

User guide

EGRS I – Wave 1-5 Panel

1. Background to the EGRS I Project

Data for the Early Grade Reading Study (EGRS I) was collected in South Africa in two districts in the North West Province, Ngaka Modiri Molema and Dr Kenneth Kaunda, in which the main home language is Setswana. The randomised controlled trial (RCT) tested three alternative programmes all aimed at improving the acquisition of home language reading and literacy. The evaluation and implementations were run by the Department of Basic Education in Partnership with the North West Department of Education, 3iE, the Human Science Research Council (HSRC), USAID, Zenex Foundation and UNICEF, and Class Act as the implementing organisation

2. Experimental Design

Three different interventions were run, with 50 schools for each intervention, with an additional 80 schools assigned to be control schools. The first intervention (implemented in 50 schools) provided teachers with lesson plans, additional reading support materials and training at centralized workshops twice a year. The second intervention (implemented in a different group of 50 schools) provides teachers with the same set of lesson plans and additional reading support materials but provides ongoing support to teachers through monthly on-site coaching and small cluster training sessions. The third intervention (implemented in a further 50 schools) holds weekly meetings with grade 1 parents to inform them of the importance of learning to read in the early grades and to empower them with knowledge and tools to become involved in their own child's reading acquisition.

Table 1: EGRS I evaluation design showing treatment and control groups

A structured learning programme & centralised training	A structured learning programme & specialist on-site coaching	Parent intervention	Control Group
50 Schools Teachers were provided with lesson plans aligned to the National Curriculum (NCS & CAPS), additional quality reading materials and two-day training at centralised workshops twice a year.	50 Schools Teachers were provided with the same lesson plans and additional quality reading materials, but professional support took the form of monthly on-site coaching and small cluster training sessions.	50 Schools Parents and caregivers invited to weekly workshops focused on the importance of learning to read and on equipping parents/ caregivers to become more involved in their child's literacy development.	80 Schools Schooling as normal: all the usual policy and support.

EGRS I started with the grade 1 class of 2015 and followed the learners into grade 2 in 2016. The programme continued for a third, year, although no assessments were held in 2017. Although, the intervention ceased, assessments were held in 2018 and 2021 with these same learners.

Table 2: EGRS I evaluation design showing treatment and control groups

Intervention	Grade 1 teachers	Grade 2 teachers	Grade 3 teachers	No intervention	No intervention
	2015	2016	2017	2018	2021
Data collection	Start of Grade 1 (Wave 1)	End of Grade 2 (Wave 3)	No data collection	End of Grade 3 and Grade 4 (Wave4)	End of Grade 7 (Wave5)
	End of Grade 1 (Wave 2)				

More information about the evaluation can be obtained from the EGRS website and the following reports and papers:

Website:

- [Early Grade Reading Study \(education.gov.za\)](https://www.education.gov.za) - Most of the information is available and nicely laid out here.

Papers

- Cilliers, J., Fleisch, B., Prinsloo, C., & Taylor, S. (2020). How to Improve Teaching Practice?: An Experimental Comparison of Centralized Training and In-Classroom Coaching. *Journal of Human Resources*, 55(3), 926–962. <https://doi.org/10.3368/jhr.55.3.0618-9538R1>
- Stern, J. M. B., Jukes, M. C. H., Cilliers, J., Fleisch, B., Taylor, S., & Mohohlwane, N. (2024). Persistence and Emergence of Literacy Skills: Long-Term Impacts of an Effective Early Grade Reading Intervention in South Africa. *Journal of Research on Educational Effectiveness*, 0(0), 1–22. <https://doi.org/10.1080/19345747.2024.2417288>

Wave 1 – Baseline Report

- [Baseline Report](#) (2015) - Baseline data collection and Year 1 programme activities

Wave 2 – Reports

- [Midline Report](#) (2016)

Wave 3 – Reports

- [Technical Report](#) (2018): Impact evaluation after two years of interventions
- [Case Studies](#) (2017) – Home Language Literacy Practices
- [Classroom Observation Study](#) (2017)
- [Policy Summary Report](#) (2017) - Results of Year 2
- [Summary Report](#) (2017) - Results of Year 2

- [Family influences on Early Grade Reading and Literacy](#) (2017)

Wave 4 – Reports

- [Technical Report](#) (Mar 2019)
- [Learning Brief](#) (Mar 2019) – Sustainability of the Early Grade Reading Study

Wave 5 - Reports

- [Report](#) (2022) - First Early Grade Reading Study – Year 7 Report

3. Previous Data Release

An earlier release of the data for Waves 1-3 can be accessed [here](#). This dataset also includes all of the variables that were used for the analysis of the two papers listed above. This would be useful in the case of replication or confirmation of findings.

4. Data collection

There is information on 4,538 learners who were assessed, surveyed and matched in the first wave, who were at evaluated schools. These learners were then tracked to Grade 4 (Wave 4) and later Grade 7 (Wave 5). The number of learners that were reassessed and matched in the following waves of data collections is recorded below, any learner for who we have data for either the oral or written test is included.

Table 3: Learners assessed and dates of assessment for the five waves of data collection

Wave	Learner observations	Data collection dates	Grade (main)
1	4 538	2015	Grade 1
2	4 137	2015	Grade 1
3	3 804	2016	Grade 2
4	3 321	2018	Grade 4
5	2 424	2021	Grade 7

In each in the first three waves learner assessments were completed, as well as a combination of contextual surveys covering principals, teachers and parents (care-givers). Table 4 gives an indication of which assessments and questionnaires were conducted in each wave.

Table 4: Overview of assessments and surveys completed and merged for each wave

Wave	Learner Assessment - Oral & written (LT)	Learner Wellness survey (LW)	Teacher Questionnaire (TQ)	Teacher Assessment (TF)	Principal Questionnaire (PQ)	School Functionality Survey (SF)	Home Background (Parent) Questionnaire (P)	Linked Census, ANA and SNAP school lev-el information (S)	Created variables – for analysis (C)	Programme implementation information (PR)
1	X		X	X	X		X	X		
2	X		X	X	X		X			
3	X		X	X	X		X		X	X
4	X				X	X				
5	X	X			X	X				

Variable from each of the waves in the dataset are named using w[X]_[YY]_[Variable name]. Here [X] indicates the number of the wave. Then [YY] indicated the survey instrument from which this variable was obtained. Lastly, the [Variable name] identifies the relevant question or variable in the survey.

A number of variables containing identifiers and learner information that is consistent across waves is included upfront. These are listed below.

5. Anonymisation of Identifiers

Learners (**learner_id**):

This is a unique learner id. It allows for the identification and tracking of learners across the five waves. Additional anonymisation took place for the data of birth. The variables (**dob_year** **dob_month**) contain the best year and month of birth across the first three waves.

Schools (**school_id**):

School names and EMIS numbers are not included in the dataset, instead, an anonymised school_id can be used to recognize schools. This school id is set to the id of the school of the learners in Wave 1, as in general children were not tracked in general when they moved schools. There are a total of 230 unique school_ids in the dataset.

Information about the location of the school is included in the form of a district_id (either Ngaka Modiri Molema and Dr Kenneth Kaunda) and the school Quintile (**quintile**) is also included in the upfront variables.

The **treatment** allocation (to treatment 1, 2, 3 or Control) was at the school level and treatment allocation is in the variable **treatment** and the three dummy variables for each treatment group (**T1**, **T2** and **T3**). The strata of the school used in the treatment assignment is also included (**strata**)

Teachers:

Teachers were assigned id's within each wave. This means that for the Teacher Questionnaire and Teachers Assessment in Wave 1, teachers with the same name (and generally same schools and class) would have the same ID's. However, even if that same teacher appears in Wave 3, they will have been assigned a different id, and cannot be matched.

In cases of inconsistencies with the class names, this was corrected eg. if there was a school with two 1B classes, but no 1A, then one of the classes was renamed to 1A, taking into account information from the Teacher survey and assessment.

Unique id's were then assigned to each teacher that either completed the survey or wrote the assessment. The total number unique teacher ids by wave are in .

Table 5: Number of unique teacher ids in each wave

Wave	In both TQ and TF	Only in TQ	Only in TF	Total unique IDs
1	268	48	46	362
2	374	7	20	401
3				355

Multi-match cases: Two different teachers for TQ and TF for a single learner within one wave

For Wave 1 there were 6 schools where the teacher who answered the questionnaire and the teacher who wrote the test (TF) for the class had different teachers assigned. The instances are shown below.

schoolid_anon	W1_TF_teacherid_anon	W1_TF_class	W1_TQ_teacherid_anon	W1_TQ_class
10013	1265	1A	1266	1A
10061	1255	1A	1256	1A
10061	1256	1B	1255	1B
10116	1267	1B	1268	1B
10029	1269	1A	1270	1A
10149	1263	1A	1264	1A
10149	1264	1B	1263	1B
10081	1271	1A	1272	1A

In Wave 2 there were 11 schools in which the teachers assigned to each student differed in the teacher's questionnaire and teacher's assessment. The classes were generally also different. The list of mismatched teachers are given below:

schoolid_anon	W2_TQ_teacherid_anon	W2_TQ_classid	W2_TF_teacherid_anon	W2_TF_class
10107	2025	C	2024	B
10107	2023	B	2024	B
10062	2068	A	2069	B
10096	2111	B	2110	A
10098	2114	B	2112	B
10110	2115	A	2116	B
10136	2199	B	2200	A
10136	2200	A	2199	B
10063	2227	A	2228	B
10100	2283	A	2282	A
10182	2331	B	2332	C
10028	2340	B	2339	B
10072	2391	B	2390	A

Principals:

The Principal id's were derived from the principal questionnaires and were matched across waves 1, 2, 3 and 5. Principal data is not available for Wave 4. It is therefore possible to identify schools where the principal (or the person who answered the principal questionnaire) changed over time.

Enumerators:

The anonymised enumerator id's for the learner tests were assigned by wave, so it is not possible to identify an enumerator across waves even if an enumerator worked in two or more waves. However, in Wave 4 enumerators the same anonymised ids were used for both the oral and written tests. The number of unique enumerators for each wave is shown in the table below.

Wave	Wave 1	Wave 2	Wave 3	Wave 4	Wave 4 -Writ.	Wave 5
# of enumerators	64	55	54	51	48	110

6. Test Information

In each wave, a set of age-appropriate assessments were selected for inclusion in the learner assessment. The EGRS I evaluation focussed mainly on supporting teachers in Setswana. Questions for English as a First Additional Language (EFAL) were included for the first time in the learner assessments in Wave 3.

Where possible, item-level information was included. But, particularly for the ORF activities. These were initially completed using paper and pen, and as such for some of the items in Waves 1-3 only the final score is available.

Table 6: Learner Assessment tasks in the various waves of data collection

Skill	Construct	Baseline (W1)		Year 1 (W2)		Year 2 (W3)		Year 4 (W4)		Year 7 (W5)	
		Start of Gr 1		End of Gr 1		End of Gr 2		End of Gr 4		End of Gr 7	
		HL	EFAL	HL	EFAL	HL	EFAL	HL	EFAL	HL	EFAL
Language comp.	Vocabulary	A								5	7
	Listening comprehension										
	Receptive proficiency										
Decoding	Phonological working mem.										
	Phonological awareness	D		G		E					
	Rapid letter naming							1.2			
	Letter sound recognition	B		A		A		2			
	Word reading fluency	E/F		B		B	G1	3	6		
	Non-word decoding			C		C					
	Sentence reading fluency	F		D			G2				
	Oral Reading Fluency (ORF)			E		D1		4.1/ 5	7.1	1.1 / 2.1	3.1
	Reading comprehension			D/ E		D2		4.2	7.2	1.2 / 2.2	3.2
	Written comprehension							8	9	4	6
Writing & spelling	Writing a CVC word										
	Writing			F		F1-3					
	Receptive vocabulary										
Visual & Cognitive	Visual perception										
	Working memory / Digit span	C									
	Rapid Object Naming							1.1			
Maths	Arithmetic					F4		10			

A summary table of the key statistics is provided in Table 7. For tasks with multiple measures (eg. ORF at 60s, at 180s and accuracy in %) only the key measure is reported in the table.

The naming of the outcome variables shown in the table and used in the dataset follows a consistent convention of the following form w[Wave]_LT_SS_[Lan]_[Type]_[Measure]_[Item], where:

- Wave – Indicates that data collection Wave
- Lan – Indicates whether the task was testing home language (H) or EFAL (E) or Mathematics (M).
- Type – This is shortened for to indicate the construct or general skill being tested
- Measure – (Optional) is used where additional information is needed to differentiate within the task, usually because there are numerous outcome measures included for a single task (eg. letter sound knowledge in 30 seconds or per minute)
- Item – This gives the number of the item in the assessment.

A note that for time tasks the optional measure to indicate that no time adjustment has been made is “15s” or “60s” even if the full allocated time was not used or required. If an adjustment has been made for the time taken to complete the assessment, then these have always been standardised to the per minute measure indicated as “pm”.

7. Notes: Data quality and data changes

Standardisation and cleaning of all outcome variables: A check was completed on all outcome variables and the naming was standardised across the 5 waves (where feasible). As the mode of testing changed from paper to digital, more detail was available at the item level in later years.

Not all of the assessment types stayed the same across waves. Even where the type of assessments remained the same, the standard or difficulty of the assessments was often increased as the learners progressed to higher grades.

Note on missing values: In cases where a learner had all missing values for a particular assessment, generally this meant that the learner could not answer and therefore was assigned a zero for this task.

Wave 1

This assessment was completed on paper. For Questions A, B and E missing items were imputed. The total for these variables, with and without imputed values, are included in the datasets.

Wave 2

This wave has information on the time take to complete assessment tasks in the data, although this was not collected in either Wave 1 or 3. As a result per minute, outcomes could be calculated for the timed tasks, including Task A (Letter recognition); Task B (Word recognition); Task C (Non-word decoding) and Task E (ORF / Paragraph reading).

The time variable was cleaned for use in the per minute measure. By default, if there was a contradiction between time taken and words correct (eg. 19 words correct in 0 seconds), the number of words correct was assumed to be the correct measure and the time was adjusted.

The per-minute measure is useful if comparing the information to other studies, however, if tracking across time within the EGRS group, the untimed (in 60 s) outcomes would be better to use as this is more comparable across waves.

Wave 4

One particular type of data entry error was found and corrected. As a result, a small number of observations (generally <15) will have different outcomes to those in previously released datasets and reports. These tended to take place in timed tasks, including rapid letter naming, letter sound recognition, word recognition (or word reading fluency) and oral reading fluency.

The most common change in Wave 4 was where an enumerator appeared to have stopped the assessment due to the learner providing an incorrect answer repeatedly, however the task was not auto-stopped, and as a result, the remaining items in the assessment were counted as correct. Such errors tended to be concentrated among a small number of enumerators.

The following tasks were affected:

- Task 2.1 – Letter Sound Knowledge (6 observations)
- Task 3 – Word recognition (11 observations)
- Task 4.1 – ORF Text 1 (5 observations)
- Task 5.1 – ORF Text 2 (9 observations)
- Task 6 – EFAL Word recognition (20 observations)
- Task 7 – EFAL ORF (16 observations)

Table 7(a): Summary Statistics of Learner Outcome Variables for Home Language

Variable	Wave	N	Mean	SD	Min	Max	p10	p25	p50	p75	p90
Letter Sound Recognition											
W1_LT_SS_H_lsr_60s_B	W1	4451	5.08	9.86	0.0	99.0	0.0	0.0	1.0	6.0	13.0
W2_LT_SS_H_lsr_60s_A	W2	4143	22.68	21.76	0.0	110.0	0.0	4.0	16.0	38.0	54.0
W3_LT_SS_H_lsr_60s_A	W3	3781	39.51	26.37	0.0	110.0	2.0	16.0	41.0	60.0	74.0
W4_LT_SS_H_lsr_60s_21	W4	3300	40.89	19.89	0.0	105.0	14.0	28.0	41.0	55.0	65.0
Word recognition											
W1_LT_SS_H_wre_60s_E	W1	4446	1.91	5.29	0.0	50.0	0.0	0.0	0.0	2.0	5.0
W2_LT_SS_H_wre_60s_B	W2	4143	6.89	9.88	0.0	50.0	0.0	0.0	3.0	9.0	22.0
W3_LT_SS_H_wre_60s_B	W3	3781	19.43	16.96	0.0	50.0	0.0	3.0	17.0	34.0	45.0
W4_LT_SS_H_wre_60s_31	W4	3297	30.33	17.63	0.0	70.0	2.0	16.0	34.0	44.0	51.0
Sentence reading											
W1_LT_SS_H_sre_F	W1	4538	1.22	3.38	0.0	15.0	0.0	0.0	0.0	0.0	3.0
W2_LT_SS_H_sre_D	W2	4143	4.09	4.58	0.0	11.0	0.0	0.0	1.0	9.0	11.0
Sentence Reading comprehension											
W1_LT_SS_H_src_F	W1	4538	0.73	1.24	0.0	3.0	0.0	0.0	0.0	1.0	3.0
Oral Reading Fluency											
W2_LT_SS_H_orf_pm_E	W2	4117	8.08	14.24	0.0	90.0	0.0	0.0	0.0	11.0	30.0
W3_LT_SS_H_orf_D1	W3	3781	25.59	24.64	0.0	66.0	0.0	0.0	23.0	50.0	64.0
W4_LT_SS_H_orf_pm_4_12	W4	3299	48.84	33.34	0.0	206.1	0.0	23.0	50.0	68.0	92.0
W4_LT_SS_H_orf_pm_52	W4	3293	55.97	38.67	0.0	293.3	0.0	25.0	55.0	82.0	107.0
W5_LT_SS_H_orf_pm_1_13	W5	2198	54.64	23.53	0.0	158.7	26.6	37.7	54.5	68.3	86.1
W5_LT_SS_H_orf_pm_2_13	W5	2186	82.62	33.05	0.0	201.6	36.7	62.0	83.3	104.3	123.9
Oral reading Comprehension											
W2_LT_SS_H_orc_DE	W2	4143	0.99	1.66	0.0	6.0	0.0	0.0	0.0	1.0	4.0
W3_LT_SS_H_orc_D2	W3	3781	1.27	1.34	0.0	4.0	0.0	0.0	1.0	2.0	3.0
W4_LT_SS_H_orc_4_2	W4	3246	2.55	1.80	0.0	8.0	0.0	1.0	3.0	4.0	5.0
W5_LT_SS_H_orc_per_1_23	W5	2357	0.29	0.27	0.0	1.0	0.0	0.0	0.3	0.4	0.7
W5_LT_SS_H_orc_per_2_23	W5	2356	0.41	0.30	0.0	1.0	0.0	0.1	0.4	0.6	0.9

Table 7 (b): Summary Statistics of Learner Outcome Variables for Home Language

Variable	Wave	N	Mean	SD	Min	Max	p10	p25	p50	p75	p90
Written Comprehension											
W4_LT_SS_H_wco_8	W4	3321	1.57	1.54	0.0	7.0	0.0	0.0	1.0	2.5	4.0
W5_LT_SS_H_wco_4	W5	2260	5.44	3.06	0.0	11.0	1.0	3.0	6.0	8.0	9.5
Writing											
W2_LT_SS_H_wri_F	W2	4143	5.88	3.40	0.0	12.0	1.0	4.0	6.0	8.0	11.0
W3_LT_SS_H_wri_F1_F3	W3	3781	5.97	2.36	0.0	9.0	3.0	4.0	6.0	8.0	9.0
Vocabulary											
W1_LT_SS_H_voc_A	W1	4538	8.58	1.40	0.0	10.0	7.0	8.0	9.0	10.0	10.0
W5_LT_SS_H_voc_5	W5	2231	6.36	3.18	0.0	10.0	1.0	4.0	7.0	9.0	10.0
Phonemic Awareness											
W1_LT_SS_H_paw_D_comp	W1	4538	0.81	1.32	0.0	4.0	0.0	0.0	0.0	1.0	3.0
W2_LT_SS_H_pat_G	W2	4143	0.70	1.18	0.0	4.0	0.0	0.0	0.0	1.0	3.0
W3_LT_SS_H_pat_E	W3	3781	1.82	1.07	0.0	3.0	0.0	1.0	2.0	3.0	3.0
Rapid Object Naming											
W4_LT_SS_H_ron_pm_1_12	W4	3300	58.52	16.68	0.0	162.9	40.0	48.0	56.0	68.0	80.0
Rapid Letter Naming											
W4_LT_SS_H_rln_15s_1_22	W4	3299	81.23	32.87	0.0	205.7	40.0	60.0	80.0	104.0	124.0
Working Memory											
W1_LT_SS_H_wme_wor_C1	W1	4538	2.45	1.31	0.0	5.0	0.0	2.0	2.0	3.0	4.0
W1_LT_SS_H_wme_num_C2	W1	4538	2.54	1.34	0.0	5.0	1.0	2.0	2.0	3.0	5.0
Non-word decoding											
W2_LT_SS_H_nwd_pm_C	W2	4138	4.51	8.23	0.0	71.4	0.0	0.0	0.0	6.0	18.0
W3_LT_SS_H_nwd_C	W3	3781	14.39	13.80	0.0	50.0	0.0	0.0	13.0	26.0	34.0

Table 8: Summary Statistics of Learner Outcome Variables for English as a First Additional Language & Mathematics

Variable	Wave	N	Mean	SD	Min	Max	p10	p25	p50	p75	p90
Vocabulary											
W5_LT_SS_E_voc_7	W5	2178	8.31	9.24	0.0	50.0	0.0	1.0	5.0	12.5	22.5
Word recognition											
W3_LT_SS_E_wre_G1	W3	3781	1.52	1.50	0.0	4.0	0.0	0.0	1.0	3.0	4.0
W3_LT_SS_E_wre_G	W3	3781	3.14	3.16	0.0	8.0	0.0	0.0	2.0	6.0	8.0
W4_LT_SS_E_wre_pm_62	W4	3296	29.83	24.03	0.0	170.0	0.0	7.0	27.0	48.0	63.0
Sentence reading											
W3_LT_SS_E_sre_G2	W3	3781	1.62	1.79	0.0	4.0	0.0	0.0	0.0	4.0	4.0
Oral Reading Fluency											
W4_LT_SS_E_orf_pm_7_12	W4	3293	38.84	36.41	0.0	294.3	0.0	2.2	30.0	65.0	91.0
W5_LT_SS_E_orf_pm_3_13	W5	2105	87.18	38.51	0.0	185.8	30.5	62.0	88.5	115.2	135.5
Oral Reading Comprehension											
W4_LT_SS_E_orc_7_2	W4	2968	1.47	1.89	0.0	8.0	0.0	0.0	1.0	2.0	5.0
W5_LT_SS_E_orc_per_3_23	W5	2353	0.51	0.32	0.0	1.0	0.0	0.3	0.5	0.8	0.9
Written Comprehension											
W4_LT_SS_E_wco_9	W4	3321	1.06	1.29	0.0	6.0	0.0	0.0	1.0	2.0	3.1
W5_LT_SS_E_wco_6	W5	2210	5.15	3.47	0.0	13.0	1.0	2.0	5.0	8.0	10.0
Mathematics											
W3_LT_SS_M_mat_F4	W3	3781	0.60	0.62	0.0	2.0	0.0	0.0	1.0	1.0	1.0
W4_LT_SS_M_mat_10	W4	3321	2.36	1.78	0.0	9.0	0.0	1.0	2.0	4.0	5.0