Investigation of Differential Item Functioning analyses on the Early Learning Outcomes Measure (ELOM 4&5), across all 11 South African languages

Matthew Snelling and Andrew Dawes

July 2023

1. Introduction

The purpose of conducting analyses of Differential Item Functioning (DIF)

The primary objective of this work was to execute Differential Item Functioning (DIF) analyses on the Early Learning Measure 4&5 Years Assessment tool (ELOM 4&5), across all 11 South African languages represented in the expanded dataset. The insights derived from these analyses can be used to enhance the ELOM 4&5 Technical Manual, through adding to existing analyses of DIF on the original 2016 sample, thus completing the assessment of ELOM 4&5 measurement equivalence.

Questions of bias and equivalence are fundamental issues that must be addressed in the development of tests, such as the ELOM 4&5 that are intended for use in several cultural or linguistic populations (Van der Vijver & Tanzer 2004[1]; Milfont and Fischer 2010[2]; Peña, 2007; Peña & Quinn, 1997[3]). Such tests must be assessed for their linguistic, cultural, functional and metric equivalence.

Linguistic equivalence is established when the words and linguistic meaning used in the instruments and instructions are the same across the languages of administration. Related, functional equivalence is established when the test instructions elicit the same behaviour in children from different backgrounds. The underlying constructs being measured, must be understood in the same way and demonstrate the same psychological (factor) structure, and scale items must permit all children regardless of their background, to demonstrate their true ability. Bias is evident when the items are formulated in a manner that the instrument makes children from particular ethnolinguistic backgrounds more or less likely to demonstrate their true competencies. It is therefore necessary to establish the metric equivalence of a test for the languages that test-takers speak. This is a central property of metric equivalence, and it is assessed using statistical tests, including factor analysis and item response modelling (DIF), to determine whether tools such as ELOM 4&5 measure the same construct on the same scale in each of the language groups on which it is to be used. It is for these reasons that the fairness of the ELOM 4&5 in measuring children from all South Arica's official languages, regardless of their language background, is important to establish.

The metric equivalence of the ELOM 4&5 was established for the five language groups included in the 2016 standardisation sample using Rasch analysis and DIF (Snelling, et al., 2019[4]): English, Afrikaans, isiXhosa, isiZulu and Setswana. Since then, we have been able to assess children in the remaining languages and assemble the samples used in the analyses described here. An adaptation for deaf children is under way.

It is important to note the distinction between "*benign*" and "*adverse*" DIF (Breslau et al., 2008[5]).

Benign DIF occurs when groups differ in their probabilities of endorsing an item because the item taps a dimension of the underlying trait or attribute that manifests *differently between the groups*. In this case, a finding of DIF in the statistical analysis reflects *real* differences between the groups in the manifestation of the underlying trait, for example as a consequences of cultural differences.

Adverse DIF reflects biases in the measurement process and occurs when groups differ in their probabilities of endorsing an item because, for example, of different understandings of a word or phrase used in the item instructions or task, or because of challenges in scoring the item (a form of measurement error). In sum, adverse DIF reflects biases in the measurement process which one would wish to avoid.

The results of a DIF analysis does not in itself reveal its *benign* or *adverse* nature. This is for the researcher to determine on the basis of their knowledge of the test and the languages involved.

2. Methods

ELOM 4&5

The ELOM 4&5 was utilised in this study. The ELOM is a rigorously standardised South African pre-school child assessment tool designed for use with children aged 50 - 69 months. The measure comprises 23 items divided into five domains:

1. Gross Motor Development (Items 1 to 4),
2. Fine Motor Control and Visual Motor Integration (Items 5 to 8),
3. Emergent Numeracy and Mathematics (Items 9 to 13),
4. Cognition and Executive Functioning (Items 14 to 17), and
5. Emergent Language and Literacy (Items 18 to 23).

Sample size for DIF

There are different views on the necessary sample size for DIF analyses, and this depends on the method used. Scott et al (2009)[6] see a sample size of 200 as appropriate for IRT Rasch analysis for scales of more than 2 or 3 items, while Serici & Rios (2013[7], p.177) state that "Poor DIF detection results using small sample sizes (N = 100)" and suggest that an adequate sample size would be around 250–500 in each group. Finally, Lai, Teresi & Gershon (2005[8]) state that several studies have shown that a sample size greater than 100 is required to detect DIF, but that around 200 is adequate when 1-parameter logistic (1-PL Rasch) model[9],[10] is used. As 1-PL Rasch modelling is used in the analyses conducted here, we have followed their estimation.

Study sample

The sample consisted of 15,487 ELOM assessments, including 7,510 males and 7,977 females ranging in age from 49 to 70 months (mean age = 58.32 months). Sample characteristics are shown in Tables 1, 2 and 3.

Table 1. Sample Home Languages

Language	Frequency
Afrikaans	2,329 (15%)
English	1,405 (9.1%)
Isindebele	78 (.5%)

Isixhosa	2,800 (18.1%)
Isizulu	3,256 (21.0%)
Sesotho	1,106 (7.1%)
Sesotho se leboa (Sepedi)	1,190 (7.7%)
Setswana	2,620 (1.6%)
Siswati	243 (1.6%)
Tshivenda	289 (1.9%)
Xitsonga	171 (1.1%)
Total	15,487 (100%)

For most languages sample sizes are more than adequate. However, isiNdebele in particular is < 100 and findings must be treated with some caution as they may be unreliable. Xitsonga is somewhat below our target, but following Lai et al., we regard it as acceptable and not likely to generate unreliable findings.

Table 2. Sample Age in months and sex

Age		Sex	Frequency
Mean Age (months)	58.32 (SD =5.426) Range: 49; 70	Male	7510 (48.5%)
Median Age (months)	58.00	Female	7977 (52,5%)
		Total	15487

Table 3. Early Learning Programme participation

Responses to the question: *For how many years has this child been in the programme?*				
	Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Do Not Know	153	1.0	1.2	1.2
1st year in the programme	7,080	45.7	54.3	55.4
2nd year in programme	3,289	21.2	25.2	80.6
3rd year in programme	2,526	16.3	19.4	100.0
Non-Missing Total	13,048	84.3	100.0	
Missing (Don't know)	500	3.2		
Missing (Unknown)	1,939	12.5		
Total Missing	2,439	15.7		
Grand Total	15,487	100.0		

3. Analysis using Rasch Modelling

The Rasch model, named after Georg Rasch, is a psychometric model for analysing categorical data, such as answers to questions on a reading assessment or questionnaire responses. This model transforms the raw score data into interval measures, which can then be analysed using powerful statistical modelling techniques. It provides tools for examining the measurement properties of assessment items and the performance of respondents, thereby ensuring the validity and reliability of the scores (Bond & Fox, 2015[11]).

Winsteps® Software

Winsteps® (https://winsteps.com/) is a software tool commonly used for conducting Rasch analysis and was used to carry out the DIF analyses on the ELOM 4&5 data across all 11 languages. It provides user-friendly interfaces and comprehensive output for examining item and person statistics, item fit, reliability estimates, and importantly for this study, Differential Item Functioning.

Procedure

Data Input: Our initial step in the Rasch modelling process was data input, performed using *Winsteps®* software. This involved importing a data file containing responses to the ELOM items. The data file, in a format compatible with the software, was structured so that each row represented a single respondent (in this case, a child) and each column represented a response to an item on the ELOM.

Model Specification: After loading the data, we specified the model for analysis. This study utilised a partial credit model, appropriate for items with more than two response categories.

Model Estimation: Post model specification, the *Winsteps®* software estimated the model parameters - the item and person measures. This process was iterative, and the derived parameters were saved for further analysis. A1-PL Rasch model was used. The findings permit one to consider the elimination of items that do not fit a pre-specified model and retain those that do[12].

Examining Model Fit

Fit Statistics: Evaluation of the model fit in our Rasch analysis involved examining fit statistics, providing insight into how well our data conformed to the model's expectations. We used *infit* and *outfit* mean-square statistics, with values close to 1.0 indicating a good fit. Notably, values significantly greater than 1.0 suggested that an item was less predictable than the model expected (termed *underfit*), while values significantly less than 1.0 suggested the item was more predictable than the model expected (*overfit*). Underfitting the model is more problematic than overfitting (Tessio, et al., 2023[13]) as it indicates that the item is less able to discriminate between children of high and low ability (it has low 'person' reliability).

Reliability: We also analysed reliability to gauge model fit, estimating the reproducibility of item and person measures (the consistency of the ELOM to discriminate between more and less able children, and consistency in item difficulty). Both forms of reliability are equivalent to Cronbach's alpha and are indicative of the measure's internal consistency.

- The *person reliability index* quantifies the extent to which the responses of a given child to an item conform to the Rasch model expectation that a more able child should have a higher probability of passing an item than a less able child.

  Interpretation: Person reliability values $>.50$ are regarded as acceptable for a A1-PL Rasch model [14]. Those below that value indicate that either there were not enough individuals in the sample with more extreme abilities (both high and low), or there are too few items in the measure to provide a valid assessment of person reliability (consistency in response).

- The item reliability index quantifies the level of difficulty of each item (by ranking the probability of passing).

  Interpretation: Higher reliability indices (closer to 1.0) suggest greater consistency Low item reliabilities indicate that the sample was too small to confirm the item difficulty of the instrument (Cordier, et al., 2018)[15]. Reliability values $>.50$ are regarded as acceptable for a 1-PL Rasch model.

- Lastly, we used the Root Mean Square Error (RMSE) as another measure to assess model fit. The RMSE reflects the standard deviation of residuals, defined as differences between the observed and predicted responses. Lower RMSE values indicated a better model fit (the predicted and observed responses are closer).

  Interpretation of RMSEA values: Good: RMSEA < 0.05: Acceptable: RMSEA between 0.05 and 0.08; Marginal: RMSEA between 0.08 and 0.1; Poor fit: RMSEA > 0.1.

4. Findings for ELOM 4&5 domains

The results are presented in Table 4 below.

Table 4. ***Winsteps***^{R} **Rasch Modelling Results**

Domain	Person Reliability	Item Reliability	Item Infit	Item Outfit	RMSE
Gross Motor Development	0.58	0.99	1.01	0.98	0.01
Fine Motor Control and Visual Motor Integration	0.62	1.00	0.8	0.9	0.01
Emergent Numeracy and Mathematics	0.54	1.00	1.01	0.97	0.01
Cognition and Executive Functioning	0.48	0.98	0.99	0.98	0.01
Emergent Language and Literacy	0.63	1.00	0.96	0.94	0.01

1. Gross Motor Development (GMD) domain: the person reliability was .58, item reliability was .99, item infit was 1.01, item outfit was .98, and RMSE was .01.

2. Fine Motor Control and Visual Motor Integration (FMC - VMI) domain: the person reliability was .62, item reliability was 1.00, item infit was .80, item outfit was .90, and RMSE was .01.

3. Emergent Numeracy and Mathematics (ENM) domain: the person reliability was .54, item reliability was 1.00, item infit was 1.01, item outfit was .97, and RMSE was .01.

4. Cognition and Executive Functioning (CEF) domain: the person reliability was .48, item reliability was .98, item infit was .99, item outfit was .98, and RMSE was .01.

5. Emergent Language and Literacy domain (ELL): the person reliability was .63, item reliability was 1.00, item infit was .96, item outfit was .94, and RMSE was .01.

The findings may be summarised as follows:

- All domains have good fit.
- Item reliability values were high across all domains, suggesting sufficient precision of item calibrations, even in this highly heterogenous sample.
- Person reliability values varied across domains. Only the Cognition and Executive Functioning domain (.48) is marginally below the threshold for a A1-PL Rasch model. This value, albeit borderline, was considered acceptable.
- RMSEA values for all domains are below 0.05 and have very good Infit and Outfit.

Differential Item Functioning

For the purposes of this study, DIF was judged at the level of the domain (summed DIF across items in a domain (Linacre, 2016[16]) as reported in Table 6 for all languages. It is suggested by Linacre that when person reliability is relatively low, then more stringent thresholds can be set to account for increased standard error. In addition, this correction is also recommended for a small sample (in this case IsiNdebele). A threshold of .025 is therefore used for the results displayed in Table 5.

Table 5. Differential Item Functioning Results by Language for each Domain

Language	GMD DIF	FMC - VMI DIF	ENM DIF	CEF DIF	ELL DIF
Afrikaans	-0.0038	-0.0513	-0.0134	0.072	-0.0062
English	-0.0239	0.0954	0.0028	0.0543	0.021
IsiNdebele	-0.0049	0.022	-0.004	0.1076	-0.0831
IsiXhosa	-0.0042	-0.0329	0.0098	-0.0154	-0.0085
IsiZulu	0.0001	-0.0033	0.0046	0.0007	-0.0146
Sesotho	0.0053	-0.0466	-0.0062	0.0252	-0.0102
Sepedi	0.0136	-0.0101	0.0101	0.0472	-0.022
Setswana	-0.004	0.0019	0.0092	0.0003	-0.0142
Siswati	0.0053	-0.0255	-0.0074	0.0794	0.0119
Tshivenda	0.0281	-0.0161	0.0123	0.0344	0.004
Xitsonga	0.0046	-0.0293	-0.0084	0.0413	-0.0012

In summary:

- DIF values for the domains are all acceptable.
- Measurement (metric) equivalence is evident across all eleven languages.
- The ELOM 4&5 does not discriminate between children on the basis of their language.

Overall, these domain level findings support the use of the ELOM as a reliable and fair tool for assessing child development outcomes across all 11 official languages.

5. Findings for ELOM 4&5 item DIF.

It is of interest to examine DIF for specific items. Heat maps of DIF for each item and language are provided in Tables 6.1-6.5. We use the criterion of =>.50 to indicate DIF at the item level. Values exceeding .50 are indicated in **red**.

DIF is evident for **4** of the **23 items.**

- FMC-VMI item 6 *Copy Triangle*: All languages**;** it is possible that this due to assessor challenges in scoring triangle drawings (adverse bias attributable to measurement error).
- CEF: item 17 *Picture puzzle completion*: Siswati; The most likely reason for DIF could be low sample ability variance (the ability range in the sample of Siswati speakers is narrow).
- ELL: Item 18 Expre*ssive language: empathy*: IsiXhosa, Sesotho, Tshivenda and Xitsonga;
- ELL: item 23 Initial sound discrimination: all languages except English, and Sesotho.

  In case of these ELL items, it is plausible that these are examples of benign bias occasioned by the nature of these languages. We note that Language tests produce particular challenges in translation (Carter et al. 2005[17]).

It is important to note that single items are not used in comparisons between groups (domains are compared), so one should not be overly concerned about these few observations. When combined in domains (Table 5), it is evident that the individual item DIF findings do not affect domain DIF.

Table 6.1. Gross Motor Development

ITEM	Afr.	Eng.	Isind.	Isixhosa	Isizulu	Sesotho	Sepedi	Setsw.	Siswati	Tshiv.	Xitsonga
1	0.0493	-0.113	-0.3913	0.0013	-0.113	-0.2411	-0.323	-0.2007	-0.0509	-0.3604	-0.1594
2	-0.0329	-0.0329	0.1086	-0.0329	-0.0329	0.0003	-0.063	-0.0329	-0.0329	-0.2081	0.0274
3	0.0847	0.1255	0.3813	0.0549	0.1255	0.0942	0.2597	0.1255	0.2024	0.3491	0.2299
4	-0.1049	-0.0035	-0.1035	-0.0275	0.0205	0.1519	0.1399	0.1041	-0.1133	0.2475	-0.0933

No DIF is evident for any GMD item (no Values higher than .50).

Table 6.2. Fine Motor Coordination and Visual Motor Integration

Item	Afr.	Eng.	Isind.	Isixhosa	Isizulu	Sesotho	Sepedi	Setsw.	Siswati	Tshiv.	Xitsonga
5	-0.1874	-0.1619	-0.3955	-0.3054	-0.3054	-0.4782	-0.4178	-0.3054	-0.354	-0.4172	-0.4452
6	0.7831	0.5927	0.6813	0.7831	0.7831	0.7831	0.8429	0.8354	0.8384	0.8465	0.7579
7	-0.09	-0.1877	-0.1849	-0.1773	-0.1094	-0.0182	-0.1444	-0.2303	-0.1074	-0.3252	-0.2591
8	-0.557	-0.1477	-0.0789	-0.3333	-0.3716	-0.3333	-0.2908	-0.2978	-0.4025	-0.1202	-0.0829

DIF is evident for: item 6 *Copy Triangle*: All languages.

Table 6.3. Emergent Numeracy and Mathematics

Item	Afr.	Eng.	Isind.	Isixhosa	Isizulu	Sesotho	Sepedi	Setsw.	Siswati	Tshiv.	Xitsonga
9	-0.0692	-0.3161	-0.0672	-0.1258	-0.0364	0.2101	0.2129	0.1959	-0.1486	0.1578	0.1424
10	0.0508	-0.0917	-0.1094	0.2732	0.3433	0.1603	0.3306	0.3181	0.0768	0.1836	0.0824
11	0.2073	0.0709	-0.1258	0.0709	0.0916	0.0105	-0.1261	0.05	0.1698	-0.1245	0.1065
12	-0.0725	0.2785	0.3311	0.0864	-0.037	-0.0587	0.1031	-0.1223	0.1409	0.0903	0.1025
13	-0.1298	0.0612	-0.0327	-0.2949	-0.3569	-0.3284	-0.5104	-0.4325	-0.2463	-0.2949	-0.4422

No DIF is evident for any ENM item (no values higher than .50)

Table 6.4. Cognition and Executive Functioning

Item	Afr.	Eng.	Isind.	Isixhosa	Isizulu	Sesotho	Sepedi	Setsw.	Siswati	Tshiv.	Xitsonga
14	0.0652	0.1802	-0.0921	0.0028	-0.0262	-0.1104	-0.1618	-0.0195	-0.0817	0.0279	-0.0601
15	-0.1132	-0.0675	0.302	-0.1546	-0.0675	0.0059	-0.0059	-0.0302	-0.144	0.1677	-0.1731
16	0.4094	0.324	-0.4372	0.0162	-0.148	-0.1504	-0.1473	-0.099	-0.2008	-0.2983	-0.0866
17	-0.2894	-0.3824	0.3349	0.1202	0.2424	0.2801	0.3622	0.149	0.5059	0.1371	0.3611

DIF is evident for: item 17 *Picture puzzle completion* for Siswati

Table 6.5. Emergent Language and Literacy

Item	Afr.	Eng.	Isind	Isixhosa	Isizulu	Sesotho	Sepedi	Sets.	Siswati	Tshiv.	Xitsonga
18	0.4932	0.4638	0.1984	0.8586	0.4059	0.6864	0.4277	0.5233	0.3101	0.5462	0.5462
19	-0.1757	-0.4161	-0.4743	-0.1003	-0.2461	0.0454	-0.0751	-0.0194	-0.139	-0.17	-0.0522
20	-0.0261	-0.0437	0.0369	-0.2721	-0.3644	-0.2958	-0.3452	-0.4191	-0.3262	-0.3447	-0.3492
21	-0.2393	-0.0946	-0.098	-0.4034	-0.16	-0.2963	-0.2963	-0.4898	-0.198	-0.3636	-0.381
22	-0.3415	-0.2714	-0.6003	-0.5364	-0.4243	-0.4039	-0.2787	-0.1118	-0.5429	-0.2493	-0.2676
23	0.2832	0.383	0.8542	0.4451	0.7743	0.254	0.5456	0.5026	0.9079	0.5854	0.5026

DIF is evident for:

- Item 18 Expre*ssive language: empathy*: IsiXhosa, Sesotho, Tshivenda and Xitsonga
- item 23 I*nitial sound discrimination*: all languages except English, and Sesotho.

DIF is more likely in these items given significant language differences in expression and initial sounds. DIF in these areas is to be expected and does not signal problems with the items.

References:

[1] 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.

[2] 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.

[3] Peña, E.D. (2007). Lost in Translation: Methodological Considerations in Cross-cultural Research. *Child Development 78*, 1255-1264. Peña, E.D., and R. Quinn (1997). Task Familiarity: Effects on the Test Performance of Puerto Rican and African American Children. *Language, Speech, and Hearing Services in Schools 28*.4: 323-332.

[4] Snelling, M., Dawes, A., Biersteker, L., Girdwood, E., Tredoux, C.G. (2019). The development of a South African Early Learning Outcomes Measure: A South African instrument for measuring early learning program outcomes. *Child Care, Health and Development, 45*, 257–270, doi:10.1111/cch.12641.

[5] Breslau, J., Javaras, K. N., Blacker, D., Murphy, J. M. & Normand, S. T. (2008). Differential Item Functioning Between Ethnic Groups in the Epidemiological Assessment of Depression. *The Journal of Nervous and Mental Disease, 196* (4), 297-306. doi: 10.1097/NMD.0b013e31816a490e.

[6] Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., ... & Quality of Life Cross-Cultural Meta-Analysis Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of clinical epidemiology*, *62*(3), 288-295.

[7] Sireci, S.G. & Rios, J.A. (2013) Decisions that make a difference in detecting differential item functioning, Educational Research and Evaluation, 19:2-3, 170-187, DOI: 10.1080/13803611.2013.767621

[8] Lai, J. S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & the health professions*, *28*(3), 283-294.

[9] https://irrc.education.uiowa.edu/blog/2020/09/technically-speaking-determining-test-effectiveness-item-response-theory

[10] Linacre, J. M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. Rasch Measurement *Transactions, 19*, 1032. Linacre, J. M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. Rasch Measurement *Transactions, 19*, 1032. https://www.rasch.org/rmt/rmt193h.htm

[11] Bond, T., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences. New York, NY: Routledge.

[12] https://irrc.education.uiowa.edu/blog/2020/09/technically-speaking-determining-test-effectiveness-item-response-theory

[13] Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2023). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, 1-14. DOI: 10.1080/09638288.2023.2169772

[14] https://www.winsteps.com/winman/reliability.htm

[15] Cordier, R., Speyer, R., Schindler, A., Michou, E., Heijnen, B. J., Baijens, L., ... & Joosten, A. V. (2018). Using Rasch analysis to evaluate the reliability and validity of the Swallowing Quality of Life Questionnaire: an item response theory approach. *Dysphagia*, *33*, 441-456. DOI 10.1007/s00455-017-9873-4(0123456789().,-volV)(0123456789().,-volV)

[16] https://winsteps.com/winman/difconcepts.htm

[17] Carter, J. A., Lees, J. A., Murira, G. M., Gona, J., Neville, B. G., & Newton, C. R. (2005). Issues in the development of cross-cultural assessments of speech and language for children. *International Journal of Language & Communication Disorders*, *40*(4), 385-401.