Ilifa Labantwana & Resep
ECD Working Paper Series

No. ECD WP 001/2021

# Estimating the impact of five early childhood development programmes against a counterfactual

**Servaas van der Berg**

October 2021

UNIVERSITEIT iYUNIVESITHI STELLENBOSCH UNIVERSITY 100 1918·2018
*saam vorentoe · masiye phambili · forward together*

RESEP Research on Socio-Economic Policy

ilifa labantwana

# Estimating the impact of five Early Childhood Development programmes against a counterfactual[1]

Servaas van der Berg[2]

Resep, University of Stellenbosch

**Abstract**

The recent development of the Early Learning Outcome Measure (ELOM), a culturally appropriate test for children aged 50 to 69 months for all eleven South Africa's official languages, offers benchmarks against which children's development can be tested. This paper compares the programme gains (value added) of participants in five early childhood development (ECD) programmes against the age-performance gradient of the ELOM test in two large scale cross-sectional studies to assess the possible influence of such programmes on early learning and learning outcomes. This goes beyond what was attempted in the programme evaluation, which did not have a control group against which to compare score gains. The cross-sectional relationship between age and learning gains is thus used as a counterfactual that can be regarded as a control group against which to evaluate programme gains. 'Effect sizes' estimated in this way range between 9% and 72% of a standard deviation under conservative assumptions. These are surprisingly large, considering that these five programmes served mainly children who would typically attend no-fee schools, and that three of the programmes are playgroups with only 2½ to 8 hours of contact time per week.

The Ilifa-Resep ECD Working Paper Series is a collaboration between Ilifa Labantwana and Research on Socio-Economic Policy (Resep) at Stellenbosch University. The working paper series aims to promote research that addresses the major systemic issues facing the ECD sector in South Africa. Key themes of the series include: financing and funding, labour, nutrition, ECD governance, regulation, economics of ECD, the household environment, and developmental outcomes of children. The series will contain research papers that address any of the components of the ECD essential package - early learning, parent and caregiver support, nutrition, maternal and child health, and social protection.

1 This paper forms part of a series of three papers that have been produced for Ilifa Labantwana on issues relating to the demand and supply of ECD services in South Africa. The author wishes to thank members of the ELOM team, in particular Andy Dawes, Linda Biersteker and Elizabeth Girdwood, as well as Colin Almeleh, Laura Brooks and Zaheera Mohamed of Ilifa Labantwana, participants in a Resep internal seminar, and Resep colleagues Eldridge Moses, Gabrielle Wills and Jesal Kika-Mistry for inputs to the development of this paper and of the wider project that it forms part of.

2 Prof Servaas van der Berg. Resep, Stellenbosch University Corresponding author: svdb@sun.ac.za

# 1. Introduction

Little is currently known about the quality of early childhood development programmes in South Africa and about their contribution in improving learning outcomes and academic readiness for school. Until recently, there was also no widely available and utilised early learning outcome measure suitable to test children in South African on an age and culturally appropriate test that was also validated across different languages. As a result, existing efforts to model expansion scenarios and their cost are almost entirely based on access to programmes, rather than potential outcomes associated with different scenarios. Fortunately, this gap has now been filled by the development of the Early Learning Outcomes Measure (ELOM), which provides benchmarks for performance of children in the age range 50 to 69 months. In 2016, this test *"was age-validated on a sample that is very likely to be representative of the range of socio-economic backgrounds of South African children and for children speaking different languages"* (DataFirst, no date) through careful assessments of 1 331 Grade R children in 173 schools (Innovation Edge 2019: 10). Subsequently, this test has also been used in an evaluation of five early childhood development (ECD) programmes in 2018, with baseline and endline data collected and compared (Dawes *et al.* 2020a). In addition, the team that undertook the benchmarking also undertook another study in 2019 to evaluate performance of Grade R learners against these benchmarks.

Some further assessments of individual programmes have also been undertaken (see the ELOM website for details), but these are not considered in this study. Instead, this paper places the emphasis on the five ECD programmes referred to above, to generate some tentative evidence about the possible influence of ECD programmes on early learning and learning outcomes. That requires going beyond what has been attempted in the programme evaluation, which did not have a control group against which to compare gain scores. This paper investigates the cross-sectional relationship between age and learning gains to create a counterfactual that can be regarded as a control group against which to evaluate the programme gains. For this purpose, data from the data sets referred to above are used in a pooled sample: ELOM2016, as we shall refer to it, the initial sample on which the benchmarks were validated; ELOM2018, the evaluation of the five ECD programmes; and ELOM 2019, the subsequent Grade R sample to evaluate Grade R performance. All three these samples are available on the DataFirst website[3].

# 2. The impact of ECD

It has become conventional wisdom that ECD is beneficial to children across a wide variety of outcomes and that its effects are large and also often lasting. The literature, especially the work of Heckman and his co-authors, points to early childhood being the most appropriate time to intervene, with early interventions seen as being less costly than trying to intervene at later ages.

Evidence of the benefits of specifically early learning interventions is a little more mixed. Reasons include that quality of interventions may sometimes be deficient, or that benefits of early learning may fade after a while, or that other children may later catch up in schools. Some authors even provide evidence of detrimental psycho-social effects for those with more exposure to early learning programmes that may dampen their academic advantage in school (Ansari 2018; Ansari et al 2019). However, this is clearly not the dominant view, which rather posits an academic advantage that only fades partially over subsequent years.

Ansari *et al.* (2019: 1497) state that the extensive literature on the effects of contemporary and large-scale Early Childhood Education (ECE) programmes indicates that it has a quite positive effect on children's short-term academic development, particularly for children from low income households.

3 https://www.datafirst.uct.ac.za/

They quote Bailey *et al.* (2017) who found that those who attended high quality early childhood programs at age 4 experienced an average treatment effect of approximately 0.25 standard deviation units.

In South Africa, there has until recently been very little measurement of early learning. In 2013, Van der Berg *et al.* (2013) undertook an investigation into the effect of the introduction of Grade R on subsequent learning outcomes in schools. They used quasi-experimental evidence from the uneven rollout of Grade R across schools as a measure of treatment intensity and performance in the Annual National Assessments (ANAs) in Grades 1 to 6 to estimate the effect of Grade R on subsequent learning. The effect sizes they found in their study and those they mentioned in the literature review on the impact of ECD interventions will be used to provide some context to the effect sizes estimated for the five programmes in this paper.

# 3. Data and methodology

Evaluating the results of the five ECD programmes in ELOM2018 is the central concern in this paper. To achieve that, the first concern is how to evaluate the gains in this dataset over the eight months between the baseline and the endline, considering that there is no counterfactual. The method applied is to construct a counterfactual by considering the effect of ageing, holding other factors constant. The question then is how large the gains between baseline and endline in ELOM2018 are compared to the gains to be expected simply from children growing older. These gains are then converted to a metric commonly used for impact evaluations, namely relative to a standard deviation (SD) in test scores.

## Data and variables used

Not one of the three ELOM datasets is fully representative of the South African population in this age group. Yet the samples in both ELOM2016 and ELOM2019 across Department of Basic Education national school quintiles[4] is broadly in line with the national distribution of learners across school quintiles. The same cannot be said for the ELOM2018 sample, a purposive sample intended to evaluate learning gains over almost an academic year in out-of-school ECD programme serving largely children likely to attend so-called non-fee schools, i.e. quintile 1-3 schools, once they enter school.

Average scores in these three tests vary significantly, but a large part of the differences result from age differences between the samples (see Table 1). ELOM2016 and ELOM 2019 were undertaken in Grade R classes, while ELOM2018 evaluated five ECD programmes, pre-Grade R. The weaker performance in ELOM2018 reflects a younger and poorer sample, drawn from ECD programmes aimed at poorer children (children likely to attend mainly attend DBE Quintiles 1 to 3, so-called non-fee paying schools, when they enter school). The score difference of 8.4 ELOM points between ELOM2019 and ELOM2016 is also considerable. In the regression analysis discussed later, a difference of about 6½ ELOM points (about 0.40 SDs) remains even after controlling for age and gender. Part of this difference is probably ascribable to the later start (one term later) of the 2019 data collection, implying that children had longer exposure to Grade R before they were tested.

The outcome variable in this study is the ELOM total score, which is the sum of scores on of 23 direct assessment items, clustered in five domains, namely Gross Motor Development, Fine Motor Development and Visual Motor Integration, Emergent Numeracy and Mathematics, Cognition and Executive Functioning and Emergent Literacy and Language. It was created to provide "a reliable, age valid tool that provides a fair assessment of children from across ethnolinguistic groups" (Dawes *et al.* 2020c: 16).

4 DBE 'quintiles' are not equally sized, despite the original intention and the nomenclature

Note that only baseline scores are used for ELOM2018. Standard deviations in these scores were 14.07 for ELOM2016, 15.2 for ELOM2018, 14.5 for ELOM 2019, and 15.9 for the pooled sample, reflecting the large differences between the samples.

***Table 1: ELOM total scores and age in the three samples***

| | ELOM score | Age (months) |
|---|---|---|
| ELOM2016 | 49.4 | 62.9 |
| ELOM2018 | 38.5 | 54.5 |
| ELOM2019 | 57.8 | 64.5 |

**Source:** Own calculations from ELOM2016, ELOM 2018 and ELOM2019

As ELOM has been benchmarked for children aged 50 to 69 months, age in the three samples falls largely in this range. Gender is often found to be a significant predictor of scores, with a female advantage also quite common on such measures of learning.

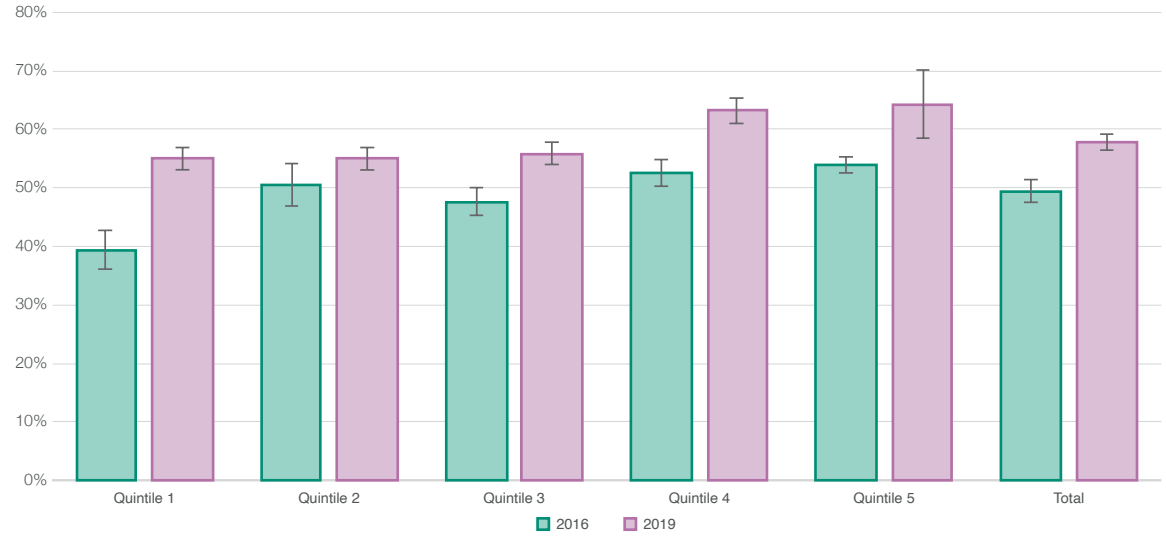
There are two potential indicators of socio-economic background:

- ELOM2016 and ELOM2019 contain information on the school quintiles of the sampled schools. As Figure 1 shows, ELOM scores are generally better for children in higher school quintiles, but this relationship is not very strong in the absence of other controls. ELOM2018 programmes were not conducted in schools, and constructed variables to provide a proxy for quintile or socio-economic status could not be converted to be consistent with the school quintiles.
- Another variable linked to socio-economic status is the height-for-age of children. This is used in some of the regressions, but as this measure was not available for all children, including it reduces the size of the pooled sample.[5] For comparison purposes, a regression including height-for-age as variable is compared with one for the same reduced sample that excludes this variable. The height-for-age z-scores in these datasets show 7.4% of the 2016 and 5.3% of the 2019 samples were stunted in terms of being at least 2 SDs below the norm, while this applied to 19.7% in the 2018 data, which covered mainly children expected to enter one of the three lowest DBE quintiles of schools.[6] To place these height-for-age measures in perspective, the most recent measures national estimate of stunting is from the Demographic and Health Survey (DHS) of 2016, which indicated that 27% of South African children were stunted. It would therefore seem that stunting levels were considerably lower in all three these studies than in the DHS.

5 This is often a very useful indicator of poverty, not so much as a continuous indicator but rather when using a dummy indicator to indicate height-for-age z-scores two standard deviations below the norm, a measure of stunting. Surprisingly, the continuous indicator gave a better fit than the dummy for stunting, even though one would not expect the relationship between heigh-for-age and learning outcomes to hold across the full range of the former.

6 In a normal population with no stunting, around 2.5% would fall into the category of being 2 standard deviations below the height-for-age norm, given that the 95% confidence interval is plus or minus 1.96 standard deviations from the mean.

*Figure 1: Performance of Grade R learners in ELOM's 2016 and 2019 surveys by school quintile*

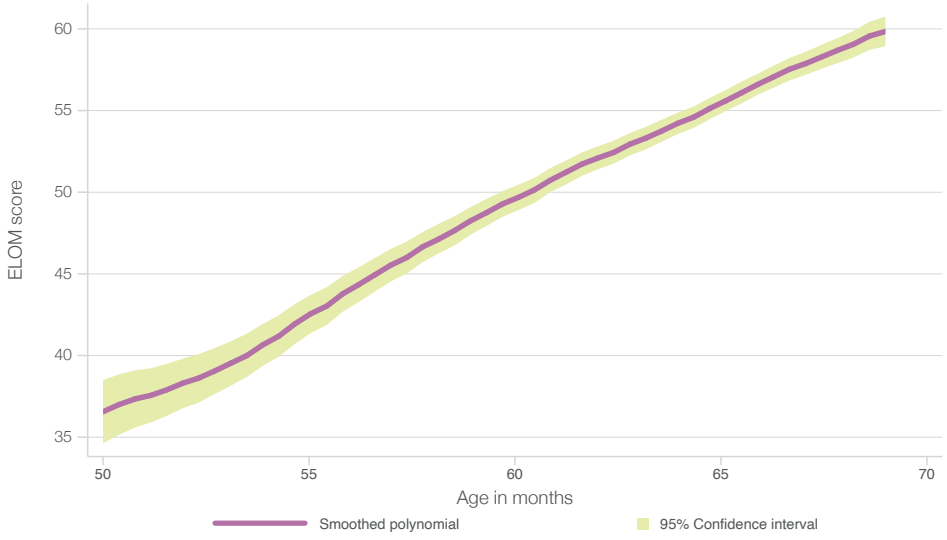



95% confidence levels. Standard errors clustered at school level.
**Source:** Calculated from ELOM 2016 & ELOM 2019

## Methodology

Neither the samples individually nor the pooled sample provides a representative sample, but together they cover a large number of children that have been carefully evaluated on a well-designed benchmarking test developed for South African circumstances, context and languages. The pooled sample contains enough children across the whole socio-economic spectrum to have some confidence in it for the purpose at hand. The focus is mainly on the relationship between age and performance. Figure 2 shows that even before controlling for potential explanatory factors, the confidence bands around this bi-variate relationship between age and test score are quite tight, while the slope is quite steep.[7]

*Figure 2: Relationship between age and total ELOM score*




**Note:** Local polynomial of ELOM total score by age in months for children aged 50-69 months on pooled dataset for ELOM2016, ELOM2018 & Elom2019.

7 In terms of the regressions discussed later, if age were to be dropped from Regression 3, the R-squared would decline from 0.27 to 0.18.

Regression analysis is undertaken mainly on the pooled sample. Using interaction variables between age and the different samples, it is possible to extract the implicit coefficient for each sample for age in the pooled data. We also test individual samples (not shown), in order to investigate the range of coefficients on the age variable that should be considered for constructing the counterfactual.

# 4. Results

Four different regression models were applied to the pooled dataset, as shown in Table 1. The basic model (Regression 1) investigates the relationship between the total ELOM scores and gender, age and year dummies (effectively dummies for the three samples). This regression statistically 'explains' just over one quarter (27%) of the variance in scores between individuals, a quite respectable fit, considering the sample's heterogeneity in home background, parental education, exposure to school or previous ECD programmes or to the type of activities that may improve learning (such as being read or told stories).

Girls have a learning advantage, scoring 3 ELOM points higher than boys, conditional on other variables. Note the coefficient of 1.2 on age variable. As referred to earlier, scores in ELOM2019 are significantly higher than in the other two tests, probably mainly because this test was administered about a term later than in ELOM2016. The greater exposure to formal Grade R may have benefited such children more than only reflected in their becoming older.

***Table 2: Regressions of various explanatory factors on ELOM scores in pooled sample***

| | Regression 1 | Regression 2 | Regression 3 | Regression 4 |
|---|---|---|---|---|
| | Basic regression | Adding interaction effects | Adding height-for-age | Reduced sample without height-for-age |
| Female | 3.260***<br>(7.26) | 3.252***<br>(7.23) | 2.885***<br>(5.83) | 2.979***<br>(5.82) |
| Age in months | 1.202***<br>(13.71) | 1.020***<br>(5.57) | 1.059***<br>(-6.67) | 1.020***<br>(-5.58 |
| Age x 2018 | | 0.574*<br>(2.05) | 0.463<br>(1.66) | 0.538<br>(1.75) |
| Age x 2019 | | 0.207<br>(1.00) | 0.163<br>(0.85) | 0.235<br>(1.08) |
| Year dummy: 2018 | -0.702<br>(-0.22) | -33.510*<br>(-2.20) | -25.970<br>(-1.71) | -32.050<br>(-1.96) |
| Year dummy: 2019 | 6.572***<br>(-6.29 | -6.484<br>(-0.49) | -3.924<br>(-0.32) | -8.189<br>(-0.59) |
| Height-for-age (z-score) | | | 2.901***<br>(10.94) | |
| Constants | -27.88***<br>(-4.91) | -16.44<br>(-1.39) | -17.18<br>(-1.68) | -16.29<br>(-1.39) |
| N | 3 867 | 3 867 | 3 371 | 3 371 |
| R-squared (adjusted) | 0.270 | 0.272 | 0.272 | 0.233 |

t statistics in parentheses. * p<0.05, ** p<0.01, *** p<0.001
**Note** that an interaction between gender and age is not statistically significant.
**Source:** Calculated from ELOM2016 & ELOM2019.

The second regression allows for interaction between the year dummies and age. While the female advantage remains almost unchanged, the effect of age now differs significantly from that for the reference group (the 2016 sample) and ELOM2018, with a statistically significant and large difference in the interaction term. This implies age effects of 1.020 for ELOM2016, 1.594 for ELOM2018, and 1.227 for ELOM2019, although this latter coefficient is not statistically different from the 2016 one. Naturally, the stronger growth in the age effect in the last two samples has a downward effect on the coefficients on the year dummies.[8]

The third regression model investigates whether height-for-age, as a measure of socio-economic status, is also conditionally correlated with test scores. It is indeed statistically significant, but the coefficient of 2.901 for this variable is quite small, implying that an improvement of a full SD in height-for-age is associated with only a 2.9 points higher test score. The coefficient on the female dummy is slightly reduced, implying that a small part of the female advantage in performance is explained by the small socio-economic status advantage on average of the girls in this sample. Note that height-for-age scores are not available for all children in the three samples, so the fourth regression repeats the second one, but for the reduced sample from Regression 3 to allow comparison. Although height-for-age is significant in Regression 3, it does not change other coefficients in a major way, as can be seen from the comparison of Regressions 3 and 4.

The regressions above point to an ageing "effect" that varies between 1.02 and 1.59 points per month. The 1.59 seems to be an outlier, even though further regressions on individual samples for ELOM2018 found an even slightly higher coefficient of 1.62, once other SES variables such as constructed quintiles enter the regression. However, regressions on the other individual samples confirm that the ageing effect is unlikely to lie outside this range. In the next section, these age-score gradients, or the effect per month of ageing, will be compared to those observed between the baseline and endline of the ELOM2018 data, as reported in Dawes *et al.* (2020a).

## 5. Gains in ELOM scores in five intervention programmes

The 2018 study by the team that developed ELOM was undertaken to determine "the relative effectiveness of different programmes in improving early learning outcomes for young children" who were on average 4½ years old at baseline in March 2018 (Dawes et al 2020a: ii). The study involved 369 children from low income households (roughly quintiles 1 to 3) in three playgroups and two centre-based programmes. Children in the playgroups were exposed to an early learning programme between 2½ and 8 hours per week, those in the two centre-based programmes 15 to 22 hours per week. Table 3 summarises the main features of the five groups.

8 Implicitly, the constant refers to age 0 months, and the year dummies have to be added to this.

*Table 3: Participating programmes: Main features*

| | Playgroup 1 | Playgroup 2 | Playgroup 3 | Centre development 1 | Centre development 2 |
|---|---|---|---|---|---|
| Delivery model | Playgroup model directly managed by PG1 | Mobile playgroup model directly managed by PG2 | Playgroup franchise model designed for scale (minimum critical specification for efficient replication) | Centre development programme for practitioners in independent ECD sites; no direct intervention with children | Centre development programme for practitioners in independent ECD sites; no direct intervention with children |
| Program target | 2 to 4 year old children | 3 to 5 year old children | 3 to 4-year old children | Practitioners of Pre-Grade R children (4 to 5 years) | Practitioners of Pre-Grade R children (4-5 years) |
| Child sessions per week | 2 to 3 sessions per week of 4 hours each | 1 session per week of 2.5 hours | 2 sessions per week of 3 hours | 5 sessions per week of 4.5 hours | 5 sessions per week of 3 to 4.5 hours |
| Practitioner qualifications | Minimum NQF Level 4 ECD qualification | Minimum NQF Level 4 ECD qualification | Minimum: 5-day training and accreditation; some have NQF Level 4 ECD qualification | Depends on the site | Depends on the site |

**Source:** Dawes *et al.* 2020c: Table 1, 10

In October 2018, 8 months after the baseline, children were again tested on the ELOM benchmark test. Multi-level modelling that inter alia allowed for analysis of the effect of frequency of attendance was undertaken. Statistically significant gains in ELOM scores of between 13 and 20 ELOM points were observed in all the programmes, and the authors concluded that these changes *"…are largely attributable to programme participation rather than to opportunities for stimulation at home"*. They found that a large proportion of parents *"never engaged in activities likely to improve early learning outcomes (reading, telling stories, or singing to children)"* (Dawes *et al.* 2020a: ii), often due to lack of time. This is closely tied to a second observation, that children who attended more sessions showed most improvement in ELOM scores. They concluded that *"(w)ell designed and closely monitored playgroup programmes can perform as well as more expensive centre-based models."*

The programme evaluations were thoroughly done, yet logistic and ethical reasons made it impossible to have a control group. This paper constructs a pseudo-control by estimating the gains that may have occurred in the absence of the intervention programmes, i.e. purely from children growing older, and evaluating programme gains against that. Based on the regressions in Table 2 and those of the individual samples, as discussed earlier, the gain in ELOM points ascribable to age alone, controlling for other factors, varies between 1.02 and 1.62 ELOM points per month. However, a coefficient of 1.62 in the regression based on ELOM2018 seems excessive and is probably an artefact of the selection criteria for the relatively small sample of programmes observed here. It would appear more sensible to accept values of 1.02 (as in the 2016 dataset) and 1.23 (as in the 2019 dataset). Even these are quite large: Taking a conservative value of 16 ELOM points as a SD, they would imply a large annual improvement in learning of 0.77-0.92 of a SD simply from children growing older., while the 1.62 coefficient would imply an implausibly large learning gain from ageing alone of 1.22 SD in a year. To put this in perspective, Evans and Yuan (2019: 35) estimated a pooled effect in five developing countries of normal learning and ageing of about 0.22 SD for children aged 5, implying that the ageing effect measured in these five South African programmes is quite high.

Accepting the lower coefficients on ageing, over the eight months from baseline to endline, a child's gain from growing older would then be somewhere between 8 and 10 ELOM points. The gains actually achieved in the interventions by participants in both the baseline and endline test were between around 11 to 21 points (Table 4), implying a programme effect ranging between 1.36 and 13.14 ELOM points (total effect minus the expected effect of ageing), or 8.5% to 82% of a SD gains more than the effect of children growing older. These are large impacts, implying programme effects of between 11 and 20 months. For Playgroup 1, that showed the largest gains and really consists of two playgroups in different provinces, the gains were significantly larger in a site in Mpumalanga where children attended three times a week for 4 hours, than in a site in the Western Cape where only two sessions of the same length are offered per week.

***Table 4: Estimating impact from five ECD programmes***

| | Playgroup 1 | Playgroup 2 | Playgroup 3 | Centre Development 1 | Centre Development 2 |
|---|---|---|---|---|---|
| Baseline score | 30.3 | 40.0 | 31.6 | 49.4 | 35.2 |
| Endline score | 51.6 | 51.2 | 46.1 | 63.9 | 55.9 |
| Gain (ELOM points) | 21.3 | 11.2 | 14.5 | 14.5 | 20.7 |
| **Gain (in months) achieved in 8 months:** | | | | | |
| Low estimate | 17 | 9 | 12 | 12 | 17 |
| High estimate | 21 | 11 | 14 | 14 | 20 |
| **Programme 'effect' (in ELOM points) given gains from ageing of 8.16 to 9.84 ELOM points** | | | | | |
| Low estimate | 11.46 | 1.36 | 4.66 | 4.66 | 10.86 |
| High estimate | 13.14 | 3.04 | 6.34 | 6.34 | 12.54 |
| **Programme 'effect' (in % of a standard deviation)** | | | | | |
| Low estimate | 71.6% | 8.5% | 29.1% | 29.1% | 67.9% |
| High estimate | 82.1% | 19.0% | 39.6% | 39.6% | 78.4% |

**Note:** Programmes arranged by ELOM point gains. Low estimates assume monthly gains from ageing alone to be 1.23 ELOM points (9.84 in 8 months), while high estimates assume this coefficient to be only 1.02 ELOM points (8.16 in 8 months). Conversion to standard deviations assumes a standard deviation of 16 ELOM points.[9]
**Source:** Own calculations from ELOM2018 using a balanced panel, i.e. children tested both at baseline and at endline.]

9 This relatively high estimate is chosen to ensure that 'programme effects' are not over-estimated. Standard deviations for the individual programmes in ELOM2017 ranged between 10 and 16 and were 14.1 in ELOM2016, 14.4 in ELOM2019 and 15.5 in the pooled sample.

# 6. Evaluating impact in context

## Why are these programme impacts so much larger than those estimated for Grade R?

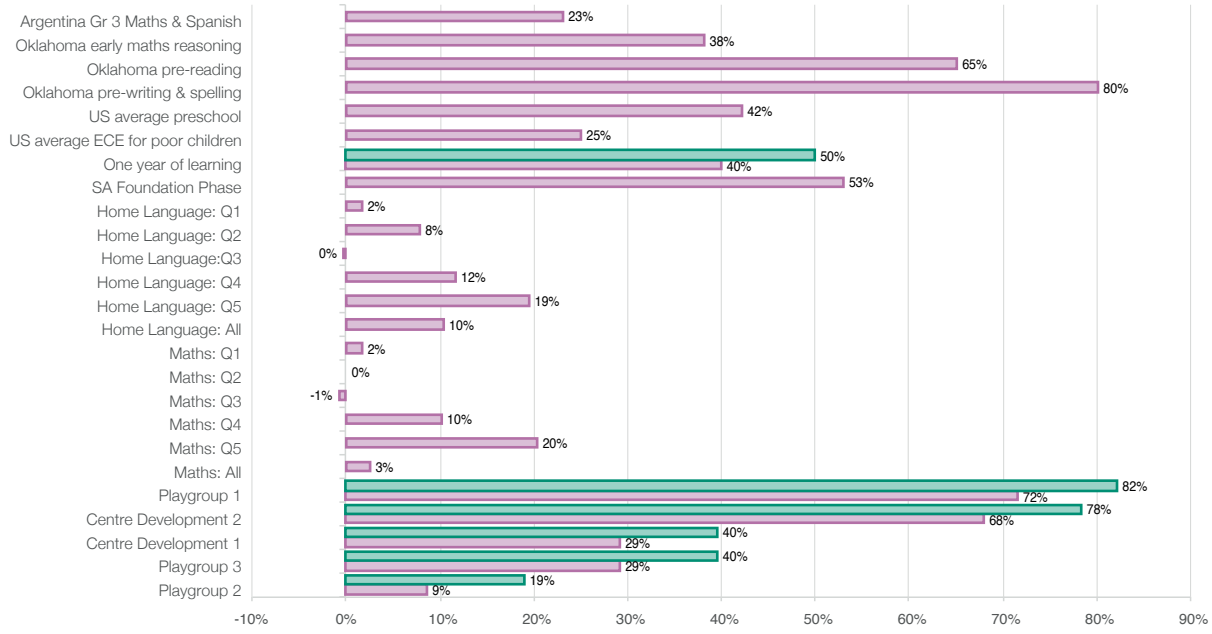
Figure 3 compares the estimates of the impact of the five programmes discussed above with other estimates, including those for Grade R in Van der Berg *et al.* (2013). The programmes evaluated in ELOM2018 have much bigger effect sizes, ranging between 9% and 72% under the more conservative assumptions, as against between 2% and 20% in the Grade R study. They appear to be best comparable to the Oklahoma study, also depicted in the figure. Considering that these five programmes served mainly children who would typically attend Quintile 1 to 3 schools, it is necessary to interrogate these large effects compared to the Grade R study. A number of factors may account for these differences:

- Firstly, the Grade R study investigated the effect on learning some years after children had been exposed to Grade R, thus some of the effects could have faded. This is often found in follow-up testing undertaken some time after the intervention.
- The Grade R study made use of a natural experiment, the uneven rollout of the Grade R programme. Treatment (exposure to Grade R, in this case) is naturally poorly measured in such a situation. Grade R attendance in a particular year in a school was expressed as a proportion of the number of children in a particular grade in the same school in the relevant year. Movement between schools, the fact that some children attended Grade R in community based ECD centres and not in schools, and grade repetition (particularly in poor schools) could mean mismeasurement of treatment effects for a specific cohort. Also, some observers question the accuracy of the ANAs. Mismeasurement of variables causes attenuation bias, meaning that the full effect would likely be under-estimated.
- There has been some attrition in the ELOM2018 sample, where some children dropped out and consequently were not tested at endline. It is possible that they may have gained less than average from the programmes they participated in, and thus would have shown lower learning gains. (On the other hand, modelling in the initial study found larger gains for those who started from a lower base.)
- Evan and Yuan (2020: 1) as well as Kraft (2020) mention that effect sizes estimated in small studies tend to be larger than those in bigger studies – typically twice as large in a large number of studies they considered.
- In a meta-analysis of South African education interventions that did not evaluate pre-school interventions, Besharati *et al.* (2021) found an average effect of 0.53 of a standard deviation for the Foundation Phase, as shown in the figure. They also confirmed that South African interventions seem to have larger effects at lower age ranges, which may also speak to ECD. They also found that effects sizes measured at school level (as in the Grade R evaluation) tend to be smaller than those based on individual data, as in the five programmes discussed here.
- Dawes *et al.* (2020a:9) note that their sample is focused on good programmes, so one may not be able to generalise from them.

> *"It is essential to repeat that the programmes studied are not representative of the South African ECD programme population. Study children were attending ECD playgroups and centres where the practitioners had been rated as well-functioning by their parent organisations. The same programmes, if poorly delivered, could not be expected to show the same outcomes as those observed here."*

To the extent that the programmes that were measured were selected to be well functioning, one may even see this as similar to the Perry Pre-school study in the USA, that also estimated effects on a high end intervention.

***Figure 3: Comparing effect sizes of five programme interventions with other estimated effects of ECD interventions***




**Notes:** The blue bars show higher estimates where more than one impact was estimated.
**Sources:** The Argentina estimate, the various Oklahoma estimates and the US average pre-school estimate are taken from Van der Berg et al (2013). US average ECE estimate from Antrabi *et al.* 2019, quoting Bailey et al (2017). The SA Foundation Phase results are those summarised by Besharati *et al.* (2021), as discussed in the text. Home Language and Maths results shown are Van der Berg *et al.*'s estimates of the effect of Grade R. The programme estimates (both high and low estimates) are from Table 4.

So while there were some selection effects in terms of the ECD programmes observed, the measurement of Grade R impact extended over the whole school system.

## How large are these impacts in international context?

Summarising data from 138 randomized controlled trials (RCTs) and 260 quasi-experimental studies in low and middle income countries, Evans & Yuan (2020: 1, 12) found an effect size of respectively 0.10 and 0.08 standard deviations in learning outcome at the median, and 0.38 and 0.63 at the 90th percentile. Kraft (2020: 247) also found an effect size of 0.10 at the median for 747 randomized control trials of education in high income countries.

Historically, Cohen (1988) and later Hattie (1994) proposed criteria by which to judge effect sizes in education. However, these effect sizes were quite large, perhaps because they included correlational estimates that were not causal impacts, something that both Kraft (2020) and Evans & Yuan (2020) frown upon. Kraft (2020: 247) proposed, for pre-school (pre-K-12, in American terminology) interventions that effect sizes up to 0.05 SD be considered as small, 0.05-0.20 SD as medium, and above 0.20 SD as large.

The estimates presented on the 'programme effect' in the five studies evaluated in this paper cannot really be considered causal in the same way as RCTs. Yet neither are they simply correlational, as would be obtained by simply evaluating gains between baseline and endline. By measuring gains relative to a counterfactual or control group, based on cross-sectional estimates of the gains from ageing, they can probably be considered as closer to causal estimates than would be obtained from simple correlational analysis alone.

Thus we may conclude that the effects observed in the five programmes evaluated in 2018 are exceptionally large.

## Do these results make a case for scaling up ECD?

Do these estimates make a case for similar interventions at scale in South Africa? In another context, Kraft believed one should consider a broader context when considering scaling up:

> *Ask, would the effects be similar if the intervention were offered to a large, diverse population of students? Is it likely the intervention would be implemented with fidelity by others? Is it politically feasible to scale the intervention? Reasonable people will disagree about the answers to these questions. The larger point is to introduce scalability into the process of interpreting effect sizes… Assessing scalability helps to provide a measure of the challenges associated with expanding a program so that these challenges are considered and addressed. (Kraft 2020:248)*

Again, it is important that the authors of the study on the five programmes emphasised that:

> *…the programme sites where children were assessed for this study were all selected because they were rated as well functioning… They are therefore neither representative of the range of quality within each programme, nor of South African playgroup and centre-based provision. The same programmes as those studied here, poorly delivered, cannot be expected to deliver similar outcomes. (Dawes et al 2020a:105)*

The evidence evaluated in the current study can thus only be considered as an indication that the five promising ECD programmes appear to be associated with relatively large benefits for children in poorer contexts, who would otherwise start Grade R with considerable deficits. These are relatively low cost programmes, within the landscape of formal ECD provision to 4 or 5 year olds in South Africa, and the 'treatment' that participants are exposed to is quite limited, at 2.5 to 8 hours per week in the playgroups and 15 to 22.5 hours per week in the two centre-based programmes.

The two tables and the figure presented next may assist a more nuanced evaluation. Table 4 sets out the benchmarks for performance on the ELOM scale for children aged 50-59 months and 60-69 months respectively. Table 5, taken from Dawes et al (2020a), evaluates programmes baseline and endline performance, using first the younger and then the older age group's criteria. In three of the five programmes, the average endline performance lies in a higher ELOM performance band than at baseline, even though the eight month of treatment is less that the ten month gap between the two age bands: Playgroup 3's and Playgroup 1's averages progress from the 'at risk' to the 'falling behind' category, and Centre Development 2's from the 'at risk' to the 'achieving the standard' category.

***Table 5: ELOM scores reflecting standards and performance bands for children aged 50-59 and 60-69 months respectively, with ELOM colour code***

| Age group | At risk | Falling behind | Achieving the standard |
|---|---|---|---|
| 50-59 months | 0 to 36.01 | 36.02 to 46.31 | 46.32 to 100 |
| 60 to 69 months | 0 to 43.23 | 43.24 to 54.37 | 54.38 to 100 |

**Note:** Colour code based on the 50-59 and the 60-69 month old ELOM performance standards. Red indicates children at risk, orange children falling behind the standard, and green children achieving the ELOM standard.
**Source:** Dawes *et al.* 2020a, Table 22

*Table 6: Changes in ELOM scores between baseline and endline in five ECD programmes, using ELOM colour code*

| | Baseline score (colour based on 50-59 months standard) | Endline score (colour based on 60-69 months standard) | Extent of change (in ELOM point) |
|---|---|---|---|
| Centre Development 1 | 49.4 | 63.9 | 14.5 |
| Playgroup 2 | 40.0 | 51.2 | 11.2 |
| Centre Development 2 | 35.2 | 55.9 | 20.7 |
| Playgroup 3 | 31.6 | 46.1 | 14.5 |
| Playgroup 1 | 30.3 | 51.6 | 21.3 |

**Note:** Red indicates children at risk, yellow children falling behind the standard, and green children achieving the ELOM standard. Colour code based on ELOM standards for respectively the 50-59 month age group at baseline and the 60-69 month group at endline. As the interval of eight months between the baseline and endline is shorter than the ten months difference between the ELOM benchmark age categories, the endline is measured against a slightly more onerous standard. Note that the results shown here are for the balanced sample, i.e. only respondents tested in baseline and endline are included.
**Source:** Table 4

The effect of the different programmes on individual children's performance relative to the ELOM standards is shown for the balanced panel (those that participated in both the baseline and the endline tests) in Figure 4. Those 'at risk' declined from 49% to 23% of participants across the five programmes, and those 'achieving the ELOM standard' increased from 28% to 48%. The latter is especially meaningful: those ready to enter Grade R increased from far less than one third to more than one half amongst these relatively poor children, simply through attending these relatively low cost and low intensity programmes.

*Figure 4: Changes in meeting ELOM standards between baseline and endline in ELOM2018*


**Note:** Baseline measured against ELOM standard for 50-59 months, endline against 60-69 months standard. Balanced panel used, i.e. only children with both a baseline and an endline score, irrespective of whether they met the age criteria in either baseline or endline. The interval of eight months between the baseline and endline is smaller than the ten months difference between the ELOM benchmark age categories, which implies that the endline is measured against a relatively slightly more onerous standard.
**Source:** Own calculations from ELOM2018.

We may then conclude that these five ECD programmes that were evaluated did extremely well in raising children's learning outcomes to levels that are more appropriate for children who would soon be entering Grade R. Even though the authors are careful in pointing out that these programmes may be better functioning than average, the findings nevertheless offer some cause for optimism.

The low intensity (limited hours) of all the programmes means that treatment was limited, yet what appears to be the causal impacts are large. Moreover, they were achieved in programmes aimed at children from the poorer segments of our society.

The developers and funders of ELOM have done the country a great service by developing and testing these benchmarks, and by also applying these benchmarks to programme evaluation of five ECD programmes. Moreover, by putting this data in the public domain, they have made this study possible, and perhaps also many more future studies utilising these data and new data that get generated. Much more research is required before one can state with great confidence how large the causal effects of South African ECD provision are. This study can only be a small step along that road. Yet the results provide some cause for optimism that ECD programmes could, under the right circumstances, lead to more children entering Grade R and beyond with fewer of the backlogs that still severely hamper so many children from poor home circumstances.

For policy makers and researchers working in this field, these results now offer the opportunity to start making assumptions about possible gains in cognitive outcomes that may be associated with various scenarios for ECD expansion, effectively creating the opportunity to consider both costs and benefits of ECD scale-up. But this needs to be done with careful consideration of the limits to what we know. Thus, for instance, international examples teach us that the beneficial effects of scale-up are likely to be smaller than those observed across individual programmes. There is a strong case for further expanding our knowledge base though investigating the effects of more programmes. Similarly, it would be good to scale up these or similar programmes while investigating their impact, to test the external validity of the findings reported in this paper.

## REFERENCES


Ansari, A. 2018. The persistence of preschool effects from early childhood through adolescence. J*ournal of Educational Psychology 110*(7), 952–973.

Ansari, Arya, Robert C. Pianta, Jessica V. Whittaker, Virginia E. Vitiello, Erik A. Ruzek. 2019. Starting early: The benefits of attending early childhood education programs at age 3. *American Educational Research Journal* 54(4): 1495-1523.

Bailey, D., Duncan, G. J., Odgers, C. L., Yu, W. 2017. Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness 10,* 7–39.

Besharati, Neissan Alessandro, Brahm Fleisch & Khotso Tsotsotso. 2021. Interventions to improve learner achievement in South Africa: A systematic meta-analysis. Chapter 2 in Maringe, Felix. 2021. *Systematic Reviews of Research in Basic Education in South Africa.* African Sun Media.

Cohen. J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Dawes, A., L.Biersteker, E.Girdwood, M, Snelling & J.Horler. 2020a. *Early Learning Programme Outcomes Study Technical Report.* Claremont Cape Town: Innovation Edge & Ilifa Labantwana.

Dawes, A., L.Biersteker, E.Girdwood & M.Snelling. 2020b. *Early Learning Outcomes Measure*. Technical Manual 3rd edition (update of 2016 edition).

Dawes, A., L. Biersteker, E.Girdwood, M.Snelling, & J.Horler. 2020c. *The Early Learning Outcomes Study: Research Insights*. Claremont, Cape Town: Innovation Edge and Ilifa Labantwana.

Early Learning Outcomes Measure Project. E*arly Learning Outcomes Measure 2016* [dataset]. Version 1.1. Cape Town: Innovation Edge [producer], 2017. Cape Town: DataFirst [distributor], 2017. DOI: https://doi.org/10.25828/m8hh-t883.

Evans, David K. & Fei Yuan. 2020. *How big are effect sizes in international education studies?* CGD Working Paper 545. Washington, DC: Center for Global Development.

Evans, David K. & Fei Yuan. 2019. *Equivalent Years of Schooling : A metric to communicate learning gains in concrete terms*. World Bank Policy Series WPS8752. Washington, DC: World Bank.

Hattie, J. 1992. Measuring the effects of schooling. *Australian Journal of Education* 36(1): 5-13.

Innovation Edge. 2019. *The South African Early Years Index*. Innovation Edge: Cape Town.

Kraft, M. A. 2020. Interpreting effect sizes of education interventions. *Educational Researcher* 49(4): 241–253.

Van der Berg, Servaas, Elizabeth Girdwood, Debra Shepherd, Chris van Wyk, John Kruger, Janeli Viljoen, Olivia Ezeobi & Poppie Ntaka. 2013. *The impact of the introduction of Grade R on learning outcomes*. Final Report for the Department of Basic Education and the Department of Performance Monitoring and Evaluation in the Presidency. Stellenbosch University: Department of Economics.


UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY
100
1918 · 2018
saam vorentoe · masiye phambili · forward together


RESEP
Research on Socio-Economic Policy


ilifa
labantwana