elom

early learning measurement tools

JUNE 2025

ELOM-R (v1)
TECHNICAL MANUAL 3
Mathematics Assessment

1ST EDITION 2025

Developed on behalf of DataDrive2030 by Matthew Kleineibst (Psychology Department University of Cape Town and Ax Consult), Jürgen Becker (Industrial Psychology Department, University of the Western Cape and Ax Consult) 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 Technical Manual 1: The Development of the ELOM-R Language (v1) and Mathematics Assessments and ELOM-R Technical Manual 2: ELOM-R Language (v1) Assessment.

TO CITE THIS MANUAL:
Kleineibst, M., Becker, J. & Dawes, A. (2025). ELOM-R (v1) Technical Manual 3: Mathematics 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
Professor Hamsa Venkat. Dublin City University; Formerly South Africa Numeracy Chair, Mathematics Education, University of the Witwatersrand Cally Kuhne, RedInk (redink.org.za). Formerly Senior Education Specialist ECD/FP R-Mathematics Programme Leader, Schools Development Unit, University of Cape Town	Specialist consultants on Foundation Phase Mathematics
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*
MHχ^2	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 MATHEMATICS (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**
3. **ELOM-R (v1) Technical Manual 3: Mathematics Assessment** *(this Manual)*

All are available on the DataDrive2030 website. Before 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 Pilot study designed to test and adjust items before finalisation for the analyses.

This chapter briefly outlines the purpose, content and use of the Mathematics 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 determine whether the ELOM-R Mathematics (v1) Assessment shows test bias in any of the languages. Here we also report on the standardisation and norms of the measure.

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

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

PURPOSE

ELOM-R Mathematics (v1) Assessment is primarily intended for research studies, surveys, and evaluations of numeracy and mathematics interventions designed to enable readiness for Grade 1. It is, therefore, appropriate for the assessment and descriptions of groups of children and is not used as a diagnostic test of individual child school readiness.

The Mathematics 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 function toward the end of the Grade R year. The tool may, therefore, be regarded as a summative assessment of children's numeracy and mathematics knowledge and skills. 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 numeracy and mathematics 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 inform strategies to enhance preschool, Grade R and Grade 1 curriculum, quality and training in the CAPS mathematics area.

This test can, therefore, be used in population surveys to estimate the proportion of children who are on Track for Grade 1 in numeracy and mathematics 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 Mathematics (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.

ELOM-R MATHEMATICS (v1) ASSESSMENT ITEMS

The item (and number of trials) set for psychometric analysis and norming is presented in Table 1.

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

GRADE R CAPS AREA	ITEM	NUMBER OF TRIALS
NUMBER SENSE AND OPERATIONS	1. Count forwards to 20	1
	2. Count backwards from 10	1
	3. Counting from a given number	2
	4. Skip counting in twos to 10	1
	5. Count with 1:1 correspondence	1
	6. Number order	2
	7. Number recognition	6
	8. Subitise to 5	5
	9. Knowledge of ordinal Numbers	6
	10. Compare two collections of objects	4
	11. Show a collection without counting	5
	12. Solving addition and subtraction problems	4
	13. Solving sharing and grouping problems	3
SHAPE AND SPACE	14. Shape and space construction (copy shape from models)	2
MEASUREMENT	15. Sorting & Grouping	4
SHAPE AND SPACE	16. Shape identification and understanding	6
PATTERNS, FUNCTIONS, AND ALGEBRA	17. Pattern extension	7
	18. Pattern completion	1

**Appendix 1 provides item scoring.*

Raw scores on each ELOM-R Mathematics (v1) item have different scales. For example, item 7 (number recognition) has six trials and a child can obtain a score from 0-6; Compare collection of objects (item 11) has five trials and a child can score from 0-5. It is clear 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 % 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 Mathematics (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 V.1, it is necessary to undertake procedures to establish whether the test's psychometric properties 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 testing languages have the same factor structure (each measures the same underlying trait). Further, 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 have not been 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 ELOM 4&5 to assess the cross-language equivalence of that instrument (Dawes et al., 2025[7]; Snelling et al., 2019[8]).

A taxonomy of bias and equivalence issues relevant to the ELOM-R (v1) assessments is drawn from Van de Vijver and Rothmann (2004)[9] pp. 2-3} and Poortinga (1998)[10] and is presented in Table 2. In their papers, the above authors refer to cross-cultural bias and equivalence. Our 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

[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.

Table 2. Types and sources of bias

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	The same item and measurement unit is used to assess all groups.

INVESTIGATING BIAS IN THE ELOM-R MATHEMATICS (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 Mathematics (v1) Assessment 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's 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).

[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 section, we summarise preliminary psychometric analyses undertaken using Classical Test Theory (CTT) and Item Response Theory (IRT) modelling procedures to investigate the factor structure and reliability of the ELOM-R Mathematics (v1) in isiZulu, isiXhosa, Sepedi, Sesotho, Setswana, Tshivenda, English and Afrikaans. Complete psychometric reports for each language are available from DataDrive2030.

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 3 to 20 times the number of variables and absolute ranges from 100 to over 1,000. Mostly, there is little empirical evidence to support these recommendations".

In their paper, Mundfrom et al. 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 Mathematics (v1) Assessment, we have 18 items and tested a single factor solution (18 items and 1 factor, i.e. a ratio of 18:1).

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 in 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 methods as well as 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 later, language and socio-economic status are 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 income 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	MINIMUM	MAXIMUM
2564	77.4	3.88	77.3	70.0	89.0

Figure 1. Distribution of Child Age (Months)

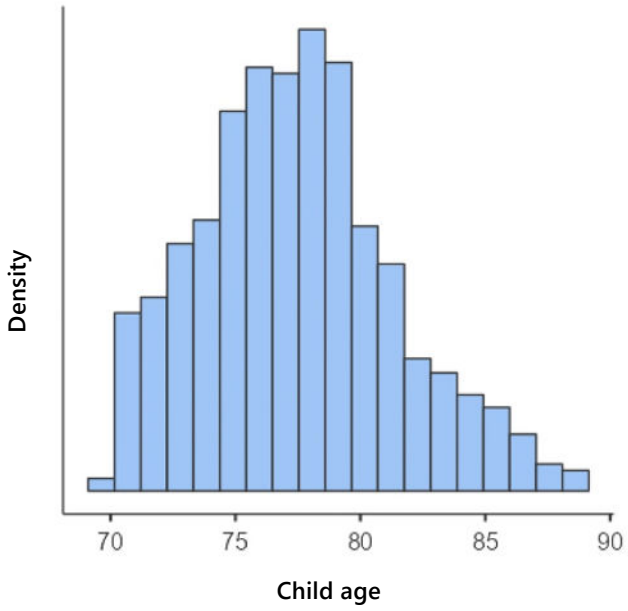
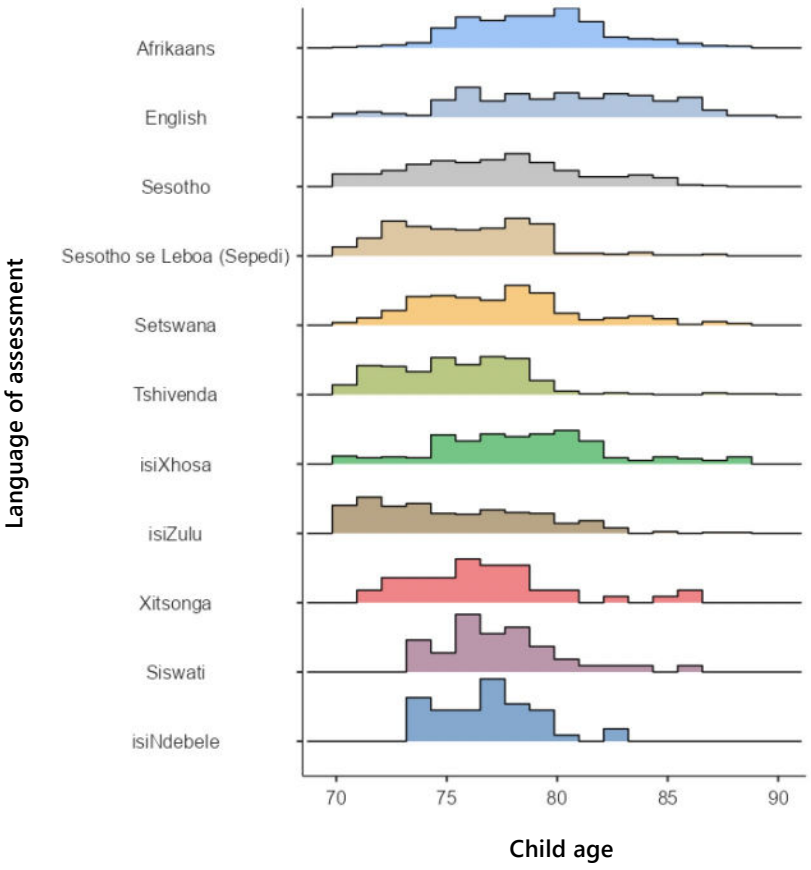


Table 4. Descriptive Statistics for Child Age (Months) Per Language Group

	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** Approval was granted by the Provincial Departments of Education of schools where 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 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).

DATA COLLECTION

Data for sample 1 was provided by the various research and evaluation study teams. Sample 2 fieldwork 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 (v1) and Language 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)[22], 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[23]). 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[24]) notes that in applied settings where *important high-stakes decisions* are made about *individuals* based on their test scores, a reliability of **0.90** 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]Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. Communication Methods and Measures, 14(1), 1–24. **https://doi.org/10.1080/19312458.2020.1718629**
[23]Kline, P. (2000). Handbook of Psychological Testing. London, United Kingdom: Routledge.
[24]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[25]; Hu & Bentler, 1999[26]; Tavakol & Wetzel, 2020[27]).

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 nonsignificant 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.**

Exploratory Factor Analyses (EFA) were also undertaken in each language to explore possible subfactor structures evident in the data. This was not the case.

[25]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**
[26]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**
[27]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[28]; Bond & Fox, 2015[29]; Fan, 1998[30]).

Based on the item scoring in the ELOM-R Mathematics (v1) Assessment and the presumption of a unidimensional factor structure (necessary for Rasch analysis), a *dichotomous* one-parameter logistic model (1PL) Rasch model was 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 Mathematics (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[31].

Scores on the ELOM-R Language (v1) were subjected to Rasch modelling to determine item difficulty and a more accurate assessment of the validity and reliability of the test. Important metrics to consider in Rasch analysis are described in Table 6 below, along with guidelines for their interpretation (Bond & Fox, 2015; Linacre, 2002[32]; Yen, 1993[33]).

[28]Baker, F. (2001). The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.
[29]Bond, T., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences. New York, NY: Routledge
[30]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.
[31]Polytomous scales have more than two possible scores for an item. This is the case in ELOM-R (v1) Assessments where item trials are individually scored and summed to derive the item score.
[32]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.**
[33]Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30(3), 187-213.

Table 6: Rasch 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 >0.5 considered acceptable.**
POINT-MEASURE CORRELATION	Correlation between raw item or scale score and Rasch ability estimates. **Considered acceptable >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) MATHEMATICS ASSESSMENT

Language group sample sizes for these analyses and findings are provided in Table 7. Detailed psychometric reports are available for each language from DataDrive2030.

As will be evident, test reliability in all languages is higher than the *acceptable* value in both CTT ($\omega > 0.7$) and Rasch approaches to analysis (person reliability > 0.5) (Kline, 2005[34]). Regarding the comparison of factor structures across languages (MGCFA), it is notable that in all languages, CFA for the ELOM-R Mathematics Assessment (V1) produced a *unidimensional* factor model with good fit. In all languages, Rasch model fit is acceptable as well. Note: this is a *much more* conservative procedure than CFA.

[34]Kline, T. J. (2005). Psychological Testing: A practical approach to design and evaluation. Sage Publications, Inc

Table 7. ELOM-R Mathematics (v1): Summary of findings by Language

LANGUAGE	SAMPLE	RELIABILITY[35]	CFA[36]	RASCH[37]
ENGLISH	N=282 Q1,2 &3 = 59% Q4&5 =41%	ω =0.841	Model fit acceptable (RMSEA = 0.041) Unidimensional	Model fit acceptable. point-measure correlation (0.928)[38]; person reliability (0.769)[39]
AFRIKAANS	N=448 Q1,2 &3 = 46% Q4&5 =54%	ω =0.882	Model fit good (RMSEA = 0.055) Unidimensional	Model fit acceptable. point-measure correlation (0.943); person reliability (0.807)
ISIXHOSA	N=290 Q1,2 &3 = 68% Q4&5 =32%	ω =0.851	Model fit good (RMSEA = 0.045) Unidimensional	Model fit acceptable. point-measure correlation (0.922); person reliability (0.761).
ISIZULU	N=283 Q1,2 &3 = 57% Q4&5 =43%	ω =0.862	Model fit good (RMSEA = 0.054) Unidimensional	Model fit acceptable. point-measure correlation (0.926); person reliability (0.711).
SETSWANA*	N=276 Q1,2 &3 = 95% Q4&5 =5%	ω =0.869	Model fit good (RMSEA = 0.051) Unidimensional	Model fit acceptable. point-measure correlation (0.894); person reliability (0.592).
TSHIVENDA*	N=290 Q1,2 &3 = 93% Q4&5 =7%	ω =0.850	Model fit good (RMSEA = 0.057) Unidimensional	Model fit acceptable. point-measure correlation (0.888); person reliability (0.531).
SESOTHO*	N = 283 Q1,2 &3 = 72% Q4&5 =28%	ω =0.839	Model fit good (RMSEA = 0.055) Unidimensional	Model fit acceptable. point-measure correlation (0.907); person reliability (0.616).
SEPEDI*	N=286 Q1,2 &3 = 89% Q4&5 =11%	ω =0.820	Model fit good (RMSEA = 0.041) Unidimensional	Model fit acceptable. point-measure correlation (0. 893); person reliability (0. 580) (just above criterion)

(*Note that in four languages (highlighted in red), very high proportions 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.)

[35]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.
[36]Dichotomous Rasch modelling was used here and takes into account both item difficulty and person ability. Point measure correlation should >0.2
[37]Rasch considers both item difficulty and a person's ability. Point measure correlation should be >0.2
[38]Relationship between raw scores and Rasch ability estimates; a measure of the reliability of the test in assessing ability.
[39]Measure of the consistency in scoring; children with a high ability estimate on Rasch should score consistently high across items (and vice versa) Person Reliability should be >0.5.

MULTIPLE GROUP FACTOR STRUCTURE CONCLUSION

A unidimensional factor structure is evident for all languages on the ELOM-R Mathematics (v1) Assessment. It is worth noting that the person reliability estimates produced in multi-group Rasch analyses indicate a split in the samples that is roughly consistent with quintile groupings. For language groups with strong representation within Q4 and Q5 schools (English, Afrikaans, isiXhosa, isiZulu), person reliability ranges from 0.711 to 0.807. For language groups with low Q4 and Q5 representation (Setswana, Tshivenda, Sesotho, and Sepedi), person reliability ranges from 0.531 to 0.616. This suggests that the ELOM-R Mathematics (v1) Assessment measures ability with greater precision in quintile 4 and 5 schools when compared to the lower quintiles and may represent the effect of the children's languages. However, precision is still considered acceptable in all language groups, and aggregated group-level estimates will be even more precise.

Assessment of Bias: Differential Item Functioning in the ELOM-R Mathematics (v1) Assessment

Following IRT Guideline TD-4 (7) requiring test developers to provide evidence that items are suitable for all intended populations, we assessed the extent to which the ELOM-R Mathematics (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[40]). In this case, DIF is used to assess whether latent ability scoring on the ELOM-R Mathematics (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[41]; Linacre, 2016[42]; Magis et al, 2010).

[40]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.
[41]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**
[42]Linacre, J. M. (2024). Mantel and Mantel-Haenszel DIF statistics. Retrieved from **https://www.winsteps.com/winman/mantel_and_mantel-haenszel_dif.htm**

Table 8. DIF Statistics and Their Interpretation

STATISTIC	INTERPRETATION
$MH\chi^2$	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 MATHEMATICS

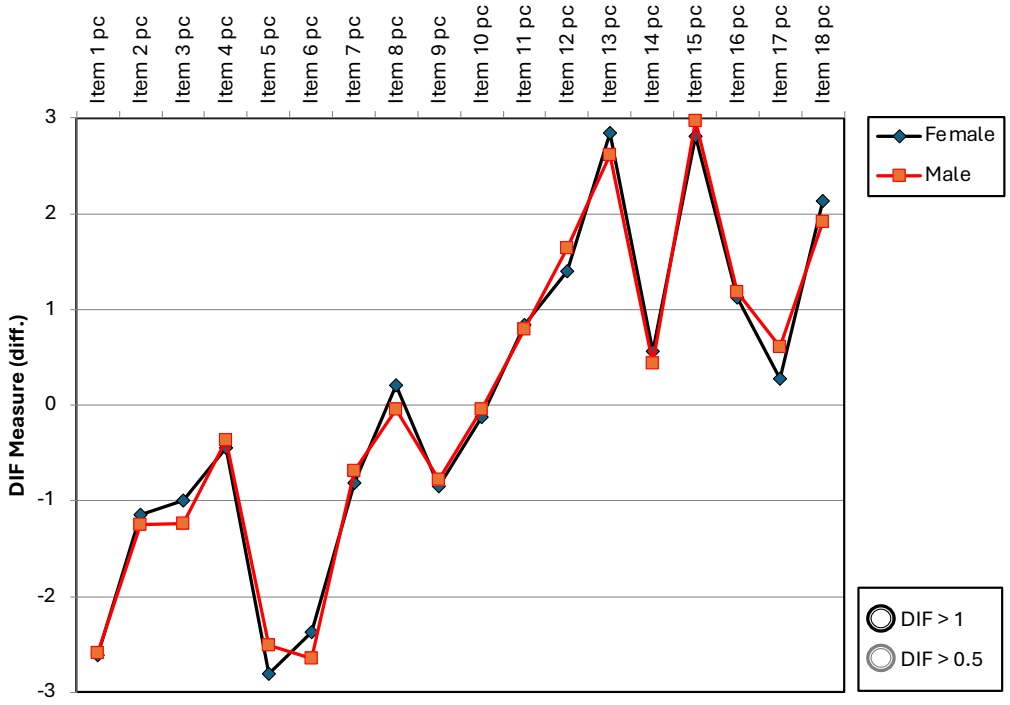
Full reports of DIF analyses are available from DataDrive2030 on request. A summary of the findings is presented below.

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 Mathematics (v1) Assessment.

A plot (item response logits per group) across items is presented in Figure 3. DIF Measures represent item difficulty estimates (on the vertical axis), across items (along the horizontal axis), for males and females. Item equivalence is indicated if item difficulties between these two groups are consistent (difference <0.5 logits).

Figure 3. ELOM-R (v1) Mathematics Sex/Gender Plot



DIF is not indicated for sex.

LANGUAGE GROUP DIF

A full report on DIF analyses for possible language bias in the ELOM-R Mathematics (v1) Assessment is available from DataDrive2030 on request. The sample for these analyses is provided in Table 9.

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

LANGUAGES*								
ENGLISH	AFRIKAANS	ISIZULU	ISIXHOSA	SESOTHO	SETSWANA	SEPEDI	TSHIVENDA	TOTAL
282	448	283	290	287	276	286	290	2442

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

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.

TESTING FOR A LANGUAGE – QUINTILE CONFOUND

As noted previously, one cannot assume the non-equivalence 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 where item performance is modelled with child ability.

Testing for interaction between language and quintile was considered using MANOVA. However, as is evident in Table 10 below, language group sample sizes were too small (0 in one group) in the higher quintile for four of the African languages and too small in the bottom two quintiles for Afrikaans and English. An ANOVA testing for quintile effects alone indicated that, overall, the school quintile groups were significantly different ($F(4, 1033.62) = 52.15$, $p < 0.001$). On post-hoc testing, statistically significant differences were evident between the mean Mathematics score for quintile 5 and all other quintile groups. Additionally, a social gradient was evident for comparisons between other quintiles (higher quintiles had higher mean scores). Overall, we can conclude that language and SES (quintile) are highly likely to be confounded for the Mathematics Assessment.

Table 10. Quintile Distributions of Children for Each Language Subsample*

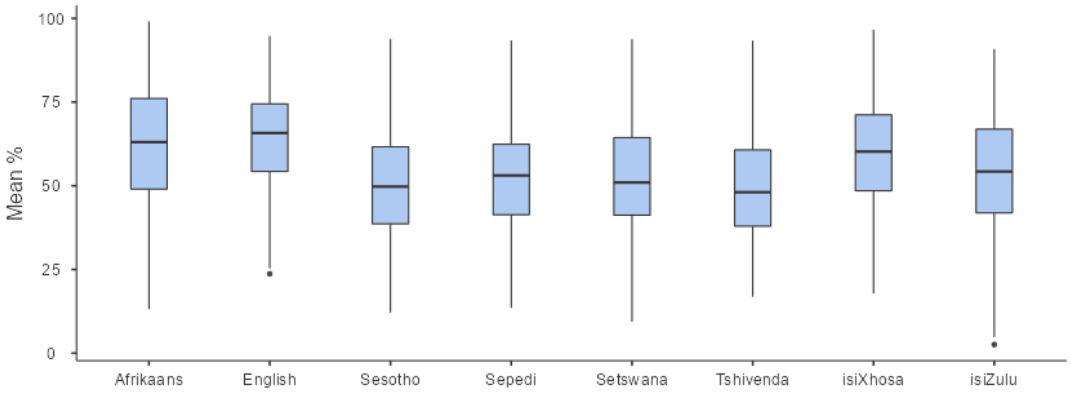
SCHOOL QUINTILE	ENGLISH	AFRIKAANS	SESOTHO	SEPEDI	SETSWANA	TSHIVENDA	ISIXHOSA	ISIZULU
1	14	84	68	218	240	98	23	43
2	33	84	63	13	0	63	74	58
3	118	38	75	23	21	109	101	61
4	48	141	46	16	8	20	57	82
5	69	101	35	16	7	0	35	39

**Modal values indicated in red text*

Table 10 indicates that the Sepedi and Setswana subsamples are predominantly within quintile 1 schools. The modal quintile for the Tshivenda sample was also 3, but the proportion of quintile 1 and 2 schools heavily outweighed the proportion of quintile 4 and 5 schools in this sample. Quintile distributions for the Sesotho, isiXhosa, and isiZulu samples were less remarkable, with each quintile represented.

High proportions of Sesotho (72%), Sepedi (89%), Tshivenda (93%) and Setswana (95%) children are in quintiles 1-3. In contrast, most Afrikaans respondents attended quintile 4 or 5 schools, and while most English schools were in the third quintile, a greater proportion were in quintile 4 or 5. These variations in the sample quintile are likely reflected in the ELOM-R Mathematics (v1) test performances of children in each language (Figure 4 and Table 11).

Figure 4. ELOM-R Mathematics (v1) Mean Score by Language*



**(bars indicate confidence intervals).*

Table 11. ELOM-R Mathematics (v1) percent correct statistics by language

LANGUAGE	MEAN (%)	SD (%)	MIN (%)	MAX (%)	DIFFERENCE (TO ENGLISH)
ENGLISH	64.1	15.3	23.7	94.8	-
AFRIKAANS	62.8	17.8	13.2	99.1	-1.3
SESOTHO	50.2	16.0	12.1	93.8	-13.9
SEPEDI	51.8	15.2	13.5	93.4	-12.3
SETSWANA	51.9	16.9	9.4	93.8	-12.2
TSHIVENDA	49.2	16.2	16.9	93.3	-14.9
ISIXHOSA	59.6	16.0	17.9	96.6	-4.5
ISIZULU	53.5	17.1	2.6	90.8	-10.6

Mean Mathematics scores range from 49.2% for Tshivenda (93% of the sample is in quintile 1-3) to 64.1% (English), amounting to a range of 14.9%. Mean scores for Sesotho (50.2%), Sepedi (51.8%), Setswana (51.9%), and isiZulu (53.5%) were all over 10% lower than the English sample mean.

While the box plots and distribution characteristics indicate differences at the raw score level, DIF analysis reveals whether these are due to genuine differences in ability level or differential item functioning indicative of bias in the measurement process, which is a form of measurement error.

DIF FINDINGS FOR LANGUAGE GROUPS

Each of the other languages is compared to English. Plots (Figures 5-11), 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 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 focal language with which it is compared.

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

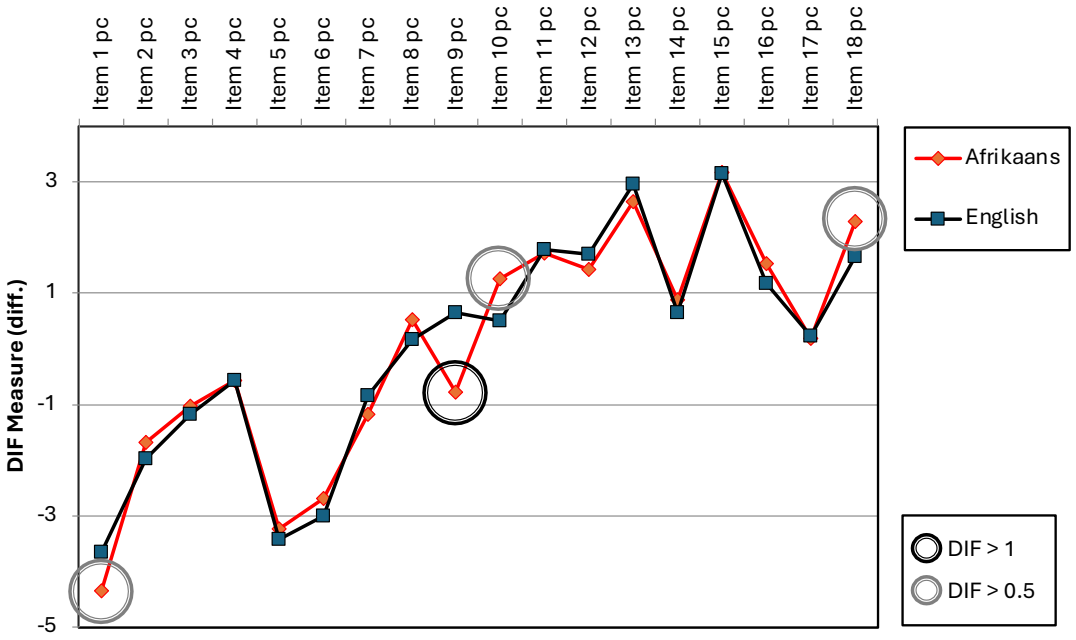


Figure 6. English – isiXhosa DIF Plot for ELOM-R Mathematics (v1)

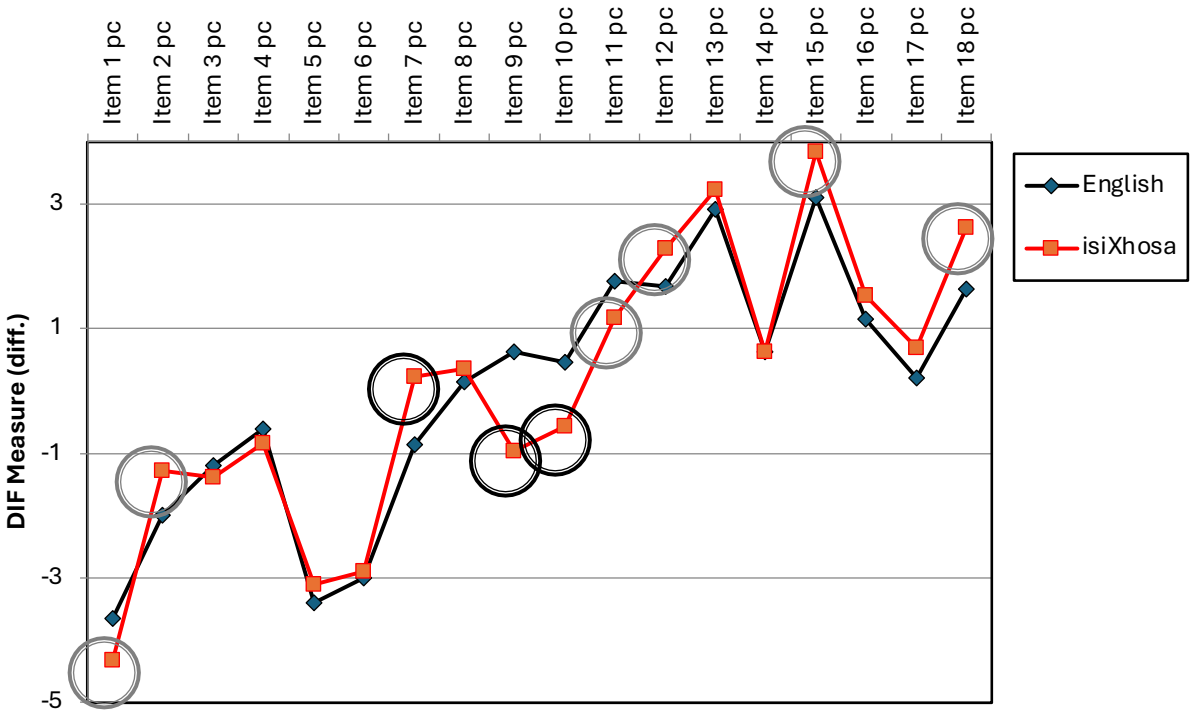


Figure 7. English – isiZulu DIF Plot for ELOM-R Mathematics (v1)

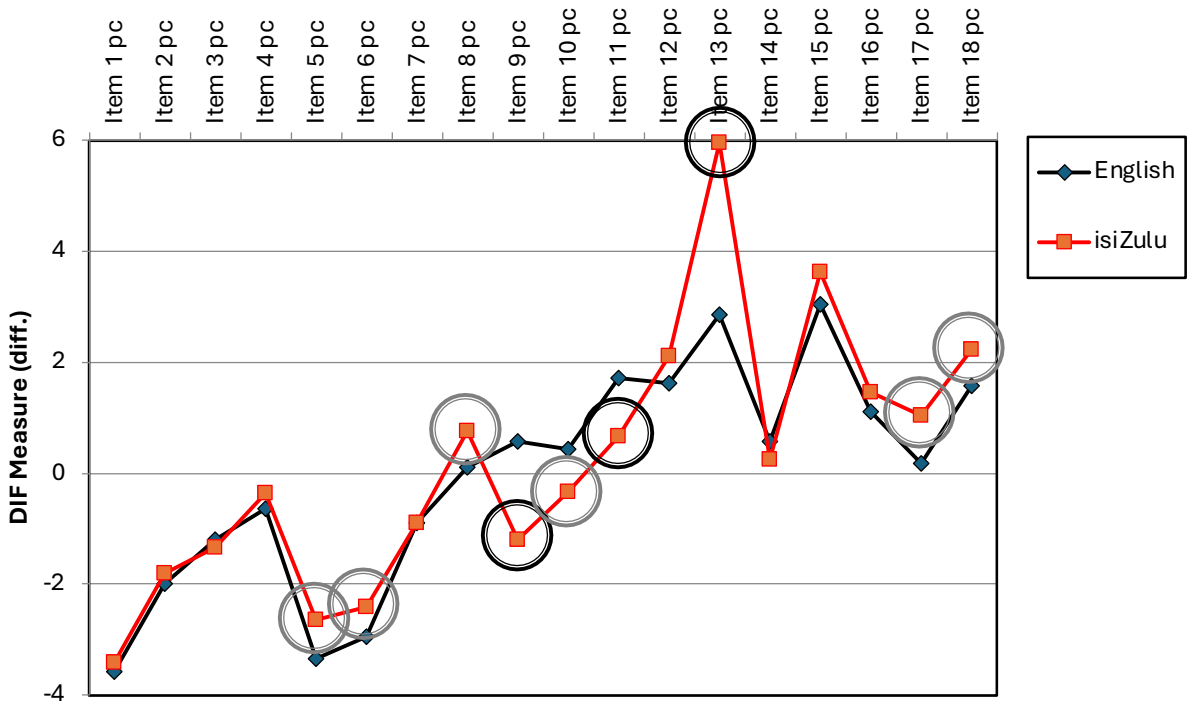


Figure 8. English – Setswana DIF Plot for ELOM-R Mathematics (v1)

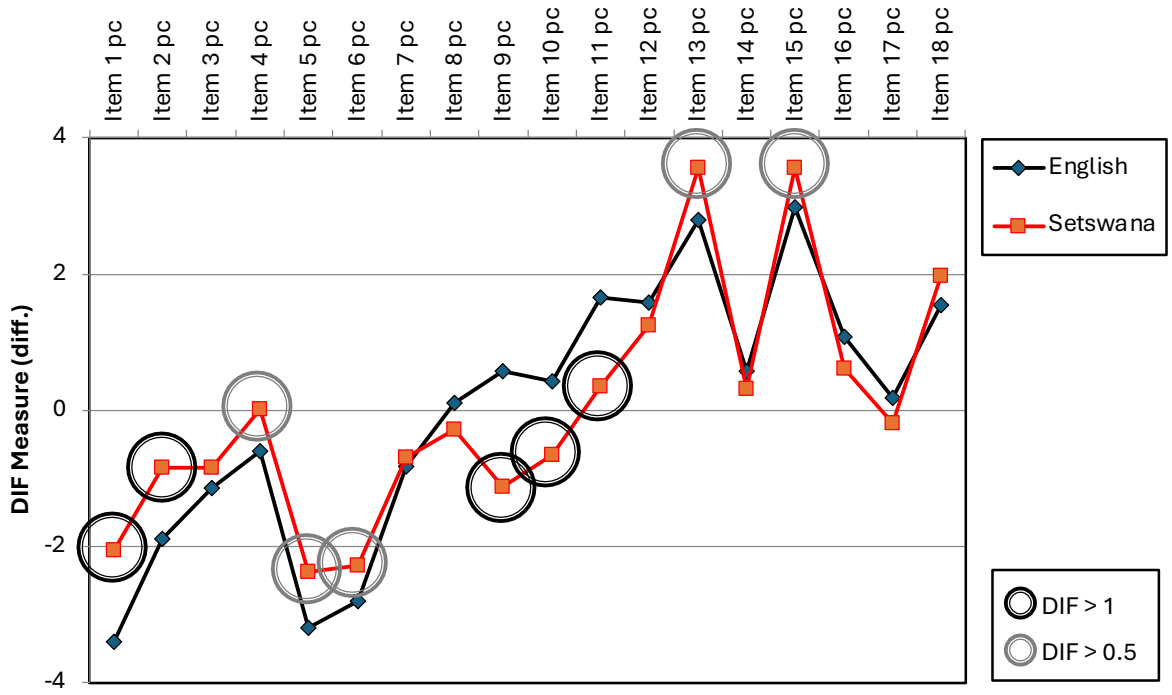


Figure 9. English – SeSotho DIF Plot for ELOM-R Mathematics (v1)

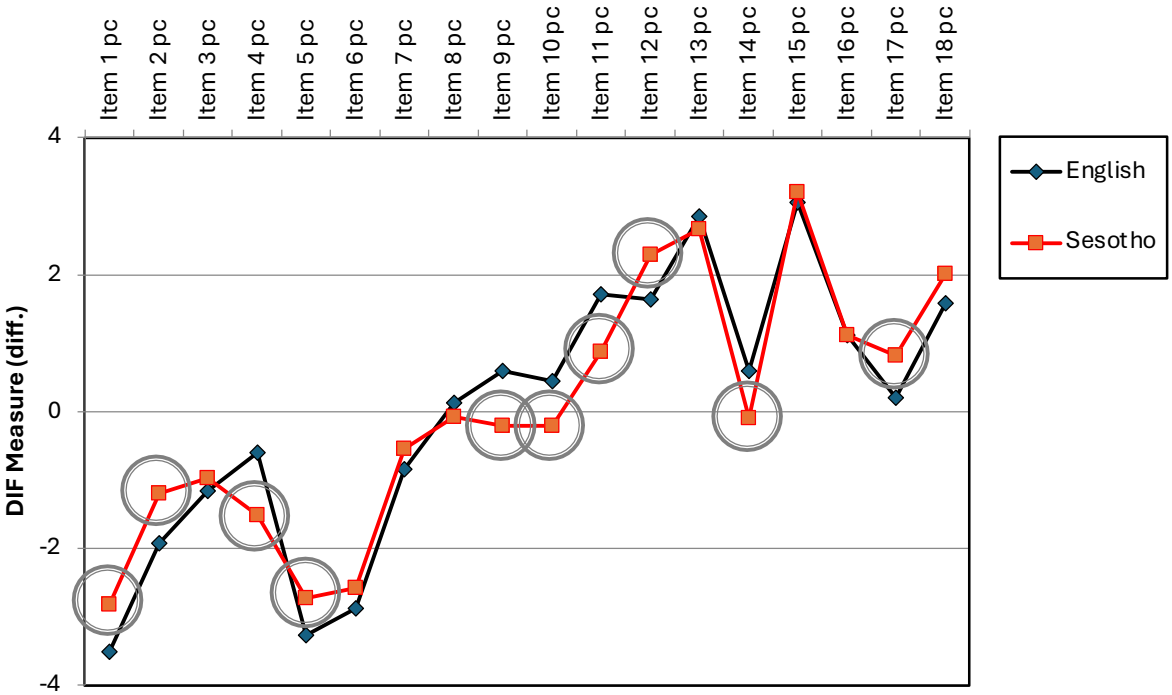


Figure 10. English – Sepedi DIF Plot for ELOM-R Mathematics (v1)

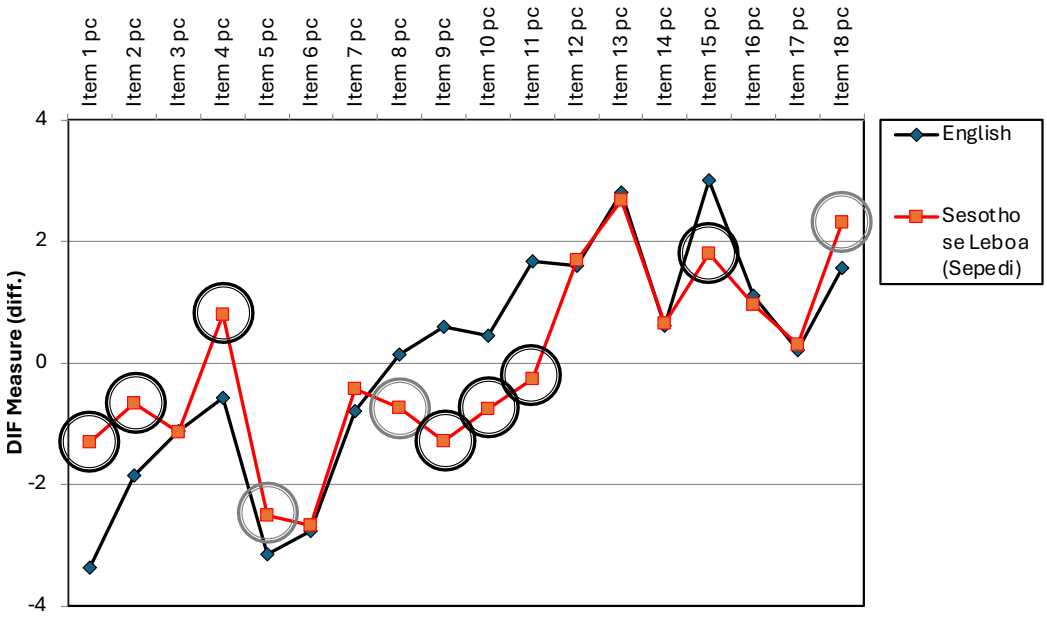
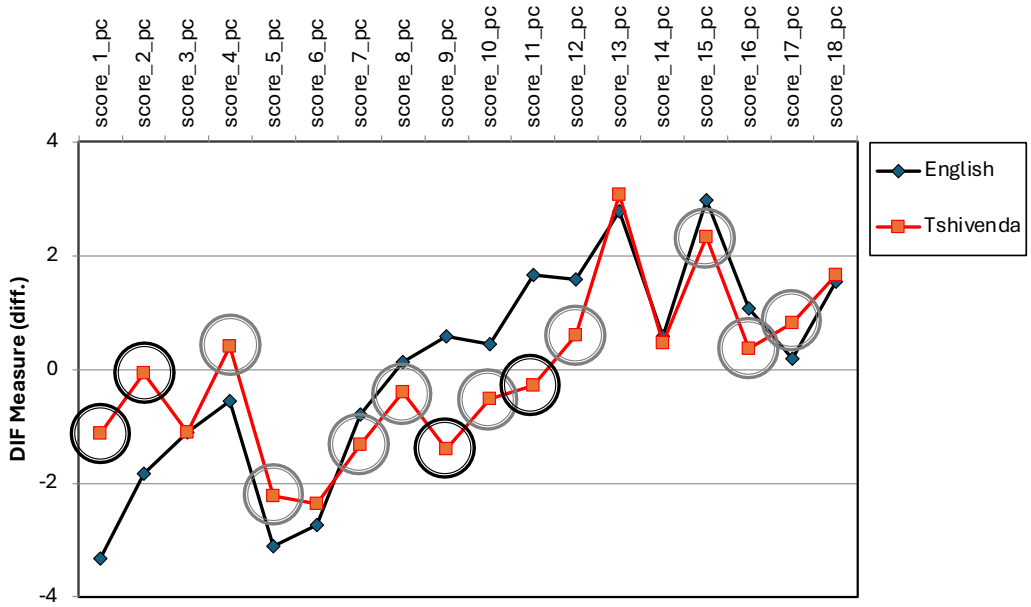


Figure 11. English – Tshivenda DIF Plot for ELOM-R Mathematics (v1)



Inspection of all plots shows considerable comparability of the languages on certain items with the English reference language, while others show bias toward English (adverse DIF) or to the focal language in question (benign DIF).

Significant DIF effects for each language against the English cohort on the Rasch-Welch t-test[43] are summarised in Table 12. Omitted values were not statistically significant. Positive effects indicate higher item difficulty in the non-English group (bias in favour of English). Negative effects indicate lower item difficulty in the non-English group (bias in favour of non-English). Values printed in **black** represent small to moderate DIF effects, while values printed in **red** represent moderate to large DIF. Cumulative DIF effects on balance bias towards the non-English language groups (benign DIF).

[43]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 Mathematics (v1) DIF Contrasts by focal Language

ITEM		AFRIKAANS	ISIXHOSA	ISIZULU	SETSWANA	SESOTHO	SEPEDI	TSHIVENDA
	DIF ACCUMULATION	**-0.76**	**-0.61**	**-0.82**	**0.24**	**-1.29**	**-1.89**	**-1.61**
1	Count forwards to 20	-0.70	-0.67		1.35	0.69	2.06	2.21
2	Count backwards from 10		0.71		1.03	0.73	1.17	1.76
3	Counting from a given number							
4	Skip counting in twos to 10				0.61	-0.91	1.36	0.98
5	Count with 1:1 correspondence			0.70	0.82	0.55	0.63	0.88
6	Number order			0.55	0.53			
7	Number recognition		1.09					-0.52
8	Subitise to 5			0.64			-0.89	-0.54
9	Knowledge of ordinal numbers	-1.42	-1.58	-1.78	-1.7	-0.80	-1.88	-1.99
10	Compare two collections of objects	0.75	-1.04	-0.77	-1.09	-0.65	-1.2	-0.97
11	Show a collection without counting		-0.58	-1.03	-1.31	-0.83	-1.94	-1.94
12	Solving addition and subtraction problems							-0.98
13	Solving sharing and grouping problems			3.10				
14	Shape and space construction					-0.70		
15	Sorting and grouping						-1.2	
16	Shape identification and understanding							-0.72
17	Pattern extension			0.87		0.63		0.62
18	Pattern completion		0.99					

Higher DIF means higher item difficulty (and lower ability) in the non-English (focal) group. DIF effects for all items that have *statistically* significant DIF on all metrics are reported below.

- **Afrikaans group:** Children score comparably to English children, amounting to an average raw score difference of 1% across items in favour of English children (range = -9% to 15%). Overall, DIF effects do not accumulate in favour of either English or Afrikaans children across items (accumulated DIF = -0.05).

- **isiXhosa group:** Children score comparably to English children, amounting to an average raw score difference of 1% across items in favour of English children (range = -9% to 15%). Overall, DIF effects do not accumulate in favour of either English or Afrikaans children across items (accumulated DIF = -0.05).

- **isiZulu group:** DIF effects are mixed, although the balance favours English. Items favouring English (adverse DIF) include: 5 (Count with 1:1 correspondence), 6 (Number order), 8 (Subitise to 5), 13 (Solving sharing and grouping problems) and 17 (Pattern completion). Benign DIF effects favouring isiZulu include 9 (Knowledge of ordinal numbers), 10 (Compare two collections of objects) and 11 (Show a collection without counting).

- **Setswana group:** DIF effects appear to balance out, but a slight bias in favour of English is observed. Items favouring English (adverse DIF) include items 1 (Count forwards to 20), 2 (Count backwards from 10), 4 (Skip counting in twos to 10), and 5 (Count with 1:1 correspondence). Benign DIF effects favouring Setswana include items 9 (Knowledge of ordinal numbers), 10 (Compare two collections of objects) and 11 (Show a collection without counting).

- **Sesotho group:** While DIF effects are evident for several items, these accumulate to a negligible effect in favour of the English reference group. Items (adverse DIF) favouring English include: 1 (Count forwards to 20), 2 (Count backwards from 10), and 5 (Count with 1:1 correspondence). Benign DIF effects favouring Sesotho include: 4 (Skip counting in twos to 10) and 9 (Knowledge of ordinal numbers).

- **Sepedi group:** Substantial adverse DIF effects are evident across many items. For instance, items 1 (Count forwards to 20) and 10 (Compare two collections of objects) appear to approach two-logit DIF contrasts. Other adverse DIF items favouring English include 1 (Count forwards to 20), 2 (Count backwards from 10), 4 (Skip counting in twos to 10), and 5 (Count with 1:1 correspondence). Items favouring Sepedi include Items 8 (Subitise to 5), 9 (Knowledge of ordinal numbers), 10 (Compare two collections of objects), and 11 (Show a collection without counting).

- **Tshivenda group:** The majority of items on the Tshivenda Mathematics assessment show adverse DIF and favour English, including: 1 (Count forwards to 20), 2 (Count backwards from 10), and 4 (Skip counting in twos to 10) and 5 (Count with 1:1 correspondence). Items favouring Tshivenda are Items 7 (Number recognition) and 11 (Show a collection without counting).

In sum, Afrikaans shows no DIF with English.

For the other South African languages:
- Items 1 (Count forwards to 20) and 2 (Count backwards from 10) show bias to English in all African languages except isiZulu and isiXhosa.
- Item 5 (Count with 1:1 correspondence) shows bias to English in all African languages except isiXhosa.
- Item 4 (Skip counting in twos to 10) shows bias to English is evident for Sepedi, Tshivenda and Setswana.

Given the mix of likely quintile and language effects, interpretation is a challenge. However, it is evident that African language groups with fewer low quintile children have isiZulu and isiXhosa have less DIF in relation to English, suggesting once again the confounding effect of the children's socio-economic backgrounds in these analyses.

ITEM DIFFICULTY COMPARISON ACROSS LANGUAGES

All languages were combined to assess Mathematics item difficulty (known as Omnibus DIF). Estimates are compared 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 Mathematics (v1) Omnibus DIF Measures

	ITEM	ENG	AFR	XHO	ZUL	SET	SES	SEP	TSH	RANGE
1	Count forwards to 20	-3.42	-4.05	-4.03	-3.24	-2.07	-2.71	-1.36	-1.2	2.85
2	Count backwards from 10	-1.89	-1.61	-1.22	-1.7	-0.86	-1.14	-0.72	-0.14	1.75
3	Counting from a given number	-1.15	-1.01	-1.32	-1.25	-0.86	-0.93	-1.2	-1.18	0.46
4	Skip counting in twos to 10	-0.61	-0.58	-0.8	-0.32	0	-1.44	0.75	0.35	2.19
5	Count with 1:1 correspondence	-3.2	-3.03	-2.91	-2.5	-2.38	-2.62	-2.57	-2.33	0.87
6	Number order	-2.82	-2.53	-2.72	-2.27	-2.29	-2.49	-2.73	-2.46	0.55
7	Number recognition	-0.84	-1.15	0.22	-0.84	-0.7	-0.51	-0.49	-1.4	1.62
8	Subitise to 5	0.11	0.44	0.34	0.76	-0.3	-0.05	-0.8	-0.47	1.56
9	Knowledge of ordinal numbers	0.56	-0.79	-0.92	-1.12	-1.14	-0.18	-1.34	-1.49	2.05
10	Compare two collections of objects	0.42	1.14	-0.55	-0.29	-0.67	-0.18	-0.8	-0.6	1.94
11	Show a collection without counting	1.66	1.6	1.13	0.69	0.34	0.88	-0.31	-0.35	2.01
12	Solving addition and subtraction problems	1.58	1.31	2.21	2.11	1.23	2.28	1.66	0.55	1.73
13	Solving sharing and grouping problems	2.8	2.5	3.14	5.94	3.55	2.64	2.66	3.05	3.44

	ITEM	ENG	AFR	XHO	ZUL	SET	SES	SEP	TSH	RANGE
14	Shape and space construction	0.56	0.78	0.6	0.27	0.3	-0.08	0.6	0.4	0.86
15	Sorting and grouping	2.99	3.03	3.74	3.6	3.55	3.18	1.76	2.3	1.98
16	Shape identification and understanding	1.08	1.41	1.48	1.46	0.6	1.12	0.92	0.31	1.17
17	Pattern extension	0.17	0.12	0.65	1.05	-0.2	0.83	0.26	0.77	1.25
18	Pattern completion	1.54	2.14	2.54	2.21	1.96	2.01	2.28	1.62	1

Difficulty estimates range widely between language groups, although rough consistency in item difficulty estimates is indicated by the monotone shading for most items. The summary findings are as follows:

- It is evident that for all languages, the item difficulty generally increases from the first to the final items. Items 14 and 17 could come earlier in the series, but overall, the order of presentation is generally appropriate.
- The smallest variation in difficulty (0.46 logits) between languages is evident for item 3 (Counting from a given number).
- The largest variation in difficulty (3.44 logits) is evident for item 13 (Solving sharing and grouping problems). This item is difficult in all languages (particularly isiZulu), with a minimum difficulty logit of 2.5 for the Afrikaans cohort.
- Attention to translation may be indicated for item 13 in isiZulu as while it is difficult in all languages, Zulu speakers find it particularly challenging. No child achieved 3/3 (100%) for this item, and 192 scored 0. Dichotomous Rasch was used in analyses, and this is why the item comes out as so difficult in this group. (101 children obtained correct scores on either trial 1 or 2, but these values are ignored in the dichotomous Percent Correct calculation.)
- No language group appears to produce systematically higher or lower difficulty estimates across items, when all are combined for omnibus DIF.
- Items with a large range in their difficulty estimates across languages are likely to indicate bias in DIF analysis. It is worth noting that items flagged for DIF against the English reference group (items 1, 2, 9, 10, and 11) all produce item difficulty estimate ranges approximating or exceeding two logits. We noted previously that items 1 (Count forwards to 20) and 2 (Count backwards from 10) showed bias to English in some African languages (not isiZulu and isiXhosa) This observation is confirmed.
- Item 5 (Count with 1:1 correspondence) shows bias to English in all African languages except isiXhosa.
- Item 4 shows a degree of bias to English for Sepedi, Tshivenda and Setswana.

CONCLUSION

Psychometric analyses of the *ELOM-R Mathematics (v1) Assessment indicate that it is a reliable and valid measure of children's abilities*. Item difficulty increases largely similarly over the course of the test in all languages.

Attention to translation may be indicated for item **13** in isiZulu as while it is difficult in all languages, Zulu speakers find it particularly challenging. As noted above, this finding is likely influenced by the binning method as well as the trial order for the item in isiZulu. With this possible exception, the ELOM-R Mathematics (v1) Assessment is suitable for use without changes to items. It also means that a single sample combining all the language groups can be constructed for final psychometry and norming. This is covered in the final section of the Manual.

Furthermore, efforts are currently underway to establish the criterion validity of the ELOM-R Mathematics (v1) Assessment by examining the regression between ELOM-R Mathematics (v1) Scores collected in Grade R with Grade 1 Early Grade Mathematics Assessment (EGMA). Theoretically, high scores on ELOM-R (v1) should translate to higher scores on EGMA in Grade 1.

Analyses conducted for standardisation combine all languages into a much larger sample. Psychometry in that group is reported next in Chapter 3 below.

CHAPTER 3. STANDARDISATION AND NORMS

Here, we present psychometric analyses undertaken on a combined sample of eight languages to standardise the ELOM-R Mathematics (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 14.

Table 14. ELOM-R Mathematics (v1) Sample for Standardisation and Norms

Home Language	Total	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
1. English	282	14	33	118	48	49
2. Afrikaans	448	84	84	38	141	101
3. isiZulu	281	43	56	61	82	39
4. isiXhosa	290	23	74	101	57	35
5. Sesotho	287	68	64	76	46	35
6. Setswana	276	240	0	21	8	7
7. Sepedi	286	218	13	23	16	16
8. Tshivenda	290	98	63	109	20	0
TOTAL after exclusion of outliers	**2440**	**788**	**386**	**546**	**418**	**302**

It was established that the isiZulu group had two cases with markedly low scores. These were removed. The final standardisation and norming sample includes **2440** cases. The poor representation of quintile 4 and 5 children in some languages will affect findings. Language and quintile are confounded.

First, the distribution of total scores on the ELOM-R Mathematics (v1) 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, and descriptive statistics describing the range, central tendency, and shape of the distribution are presented in Table 15.

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

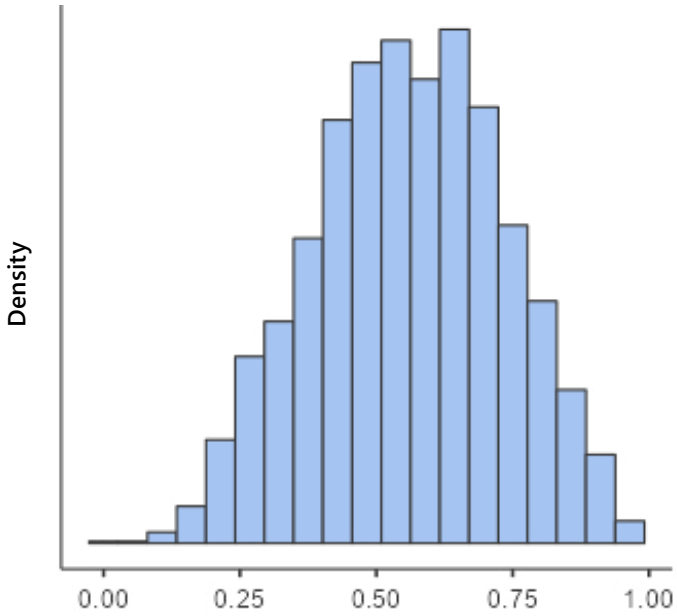


Table 15. ELOM-R Mathematics (v1) Total Percent Correct Score Descriptive Statistics

N	MISSING	MEAN	MEDIAN	SD	MINIMUM	MAXIMUM
2440	0	0.559	0.560	0.174	0.026	0.991

SKEWNESS		KURTOSIS	
SKEWNESS	SE	KURTOSIS	SE
-0.083	0.050	-0.513	0.099

PC scores on the ELOM-R Mathematics (v1) assessment appear normally distributed on the total score histogram. The skewness value is both small and statistically nonsignificant, and the difference between the mean and median is negligible, indicating that the distribution is symmetrical. The distribution does not present with the typical flat peak of a platykurtic distribution and can be considered normal.

ELOM-R MATHEMATICS (v1) STANDARDISATION SAMPLE SCHOOL QUINTILE DISTRIBUTIONS[44]

As scores are normalised across South Africa's diverse population, language groups and socioeconomic status are reported. While both group designations are important to consider in their own right, as previously noted, they are heavily confounded in South Africa. The school quintile composition of each language group is reported in Figure 13 to provide context for consideration of confounding effects. SES is operationalized in terms of the quintiles assigned to the schools from which children were sourced. These are further collapsed into schools that do not require the payment of fees (quintiles 1, 2, and 3), and those that do (quintiles 4 and 5).

[44]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 of households in the school catchment area, as obtained from South African national census data.

Figure 13. 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. Sample school quintile composition is reported in Table 16.

Table 16. 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 16 reveals very different school quintile distributions across the language groups. Sepedi, Setswana, and Tshivenda fee-paying schools are poorly represented, and SES effects are likely to heavily influence test performance in these groups. The Sesotho cohort may be less affected than Sepedi or Setswana, as they possess far greater numbers of quintile 2 and 3 schools. Subsamples for paying and non-paying schools for all other language groups appear reasonably well populated. Next, the psychometric properties of the ELOM-R Mathematics (v1) Assessment norm sample are assessed to establish the reliability and validity of its scale scores.

Psychometric properties of the ELOM-R Mathematics (v1) Standardisation sample

To assess whether the ELOM-R Mathematics (v1) items are consistent in their measurement of mathematical ability across all the subsamples included in the norms, reliability testing procedures were undertaken. Reliability was tested using McDonald's omega (ω), which assesses the internal consistency of assessment scores. Results are presented in Table 17.

Table 17. ELOM-R Mathematics (v1) Reliability Statistics

	Item-rest correlation	ω
ELOM-R (v1) Language Assessment		**0.864**
When item excluded...		
1 Count forwards to 20	0.400	0.860
2 Count backwards from 10	0.586	0.853
3 Counting from a given number	0.609	0.851
4 Skip counting in twos to 10	0.448	0.858
5 Count with 1:1 correspondence	0.394	0.860
6 Number order	0.609	0.851
7 Number recognition	0.563	0.853
8 Subitise to 5	0.574	0.853
9 Knowledge of ordinal numbers	0.525	0.855
10 Compare two collections of objects	0.318	0.863
11 Show a collection without counting	0.399	0.860
12 Solving addition and subtraction problems	0.561	0.853
13 Solving sharing and grouping problems	0.364	0.861
14 Shape and space construction	0.418	0.859
15 Sorting and grouping	0.389	0.860
16 Shape identification and understanding	0.571	0.853
17 Pattern extension	0.421	0.859
18 Pattern completion	0.323	0.863

The ELOM-R Mathematics (v1) Assessment demonstrates a very acceptable level of reliability (ω = 0.864). No items produce sub-threshold item-rest correlations (r < 0.3) or detract from scale reliability (ω when item removed < 0.864). The ELOM-R Mathematics (v1) Assessment can be considered a reliable measure within the norm group.
Next, a confirmatory factor model is fitted to the norm sample to establish construct validity for the mathematics assessment within this cohort.

CONFIRMATORY FACTOR ANALYSIS (CFA)

As in earlier sections, a unidimensional factor model was specified in which all items contribute towards the same underlying Mathematics construct. Fit statistics in Table 18 describe the fit of this model to the observed data. Factor loadings of individual items to the single factor are evaluated to assess potential misfit at the item level. The tables below show that Model fit statistics (RMSEA < 0.05, CFI = 0.921, TLI = 0.910) and factor loadings (λ > 0.3, p < .001) are all acceptable, supporting construct validity of the ELOM-R Mathematics (v1) Assessment.

Table 18. ELOM-R Mathematics (v1) CFA Model Fit

χ^2	Df	P	CFI	TLI	RMSEA	Lower CI	Upper CI
920.26	135	< .001	0.921	0.910	0.049	0.046	0.052

Table 19 reports CFA model loadings using percent correct scores for each item. It will be evident that all loadings exceed the criterion (λ>0.3).

Table 19. ELOM-R Mathematics (v1) CFA Model Factor Loadings

Item	Estimate	SE	Z	P	λ
1 Count forwards to 20	0.103	0.005	20.564	< .001	0.423
2 Count backwards from 10	0.284	0.009	33.143	< .001	0.636
3 Counting from a given number	0.282	0.008	34.649	< .001	0.657
4 Skip counting in twos to 10	0.208	0.009	23.682	< .001	0.479
5 Count with 1:1 correspondence	0.131	0.006	20.573	< .001	0.423
6 Number order	0.256	0.007	35.525	< .001	0.670
7 Number recognition	0.17	0.005	32.099	< .001	0.619
8 Subitise to 5	0.18	0.006	32.125	< .001	0.619
9 Knowledge of ordinal Numbers	0.162	0.006	29.176	< .001	0.573
10 Compare two collections of objects	0.066	0.004	16.331	< .001	0.342
11 Show a collection without counting	0.126	0.006	21.297	< .001	0.436
12 Solving addition and subtraction problems	0.19	0.006	30.543	< .001	0.596
13 Solving sharing and grouping problems	0.1	0.005	18.79	< .001	0.390
14 Shape and space construction (copy shape from models)	0.128	0.006	22	< .001	0.449
15 Sorting & Grouping	0.094	0.005	19.981	< .001	0.412
16 Shape identification and understanding	0.15	0.005	31.668	< .001	0.612
17 Pattern extension	0.173	0.008	22.051	< .001	0.450
18 Pattern completion	0.093	0.006	16.286	< .001	0.341

CONCLUSION

The ELOM-R Mathematics (v1) Assessment is construct-valid in the total sample. With the normality, reliability, and construct validity of the mathematics assessment established, we proceed to standardise and construct norms for the South African ELOM-R Mathematics (v1) Assessment.

Standardisation

As the ELOM-R Mathematics (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[45]; Kline, 2005).

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 such that they 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.

Setting the ELOM-R Mathematics (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 Mathematics (v1) and Language (V1) 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 of the ELOM 4&5 Years Assessment tool. As noted in the ELOM 4&5 Technical Manual, it is international practice to set early learning standards between the 50th and 60th percentile of the norm sample standardized 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 Mathematics (v1) 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.

[45]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.

STANDARDISED SCORE DISTRIBUTIONS

Figure 14[46] presents the standardised distributions of both raw and normalised ELOM-R Mathematics (v1) scores. Raw scores across the full sample of 2440 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 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 are substantial differences in the performance of students within these quintile groups, with scores increasing monotonically as a function of students' quintile grouping. It is worth noting the disproportionate increase in median (Z = 0) scores between quintile 4 and 5 subsamples (8%), which is greater than the combined score difference across quintiles 1 through 4 (7.4%).

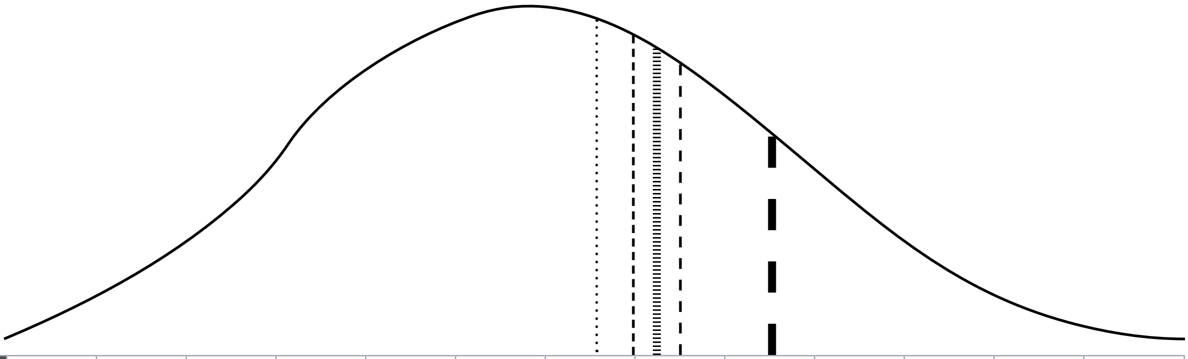
Figure 14 indicates that ELOM-R Mathematics (v1) scores for lower quintile schools are consistently lower compared to higher quintile schools, and as a result, we expect significant group effects based on quintile, which will translate to language groups as well. Generally, scores for lower quintile schools are consistently lower compared to children in higher quintile groups. It should be reiterated that these differences cannot be purely ascribed to differences between the SES of children within these groups.

In summary, the ELOM-R Mathematics (v1) Assessment is considered a reliable and construct valid tool for the assessment of abilities covered in the Grade R CAPS. 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 standards applicable across quintile groups with known ability distribution differences, the observed median score differences are acceptable. No accumulation of DIF effects was observed, meaning score differences likely reflect the difference in underlying numeric ability (as opposed to bias), and are meaningful indicators of the need for greater support with regard to lower-scoring language cohorts.

The next step in this process is to set the standard of expected performance on the ELOM-R Mathematics (v1) Assessment for children entering Grade 1. This will be done by defining the appropriate cut score (for example, the 60th percentile) in consultation with appropriate stakeholders.

[46]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 Mathematics (v1) Standardised Score distribution



	Z-Score	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3
	Norm Score (Percentile)	**0**	**0.1**	**2**	**7.8**	**17**	**31.9**	**49.7**	**66.8**	**82.9**	**93.3**	**98.4**	**100**	**100**
Raw Percentage Correct Scores	**Norm Sample**	**4.0**	**12.6**	**21.3**	**30.0**	**38.6**	**47.2**	**55.9**	**64.6**	**73.2**	**81.8**	**90.6**	**99.1**	**100**
	Quintile 1	1.7	9.9	18.1	26.4	34.6	42.8	51.1	59.3	67.5	75.8	84.0	92.3	100
	Quintile 2	0.5	9.4	18.2	27.1	35.9	44.8	53.7	62.5	71.4	80.2	89.1	97.9	100
	Quintile 3	6.3	14.7	23.1	31.5	39.9	48.3	56.7	65.1	73.5	81.9	90.3	98.7	100
	Quintile 4	11.7	19.5	27.3	35.1	42.9	50.7	58.5	66.3	74.1	81.8	89.6	97.4	100
	Quintile 5	16.7	25.0	33.3	41.6	49.9	58.2	66.5	74.8	83.1	91.4	99.7	100	100

Key: Quintile Median scores are indicated in lines in the Figure.

Quintile median	**1**	....
Quintile median	**2**	--
Quintile median	**3**	||||||||||||||||||||
Quintile median	**4**	- - - - - - - - - - - - - - - - - - - -
Quintile median	**5**	▬ ▬ ▬ ▬ ▬ ▬ ▬ ▬ ▬ ▬ ▬ ▬ ▬

NORMS

Table 20 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 20. ELOM-R Mathematics (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 % 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 % 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) RAW SCORES

Steps
1: Calculate the mean % 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 = 47.3, 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 47,3 or less on this test. The corresponding Z score in the table tells you how many standard deviations above (+) or below (-) your sample's score is from the mean. The corresponding Z-score in the table tells you how many standard deviations above (+) or below (-) your sample % 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
18.2	-2.18	1	47.3	-0.50	32	61.3	0.31	60
21.3	-2.00	2	47.8	-0.47	33	61.8	0.34	61
23.4	-1.88	3	48.3	-0.44	34	62.3	0.37	62
25.3	-1.77	4	48.8	-0.41	35	62.6	0.39	63
26.8	-1.68	5	49.3	-0.38	36	63.1	0.41	64
28.1	-1.61	6	49.7	-0.36	37	63.6	0.45	65
29.1	-1.55	7	50.4	-0.32	38	64.1	0.47	66

FALLING FAR BEHIND			FALLING BEHIND			ON TRACK		
Raw Score	Z	Percentile	Raw Score	Z	Percentile	Raw Score	Z	Percentile
30.4	-1.48	8	50.9	-0.29	39	64.7	0.51	67
31.4	-1.42	9	51.4	-0.26	40	65.2	0.54	68
32.6	-1.35	10	51.8	-0.24	41	65.6	0.56	69
33.6	-1.29	11	52.2	-0.21	42	66.1	0.59	70
34.5	-1.24	12	52.9	-0.17	43	66.6	0.62	71
35.3	-1.19	13	53.2	-0.16	44	67.0	0.64	72
36.4	-1.13	14	53.7	-0.13	45	67.6	0.67	73
37.2	-1.08	15	54.3	-0.09	46	68.1	0.70	74
38.0	-1.04	16	54.7	-0.07	47	68.6	0.73	75
38.6	-1.00	17	55.1	-0.05	48	69.0	0.76	76
39.2	-0.97	18	55.6	-0.02	49	69.5	0.78	77
40.1	-0.92	19	56.1	0.01	50	70.1	0.82	78
40.9	-0.87	20	56.6	0.04	51	70.5	0.84	79
41.5	-0.83	21	57.2	0.07	52	71.2	0.89	80
41.9	-0.81	22	57.6	0.10	53	71.9	0.92	81
42.5	-0.78	23	58.0	0.12	54	72.5	0.96	82
43.0	-0.75	24	58.4	0.14	55	73.3	1.00	83
43.5	-0.72	25	59.0	0.18	56	73.9	1.04	84
44.0	-0.69	26	59.6	0.21	57	74.6	1.08	85
44.6	-0.65	27	60.1	0.24	58	75.2	1.11	86
45.2	-0.62	28	60.6	0.27	59	76.1	1.16	87
45.8	-0.59	29				76.9	1.21	88
46.3	-0.56	30				77.9	1.27	89
46.9	-0.52	31				78.8	1.33	90
						79.8	1.38	91
						80.6	1.43	92
						81.4	1.47	93
						82.9	1.56	94
						84.1	1.63	95
						85.5	1.71	96
						87.1	1.80	97
						89.5	1.94	98
						92.6	2.12	99
						99.1	2.50	100

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

ITEM	TRIALS	SCORING
1 Count forwards to 20	1	**Task:** The child is asked to "see how far you can count". **Scoring:** Record the total number counted in the correct order (e.g.: 1,2, 3, 4, 5, 6, 9: in this example, the correct score is 6. Total possible score = 20.
2 Count backwards from 10	1	**Task:** The child is asked to count backwards from 10 to zero/nought. **Scoring:** 1 point per number counted in the correct order. Total possible score = 11.
3 Counting from a given number	2	**Task:** The assessor asks the child to count from 5 to 10 and 9 to 14. **Scoring:** 1 point for each correct response. Total possible score = 2.
4 Skip counting in twos to 10	1	**Task:** The child is asked to count in twos, from four up to ten. **Scoring:** Total possible score = 3.
5 Count with 1:1 correspondence	1	**Task:** Assessor asks the child to count the (20) counters placed on the table. **Scoring:** The total number counted. Total possible score = 20.
6 Number order	1	**Task:** The child is given cards numbered 0-10 and is asked to put them in the correct order. **Scoring:** The last number placed in the correct order is the score (e.g. if card with number 6 is last card, score = 6). Total possible score = 10. Scoring: 0 = 1, 1-5 = 1; 6-10 =1. Total possible = 3.
7 Number recognition	2	**Task:** The child is asked to give the name of numbers presented on cards. **Scoring:** a point for each correctly named number. Total possible score = 2.
8 Subitize to 5	5	**Task:** The assessor's tablet flashes a gif card with dots on it for 2 seconds (there are 5 trials); the child is asked to say how many dots they see. **Scoring:** One point for a correct answer to each trial. Total possible score = 5.
9 Knowledge of ordinal Numbers	6	**Task:** The child is shown a picture of children running in a race and is asked to name to the position of the child relative to others in the race (first, fifth, last, etc) **Scoring:** One point for a correct answer to each trial. Total possible score = 6.
10 Compare two collections of objects	4	**Task:** The assessor places two piles of counters before the child: one has five and the other nine. The child is asked which of the two groups has more counters, which has fewer, and whether they are able to make the piles equal to each other. **Scoring:** For questions 1 and 2, one point is awarded for a correct answer. For question 3 (equalizing), the child receives 1 point for using 1:1 correspondence and 2 points if the piles are counted. Total possible score = 4.
11 Show a collection without counting	5	**Task:** The child is given a string of beads and asked to show the assessor 3, 7 and 5 beads. **Scoring:** One point for each correct answer. Total possible score = 3.
12 Solving addition and subtraction problems	4	**Task:** The child is asked 4 addition and subtraction word problems. **Scoring:** One point per correct answer. Total possible score = 4.
13 Solving sharing and grouping problems	3	**Task:** the assessor asks the child three sharing and grouping word problems. For example: "Granny has 10 bananas. She wants to share them equally between 5 children. How many bananas will each child receive?" **Scoring:** One point per correct answer. Total possible score = 3.
14 Shape and space construction (copy shape from models)	2	**Task:** The child is given shapes (black and red triangles, circles and squares) and is asked to use them to copy two designs under different instruction conditions (aeroplane and man). **Scoring:** Depends on the number of counters correctly positioned for each construction. Scoring 6 for aeroplane, 7 for man. Total possible score = 13.
15 Sorting & Grouping	4	**Task:** The child is given 9 cards (3 triangles, squares and circles); each shape has three colours and three sizes. The child is asked to sort the cards into three groups on the basis of a common attribute. The child is then asked why they sorted in that manner. In the second and third trials the child is asked to sort in two different ways and give reasons. **Scoring:** 3 points for correctly sorting the shapes into groups, and another 3 points for correctly describing the attributes of the shapes that they used for grouping. Total possible score = 6.
16 Shape identification and understanding	6	**Task:** The assessor places a shape chart in which the circles, squares, rectangles and triangles are presented in different colours, orientations and sizes. For triangles, different types of triangles are presented. After ensuring the child recognizes the differences between the shapes, they are asked to describe the attributes of each. **Scoring:** Naming a shape correctly = one point each. Describing attributes of a circle and triangle = 1 point each. Explaining the difference between a rectangle and a square = 2 points. Total possible score = 8.
17 Pattern extension	7	**Task:** The assessor presents a strip of shapes making up a pattern (ABC AB) with seven blank spaces. The child is provided with a variety of coloured circles and squares that they are told to use to fill the missing spaces with the correct shape of the right colour and size. **Scoring:** One point for every correctly placed shape. Total possible score = 7.
18 Pattern completion	1	**Task:** The child is shown three pattern cards, with blank spaces in between the pattern. The assessor asks the child to point to the shape that they think should complete the pattern. **Scoring:** One point for every correct answer; the last trial has two blank spaces and counts for 2 points. Total possible score = 4.