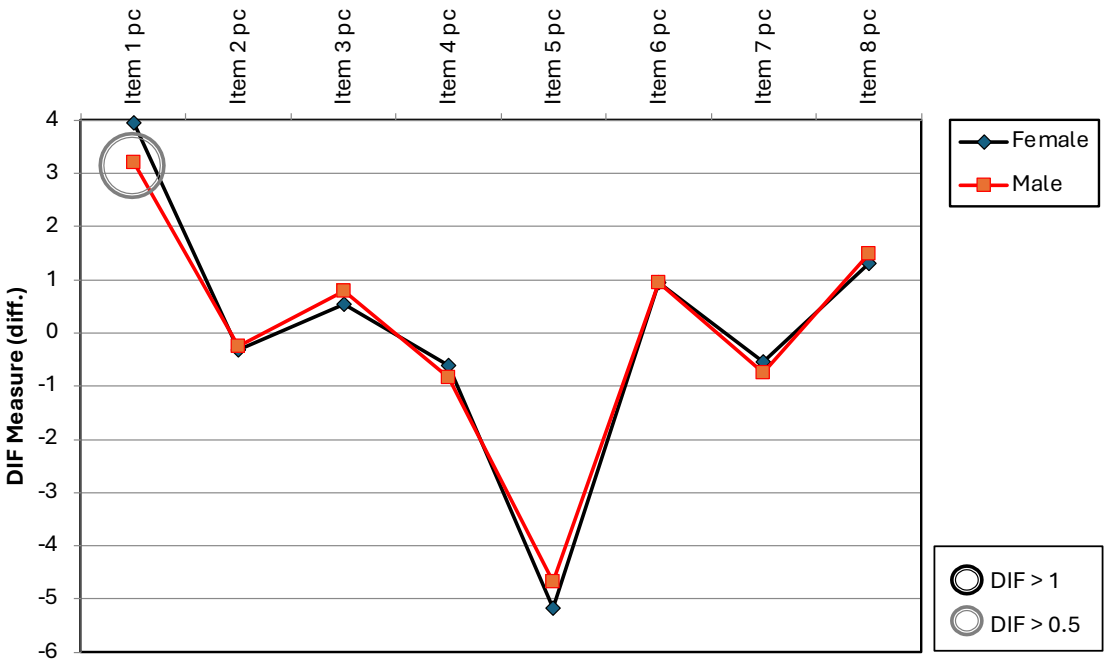
SEX/GENDER DIF

For these DIF analyses, males (n=1237) are used as the reference group, and females (n=1327) as the focal group. The sum of DIF effects across items amounts to a logit value of - 0.02, indicating that DIF does not accumulate in favour of either gender across the scale. That is, sex/gender has a negligible effect on the fairness of the ELOM-R Language (v1) Assessment. On the ELOM-R Language (v1) Assessment, Males (N = 1234, μ = 62.8%) score lower than Females (N = 1327, μ = 67.3%) amounting to a mean score difference of 4.5%.

For Sex/Gender DIF analyses, Males (n=1234) are used as the reference group, and females (n=1327) as the focal group. DIF Measures represent item difficulty estimates (on the vertical axis), across items (along the horizontal axis), for **male and female** groups. Item equivalence is indicated if item difficulties between these two groups are consistent (difference <0.5 logits).

The sum of DIF effects across items amounts to a logit value of 0.21, indicating that there is negligible accumulation of DIF across the scale, meaning that no advantage accrues to males or females on Language assessment. This is illustrated in Figure 3 which indicates that only Item 1 (productive vocabulary) is moderately easier for girls as indicated by the light circle in the figure where DIF >0.5 logits for this item.

Figure 3. ELOM-R Language Sex/Gender Plot



LANGUAGE GROUP DIF

As noted above, misfit was evident in the ELOM-R Language (v1) Assessment in both CFA and Rasch models. In consequence, DIF results for languages need to be interpreted with great caution, as the construct validity and configural invariance of the ELOM-R Language (v1) Assessment are not established. It is specifically worth noting that since the slopes of the DIF model are constraint equal for language groups – only uniform DIF[40] can be diagnosed in the current version of the tool. To diagnose non-uniform DIF, a 2-parameter IRT model is needed. This analysis will be undertaken in the next iteration of the measure. The sample for Language group DIF analyses is presented in Table 9.

Table 9. ELOM-R Language (v1) DIF: Language Group Samples

LANGUAGES*								
ENGLISH	**AFRIKAANS**	**ISIZULU**	**ISIXHOSA**	**SESOTHO**	**SETSWANA**	**SEPEDI**	**TSHIVENDA**	**TOTAL**
281	447	280	291	289	277	282	292	**2439**

*isiNdebele, Siswati, and Xitsonga samples were excluded due to inadequate sample size.

[40]Uniform DIF occurs when all children in one language group perform very similarly on an item.

As the English language versions of the Mathematics and Language assessments were the originally developed forms, English is the reference group for DIF analyses. Focal groups are the Afrikaans, isiZulu, isiXhosa, Sesotho, Setswana, Sepedi, and Tshivenda samples whose ELOM-R Language (v1) assessments are translations of the original English version. Each focal group is contrasted against the English reference group separately to offer clear and comprehensive estimates of DIF for each focal language group. ELOM-R Language (v1) Assessment quintile samples are presented in Table 10.

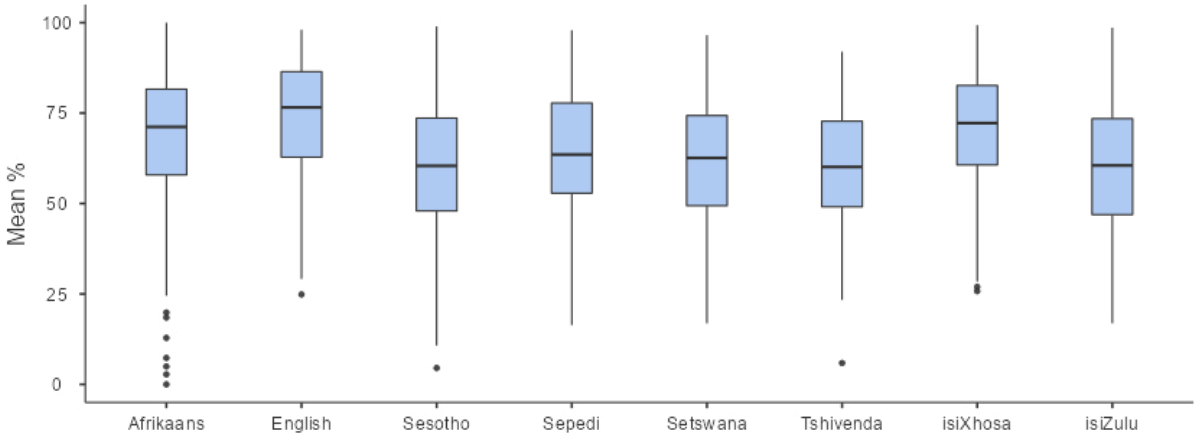
Table 10. ELOM-R Quintile Distributions in Each Language Sample for DIF Analysis

	LANGUAGES*							
SCHOOL QUINTILE	**ENGLISH**	**AFRIKAANS**	**ISIZULU**	**ISIXHOSA**	**SESOTHO**	**SETSWANA**	**SEPEDI**	**TSHIVENDA**
1	13	86	43	23	68	241	214	100
2	34	82	55	74	64	0	13	63
3	118	37	62	102	76	21	23	109
4	47	141	81	57	46	8	16	20
5	69	101	39	35	35	7	16	0

The Sepedi and Setswana subsamples were predominantly within quintile 1 schools. In contrast, most Afrikaans respondents attend schools in quintiles 4 or 5, while most English schools were in the third quintile, a greater proportion were in quintiles 4 or 5 rather than quintiles 1 or 2. The modal quintile for the Tshivenda sample was also 3, but the proportion of quintile 1 and 2 schools heavily outweighed the proportion of schools in quintiles 4 and 5 in this sample. Quintile distributions for the Sesotho, isiXhosa, and isiZulu samples were less remarkable, with each quintile represented.

The English subsample is the reference group for language group DIF analyses. The Afrikaans, isiZulu, isiXhosa, Sesotho, Setswana, Sepedi, and Tshivenda versions are translations of the original English version, thus the subsamples representing these languages are considered focal groups, each of which is contrasted against the English reference group separately to offer clear and comprehensive estimates of DIF for each focal language group. These variations in the sample quintile are likely reflected in the ELOM-R Language (v1) test performances of children in each language. ELOM-R Language (v1) mean percent correct scores for each language are provided in Figure 4, and Total score statistics are in Table 11.

Table 10. ELOM-R Quintile Distributions in Each Language Sample for DIF Analysis



Confidence intervals (indicated by bars) are large for several languages indicating that these samples do not provide a precise representation of the language population mean. This high degree of variance may lead to less precise modelling estimates across the board, particularly those relying on variance partitioning methods such as omega (CTT reliability) and person reliability (IRT reliability).

Table 11 ELOM-R Language (v1) Percent Correct Statistics by Language

LANGUAGE	MEAN (%)	SD (%)	MIN (%)	MAX (%)	DIFFERENCE (TO ENGLISH)
ENGLISH	74.0	15.4	24.9	98.1	-
AFRIKAANS	69.8	17.7	0	100	-4.2
SESOTHO	60.1	18	4.5	99	-13.9
SEPEDI	64.3	16.8	16.4	97.9	-9.7
SETSWANA	62.1	17	16.9	96.5	-11.9
TSHIVENDA	59.9	15.6	5.9	92	-14.1
ISIXHOSA	70.4	16	25.8	99.3	-3.6
ISIZULU	59.8	17.1	16.9	98.6	-14.2

TESTING FOR A LANGUAGE – QUINTILE CONFOUND

As noted previously, where observed, one cannot assume the non-equivalence of the Language assessment across groups is due to language alone, as in some groups, it is confounded with our proxy measure of socio-economic status – school quintile. This will have an influence on DIF analyses particularly where item performance is modelled with child ability.

Testing for interaction between language and quintile was considered using MANOVA. However, as evident in Table 10 above, language group sample sizes were too small in the higher quintiles for four of the African languages and too small in the bottom two quintiles for Afrikaans and English, this was not undertaken. An ANOVA testing for quintile effects alone indicated that, overall, school quintile groups were significantly different ($F(4, 1033.67) = 25.80$, $p < 0.001$). However, post-hoc tests only reveal statistically significant differences between the mean Language score for quintile 5 and all other quintile groups). Overall, we can conclude that language and SES (quintile) are likely to be confounded for the Language Assessment, with quintile five children being particularly advantaged relative to others regardless of home language.

While the box plots and distribution characteristics in Figure 4 indicate differences at the raw score level, the DIF analysis that follows is intended to show whether these are due to genuine differences in ability level or differential item functioning. We reiterate our observation that as a single factor model for the ELOM-R Language (v1) Assessment has not been established, meaning DIF findings must be treated with caution.

DIF FINDINGS FOR LANGUAGE GROUPS

Each of the other languages is compared to English. Plots (Figures 6-12), based on percentage correct responses (PC scores), are provided for each language. DIF measures are shown on the vertical axis and represent item difficulty within the indicated language group. These language-specific difficulty estimates are shown per item reported across the

the horizontal axis in Rasch logit units. DIF effects over 1 logit (large DIF) are circled in black, and effects between 0.5 and 1 (Moderate DIF) are circled in grey. In all Plots, **blue** is the English reference language and **red** represents the compared focal language with which it is compared.

Figure 5. ELOM-R Language (v1) English – Afrikaans DIF Plot

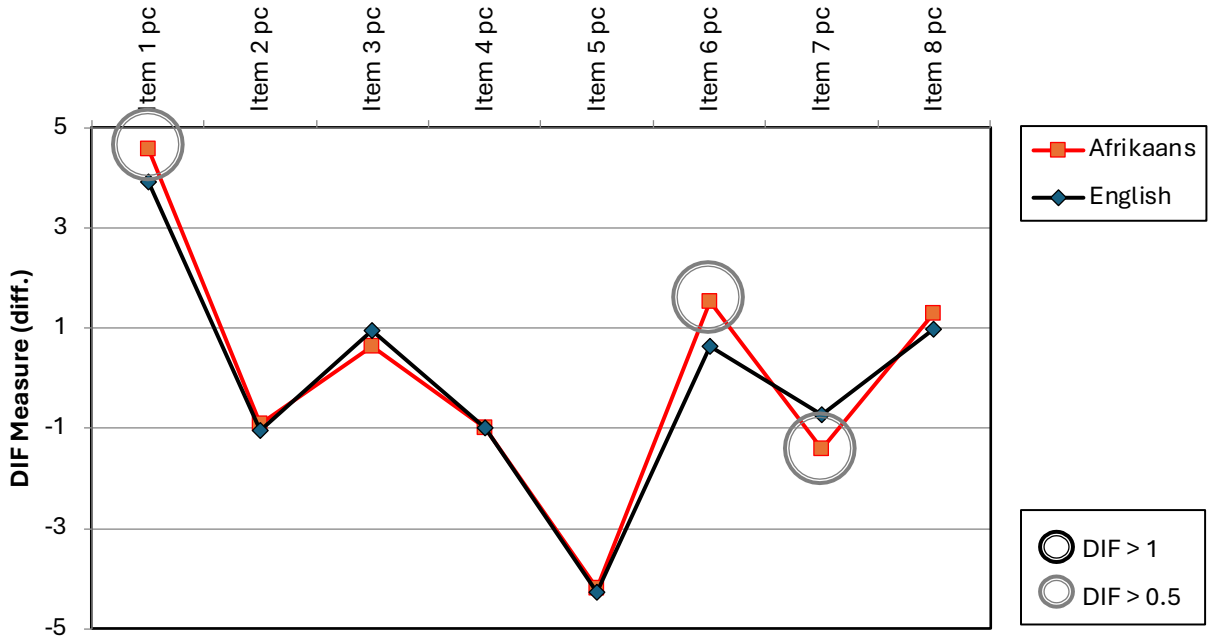


Figure 6. ELOM-R Language (v1) English – isiXhosa DIF Plot

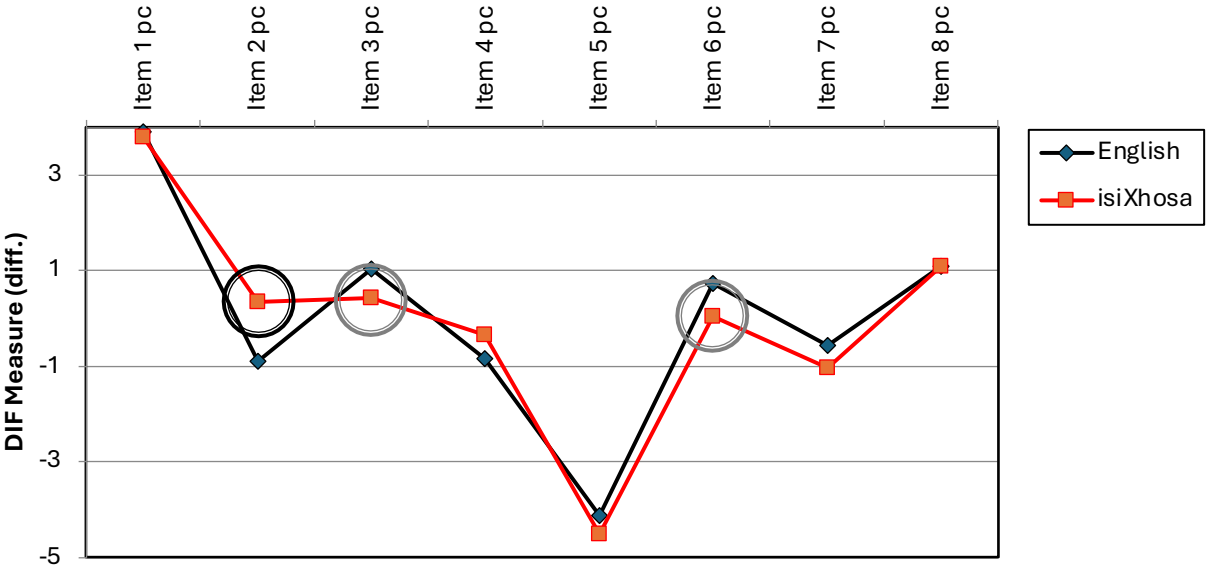


Figure 7. ELOM-R Language (v1) English– isiZulu DIF Plot

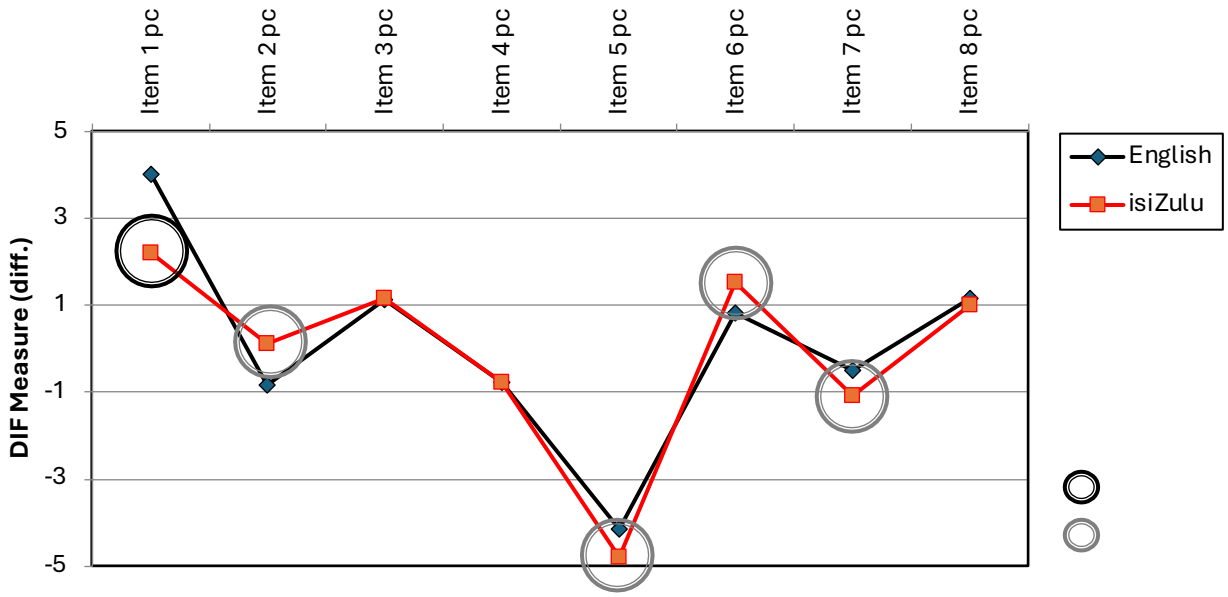


Figure 8. ELOM-R Language (v1) English – Setswana DIF Plot

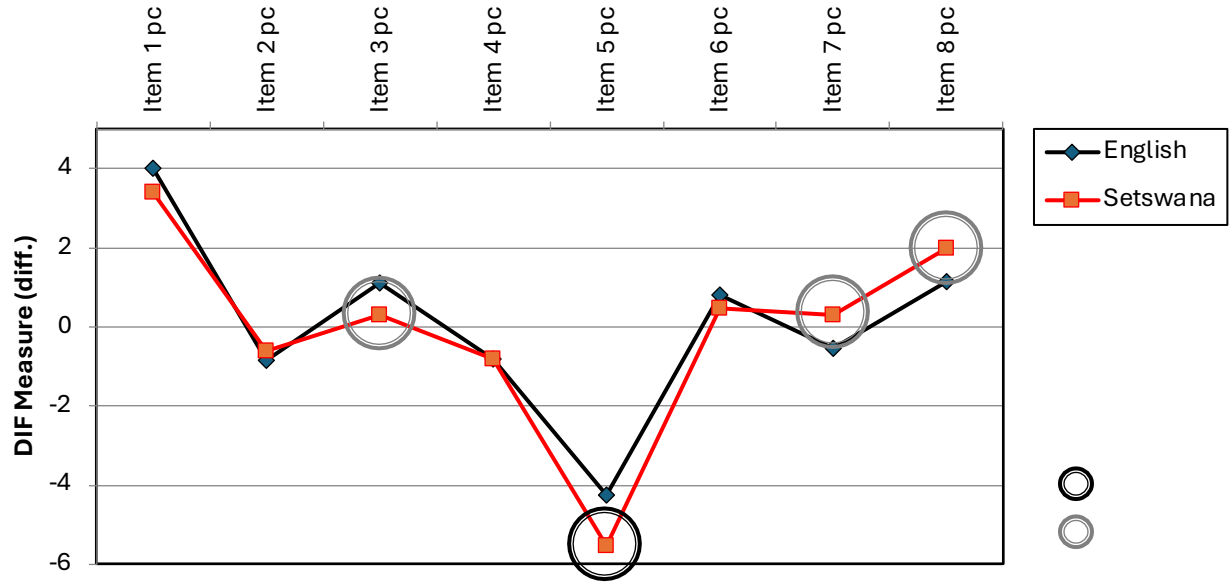


Figure 9. ELOM-R Language (v1) English – Sesotho DIF Plot

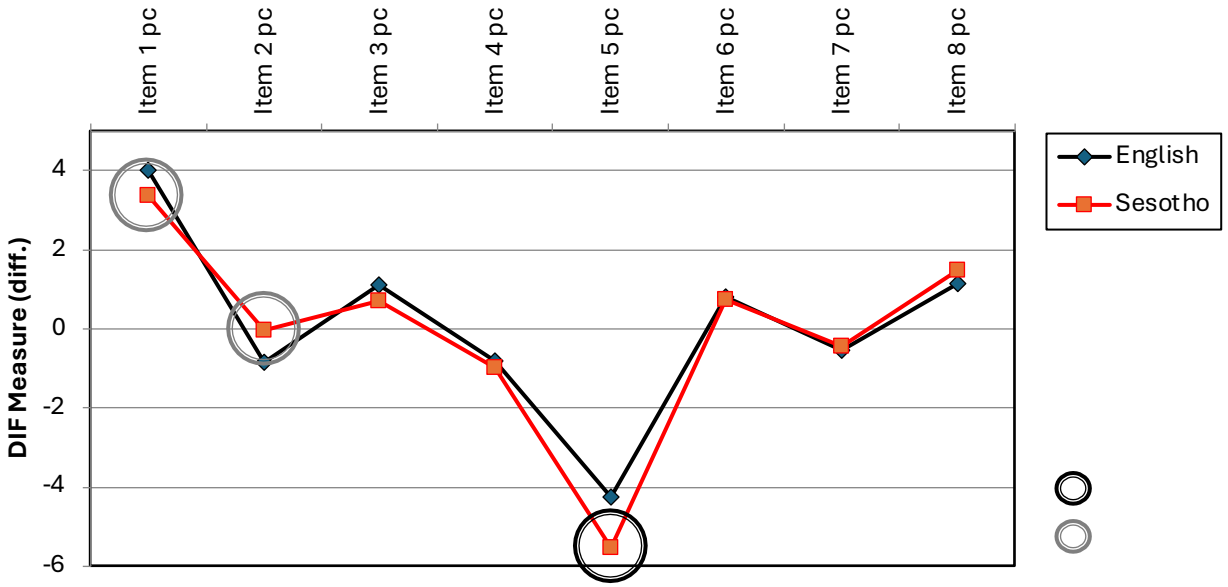


Figure 10. ELOM-R Language (v1) English – Sepedi DIF Plot

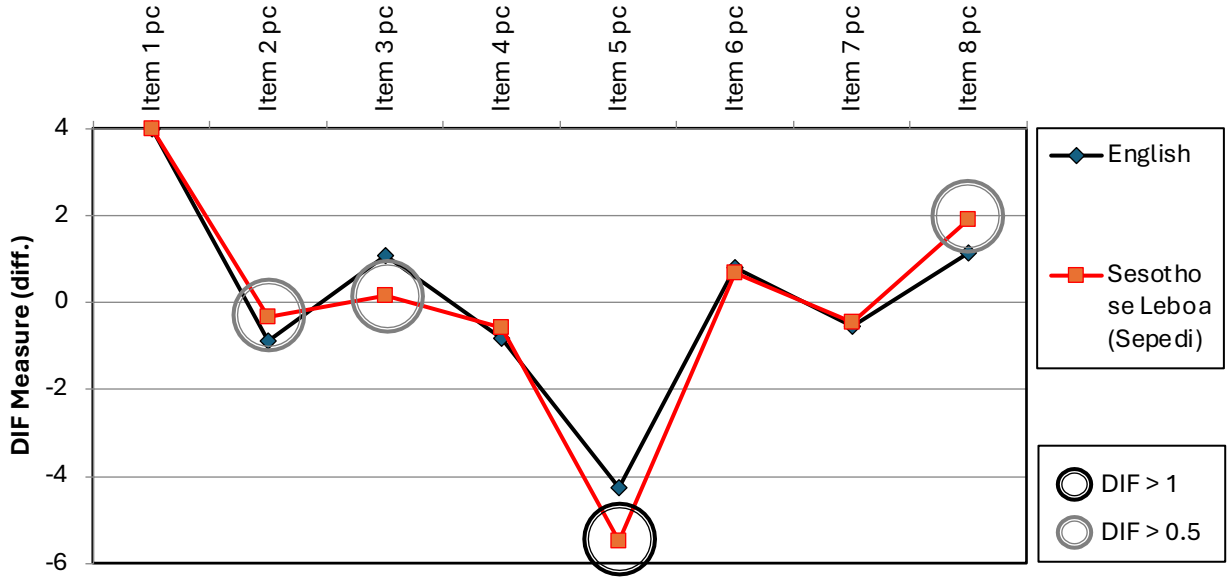
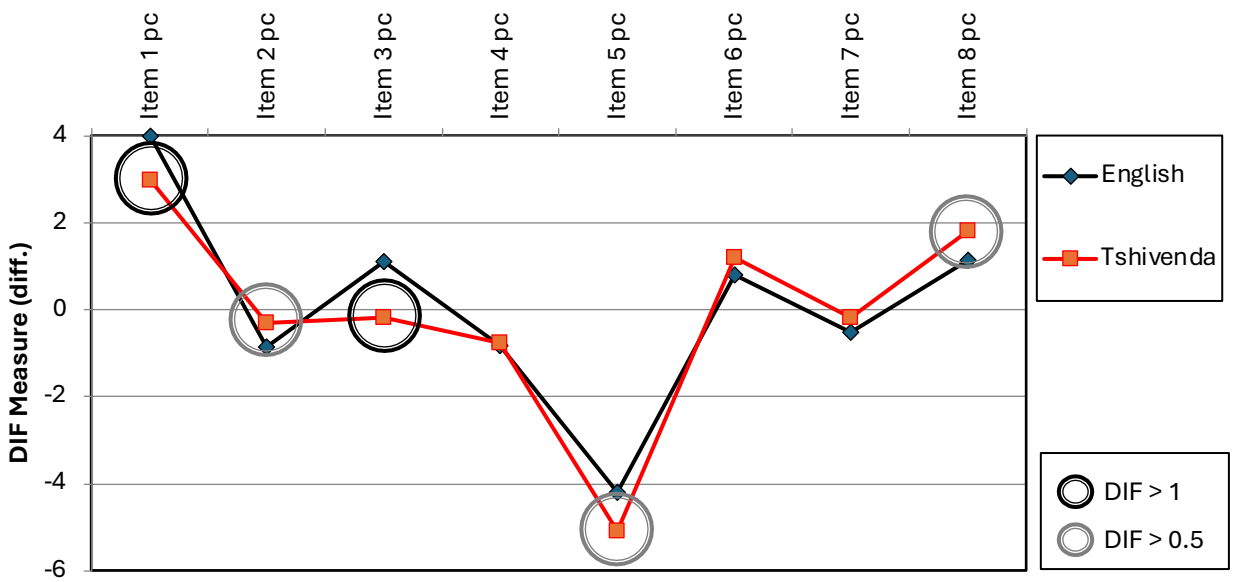


Figure 11. ELOM-R Language (v1) English – Tshivenda DIF Plot



Evaluation of ELOM-R Language (v1) DIF contrasts between each language and English revealed numerous mixed DIF effects (both positive and adverse bias at the item level), but these tend to balance out, culminating in predominantly small to moderate DIF effects at the level of the test as a whole (Test DIF). While these results may be skewed by the diffuse factor structure and lack of Rasch model fit, the findings suggest that DIF effects are minor when accumulated over the scale and favour non-English respondents on balance despite their much lower raw scores.

These results indicate a relatively fair test, considering the large differences in raw scores between language groups, although significant effects are identified at the item and test levels.

Significant DIF effects for each language against the English cohort on the Rasch-Welch t-test[41] are summarised in Table 12. Omitted values were not statistically significant. Positive effects indicate bias in favour of English. Negative effects indicate bias in favour of non-English groups.

Values printed in **black** represent small to moderate DIF effects, while values printed in red represent moderate to large DIF. DIF estimates are reported according to their original item difficulty values - higher DIF means higher item difficulty (and lower ability) in the non-English groups.

[41]The Rasch-Welch t test compares Rasch model difficulty estimates between groups. Higher negative numbers indicate potential bias in favour of each language; higher positive numbers indicate potential bias in favour of English (the reference group).

Table 12. ELOM-R Language (v1) DIF Contrasts by Focal Language

ITEM		AFRIKAANS	ISIXHOSA	ISIZULU	SETSWANA	SESOTHO	SEPEDI	TSHIVENDA
	DIF ACCUMULATION	**0.21**	**-0.05**	**-0.76**	**-0.39**	**-0.44**	**-0.86**	**-1.61**
1	Productive vocabulary			-1.81				
2	Beginning sounds		1.22	0.94		0.83	0.53	0.54
3	Letter sounds		-0.60		-0.80		-0.93	-1.27
4	Copying shapes		0.49					
5	Write name			-0.67	-1.27	-1.27	-1.23	-0.88
6	Writing with encouragement	0.90	-0.69	0.69				
7	Listening comprehension	-0.69	-0.47	-0.58	0.84			
8	Book concept, orientation, and word concept				0.84		0.77	

When compared with English, findings that stand out are:

- Productive vocabulary is particularly easy for isiZulu children. However, as has been noted in the discussion of Table 12, it has been necessary to adjust the trial order for this item.
- Beginning sounds is more difficult for isiXhosa children (and to a lesser extent in most other African languages).
- Letter sounds are particularly easy for Tshivenda speakers who find this assessment less difficult overall than children speaking other languages.
- Write name is more difficult in English than in the Setswana, Sepedi and Sesotho groups (notably more lower quintile children).

Overall, these observations suggest differences in phoneme awareness ability between English and some African languages. This may be a function of the differences in phoneme structure between English and these languages as well as the manner in which this is measured in the ELOM-R Language (v1) Assessment. This warrants further investigation.

ITEM DIFFICULTY COMPARISON ACROSS LANGUAGES

All languages were combined to assess ELOM-R Language (v1) item difficulty (known as Omnibus DIF). Estimates are reported in Table 13, with highlighting to indicate their relative difficulty. Red highlighting indicates that the item is more difficult for respondents within the language (in the top row), while blue highlighting indicates that the item is easier.

Table 13. ELOM-R Language (v1) Omnibus DIF Measures

ITEM	ENGLISH	AFRIKAANS	ISIXHOSA	ISIZULU	SETSWANA	SESOTHO	SEPEDI	TSHIVENDA	RANGE
1. Productive vocabulary	4.03	4.57	3.89	2.21	3.43	3.38	4.02	3.01	2.36
2. Beginning sounds	-0.83	-0.70	0.41	0.10	-0.59	0.00	-0.30	-0.30	1.24
3. Letter sounds	1.13	0.84	0.50	1.17	0.33	0.72	0.20	-0.16	1.32
4. Copying shapes	-0.79	-0.79	-0.30	-0.81	-0.79	-0.97	-0.53	-0.75	0.67
5. Write name	-4.23	-4.14	-4.68	-4.93	-5.50	-5.50	-5.46	-5.14	1.36
6. Writing with encouragement	0.83	1.70	0.11	1.51	0.49	0.77	0.73	1.23	1.59
7. Listening comprehension	-0.51	-1.20	-1.02	-1.11	0.33	-0.42	-0.42	-0.16	1.53
8. Book concept, orientation, and word concept	1.16	1.46	1.19	0.99	2.00	1.49	1.94	1.83	1.01

Most of the item difficulty estimates range over a logit across language groups, although rough consistency in item difficulty estimates is indicated by the monotone shading for most items. The smallest range amounts to 0.67 logits for item 4 (Copying shapes), and the largest to 2.36 logits for item 1 (Productive vocabulary). This was the most difficult item across all language groups and it is likely due to trial order (since adjusted; see Table 12 below). The easiest in all is item 5 (Write name). In contrast to the ELOM-R Mathematics (v1) assessment, there is no graduation in item difficulty from items 1 to 8.

MODIFICATIONS TO ITEM 1: PRODUCTIVE VOCABULARY

Regarding item 1 (Productive vocabulary), we believe that the ordering of its composite trials may have contributed to its inflated difficulty estimate. Children are required to name the objects presented in a series of 36 pictures, and a stop rule is applied after eight incorrect responses. The item's difficulty will, therefore, be inflated if more difficult trials appear early (causing early stoppage), and this effect may differ between languages depending on the challenge level presented by trial word translations. DIF analyses on the 36 trials that comprise this item found considerable variations in trial difficulty across the languages which very likely contributed to the high difficulty estimates for this item. Trial DIF was investigated, and the results are shown in Table 14. It displays trial difficulty in each language based on the percentage of children who passed the trial. The numbers in each cell refer to the rank order of the trial (percentage correct) in each language. For example, Trial Picture 4 (Red) is the easiest in English (Rank 1) and no others, while Trial Picture 9 (Digging) is easiest in isiXhosa and Sepedi only.

Table 14. ELOM-R Productive Vocabulary Trial Difficulty in Each Language

TRIAL PICTURE PRESENTATION ORDER		ENGLISH	AFRIKAANS	ZULU	XHOSA	SESOTHO	SETSWANA	SEPEDI	TSHIVENDA
		281	442	280	291	287	277	282	291
		PICTURE	PICTURE	PICTURE	PICTURE	PICTURE	PICTURE	PICTURE	PICTURE
1	Bus	4	7	27	30	21	7	7	7
2	Flower	2	30	20	27	2	27	27	27
3	Feather	11	19	25	19	20	1	30	25
4	Red	1	4	30	12	7	2	25	30
5	Jumping/ leaping	7	2	7	20	27	17	20	1
6	Fly	27	11	1	4	19	30	2	20
7	Box	22	18	21	18	1	13	17	21
8	Happy	17	1	2	25	12	19	21	2
9	Digging	30	24	17	1	17	25	1	12
10	Throwing	20	17	19	21	22	28	19	24
11	Yellow	19	20	4	16	24	11	14	36
12	Radio	18	22	13	11	9	18	13	9
13	Cow	25	21	11	13	30	24	22	35
14	Drawing	21	28	18	7	13	21	24	17
15	Duck	24	25	22	2	28	34	9	14
16	Drum	13	27	32	17	32	5	12	19
17	Carrot(s)	8	5	9	28	25	20	18	13
18	Carpet	15	8	24	32	18	36	23	22
19	Mouse	5	13	15	5	4	4	34	4
20	Aeroplane	16	29	23	15	5	14	35	18
21	Swinging	28	33	14	24	11	35	4	5
22	Monkey	29	10	5	26	31	8	11	32
23	Rubbish	9	31	12	33	15	9	8	8
24	Cloud(s)	33	14	35	22	6	15	32	28
25	Pencil	10	6	8	31	16	12	36	23
26	Elbow	12	23	16	6	36	23	5	34
27	Umbrella	26	26	33	23	35	6	3	11
28	Tortoise	6	32	28	9	23	22	16	10
29	Giraffe	32	12	6	29	8	10	10	3
30	Swimming	14	9	31	10	3	33	31	16
31	Buton	31	35	10	35	14	16	28	15
32	Ladder	3	3	29	36	33	32	15	31
33	Curtain	35	16	36	14	29	3	33	6
34	Thorn(s)	23	36	3	8	34	31	6	29
35	Peeling	36	15	34	34	10	29	29	33
36	Raking	34	34	26	3	26	26	26	26

PRODUCTIVE VOCABULARY ADAPTATION

These variations in trial difficulty have been considered in adjustments to the ELOM-R Language (v1) Assessment by reordering trials from easiest (Rank 1) to most difficult (Rank 36) in each language.

ALTERNATIVE SYNONYMS FOR PRODUCTIVE VOCABULARY (ITEM 1)

To be scored correctly on a productive vocabulary trial, a child has to use the specific word shown in Table 14. However, we observed that some children used different words to describe the object or action depicted in the productive vocabulary trial pictures. In such cases, they would score 0 on the trial. We were concerned that this might lead to bias in the measurement of productive vocabulary in some languages.

To explore the frequency with which alternative words were produced, assessors were given three options to score the child's response when presented with an image: a) correct - the target word for the image, b) no response, c) an alternative word to the target was used. These alternative words were then recorded during the assessment.

We compiled a comprehensive list of alternative words provided by children in response to trial pictures. Where alternatives occurred in between 5-10% of child responses to a specific trial image in one of the languages, we reasoned that the alternative might be in relatively common usage among speakers of that language. We consulted language experts to ascertain whether the alternative word used would be in common usage and acceptable in that language. If so, and to reduce measurement error (bias), it was decided that the alternative response to the image should be scored correctly even if it differed from the target word for a correct answer.

Careful scrutiny of alternative word usage was necessary as spelling errors by the assessor were sometimes observed. For example, in the English administration of trial 36, 'sweeping' and 'swipping' were both recorded as children's responses to the image of a person raking (the correct response). 'Swipping' was the assessor's spelling error rather than another alternative word, so not accepted. Alternative words were also interrogated for accuracy. For example, although 6% of English children and 9% of Sesotho children called an image of a giraffe, a zebra, this is incorrect, and the alternative was not credited. Across many of the languages (25% Setswana, 16% isiXhosa, 12% Tshivenda), the image of a monkey prompted children to say baboon. Given its frequency, baboon was added as an acceptable alternative. Stretch words were maintained. One example is item 36 'raking'. Although 32% Afrikaans, 15% Sesotho, 12% Setswana, and 8% isiZulu children said the boy was sweeping, the target word remains raking.

Finally, an additional prompt was created for item 26 (image of an elbow), as so many children across languages responded "arm". The task was clarified in the assessor's instructions by the assessor saying to the child: "Yes, that is an arm, but what part of the arm is the arrow pointing to?" Alternative synonyms judged to be in common usage in specific languages have been incorporated in the ELOM-R (v1) tablet and are scored as correct should the child use them. These are displayed in Table 15.

Trial Order	English	Afrikaans	Sesotho	isiZulu	Tshivenda	Setswana	Sepedi	isiNdebele	siSwati	Xitsonga	isiXhosa
1	Bus	Bus	Bese	Ibhasi	Bisi	Bese	Pese	Ibhesi	Ibhasi	Bazi	Ibhasi
2	Flower, Sunflower, Daisy	Blom, Sonnebloem, Madeliefie	Palesa, Sonobolomo, Madeliefie	Imbali, Ubhekilanga	Ḽiluvha (mulivhaḓuvha)	Lelomo, sethunya	Letšoba (sonopolomo)	Ithuthumbo	Imbali	Xiluva, blomu, bilomu	Intyatyambo,yiflawa, liflawa
3	Feather	Veer, voelveer	Lesiba	Uphaphe	Muthenga, tari	Lefofa, lephuka, lebowa	Lefofa	Isiba	Lusiba	Risiva	Usiba
4	Red	Rooi	Kgubedu	Kubomvu	Tswuku	Khibidu	Khubedu	Bomvu	Ubovu	Tshwuka	Bomvu
5	Jumping, leaping	Spring, Springende	Qhoma, Tlola	Uyagxuma, jomba	Thamuwa, fhufha	O a tlola	Taboga / Tlola	Ukweqa /ukuhluza, fofa	Uyazuba	Ku tlula	Uyatsiba / uyaxhuma xhuma, uyajumpa
6	Fly	Vlieg, brommer	Tshintshi	Impukane	Thunzi	Ntsi	Ntšhi	Ipukani	Imphungane	Nhongani	Impukane
7	Box	Boks, Kartondoos	Lebokose	Ikhathoni, ibhokisi	Bogisi	Lebokoso	Lepokisi	Ibhoksi	Libhokisi	Bokisi	Ibhokisi, ikhadibhodi
8	Happy	Gelukkig, snaaks, bly, lekker	Thabile	Bajabule	Dakalo	Itumetse, o ikutlwa monate, go monate	Thabile	Bathabile	Bajabulile	Tsaka, kahle	Bonwabile, baziva kakuhle, kamnandi, bayavuya
9	Digging	Grawe, skep skoffel, spit	Moshemane wa tjheka, moshemane o etsa mokoti, epa	Uyagubha, uyemba, wenza umgodi	U bwa, u shuma nga tshipeidi, fukula	Go epa	Epa	Uyemba, lema	Uyasebenta/ Uyagubha/ Uyemba	Ku cela/ku tirha hi foxolo	Iyomba/ Isebenzisa umhlakulo, iyagrumba
10	Throwing, bowling	Gooi, boul	Akgela, lahlela, betsa	Uyaphonsa, uyalahla	Posa	Go latlhela	Foša	Ukuphosa / Uyaphosa	Uyajikijela	Ku hox	Uyajula/uyagibisela
11	Yellow	Geel	Tshehla	Umbala ophuzi	Tada, thopi	Serolwana	Serolane	Sarulana	Umtfubi	Xitshopani	Utyheli
12	Radio, speaker	Radio, luidspreker, draadloos, speeker	Seyalemoya, Sepekara	Irediyo, Umsakazo, isipikha	Radio, tshipikara	Radio/ Seyalemowa, sepekara, radiyo	Seyalemoya	Umrhatjho	Umsakato /irediyo/ iwayilesi	Rhadiyo/xiyanimoya	Iradiyo/unomathotholo/isipika
13	Cow	Koei, Bees, bul	Kgomo	Inkomo	Kholomo	Kgomo	Kgomo	Ikomo	Inkhomo	Homu	Inkomo
14	Drawing, colouring in	Teken, in kleur	Taka	Uyadweba	U ola, ḓirowa	Go taka/tshwantsha	Thala	Ukugwala/ Ukukhrayona/ Ukutlola	Inkhomo, uyadvweba, uyakhrayona	Ku dirowa/ku khalara/ ku tsala	Iyazoba
15	Duck	Eend, gans	Letata	Idada	Sekwa	Pidipidi	Lepidipidi	Idada	Lidada	Sekwa	Idada, irhanisi

Trial Order	English	Afrikaans	Sesotho	isiZulu	Tshivenda	Setswana	Sepedi	isiNdebele	siSwati	Xitsonga	isiXhosa
16	Drum, bongo	Trom, drom	Moropa, sekupu	Isigubhu, amadra-mu	Ngoma	Moropa	Moropa, sekupu	Isigubhu	Sigubhu	Xigubu	Igubu
17	Carrot(s)	Wortel(s)	Sehwete, dihwete	Izaqathi /ukherothi	Kherotsi	Segwete/Digwete	Kherotse/Segwete	Ikherothi	Ticadze/ Emakherothi	Kheroto/tikheroto	Umnqathe/imin-qathe
18	Carpet, mat, rug	Tapyt, mat	Khapete, mmata	Ukhaphethi, umata	Khaphethe, Tho-vho, Methe	Khapete/mmetshe	Khapete / Mmetse	Umada	Limethi	Khapete/mete	Imethi/ ikhaphethi
19	Mouse, rat	Muis, rot	Tweba, kgoto	Igundane /ibuzi	Mbevha	Peba /legotlo	Legotlo /Peba	Ikhondlo	Ligundvwane	Kondlo/nthanyani	Impuku
20	Aeroplane, plane	Vliegtuig, plane, jet	Sefofane	Indiza, indizamshi-ni, ibhanoyi	Ṱharabuḽeni/ Bufho	Sefofane	Sefofane	Isiphaphamtjhini	Indiza/ indizam-shini	Xihahampfhuka/jete	Inqwelo moya/ ieropleyini, playini
21	Swinging	Swaai	Bapala ka moswinki	idlala uzwingi	U dembelela, U devhuwa	Go swinka	Kadiela/raloka/ swinka	Ukujinka, se-kokoromeiye, mo/ me/ma/di/swinki	Uyajikela	Jombha/tlanga	Iyajinga
22	Monkey, baboon	Aap, Apie, bob-bejaan	Tshwene,	Inkawu, imfene	Ṱhoho, pfene	Kgabo	Kgabo, tšhwene	Ifene, tshwene, indwangu	Ingobiyane, imfene	Nkawa/Ritoho, mfenhe	Inkawu, mfene
23	Rubbish, gar-bage, trash	Rommel, Vullis, rubies, vuilgoed, gemors	Moqomo wa matlakala	Umgqomo kadoti, udoti	Mathukhwi, Bini ḽa mathukhwi	Matlakala	Matlakala/ Setšhelamatlaka-la/Tasbini	Iinzibi/Umgqomu weenzibi	Tibi/ umgcoma wetibi/idasbin	Thyaka/thini ro chela thyaka/dasbin	Inkunkuma/ Umgq-omo wenkunkuma, udothi lintwezi mdaka
24	Cloud(s)	Wolk(e)	Maru	Amafu	Gole (Sumbani kha gole hu si makoleni)	Maru	Maru	Amafu	Emafu	Mapapa/Papa	Ilifu
25	Pencil	Potlood, pottie	Pensele	Ipensela	Penisela	Phensele	Phensele	Ipensela	Ipenseli	Penisele	Ipensile
26	Elbow	Elmboog	Setsu	Indololwane	Lukuḓavhavha (kha vha vha tendele u sumba lukuḓavhavha hu si tshanḓa.)	Sekgono	Sejabana	Indololwana	Ingcosa	Xikokola	Ingqiniba
27	Umbrella	Sambreel	Sekgele/ sam-borele	Isambulela	Tshasambureni	Sekhukhu / Sekgele	Samporele	Isambreni	Sambulelo	Xiambhulele	I-ambrela
28	Tortoise, turtle	Skilpad	Kgudu/sekolopata	Ufudu	Tshibode	Khudu	Khudu	Ikghuru	Lufudvu	Xobodze/Futsu	Ufudo
29	Giraffe	Kameelperd, lang-nekke	Thuhlo	Indlulamithi	Ṱhuḓwa	Thutlwa	Thutlwa	Idlulamithi	Indlulamitsi	Nhuntlwa/Jirafu	Indlulamthi

Trial Order	English	Afrikaans	Sesotho	isiZulu	Tshivenda	Setswana	Sepedi	isiNdebele	siSwati	Xitsonga	isiXhosa
30	Swimming	Swem	Sesa	Bayabhukuda	U bambela	Go sapa/Go thuma	Rutha	Ukududa	Bayabhukusha	Ku khida/Ku hlambela	Bayaqubha/baya-dada
31	Button	Knoop	Konopo	Inkinobho	Gunubu	Konopo, talama	Konopi	Ikunubhe	Likinobho	Kunupu	Iqhosha
32	Ladder	Leer	Leri/Setepisi	Isitebhisi	Ḽeri	Llere	Llere	Ileri	Lilele/liladi/sitepisi	Lerhe/Xitepisi	Ileli
33	Curtain	Gordyn	Kgaretene	Ikhethini	Khetheni	Garetene	Garetene	Amarharideni	Likhethini	Kheteni	Ikhethini
34	Thorn(s)	Doring(s), pen-doring	Tshehlo/Meutlwa	Ameva	Mupfa	Mmitlwa	Mootlwa	Ameva	Linyeva	Mintwa/mutwa	Ameva, nkunza ne
35	Peeling	Skil, Afskil	Ebola	Uyahluba	U vhaḓa (U khou ita mini)	Go obola	Ebola	Ukukela	Ucata lihhabhula, uvula libhanana	Ku vandla	Iyachuba, ixobula ibanana
36	Raking	Hark	Haraka	Uyahhala	U haraga	Go haraka	Haraka	Kuhariga	Uyahhaliga	Ku kukula	Iyaharika, Iyahaki-sha, Iyareyikha

CONCLUSION

While reliability in all languages is sound, CFA and Rasch analyses did not establish clear construct validity in the present eight-item set of the ELOM-R Language (v1) Assessment. As dichotomised scores represent an oversimplification of ELOM Language responses, other Rasch modelling procedures were attempted, including the Partial Credit Model (PCM). However, this was not successful for the reasons provided above. For future analyses, a hybrid approach wherein items with similar scale properties are treated as separate testlets is being considered, but it may require the development of more items to ensure each testlet contains sufficient items for parameter estimation.

Differential item functioning analyses showed that some item difficulties vary across languages, so that measurement equivalence is not established. However, as we have noted, item structure (trials and stop rules) has led to challenges in these analyses. Socioeconomic status (SES), as indicated by the school quintile proxy, is also highly likely to have played a role here as it influences language development. As we have noted, SES and language are confounded, and it is impossible to separate their effects.

Furthermore, efforts are currently underway to establish the criterion validity of the ELOM-R Language (v1) Assessment by examining the regression between ELOM-R Language (v1) Scores collected in Grade R and the Grade 1 Early Grade Reading Assessment (EGRA). Theoretically, high scores on the ELOM-R Language (v1) should translate to higher Early Grade Reading Assessment in Grade 1. It would be desirable to establish concurrent validity with another language test designed for this age group.

CHAPTER 3. STANDARDISATION AND NORMS

In this chapter, we present psychometric analyses undertaken on a combined sample of eight languages to standardise the ELOM-R Language (v1) and derive norms that can be used to compare the performances of groups of children regardless of language.

Standardisation Sample

As noted previously, isiNdebele, Siswati, and Xitsonga languages have been excluded as their samples were too small. The standardisation sample is provided in Table 16.

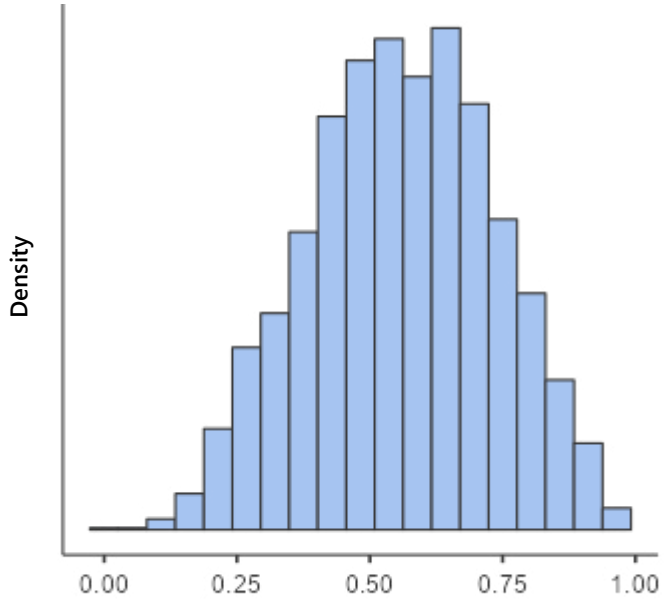
Table 16. ELOM-R Language (v1) Standardisation Sample for Standardisation and Norms

Home Language	Total	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
1. English	281	13	34	118	47	69
2. Afrikaans	447	86	82	37	141	101
3. isiZulu	280	43	55	62	81	39
4. isiXhosa	291	23	74	102	57	35
5. Sesotho	289	68	64	76	46	35
6. Setswana	277	241	0	21	8	7
7. Sepedi	282	214	13	23	16	16
8. Tshivenda	292	100	63	109	20	0
TOTAL	**2439**	**788**	**385**	**548**	**416**	**302**
Final Total after exclusion of outliers	**2431**					

As we have already noted, the underrepresentation of quintile 4 and 5 children in some languages will affect findings. And it is worth mentioning once more that language and quintile are confounded.

First, the distribution of total scores on the assessment is investigated. Note that item-level scores are reported as the percentage of correct responses to trials comprising test items (PC scores). Test scores are calculated based on these percentage scores, yielding a decimal scale ranging from 0 to 1. The histogram of total PC scores across the sample is presented in Figure 12, which reveals a symmetrical distribution.

Figure 12. ELOM-R Language (v1) Standardisation Sample Mean Percent Correct Score Distribution



Descriptive statistics including the range, central tendency, and shape of the distribution are presented in Table 17.

Table 17. ELOM-R Language (v1) Total Percent Correct Score Descriptive Statistics

N	MEAN	MEDIAN	SD	MINIMUM	MAXIMUM
2431	0.656	0.665	0.173	0.164	1.00

SKEWNESS		KURTOSIS	
SKEWNESS	SE	KURTOSIS	SE
-0.271	0.050	-0.582	0.099

Skewness is statistically significant. However, the value is below the threshold for meaningful distortion of the distribution, and it is reasonable to proceed with standardisations. The final standardisation group comprises 2431 cases.

ELOM-R LANGUAGE (v1) STANDARDISATION SAMPLE SCHOOL QUINTILE DISTRIBUTIONS[42]

As scores are normalised across South Africa's diverse population, language groups and socioeconomic status (SES) are reported. While both group designations are important to consider, as previously noted, they

[42]Quintile ranks are assigned to public schools in South Africa roughly according to the relative poverty levels of the population they serve, aggregated over an area within three kilometres of the school. Quintile 1 schools serve children in the poorest areas, while quintile 5 schools serve the wealthiest. Ranks are predominantly based on the income, education level and unemployment

are heavily confounded in South Africa (Laher et al., 2019). The school quintile composition of each language group is reported in Figure 13 to provide context for consideration of confounding effects. SES is operationalised in terms of the quintiles assigned to the schools from which children were sourced. These are further collapsed in Figure 13 into schools that do not require the payment of fees (quintiles 1, 2, and 3), and those that do (quintiles 4 and 5).

Figure 13. ELOM-R Language (v1) Standardisation Sample School Quintile Distributions

Fee paying schools predominated for the Afrikaans cohort alone, with comparable proportions of paying and non-fee-paying schools in the English and isiZulu samples. Fee paying isiXhosa and Sesotho schools are well outnumbered by non-paying schools, while very low to negligible proportions of Sepedi, Setswana, and Tshivenda schools pay fees. The standardisation sample school quintile composition is reported in Table 18 where the language / quintile confound is quite evident.

Table 18. ELOM-R Language (v1): Quintile Frequencies by Language*

School Quintile	Afrikaans	English	Sesotho	Sepedi	Setswana	Tshivenda	isiXhosa	isiZulu
1	83	13	68	214	241	100	23	43
2	82	34	63	13	0	62	74	55
3	37	118	75	23	21	109	102	62
4	139	47	46	16	8	20	57	81
5	101	69	35	16	7	0	35	39
Not Paying Fees (Q 1 – 3)	202	165	206	250	262	271	199	160
Paying Fees (Q 4 & 5)	240	116	81	32	15	20	92	120

**Modal values indicated in red text*

Table 18 reveals very different school quintile distributions across the language groups. Sepedi, Setswana, and Tshivenda feepaying ELPs are poorly represented, and SES effects are likely to heavily influence test performances in these groups. The quintile frequencies suggest that the Sesotho cohort may be less affected than the Sepedi or Setswana, as they possess far greater numbers of quintile 2 and 3 ELPs. Subsamples for paying and non-paying ELPs for all other language groups appear reasonably well populated.
Next, the psychometric properties of the ELOM-R Language (v1) Assessment within the norm sample are assessed to establish the reliability and validity of its scale scores.

Psychometric Properties of the ELOM-R Language (v1) Standardisation Sample

RELIABILITY

To assess whether the ELOM-R Language (v1) items are consistent in their measurement of numerical ability across all the subsamples included in the overall norm, reliability testing procedures were undertaken. Reliability of the assessment was tested using McDonald's omega (ω), which assesses the internal consistency of assessment scores. Results are presented in Table 19.

Table 19. ELOM-R Language (v1) Reliability Statistics

	Item-rest correlation	ω
ELOM-R (v1) Language Assessment	83	13
When item excluded...		
1 Productive vocabulary	0.315	0.761
2 Beginning sounds	0.615	0.708
3 Letter sounds	0.641	0.706
4 Copying shapes	0.352	0.756
5 Write name	0.326	0.760
6 Writing with encouragement	0.541	0.725
7 Listening comprehension	0.363	0.754
8 Book concept, orientation, and word concept	0.529	0.726

All values exceed the acceptable threshold (0.70), but items 2 and 3 are marginal. These are Phoneme Awareness items and reliability may be affected by the different phonetic structures of the African and Germanic languages (English and Afrikaans) to which we have drawn attention in ELOM-R (v1) Technical Manual 1.

No items produce sub-threshold item-rest correlations ($r > 0.3$) or detract from scale reliability (ω when item removed < 0.763). The ELOM-R Language (v1) Assessment can be considered a reliable measure within the norm group.

Next, a confirmatory factor model (CFA) analysis is fitted to the norm sample to assess construct validity for the language assessment within this cohort.

CONFIRMATORY FACTOR ANALYSIS (CFA)

As in earlier sections, a unidimensional factor model was specified, and the fit statistics in Table 20 describe the model's fit to the observed data. Factor loadings of individual items to the single factor are evaluated to assess potential misfit at the item level. CFA loadings are presented in Table 21.

Table 20 ELOM-R Language (v1) Assessment CFA Model fit

χ^2	Df	P	CFI	TLI	RMSEA	Lower CI	Upper CI
468.06	20	< .001	0.889	0.845	0.096	0.089	0.104

Table 21 ELOM-R Language (v1) Assessment CFA Model factor loadings

Item	Estimate	SE	Z	P	λ
1 Productive vocabulary	0.050	0.003	16.072	< .001	0.351
2 Beginning sounds	0.264	0.007	36.875	< .001	0.719
3 Letter sounds	0.248	0.006	39.68	< .001	0.767
4 Copying shapes	0.095	0.006	17.164	< .001	0.374
5 Write name	0.089	0.005	16.828	< .001	0.365
6 Writing with encouragement	0.240	0.007	33.301	< .001	0.665
7 Listening comprehension	0.088	0.005	17.923	< .001	0.393
8 Book concept, orientation, and word concept	0.151	0.006	26.835	< .001	0.562

Model misfit is evident for the single factor model (RMSEA = 0.096, CFI = 0.889, TLI = 0.845). And while all items load saliently ($\lambda > 0.3$, $p < .001$), model misfit indicates that the construct validity of the ELOM-R Language (v1) Assessment is not clearly established. Covariances are reported in Table 22.

Table 22. ELOM-R Language (v1) Standardisation Scale Observed/Residual Covariances

ITEMS	ITEM NUMBER							
	1	2	3	4	5	6	7	8
Productive vocabulary	—	0.006	-0.009	-0.025	0.003	-0.065	0.068	0.078
Beginning sounds	0.258	—	0.011	0.013	0.005	-0.025	0.012	-0.007
Letter sounds	0.260	0.562	—	-0.043	0.018	0.073	-0.072	-0.066
Copying shapes	0.106	0.282	0.244	—	0.106	-0.045	0.039	0.074
Write name	0.131	0.268	0.298	0.242	—	-0.038	-0.062	-0.017
Writing with encouragement	0.168	0.453	0.582	0.203	0.205	—	-0.055	-0.013
Listening comprehension	0.206	0.294	0.229	0.186	0.081	0.206	—	0.200
Book concept, orientation, and word concept	0.275	0.397	0.365	0.284	0.189	0.360	0.420	—

Weaker patterns of covariance are observed for item 1 (Productive vocabulary) and item 5 (Write name), possibly due to individual language trial order and stop rules noted previously. Residual covariances are predominantly negligible, but potential local dependence is indicated between items 7 (Listening comprehension) and 8 (Book concept, orientation, and word concept) (r = 0.200).

CFA allows for such local dependencies to be controlled for and quantified in terms of modification indices, which represent the improvement (amount of decrease) in chi-square (an absolute measure of model fit) should the local dependency between two items be accounted for (by allowing the items to covary within the CFA model). The modification index describing this relationship is 199.20, which far exceeds the threshold for statistical significance of 3.84. Model fit statistics are acceptable when this shared variance is specified (RMSEA = 0.073, CFI = 0.939, TLI = 0.910). This means that the ELOM-R Language (v1) construct fits the CFA modelling (albeit marginally).

While the ELOM-R Language (v1) Assessment items share sufficient common variance to produce reliable scale scores, we have not established that a single latent construct underlies it. Factor analytic methods revealed added dimensionality and item dependencies, suggesting that a more complex model may be needed to describe the ELOM-R Language (v1) construct. However, additional factors did not explain sufficient variance in the current set of indicators to warrant a multifactor model. Additional underlying factors could be evident if more items were added to the assessment that tap the CAPS Drawing and Emergent Writing and Understanding of Print in particular. This will be investigated further in subsequent versions of this instrument.

With these limitations, the ELOM-R Language (v1) Assessment can be considered a reliable scale, making it possible to construct norms for total scores on the measure. However, it must be noted that the construct validity of the measure was not established in this standardisation sample.

Standardisation

As the ELOM-R Language (v1) Assessment was designed to test the achievement of children exiting Grade R / entering Grade 1 across a highly diverse population, it is important to establish clear, meaningful score distributions. This was achieved using normalisation and standardisation techniques (Cohen et al., 1996[43]; Kline, 2000).

[43]Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). Psychological testing and assessment: An introduction to tests and measurement, 3rd ed (pp. xxviii, 798). Mayfield Publishing Co.

Normalisation involves transforming raw scores into standard (Z-scores) such that they are:
a) centred on 0 according to the population mean, and
b) scaled according to the spread (standard deviation) of data around the mean.

This allows scores across assessments and groups to be compared according to their distribution-relative distance from the mean.

Percentile ranking is another standardisation procedure and involves transforming raw scores to represent the performance of individuals relative to typical performance on the assessment. For a given raw score, its percentile-ranked equivalent represents the proportion of the raw score distribution that falls equal to or below it. A standardised score distribution has been derived, allowing for population-referenced, standardised scores to be calculated. As the purpose of this assessment is to evaluate the attainment of educational standards applicable across quintile groups with known ability distribution differences, the observed median score differences are acceptable.

CONCLUSION

The *evidence* presented throughout this Manual demonstrates that it is very challenging to produce a single psychometrically sound measure of Language ability that provides an equivalent assessment of children in the many and diverse languages of this country. Considering the *conceptual* breadth of the ELOM-R Language (v1) Assessment construct, it seems likely that additional factors represent subdomains within this measure that are not being adequately sampled by the current set of items. The development of items targeting them specifically may allow for a viable multifactor solution with more robust second and third factors.

New items will be developed to address this issue. They will be added to the current ELOM-R Language (v1) Assessment and trialled before inclusion in the next version of the measure. This, as well as efforts to minimize differential item functioning and other item-level sources of misfit, may be incorporated into ongoing validation efforts with the goal of establishing a cross-culturally fair and reliable single or multiple-factor measure.

Future DIF investigations will also be supported by a more fine-grained IRT approach. IRT methods focus on modelling raw responses to test prompts, so the aggregate item-level data used as input for the above analyses may have restricted modelling precision. Ongoing validation efforts will incorporate attempts to produce a structural model wherein trials are the primary unit of analysis, and items are treated as higher-order variables within the Language assessment.

The standardisation and norms established for this 8-item version of the measure must be regarded as provisional.

SETTING THE ELOM-R LANGUAGE (V1) ASSESSMENT STANDARDS

PROCESS

Performance standards describe what children should know and be able to do at particular levels – in this case, at the end of the Grade R year. As described in ELOM-R (v1) Technical Manual 1 (Dawes & Biersteker, 2025), items in both the ELOM-R (v1) Mathematics and Language tests are closely aligned with the Grade R *Curriculum Assessment Policy Statements* (CAPS) specified by the National Department of Basic Education. Their development was also informed by research on predictors of Foundation Phase learning outcomes, consultations with experts in the field of early education, Foundation Phase educators, and a review of other available measures.

The process for setting ELOM-R (v1) standards followed that for ELOM 4&5 Years Assessment tool. As noted in the ELOM 4&5 Technical Manual, it is international practice to set early learning standards at between the 50th and 60th percentile of the norm sample standardised score distribution.

- A provisional benchmark for a child or a group being "On Track" was set at the 60th percentile of the standardised score distribution (equivalent to the percent correct score achieved by the top 40% of children in the standardisation sample).

- That proposal was discussed at a standards setting consultation in December 2024 with external experts in the field and members of the DataDrive2030 psychometrics team.

- The **60th percentile** was confirmed for both the ELOM-R (v1) Mathematics and Language Assessments, and following ELOM 4&5 practice, scores **between the 32nd and 59th** percentiles were classified as "Falling Behind", while those **below the 32nd percentile** were classified as "Falling Far Behind".

These bands are used for interpretive purposes in the norms that follow.

STANDARDISED SCORE DISTRIBUTIONS

Figure 14[44] presents the standardised distributions of both raw and normalised ELOM-R Language (v1) scores. Raw scores across the full sample of 2431 respondents are transformed into Z-scores, and columns represent increments of Z, starting at -3 and ending in + 3. For each increment of Z (representing half standard deviation units), normed as well as raw Percent Correct (PC) scores corresponding to these distribution points are presented.

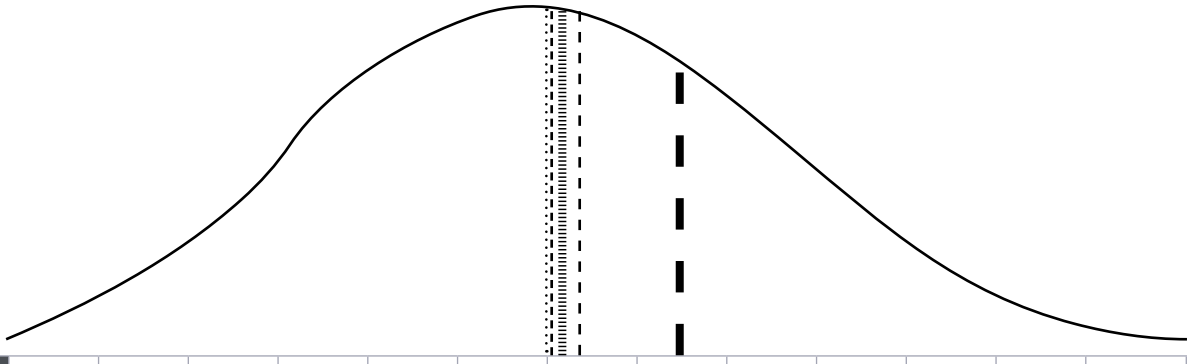
Raw score counterparts to each Z interval are also presented by quintile, representing the scores corresponding to the indicated Z value within each school quintile-specific subsample. Median raw scores per quintile group in relation to the normalised distribution are indicated with dashed lines overlaid on the distribution curve, a key for which is presented under the standardisation table.

Median score differences between quintiles across increments of Z indicate that there is little difference in performance of quintile groups 1 to 4, but that quintile 5 children perform considerably better than the other groups on ELOM-R Language (V1).

Standardised (percentile ranked) raw scores, Raw Percentage Correct scores and Z-normalised scores are provided in Table 23. For reference purposes. These are ordered by standards bands as indicated.

[44]For these calculations, each trial in each item is scored correct / incorrect. The proportion of trials correctly answered in each item is the Raw Percent Correct score for that item. The Raw Percent Correct score on the test as a whole reported in the Figure, and the Table is the average item percent correct score for all items.

Figure 14. ELOM-R Language (v1) Standard Score Distribution



	Z-Score	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	w
	Norm Score (Percentile)	0.0	0.5	2.6	7.6	17.6	30.0	47.9	66.1	81.5	95.1	99.9	100	100
Raw Percentage Correct Scores	Norm Sample	13.8	22.4	31.0	39.7	48.3	56.9	65.6	74.2	82.8	91.4	100	100	100
	Quintile 1	11.9	20.5	29.1	37.6	46.2	54.8	63.3	71.9	80.5	89.0	97.6	100	100
	Quintile 2	9.9	18.8	27.7	36.6	45.5	54.4	63.3	72.2	81.1	90.0	98.9	100	100
	Quintile 3	15.3	23.7	32.0	40.3	48.7	57.0	65.4	73.7	82.0	90.4	98.7	100	100
	Quintile 4	16.7	24.9	33.0	41.2	49.4	57.5	65.7	73.9	82.0	90.2	98.4	100	100
	Quintile 5	24.9	33.1	41.4	49.6	57.8	66.0	74.2	82.5	90.7	98.9	100	100	100

Quintile median	**1**	**64.0%**
Quintile median	**2**	**64.6%**
Quintile median	**3**	**65.6%**
Quintile median	**4**	**67.6%**
Quintile median	**5**	**76.7%**

NORMS

Table 23 provides the raw and z-score equivalents for each normalised score percentile. These can be used to compare the performance groups of children against the norms.

Table 23 ELOM-R Language (v1) Standardised Score Reference Table

KEY	
RAW SCORE	The Raw (Percentage Correct) score on the test ranging 0 to 100. Note: Raw scores on each ELOM-R (v1) item have different scales. For example, a child can obtain a score from -1 to 20 on item 1 and a score from -1 to 10 on item 2. It is obvious that these two items have different scales. When a test is standardised, all scores must be converted to the same scale. For this reason, all ELOM-R (v1) item scores are converted to percentage correct total scores on the test ranging from 0-100.
Z	Z-scores range from -3 to +3 (in a normal distribution). The Z-score shows the distance of the raw percentage correct score from the mean of the distribution in standard deviation units either above (+) or below (-) the mean (in a normal distribution such as this, the mean and median have the same value). Where two tests have Z-scores, these are then on the same scale and can be used in statistical analyses to compare scores on the two tests.
PERCENTILE	This value shows the % of the standardisation sample whose scores fall below the corresponding Raw Percentage Correct score. The percentile rank is the band of scores below the percentile.
COLOUR CODING	ELOM-R (v1) standards bands are shown on the table: *Green: On Track: => 60th percentile* *Orange: Falling Behind: 32nd-59th percentile* *Red: Falling Far Behind: <32nd percentile*

INTERPRETATION OF ELOM-R (v1) LANGUAGE RAW SCORES

Steps
1: Calculate the mean percentage correct raw score for your sample.
2: Use the norm table to look up the corresponding percentile and Z-score values for that score. This will tell you how your sample compares with the standardisation sample used to construct the ELOM-R (v1) norms.
Example:
If your sample's mean Raw score = 57.5, it falls at the 32nd percentile of the standardised distribution. This tells you that your group scored in the same range as 32% of the standardisation sample who scored 57.5 or less on this test. The corresponding Z-score in the table tells you how many standard deviations above (+) or below (-) your sample percentage correct score is from the mean of the standardisation sample, in this case, 0.50 standard deviations below the standardisation sample mean.

FALLING FAR BEHIND			FALLING BEHIND			ON TRACK		
Raw Score	Z	Percentile	Raw Score	Z	Percentile	Raw Score	Z	Percentile
16.4	-2.85	0	57.5	-0.47	32	71.7	0.35	60
25.6	-2.31	1	58.0	-0.44	33	72.2	0.39	61
28.5	-2.15	2	58.5	-0.41	34	72.6	0.41	62
31.9	-1.95	3	59.2	-0.37	35	73.1	0.44	63
34.0	-1.83	4	59.8	-0.34	36	73.6	0.47	64
36.1	-1.71	5	60.2	-0.31	37	74.1	0.50	65
37.7	-1.61	6	60.7	-0.28	38	74.5	0.52	66
38.9	-1.55	7	61.1	-0.26	39	74.8	0.54	67

FALLING FAR BEHIND			FALLING BEHIND			ON TRACK		
Raw Score	Z	Percentile	Raw Score	Z	Percentile	Raw Score	Z	Percentile
40.0	-1.48	8	61.5	-0.24	40	75.5	0.57	68
41.0	-1.42	9	62.2	-0.20	41	75.8	0.59	69
41.8	-1.38	10	62.7	-0.17	42	76.3	0.62	70
42.6	-1.33	11	63.1	-0.14	43	76.7	0.65	71
43.4	-1.28	12	63.5	-0.12	44	77.2	0.67	72
44.6	-1.22	13	64.0	-0.09	45	77.7	0.70	73
45.4	-1.17	14	64.5	-0.06	46	78.1	0.72	74
46.5	-1.10	15	65.1	-0.03	47	78.8	0.77	75
47.0	-1.07	16	65.6	0.00	48	79.2	0.79	76
48.1	-1.01	17	66.0	0.02	49	79.9	0.83	77
48.8	-0.97	18	66.5	0.05	50	80.7	0.88	78
49.5	-0.93	19	67.2	0.09	51	81.1	0.90	79
50.0	-0.90	20	67.5	0.11	52	81.6	0.93	80
50.5	-0.87	21	67.9	0.14	53	82.1	0.96	81
51.2	-0.83	22	68.4	0.16	54	82.7	1.00	82
51.7	-0.80	23	69.1	0.20	55	83.3	1.03	83
52.4	-0.76	24	69.5	0.23	56	84.2	1.08	84
52.9	-0.73	25	70.2	0.27	57	84.8	1.12	85
53.6	-0.69	26	70.7	0.30	58	85.7	1.17	86
54.4	-0.65	27	71.2	0.33	59	86.1	1.19	87
55.0	-0.61	28				86.8	1.23	88
55.8	-0.57	29				87.4	1.27	89
56.4	-0.53	30				88.1	1.30	90
57.0	-0.49	31				88.9	1.35	91
						89.7	1.40	92
						90.3	1.44	93
						90.9	1.47	94
						91.5	1.50	95
						92.5	1.56	96
						93.9	1.64	97
						95.3	1.72	98
						97.2	1.84	99
						100.0	2.00	100

APPENDIX 1: ELOM-R LANGUAGE (v1) ASSESSMENT ITEM SCORING

APPENDIX 1: ELOM-R LANGUAGE (v1) ASSESSMENT ITEM SCORING		
ITEM	**TRIALS**	**SCORING**
1. PRODUCTIVE VOCABULARY	36	**Task**: The child is shown 36 pictures of objects/actions and asked to name each in turn. **Scoring:** 1 point for each object correctly named. Total possible score = 36.
2. BEGINNING SOUNDS	8	**Task**: The child is shown pictures of objects or actions (e.g. cow or dance) and asked to say the sound that each word starts with (e.g. /c/ for cow). Words with the same initial sounds were provided for in each language. **Scoring:** 1 point for each correct answer. Only score correctly if the child is able to isolate the first phoneme in the word. Total possible score = 8.
3. LETTER SOUNDS	8	**Task:** The child is shown pictures of objects or actions (e.g. cow or dance) and asked to say the sound that each word starts with (e.g. /c/ for cow). Words with the same initial sounds were provided for in each language. **Scoring:** 1 point for each correct answer. Only score correctly if the child is able to isolate the first phoneme in the word. Total possible score = 8.
4. COPY SHAPES	4	**Task**: The child is shown a triangle, rectangle and vertical diamond, and is asked to copy these shapes by drawing them. **Scoring:** Triangle: 3 sides and one corner higher than others = 1. Rectangle: At least 3 joining corners were closed AND sharp, not rounded, no gaps; at least 2 parallel sides of equal length, less than 1 cm difference; needs to be identifiable rectangle; horizontal orientation = 1. Vertical Diamond: 4 corners, horizontal within 170 – 190˚, sides more or less equal lengths, vertical orientation = 1. Total possible score = 3.
5. WRITE NAME	5	**Task:** The child is asked to write down their first name. **Scoring:** Not able to write ame score = 0; Most letters correct, some may be reversed or missing score =1; Name is correctly written score = 2.
6. WRITING WITH ENCOURAGEMENT	6	**Task**: The child is shown pictures of two common objects, a shorter and a longer word (e.g. in English: cat, helicopter). The child is asked to write the words. **Scoring**: Child writes down the first letter of the word correctly = 1; More than one letter is correct= 2; Child's spelling includes three or more letters of which one is a vowel. Total possible score = 6.
7. LISTENING COMPREHENSION	10	**Task**: The assessor lays out 6 pictures of a scene and reads the child a story pertaining to this scene. After a warm-up question ("did you like the story?"), the assessor asks the child ten comprehension questions about the story, pointing to the relevant pictures containing the subject of the question (e.g. "why does the dog jump forward?"). **Scoring:** 1 point for each correct answer. Total possible score = 10.
8. BOOK CONCEPT, ORIENTATION, AND WORD CONCEPT	9	**Task**: The assessor gives the child a picture book and asks nine questions about how it is structured, where the front is, title, where to start and continue reading, etc. **Scoring:** Score 1 point for each correct answer. Total possible score = 9.