Impact Evaluation of a Grade R Literacy Programme: Using a Difference-in-Difference method.

Janeli Kotzé[1]

Both government departments and donor funders are pursuing more effective methods of redistributing resources in a way that will show a commensurate change in the outcomes that they are targeting. In this pursuit, the use of impact evaluations has also become a more popular means of determining whether the implementation of a targeted programme is indeed causing the desired change in the outcomes which the programme is directed at. In 2016, a Grade R literacy programme aimed at improving the early literacy skills of learners before they enter Grade 1 was implemented in one of the South African provinces. The impact of the programme is estimated by combining a difference-in-differences model design with school- and teacher-fixed effects, using data on learner assessments, classroom observations and teacher and school information. Since the programme was implemented across all school in the province in 2016, Grade R learners in the 2015 cohort were tested as the control group learners. Schools and teachers were then followed and the 2016 cohort of Grade R learners were tested in the same schools and the classes of the same teachers. Using teacher fixed effects, the evaluation found learning gains of 0.36 standard deviations in the treatment group over a six month period. On the overall learner assessment the difference-in-differences model did not find any significant differences, however, when differentiated by specific sub-tasks, it is clear that there were significant differences in the writing and reading sub-tasks.

Keyword: Impact Evaluation, Difference-in-Differences, Education, Grade R

JEL Classification: C21, I21, I28

[1] Dr. Janeli Kotzé was a researcher at ReSEP, at the University of Stellenbosch while conducting this research.

1. Introduction

In the pursuit of improving the quality of education in developing countries, both national and international researchers are shifting their focus to finding policies and programmes that will affect change in learner performance. In 2016, a Grade R literacy programme aimed at improving the early literacy skills of learners before they enter Grade 1 was implemented in one of the South African provinces. The impact of the programme is estimated by combining a difference-in-differences model design with school- and teacher-fixed effects, using data on learner assessments, classroom observations and teacher and school information. Since the programme was implemented across all school in the province in 2016, Grade R learners in the 2015 cohort were tested as the control group learners. Schools and teachers were then followed and the 2016 cohort of Grade R learners were tested in the same schools and the classes of the same teachers. Using teacher fixed effects, the evaluation found learning gains of 0.36 standard deviations in the treatment group over a six month period. On the overall learner assessment the difference-in-differences model did not find any significant differences, however, when differentiated by specific sub-tasks, it is clear that there were significant differences in the writing and reading sub-tasks.

2. Background

The Grade R literacy programme that that forms the focus of this evaluation centers on the provision of participative training and materials (story books, visual aids, activity guidelines, tips and ideas, resources, sequence pictures) to Grade R teachers to support them in the development of relevant skills and knowledge that enables them to effectively teach early language and literacy to Grade R learners. The programme is implemented through cascade training from the implementing agency to Lead Teachers and, in turn, from Lead Teachers to ordinary Grade R teachers over a period of 18 months. The approach focuses on a wider methodological framework for teaching language and literacy in Grade R, but it does also include training on the use of specific materials through monthly "cluster" meetings. The 18-month training process began in with Foundation Phase and Early Childhood Development (ECD) Advisors being trained in the methodological approach, followed by these Advisors in turn training 200 lead teachers (teachers selected for the project and/or volunteers who will provide cascade training to other teachers in 2016) in the methodological framework. At the ECD and Foundation Phase advisor training, the program was more theory-focused whereas at the lead teacher training, the use of various resources and materials was highlighted. Lead teachers received follow-on support during monthly cluster meetings after the initial training.

In 2016 the lead teachers trained all Grade R teachers in public schools across the province in which the literacy programme was implemented. The theory of change of the literacy programme is such that teachers receive the treatment, but that the effects of this should be seen on the language and literacy abilities of their learners, in Grade R and potentially into Grade 1. Ultimately, the project aims to contribute to supporting the Department of Basic Education in its efforts to improve Grade R quality and to raise early grade literacy scores.

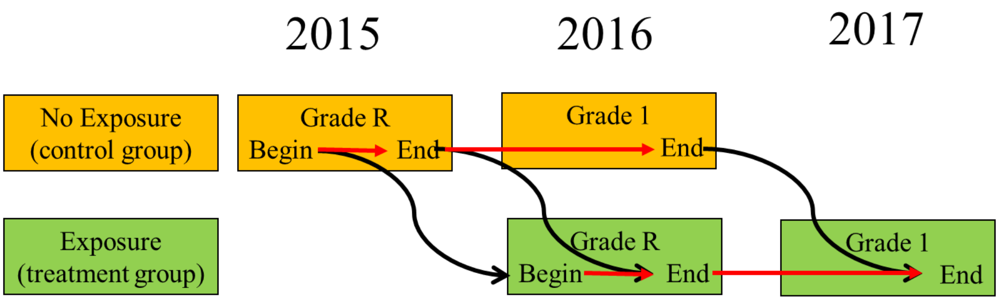
3. Method and Data

3.1. Evaluation Design

Given the universal roll-out of the literacy programme to all teachers across one province at the same time, it was not possible to obtain a credible counterfactual group of learners during the same year. To circumvent this problem, the Control Group data was collected the year before the implementation of the literacy programme, whereas the Treatment Group data was collected during the year of literacy programme implementation.

This impact evaluation is therefore designed using a difference-in-difference design. That is, the learning gains for the 2015 Grade R cohort (the Control Group) will be compared with the learning gains for the 2016 Grade R cohort (the Treatment Group) in the same schools. The approach will also track the learners from each of the Grade R cohorts longitudinally into Grade 1, comparing the performance of these cohorts over time. By making use of the same schools from which we have obtained the control and treatment cohorts, school fixed effects can be accounted for and will therefore reduce any bias in the results that may originate from unobservable variables that remain constant over the period of the evaluation.

The evaluation design is illustrated in the Figure, below.



** "Begin" is start of 3rd term. "End" is end of 4th term.

- Black arrows show comparisons between groups.
- Red arrows show comparison within groups

Figure 1. Evaluation Design

The baseline data collection could only start in August/September 2015, which only allowed the evaluation of approximately three months of the Control Group. For the counterfactual to be credible, the Treatment Group could also only be tested for the same time period. This evaluation therefore effectively evaluates the effect of the literacy programme's implementation over almost three months.

3.2. Estimation Method

Difference-in-Difference Estimate:

Due to the set-up of our experiment, we will observe both the control and the treatment groups for two time periods, in the first time period (the baseline tests), neither group is exposed to the treatment. In the second period, the treated get exposed to treatment, while the control does not. To get the Diff-in-Diff estimate, we assume that both groups will show the same level of learning gains, had the intervention not been implemented. We calculate the average gain among the control group is subtracted from the average gain among the treated. This estimate will allow us to get rid of two types of biases: the part that could be due to permanent differences between the two groups and the part that could be due to a time trend. The estimate will be run using our baseline scores for the first time period and then compare that to the mid-year and end-year scores for the learners in Grade R in 2015, and then their scores for Grade 1 in 2016. This will not only also give us an indication of the immediate effects of the intervention on language and literacy, but also the longer run effects.

Step 1: Get the gains for the treated

$$\Delta Y_{1it} = Y_{1it} - Y_{1it-1}$$

$$= (u_1 + a_i + \lambda_t + \varepsilon_{it}) - (u_0 + a_i + \varepsilon_{it-1})$$

$$= u_1 - u_0 + \lambda_t + \varepsilon_{it} + \varepsilon_{it-1}$$

Step 2: Get the gains for the control

$$\Delta Y_{0it} = Y_{0it} - Y_{0it-1}$$

$$= (u_0 + a_i + \lambda_t + \varepsilon_{it}) - (u_0 + a_i + \varepsilon_{it-1})$$

$$= \lambda_t + \varepsilon_{it} + \varepsilon_{it-1}$$

Step 3: Subtract the latter from the prior

$$\Delta Y_{1it} - \Delta Y_{0it} = (u_1 - u_0 + \lambda_t + \varepsilon_{it} + \varepsilon_{it-1}) - (\lambda_t + \varepsilon_{it} + \varepsilon_{it-1})$$

$$= u_1 - u_0$$

3.3. Sample

Given the evaluation design, the control and treatment schools are the same schools. Fifty schools were selected in 2015 using a stratified random sampling approach. School quintile was chosen as the stratification variable to ensure that the sample would be able to provide information regarding the success of the intervention across all school functionalities. Stratification by school Language of Learning and Teaching (LOLT) would have resulted in over-sampling of highly functional schools, as English schools are hugely overrepresented in quintiles 4 and 5. Ten schools per quintile were sampled and in each school 9 learners were randomly selected, which resulted in 450 learners tested in total.

During the August 2016 Treatment Group baseline assessment, one school was found to have discontinued their Grade R classes. This bought the total number of schools down to 49. In response, and given the previous year's high attrition rate, the evaluation team increased the number of learners assessed per school from 9 to 10. This raised the Treatment Group sample from N = 450 to N = 490.

As mostly the same schools and teachers were observed in 2016 as in 2015, the school characteristics are also very similar. Grade R teachers in more than half (57%) of schools in 2016 only use Afrikaans in the classroom, compared with 46% of the schools in 2015. As in 2015, Afrikaans is used more often in quintile 1 and 2 schools than in schools in other quintiles. Similar to 2015, in 14% of schools (quintiles 3, 4 and 5), teachers only use English and 18% only use Xhosa in their classrooms. In 2016, the number of schools in which teachers use a combination of English and Afrikaans is not as evenly spread across quintiles as in 2015, with only quintile 4 and 5 schools using an English/Afrikaans combination. In 2016, no schools used a combination of English, Afrikaans and isiXhosa compared to 2015 where one quintile 3 teacher used all three languages in the classroom. Overall, the patterns of language use in the classrooms are comparable from 2015 to 2016.

Table 1: Grade 1 Home Language and LOLT

		LoLT		
		Afrikaans	English	Xhosa
Language spoken at home	Afrikaans	**95.9%**	4.1%	0.0%
	English	5.3%	**94.7%**	0.0%
	Xhosa	11.1%	19.8%	**69.1%**
	Sotho	100.0%	0.0%	0.0%

Table 1: Grade 1 Home Language and LOLT shows the learner Home Language versus the LOLT in Grade 1. Overall, the majority of learners are still being taught in the same language as that used at home (89% overall). This does represent a small decrease from the Grade R control and treatment groups, in which 94% and 95% were taught in their home languages respectively.

As in 2015, most teachers in 2016 (80%) remain qualified with a post matric diploma or certificate, across all quintiles[2]. The same number of teachers in 2016 (8%) hold a degree or higher, mainly in quintile 5 schools. One teacher in a quintile 3 school and one teacher in a quintile 1 school reported not having a matric certificate, compared with only one quintile 3 teacher in 2015. Four teachers (less than 1%) reported only having a matric certificate, compared with three in 2015. The level of teacher education remains comparable from 2015 to 2016. As in 2015, all teachers, except one, reported reading for enjoyment at home. Teachers reported reading for enjoyment for 4.7 hours on average per week, slightly more than the average number of hours reported in 2015.

3.4. Attrition

During baseline data collection, test score data was collected from 450 Control Group learners in 50 schools, and 490 Treatment Group learners in the same 50 schools. The assessments were administered on the same learners over time. The baseline and midline data collection points were not very far apart and it was therefore highly likely that the learners would have remained in the same schools and classes. Given this, very low attrition was present in the midline assessment. During the midline assessment of the Control Group there was a chicken pox outbreak in one of the districts, which led to a slightly higher attrition rate among the Control

[2] Note that only the teacher's highest level of education was captured.

Group learners. However, the attrition rate among the Treatment and Control Group in the midline assessment is negligibly small.

Table 2 below demonstrates the rate of sample attrition from baseline to midline in both the Control and Treatment Groups.

Table 2: Attrition Rate

		Control	**Treatment**
Baseline	Total Sample	450	490
Midline	Total Sample	434	481
Endline	Total Sample	405	NA
	Attrition Midline	16 (4%)	9 (2%)

Given that sampling strategy of sampling both the Control learners and the Treatment learners from the same schools, the characteristics of the Treatment Group learner sample are, as expected, very similar to the Control Group sample, as are the school characteristics (such as languages used in the classroom, teacher level of education, teacher years of experience teaching Grade R, etc.)

4. Results

4.1. Test Reliability

The Learner Assessment instrument is the primary tool designed to measure learner performance in this evaluation. At the time of the evaluation design, there were no standardised, widely used, and validated in-country Grade R language and literacy assessments available. Language and literacy, and test development expert consultants were hired to develop the tool in line with the Curriculum Assessment Policy Statement (CAPS). The tool is therefore criterion-referenced as the intervention is carried out within the CAPS framework and is 100% CAPS aligned. As the evaluation covers performance from Grade R into Grade 1, items from both curricula are included in the instrument. Some Grade 2 items are also included to measure "extension" performance.

The Learner Assessment tool measures all four CAPS Home Language skills: listening and speaking (12 items, 6 each for Listening and Speaking in a Conversation and 6 for Listening and Speaking based on a Story), phonics (12 items, split into Phonological Awareness and Phonics Letters), Reading (and viewing - 6 items) and Writing (- 6 items). Items were constructed so that answers could be anchored to as many of the following categories as possible, using a 6-point scale: Not done (0), Partial Grade R level (1), End of Grade R level (2), Partial Grade 1 level (3), End of Grade 1 level (4), Grade 2 level (5).

The overall reliability coefficient (Cronbach's Alpha) for all of the items is 0.80. Cronbach's alpha is high enough to indicate good reliability, but not so high that it indicates that the test items are redundant. This provides a good indication of the internal consistency of the Learner Assessment instrument, signifying that the tool is consistently measuring the underlying construct that it intends to measure - language and reading proficiency.

Table 3 shows the correlations between the baseline sub-test scores and the midline sub-test scores for each of the groups. Overall the correlations between the final scores are relatively high at 0.66 for the Control Group and 0.69 for the Treatment Group. This is reassuring as this increases the precision in the regressions. Among the sub-tests, the correlations are also relatively strong with the correlations among the Phonics sub-tests being the highest.

Table 3: Correlation between Baseline and Midline Scores

	Control	**Treatment**
Final Score	0.66	0.69
Writing	0.53	0.45
Listening and Speaking: Conversation	0.46	0.39
Listening and Speaking: Story	0.39	0.43
Reading	0.36	0.38
Phonics: Audio	0.59	0.56
Phonics: Letters	0.61	0.60
Cronbach's Alpha for the Midline:	0.80	

Notes: Correlation between the baseline and midline percentage scores for each of the sub-tests.

4.2. Descriptive Statistics

When interpreting the results below it is necessary to keep in mind the design of the evaluation. Due to unfortunate delays at the start of the project, the evaluation team was only able to test the control group over a short period of time at the end of 2015. To ensure the comparability of learning gains between the treatment and control groups, the delays has meant that the treatment group could also only be tested over a similar time period at the end of 2016. Given this, the treatment group has already been exposed to 6 months of the intervention when the baseline assessments were administered. It is therefore to be expected that the baseline scores of the treatment and control groups will be different, a difference that will subsequently influence the midline scores as well.

The midline learner assessment instrument was exactly similar to the baseline learner assessment, since the original design included items of Grade 1 and Grade 2 difficulty levels. This design allows a direct comparison between the baseline and midline percentage scores to assess learning gains made over the three-month period. The summary statistics of the score distributions for each sub-test are presented in table 1 for each of the treatment and control groups. Overall it is reassuring that no severe floor or ceiling effects are observed in either the treatment or control group. Among some of the sub-tests (the two Listening and Speaking sub-tests) learner performance was quite high, but the normal distribution of test scores in the other sub-tests ensures an overall balance. The higher scores for these two sub-tests has lead the evaluation team to include some additional EGRA items in the endline assessment to ensure that potential ceiling effects do not inhibit the ability to show learning gains during Grade 1.

Table 4: Midline Learner Score Distribution

		p10	**p25**	**p50**	**p75**	**p90**	**min**	**max**
Total Score	*Control*	0.300	0.379	0.443	0.536	0.614	0.043	0.921
	Treatment	0.357	0.414	0.479	0.557	0.636	0.171	0.857
Grade R	*Control*	0.310	0.379	0.448	0.534	0.621	0.069	0.948
	Treatment	0.362	0.414	0.483	0.569	0.655	0.224	0.845
Grade 1	*Control*	0.298	0.386	0.456	0.579	0.667	0.035	0.895
	Treatment	0.317	0.383	0.467	0.550	0.633	0.183	0.817
Grade 2	*Control*	0.200	0.320	0.400	0.520	0.600	0.000	0.960
	Treatment	0.320	0.360	0.400	0.560	0.680	0.000	0.920
Writing	*Control*	0.154	0.231	0.308	0.385	0.462	0.000	0.808
	Treatment	0.231	0.308	0.385	0.462	0.538	0.038	0.808
Listening: Conversation	*Control*	0.625	0.750	0.792	0.875	0.917	0.042	1.000
	Treatment	0.667	0.708	0.792	0.833	0.875	0.458	0.958
Listening: Story	*Control*	0.421	0.579	0.737	0.842	0.947	0.000	1.000
	Treatment	0.526	0.684	0.789	0.842	0.895	0.158	1.000
Reading	*Control*	0.083	0.208	0.250	0.292	0.417	0.000	0.958
	Treatment	0.167	0.208	0.250	0.375	0.500	0.000	0.917
Phonics: Audio	*Control*	0.138	0.207	0.345	0.483	0.621	0.000	0.966
	Treatment	0.138	0.207	0.345	0.517	0.621	0.000	0.966
Phonics: Letters	*Control*	0.036	0.107	0.179	0.357	0.429	0.000	0.786
	Treatment	0.107	0.143	0.250	0.393	0.500	0.000	0.714

Notes: Scores in this table are converted to percentage scores to provide a more intuitive sense of the score distributions for each of the sub-tests. The test scores are those collected during the midline data collection exercise at the end of Grade R.

Table 4 shows both the mean baseline and midline scores per sub-test for the treatment and control group. As expected, it is evident that the baseline scores in each of the sub-tests are higher in the treatment group than the control group. The control group showed significant learning gains in all of the sub-tests, except Phonics: Letters, over the period tested. The treatment group, however, did not show any learning gains over this period in the sub-tests Listening and Speaking Based on a Story, and Phonics: Audio. The lack of learning gains in the sub-test Listening and Speaking Based on a Story is likely due to the high baseline score for the treatment group. Given the design of the literacy programme and the focus on stories throughout the year, it is to be expected that the largest learning gains of this skill would have taken place during the first six months of the year and therefore would not be significant in the time period which is being evaluated. The low scores in the Phonics: Audio and Phonics: Letters subtests among both the Treatment and Control groups are indicative of either learners not having understood the task sufficiently, or a weak grasp of this skill among Grade R learners in general (despite the implementation of the literacy programme).

Table 5: Mean percentage scores per sub-test

		Control				Treatment			
		Mean	**Std. Err.**	**Learning Gains**		**Mean**	**Std. Err.**	**Learning Gains**	
Total Score	*Wave 1*	0.401	0.006	0.055	***	0.445	0.005	0.044	***
	Wave 2	0.459	0.006			0.489	0.005		
Grade R	*Wave 1*	0.415	0.005	0.045	***	0.452	0.005	0.042	***
	Wave 2	0.461	0.006			0.494	0.005		
Grade 1	*Wave 1*	0.427	0.006	0.050	***	0.451	0.005	0.022	***
	Wave 2	0.477	0.007			0.473	0.005		
Grade 2	*Wave 1*	0.321	0.011	0.094	***	0.414	0.009	0.043	***
	Wave 2	0.414	0.008			0.458	0.007		
Writing	*Wave 1*	0.289	0.006	0.031	***	0.312	0.006	0.073	***
	Wave 2	0.320	0.006			0.385	0.006		
Listening: Conversation	*Wave 1*	0.693	0.008	0.094	***	0.741	0.007	0.045	***
	Wave 2	0.787	0.005			0.786	0.004		
Listening: Story	*Wave 1*	0.608	0.011	0.101	***	0.743	0.008	0.003	.
	Wave 2	0.708	0.009			0.740	0.007		
Reading	*Wave 1*	0.236	0.006	0.042	***	0.245	0.005	0.060	***
	Wave 2	0.277	0.008			0.304	0.007		
Phonics: Audio	*Wave 1*	0.343	0.008	0.018	*	0.364	0.008	0.008	.
	Wave 2	0.361	0.010			0.372	0.009		
Phonics: Letters	*Wave 1*	0.228	0.008	0.002	.	0.256	0.008	0.011	.
	Wave 2	0.231	0.008			0.267	0.007		

Notes: Scores in this table are converted to percentage scores to provide a more intuitive sense of the mean scores. The learning gains in this table are also merely the raw difference between the Baseline and Midline scores for each of the treatment groups.

Table 5 explores the relationship between each of the sub-tests in the baseline assessment with the final score in the midline assessment. Overall each of the sub-tests are strong predictors of the final score in the midline assessment, with Listening and Speaking Based on a story having the lowest correlation. This is once again a result of the high baseline scores and the relatively lower midline scores. Interestingly, reading is not a strong predictor of the final midline score among the control group, but in the treatment group there is a strong positive relationship between the reading sub-test and the final midline score.

Table 6: Baseline sub-tests predicting midline total score

	Control		**Treatment**	
	β	*s.e.*	*β*	*s.e.*
Writing	0.172***	(0.040)	0.186***	(0.036)
Listening and Speaking in a Conversation	0.096***	(0.034)	0.088***	(0.028)
Listening and Speaking based on a story	0.033	(0.025)	0.052*	(0.027)
Reading	0.031	(0.049)	0.101***	(0.037)
Phonics Audio	0.170***	(0.038)	0.194***	(0.036)
Phonics Letters	0.240***	(0.035)	0.138***	(0.036)
Constant	0.202***	(0.024)	0.196***	(0.020)
Observations	434		481	
R-squared	0.47		0.485	

*Notes: The final percentage score of the midline assessment is regressed on the baseline scores for each of the sub-tests to show the relationship and predictability of the sub-tests in the baseline on the midline results. Significance levels: * 0.1; ***

Table 7 shows the correlations between the baseline sub-test scores and the midline sub-test scores for each of the treatment groups. Overall the correlation between the final scores are relatively high at 0.66 for the control group and 0.69 for the treatment group. This is reassuring since this will increase the precision in the regressions. Among the sub-tests the correlations are also relatively strong with the correlations among the Phonics sub-tests being the highest. Cronbach's alpha for the overall midline assessment is high at 0.80. Given the difficulties in testing learners of this age group, the strength of the correlations is reassuring in that the results do seem credible.

Table 7: Correlation between Baseline and Midline Scores

	Control	**Treatment**
Final Score	0.66	0.69
Writing	0.53	0.45
Listening and Speaking: Conversation	0.46	0.39
Listening and Speaking: Story	0.39	0.43
Reading	0.36	0.38
Phonics: Audio	0.59	0.56
Phonics: Letters	0.61	0.60
Cronbach's Alpha for the Midline:	0.80	

Notes: Correlation between the baseline and midline percentage scores for each of the sub-tests.

4.3. Main Midline Results

The Model:

This impact evaluation is designed using a difference-in-difference approach, which compares the learning that takes place over a period of time between two Grade R cohorts, in the same schools. That is, the learning gains for the 2015 Grade R cohort (the control group) are being compared with the learning gains for the 2016 Grade R cohort (the treatment group) in the same schools. This approach was chosen since the intervention was rolled-out to the entire province, thereby rendering a credible counterfactual in the same time period impossible.

This approach is therefore built on the relatively strong identifying assumption that the treatment group would have similar learning trends to the control group in the absence of treatment. This assumption will hold under two conditions: (1) if were no systematic difference across the province in ability between one cohort of grade R's and the next; (2) that apart from the literacy programme, no other interventions were rolled out in the province which would influence the Grade R teachers' ability to teach literacy. Testing the credibility of the first assumption is difficult in this case, since there is no clean baseline data available for the treatment group.[3] However, based on the balance in learner characteristics between the treatment and control group, there is no reason to believe that there will be any systematic differences in the ability of the two groups. To ensure that the second assumption will hold, the conditions was made clear to the implementing province and the donor organisation upfront and the province was strongly advised not to implement any interventions that would influence the Grade R classes during this time period. In interviews with teachers and the provincial officials it appears as if the second assumption holds.

As mentioned previously, the baseline assessment for the treatment group was administered after the programme has been implemented for 6 months. It is therefore likely that much of the learning gains would have taken place during that time period. To take this into consideration two different models were run to evaluate the impact of the programme in the short run.

Firstly, a school fixed effects model was run on the baseline scores. Through having tested the treatment and control cohorts in the same schools, it is possible to control for any unobservable school characteristics which might be correlated with the treatment group. That is, any school level changes which might influence learner literacy ability in Grade R. Secondly, a difference-in-difference model was run to evaluate the difference in learning gains between the treatment and control groups over the time period observed.

The school fixed effects model is as follows:

[3] Since the baseline data could only be collected in August, the treatment group learners have been exposed to the intervention by the time of baseline data collection.

$$Baseline\ Scores_i = \alpha_i + \delta_1 T + \beta_{1,i}X_{1,i} + \cdots + \beta_{k,i}X_{k,i} + \gamma_{1,i}S_{1,i} + \cdots + \gamma_{n,i}S_{n,i} + \varepsilon_i$$

Where α_i is the constant intercept, δ_1 is the coefficient on the treatment variable, β_k are the coefficients on the learner characteristic variables (X_k), $\gamma_n$are the coefficients for the school fixed effects (S_n) and ε_i is the error term. The school fixed effects are essentially a binary regressor included for each school participating in the study.

The difference-in-difference models can be represented as follows:

$$Index\ Score_{iTw} = \beta_0 + \beta_1\gamma_{T=1} + \beta_2\delta_{w=midline} + \beta_3(\gamma_{T=1})(\delta_{w=midline}) + \varepsilon_{iTw}$$

Where the final score ($Index\ Score_{iTw}$) for learner *i,* in treatment group *T* in wave *w,* is regressed on a treatment dummy ($\gamma_{T=1}$), a dummy indicating the data collection wave ($\delta_{w=midline}$) and an interaction between the treatment dummy and the data collection wave dummy. The coefficient of interest in this case is β_3.

Table 8 shows the results of the school fixed effects model which was run on the baseline scores. The first model (A) is the raw model which only includes the treatment variable, the second model (B) includes the learner level characteristics, and the third model (C) includes the school fixed effect. Since learners in the same classrooms were tested in both years, school fixed effects will control for teacher fixed effects if the same teacher remained in the classroom in 2016. This is the case in 41 of the schools, so the fourth model (D) restricts the sample to only include these schools. The coefficient on the treatment variable in all four models is relatively constant and ranges between 0.39 and 0.36 of a standard deviation. Under the assumption that, after controlling for learner level characteristics, school fixed effects and teacher fixed effects, the only difference between the treatment and control groups is the implementation of the literacy intervention, these results suggest that the treatment group gained around 0.36 standard deviations of learning over the first 6 months of the intervention.

Table 8: Reduced form school fixed effects model on baseline scores

	A		**B**		**C**		**D**	
	β	*s.e.*	*β*	*s.e.*	*β*	*s.e.*	*β*	*s.e.*
Treatment	0.38***	0.06	0.39***	0.06	0.38***	0.05	0.36***	0.06
Female			0.23***	0.06	0.23***	0.06	0.21***	0.06
Learner age			0.22**	0.08	0.22**	0.08	0.22*	0.09
English			0.19*	0.08	-0.06	0.15	-0.07	0.19
isiXhosa			0.12	0.08	-0.42	0.23	-0.45	0.25
Other			0.46	0.36	-0.09	0.43	-0.10	0.43
Constant	0.80***	0.05	-0.63	0.47	-0.35	0.52	-0.31	0.56

School FE	No	No	Yes	.
Teacher FE	No	No	No	Yes
N	940	937	937	795
R-squared	0.036	0.07	0.353	0.316

Table 9 shows the results for the main difference in difference models. Similar to the fixed effects models, four different models were run: (A) a basic model; (B) a model including learner characteristics; (C) a model including school fixed effects; (D) a model restricted so that the school fixed effects essentially serve as teacher fixed effects. The treatment variable in this model controls for the differences between the treatment and control group, whereas the Wave variable controls for the difference between the midline and baseline scores. The coefficients on the Treatment and Wave variables are both positive and quite large. This is to be expected given the large differences between the treatment and control groups at baseline, as well as the natural learning gains that learners would obtain in the time period between the baseline and midline assessments.

The coefficient of interest in this model is the coefficient for the interaction variable (Treatment * Wave), which essentially reports the learning gains for the treatment group. Although the coefficient is positive and indicate learning gains of around 0.23 standard deviations for the full sample over the 3-month period, it is not possible to say with confidence that the learning gains of the treatment groups is statistically significantly different from zero. For the teacher fixed effects model the learning gains are around 0.41 of a standard deviation over the 3-month period, but this is only statistically significantly different at a 78% confidence level.

Table 9: Main results of the Difference in Difference Model

	A		**B**		**C**		**D**	
	B	*s.e.*	*β*	*s.e.*	*B*	*s.e.*	*β*	*s.e.*
Treatment	0.830***	(0.161)	0.855***	(0.159)	0.833***	(0.137)	0.790***	(0.147)
Wave	1.029***	(0.166)	1.046***	(0.164)	1.035***	(0.141)	0.890***	(0.153)
*Treatment * Wave*	-0.26	(0.229)	0.133	(0.323)	0.234	(0.294)	0.407	(0.325)
HL is same as Lolt			0.440*	(0.254)	0.519**	(0.244)	0.500*	(0.271)
Learner Age			0.513***	(0.113)	0.471***	(0.101)	0.468***	(0.108)
Learner Female			0.592***	(0.113)	0.585***	(0.099)	0.539***	(0.107)
	-	(0.116)	-	(0.704)	-	(0.722)	-	(0.767)
Constant	0.877***		4.531***		3.376***		3.270***	
School FE	No		No		Yes		.	
Teacher FE	No		No		No		Yes	
N	1843		1837		1837		1557	
F-stat	33.491		25.264		16.457		13.336	
R-squared	0.052		0.076		0.337		0.293	

The Difference-in-Difference Model that controls for school fixed effects were run for each of the sub-scores separately, to determine whether there were significant differences within the different skills that were assessed. As in table 7, the main variable of interest is the interaction between the treatment dummy and the wave dummy (Treatment * Wave). Each of the scores for the sub-task were standardised to have a mean of zero and a standard deviation of one, which means that the coefficients can be interpreted as standardised scores. From Table 8 it is therefore evident that the programme has led to significant increases in the learners' writing ability, in their ability to listen and speak in a conversation and in their reading ability. A significant decrease is evident in the sub-task of listening based on a story, but as discussed previously, this result is most likely driven by the very high baseline scores that both the control and treatment group achieved. No significant differences were observed in the sub-tasks testing Phonological Awareness and Phonics: Letters.

	Writing		**Conversation**		**Listening to a Story**		**Reading**		**Phonological Awareness**		**Phonics: Letters**	
	B	*s.e.*	*β*	*s.e.*	*β*	*s.e.*	*β*	*s.e.*	*β*	*s.e.*	*β*	*s.e.*
Treatment	0.178***	(0.059)	0.358***	(0.058)	0.697***	(0.059)	0.069	(0.060)	0.133**	(0.056)	0.185***	(0.057)
Wave	0.231***	(0.060)	0.705***	(0.060)	0.513***	(0.061)	0.300***	(0.062)	0.116**	(0.057)	0.024	(0.059)
Treatment * Wave	0.251**	(0.126)	0.239*	(0.125)	-0.347***	(0.127)	0.283**	(0.129)	-0.086	(0.120)	0.144	(0.123)
HL is same as Lolt	-0.05	(0.105)	0.657***	(0.104)	0.198*	(0.105)	0.186*	(0.107)	-0.009	(0.099)	0.107	(0.102)
Learner Age	0.160***	(0.043)	0.072*	(0.043)	0.082*	(0.044)	0.125***	(0.044)	0.132***	(0.041)	0.137***	(0.042)
Learner Female	0.231***	(0.043)	0.131***	(0.042)	0.033	(0.043)	0.131***	(0.044)	0.201***	(0.040)	0.195***	(0.041)
Constant	-1.071***	(0.309)	-1.674***	(0.307)	-1.065***	(0.311)	-1.111***	(0.317)	-0.032	(0.294)	-0.720**	(0.302)
School FE	Yes		Yes		Yes		Yes		Yes		Yes	
Teacher FE	No		No		No		No		No		No	
N	1837		1837		1837		1837		1837		1837	
F-stat	9.85		10.416		9.223		7.624		14.014		11.945	
R-squared	0.233		0.243		0.222		0.191		0.302		0.269	

	Bottom Tercile		Mid Tercile		Top Tercile	
	β	*s.e.*	*β*	*s.e.*	*β*	*s.e.*
Treatment	0.304*	(0.181)	0.093	(0.146)	0.213	(0.231)
Wave	1.651***	(0.152)	0.456***	(0.146)	0.672***	(0.246)
Treatment * Wave	0.806**	(0.357)	0.256	(0.333)	0.239	(0.464)
HL is same as Lolt	1.257***	(0.295)	-0.118	(0.289)	0.544	(0.392)
Learner Age	0.327**	(0.127)	-0.034	(0.125)	0.252	(0.156)
Learner Female	0.044	(0.128)	0.126	(0.106)	0.433***	(0.163)
Constant	-5.343***	(0.898)	0.613	(0.851)	0.219	(1.124)
School FE	Yes		Yes		Yes	
N	608		614		615	
F-stat	5.615		3.792		4.562	
R-squared	0.354		0.268		0.283	

In the table above thee different models were run to determine which learners benefited the most from the intervention. Learners were divided into three groups, based on their performance at the baseline. From the table above it is evident that the weakest performing learners benefited the most from the literacy programme.

Conclusion

The results shown above indicate that the literacy programme might have had a positive influence on learner reading ability over the implementation period. Although it is necessary to keep in mind the assumptions made in the school fixed effects model, the results indicate that the 2016 cohort of Grade R learners who were exposed to the literacy programme might have gained up to three-quarters of a year worth of learning additional to their counterparts the previous year. The short time period over which the difference-in-difference data collection took place, it is not possible to say with confidence the learning gains obtained over this time period, but the results corroborate the results of the fixed-effects model.

Appendix:

			Control						Treatment					
			Baseline			Midline			Baseline			Midline		
			Obs	*Mean*	*Std. Dev.*	*Obs*	*Mean*	*Std. Dev.*	*Obs*	*Mean*	*Std. Dev.*	*Obs*	*Mean*	*Std. Dev.*
School Characteristics:	**Urban:**		450	0.70	0.46	434	0.70	0.46	490	0.69	0.46	481	0.70	0.46
	LOLT:	*English*	450	0.18	0.38	434	0.14	0.35	490	0.18	0.39	481	0.19	0.39
		Afrikaans	450	0.64	0.48	434	0.67	0.47	490	0.63	0.48	481	0.63	0.48
		isiXhosa	450	0.18	0.38	434	0.17	0.38	490	0.18	0.39	481	0.19	0.39
	Ex Dept:	*CED*	450	0.16	0.37	434	0.16	0.37	490	0.16	0.37	481	0.16	0.37
		DET	450	0.10	0.30	434	0.09	0.29	490	0.10	0.30	481	0.10	0.30
		HOR	450	0.58	0.49	434	0.58	0.49	490	0.57	0.50	481	0.57	0.50
		WCED	450	0.16	0.37	434	0.17	0.37	490	0.16	0.37	481	0.16	0.37
	Quintile:	*1*	450	0.18	0.38	434	0.18	0.39	490	0.18	0.39	481	0.18	0.38
		2	450	0.20	0.40	434	0.19	0.40	490	0.20	0.40	481	0.21	0.41
		3	450	0.22	0.41	434	0.22	0.42	490	0.20	0.40	481	0.20	0.40
		4	450	0.20	0.40	434	0.21	0.40	490	0.20	0.40	481	0.21	0.40
		5	450	0.20	0.40	434	0.19	0.40	490	0.20	0.40	481	0.20	0.40
Teacher Characteristics:	**Female:**		450	1.00	0.00	434	1.00	0.00	490	1.00	0.00	481	1.00	0.00
	Experience:		450	10.50	9.55	434	10.34	9.50	490	9.41	7.13	481	9.48	7.14
	Age[J]:		450	41.10	10.78	434	41.22	10.75	.	.	.	441	41.18	11.89
	Language used:	*English*	450	0.16	0.37	434	0.16	0.37	490	0.14	0.35	481	0.14	0.35
		Eng & Afr	450	0.18	0.38	434	0.18	0.39	490	0.10	0.30	481	0.10	0.31
		Eng & Afr & isiXhosa	450	0.02	0.14	434	0.02	0.14	490	0.00	0.00	481	0.00	0.00
		Afrikaans	450	0.46	0.50	434	0.46	0.50	490	0.57	0.50	481	0.57	0.50
		isiXhosa	450	0.18	0.38	434	0.17	0.38	490	0.18	0.39	481	0.19	0.39
	Literacy Programme:	*Heard*	450	0.16	0.37	434	0.16	0.37	490	0.98	0.14	481	0.98	0.14
		If heard, used	72	0.00	0.00	71	0.00	0.00	480	0.98	0.14	471	0.98	0.14
	Grade R Resource Kit:	*Heard*	450	0.38	0.49	434	0.38	0.49	490	0.98	0.14	481	0.98	0.14
		If heard, used	171	0.42	0.50	166	0.43	0.50	480	0.96	0.20	471	0.96	0.20
		< Matric	450	0.02	0.14	434	0.02	0.14	490	0.04	0.20	481	0.04	0.19

	Qualification:	*Matric*	450	0.06	0.24	434	0.06	0.24	490	0.08	0.27	481	0.08	0.28
		Certificate / Diploma	450	0.84	0.37	434	0.84	0.37	490	0.80	0.40	481	0.79	0.40
		>= Degree	450	0.08	0.27	434	0.08	0.27	490	0.08	0.27	481	0.08	0.28
	NQF Level:	*Level 4*	378	0.07	0.26	364	0.07	0.26	470	0.57	0.49	461	0.57	0.50
		Level 5	378	0.74	0.44	364	0.73	0.44	470	0.21	0.41	461	0.21	0.41
		Other	378	0.19	0.39	364	0.20	0.40	470	0.21	0.41	461	0.21	0.41
Learner Characteristics:	**Female:**		450	0.51	0.50	434	0.52	0.50	490	0.47	0.50	481	0.46	0.50
	Age:		449	5.73	0.55	433	5.72	0.55	488	5.71	0.46	479	5.70	0.46
	Home Language:	*Afrikaans*	450	0.63	0.48	434	0.63	0.48	490	0.62	0.49	481	0.62	0.49
		English	450	0.15	0.36	434	0.15	0.26	490	0.17	0.37	481	0.17	0.37
		isiXhosa	450	0.21	0.41	434	0.21	0.40	490	0.21	0.40	481	0.21	0.41
		Other	450	0.01	0.05	434	0.01	0.11	490	0.01	0.08	481	0.01	0.08
	Home Language and Lolt different:		450	0.94	0.24	434	0.94	0.25	490	0.95	0.22	481	0.95	0.23

Notes: [J] Teacher age data is not credible for the treatment group in the baseline, and some of the fieldworkers entered the current date instead of the date of birth in the midline data collection.