Words Can Change Worlds: An impact evaluation of Shine Literacy

Charlotte-Kathrin Stollberg

February 18, 2018

Thesis presented for the degree of
MASTER OF COMMERCE (APPLIED ECONOMICS)

Supervisor:
Professor Justine Burns

University of Cape Town, School of Economics

Abstract

The recent Progress in International Reading Literacy Study and The Southern and Eastern Africa Consortium for Monitoring Educational Quality reports painted a dire picture of South Africa's literacy situation: 80% of Grade 4 learners were rendered unable to read for meaning and 27% of Grade 6 learners as functionally illiterate. These results need to be contextualized against the extensive public spending that is incurred on education. Though it appears that learners are in school, they do not seem to be learning, a phenomenon encountered repeatedly in the developing world. The production process of educational outputs is often being hidden in the black box.

With a large body of research confirming how reading literacy holds predictive validity for later child development and academic success, Shine Literacy offers an intervention that is set at lower primary school level and is intended as a swift corrective measure for those who struggle to read early on. This dissertation conducted a quasi-experimental impact evaluation by estimating the treatment effect of Shine Literacy via difference-in-differences and propensity score matching. By using the available data which included (1) Shine's diagnostic test scores, (2) attendance data and (3) Grade 3 Systemic test score data obtained from the Western Cape Department of Education, the estimation procedures arrived at average treatment effects ranging between 0.6 to 1.9 standard deviations. IsiXhosa and "At risk" learners capture the largest test score improvements, and therefore are the main beneficiaries of the programme. This renders Shine Literacy as an extremely valuable input in the production of better literacy and thereby better schooling outcomes. It helps those at the bottom end of the distribution. Furthermore, positive impact on Systemic Mathematics test scores was found as well, confirming the predictive power of literacy on numeracy repeatedly discussed in the literature.

Acknowledgements

I am indebted to Professor Justine Burns, who has not only been supportive and shaping of my professional goals but has also provided me with irreplaceable personal and professional guidance. As my supervisor and mentor, she has given me more than I could assign her credit for. I would like to express my gratitude for her valuable comments and vision during the development of this dissertation.

Thanks also goes to Shine Literacy, and specifically to Jacqui Dornbrack, Bea Volbrecht and Carrie Mashek who have been incredibly supportive and helpful with the compiling of data, and providing me immense insight on the Shine Literacy programme. I am also grateful for the assistance from Ronald Cornelissen from the Western Cape Department of Education for assisting me with acquiring the Systemic test score data.

In the pursuit of this dissertation, nobody has been more instrumental than my friends and my family. I specifically would like to thank Samantha, Emma, Kristin, Maria, Lindokuhle, and Karen. Each of you never ceases to believe in me and my abilities, your contribution to this work is enormous. I would like to also thank my father Joachim, whose unconditional love and nurture, guidance and support have been with me whatever I aspire to be. Then, my grandmother Karin who has been giving me unending inspiration, and is my greatest role model: thank you.

Table of Contents

List of Figures

List of Tables

Introduction: The South African literacy crisis

Reading literacy is a key problem among South African learners. The latest results from the 2016 Progress in International Reading Literacy Study (PIRLS) on Grade 4 reading achievements across 50 participating developed and developing countries indicated that 8 out of 10 Grade 4 learners in South Africa are unable to read for meaning (Mullis et al., 2017). Learners cannot locate clearly listed information in text or make inferences about stated events or actions (Mullis et al., 2017). The 2011 PIRLS, which considered South African Grade 5 learners, showed that even in higher grades the picture is dismal: 29% of those students did not reach the pre-PIRLS international benchmark for literacy (Pretorius & Spaull, 2016). Similarly, the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) reports have indicated that 27% of South African Grade 6 learners are functionally illiterate[1], with only very little improvement between 2000 and 2007 (Moloi & Chetty, 2011). With less than 40 correct words per minute, learners are essentially rendered as "non-readers" (Pretorius & Spaull, 2016: 1456).

The results from the SACMEQ and PIRLS reports need to be contextualized against the extensive public spending South Africa incurs on education, which has been between 5- 6% of GDP since 1994 (World Bank, 2016). This is relatively high, internationally speaking (World Bank, 2016). Given that about 99% of learners are enrolled, school enrolment itself does not seem to be the issue (World Bank, 2016). Instead, South Africa's education crisis can be linked to the lack of quality inputs, not access – a phenomenon encountered repeatedly in the developing country context (Pretorius & Spaull, 2016). As a consequence of a poorly performing education sector, the proportion of unemployable low-skilled youth is increasing, ultimately leading to a widening of the already stark income and wealth inequality, and to a further halt on the move towards a knowledge-based South African economy (McCarthy & Oliphant, 2013).

Universally, as one of the primary skills acquired at school, reading literacy is defined not only as having the ability to understand, use and reflect on text to develop reading knowledge and skills, but ultimately to participate actively and successfully in society[2] (Linnakylä et al., 2004; Howie et al., 2008; Pretorius & Spaull, 2016). Illiterate individuals are at risk of being excluded from working and studying, but also cultural or social life (Linnakylä et al., 2004). To further emphasise the importance of this topic, Overett & Donald (1998)) highlight literacy's ability to equip people with cultural power and identity, which may – particularly for individuals with marginalised background – give individuals the necessary skill set to transform their life. There is a large body of research confirming how early reading ability holds predictive validity for later child development and academic outcome (Du Plessis et al., 2003; Howie et al. 2008; The National Institute for Literacy, 2008; Spaull, 2016). It has been shown that interventions targeted at lower primary school level

[1] By being functionally illiterate, a learner is unable to read a simple, short text and infer any meaning from the written text (Spaull, 2011).
[2] Linnakylä et al. (2004) go as far as framing reading literacy not only a basic skill, but a goal in itself; providing means in individual and educational development, at and outside of school, during work but also leisure activities. That is, acquiring the ability to read lays not only the pathway to learn in one's mother tongue and a range of other subjects, but also symbolises a pre-requisite for active participation in adolescent and adult life.

can be very effective in serving as relatively swift corrective measures for those who struggle to read (The National Institute for Literacy, 2008; He et al., 2009)[3].

Undeniably, the current South African educational system has failed to deliver. Sub-standard academic performance and other education metrics such as high dropout rates and grade retention are indicative of a rather dire situation[4]. This is in large part due to the structural and systemic issues that are deeply entrenched in the post-apartheid South African schooling environment. Apartheid left its footprint on the South African educational system with unequal access to education and varying levels of educational quality – when it comes to both private and social returns to schooling, disparities along racial lines still exist. Even though the lives of African and Coloured people have generally improved, returns to education remain much higher for white people, who earn 40% more than Africans and 20% more than Coloured individuals per added year of schooling (Salisbury, 2016).

The Shine Literacy program offers an important, high-quality input to resource-poor schools with the goal of addressing some of these educational shortfalls and thus redress some of the injustices of the past early on. Shine Literacy is a South African non-profit organisation (NPO) that provides reading assistance to learners who are identified to be at literacy risk by the end of Grade 1. Assistance takes place during Grade 2 in the form of bi-weekly one-on-one sessions between volunteers and at-risk learners, teaching phonetics, word recognition and decoding, and the shared reading of whole paragraphs. If the learner does not improve sufficiently after the completion of Grade 2, they will be taken on for another year during Grade 3. Shine's program is embedded in the mutual understanding between government, business and NPOs that South Africa's literacy problem requires strong collaboration between a variety of actors, calling for a holistic approach. Shine Literacy aims to ensure that learners are equipped with their developmental age/grade adequate literacy level. Their ethos makes it clear that early interventions imply large gains in terms of academic self-esteem in later years, keeping the goal of promoting a culture of learning in mind. This mission follows the notion emphasized in the work on the returns to education by Heckman (2006; 2011) and Heckman & Masterov (2004), namely that the earlier the intervention, the more persistent and socially just the returns will be. For instance, Heckman & Masterov (2004) estimated that the rate of return of an early intervention program designed for low-ability children is 16%.

[3] Du Plessis et al. (2003) argue that the earlier the reading literacy intervention is implemented the more effective is the outcome; they stress that there should be intensive involvement in informal literacy interventions already at preschool age, with the circumstantial evidence in mind that young children have the ability to learn already complex subject matters and are generally keen and eager to learn.

[4] According to the *Education Statistics in 2014* report and with specific reference to the Western Cape, only 6.2% of Grade 9 learners achieve a score of 50% or higher in Mathematics, the average score being 13% (Department of Basic Education [DBE], 2016). Similarly, the average score for first additional language (which in most cases is English) is 38%, with only 24.5% scoring 50% or higher – slightly better than Mathematics performance; however, still dramatically sub-par. With respect to overall enrolment figures, only 50% of learners that register at the foundation phase (Grade 1 – 3) will eventually reach Grade 12, with only 12% being equipped with a qualification that renders them eligible to enter university (McCarthy & Oliphant, 2013). Undeniably, at an individual level this has drastic consequences – ultimately implying a lack of agency: there is essentially no freedom to choose which vocational path to take and no freedom to acquire the knowledge that the learner rightfully deserves.

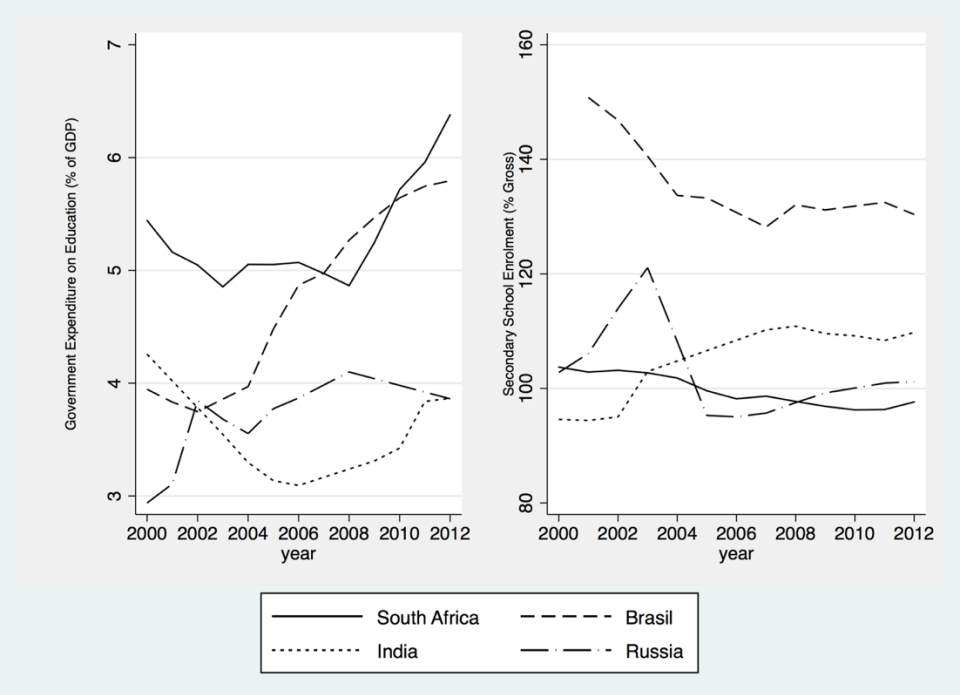
This dissertation evaluates the impact of Shine Literacy in using difference-in-differences and propensity score matching approach. The analysis can be placed amid the international discussion on which inputs to education are required to overcome the persistent lack of quality. With a potential lookout for externality effects, this study did not only aim to measure the programme impact on learners' literacy, but also provide some evidence on why Shine Literacy should be scaled up across South Africa. Large standard deviation literacy improvements of up to 1.9 were found, with the most conservative estimate starting from 0.6 standard deviations. This amounts to educationally meaningful results. Further, it appeared that those who were most deficient with regards to their literacy and thereby later schooling were assisted the most, and experienced the greatest literacy improvements. No heterogeneous effects on gender were found when part of the programme, boy learners do just as well as girl learners – this is an important insight into closing the gender gap that exists in literacy. Lastly, better literacy had predictive power on numerical ability as assessed in the Systemic Mathematics tests – Shine learners did better by 6.3%.

By way of introduction, Chapter 1 contextualizes the evaluation of Shine Literacy against the international evidence on the effect of education interventions, including reading literacy programmes. Chapter 2 discusses the available data and methods used to evaluate the programme. This chapter will also discuss the results' contextual implications. By way of conclusion, the discussion section aims to provide some linkages to the international evidence and deliver some remarks on the analysis' limitations but also future work.

Chapter 1: Re-examining the black box of education

Though the literature on the economics of education is vast, little is known about how educational output is actually produced. Yet, any government, international donor agency and most economists will agree that a well-functioning education sector is one of the most vital components of economic growth and development (Bowles, 1970; Kremer, 2003; Heckman, 2011; Glewwe & Muralidharan, 2016). Large amounts of public expenditure are funnelled into education, especially in the developing world. In South Africa, public spending on education is an increasing function of time. As can be seen in Figure 1, in 2014 South Africa's spending on education was around 6% of its GDP – in comparison, two other BRICS nations, Russia and India, spent under 4% (World Bank, 2018). Still, gross primary school enrolment remains lower in South Africa than in its comparison countries (or at the very least, is at a similar level to countries who spend far less).

Figure 1: Government expenditure on education and school enrolment in BRI(C)S



Source: World Bank (2018)
Note: No government expenditure data were available for China, so it was not included the dataset used to compute this graph.

Nonetheless, South Africa's increasing financial commitment to improving education has paid off in terms of enrolment: since the early 1990s, more and more children attend primary schools with the proportion of eligible learners that are enrolled peaking at 95.8% in 2004 (World Bank, 2018). Yet, while much has been achieved in terms of universal primary schooling – an ideal further supported by the Millennium Development Goals – it appears that learners are in school but are not learning (Pretorius & Spaull, 2016). Primary school completion is increasing, but lower secondary school completion rates remain on a sub-optimal trend with 77.2% during 2014 (World Bank, 2018). Even though more and more children are now able to go to school, it appears that many either do not complete it, or do not learn much.

If one models schools as firms that use various inputs to produce education output, the puzzle of more and more spending yet no increase in educational attainment could either be seen as spending that is directed at the wrong inputs; or as a lacking understanding of the process of *how* inputs get utilised and applied to produce educational outputs (Bowles, 1970). Unpacking this technology and thereby discovering what the most effective inputs are is the aim of estimating the structural parameters of the production function, which is often specified as consisting of school and family characteristics, as well as student innate ability (Bowles, 1970).

Given the evidence of a strong relationship between labour market prospects and school achievement well embedded in the literature[5], proxies for individual economic potential in the labour market are often based on indices of school outputs, which in turn are based on test score achievements (Bowles, 1970). Therefore, test scores are used as robust indicators of educational output. The production of test scores - a structural relationship – is then depicted as per Bowles (1970):

$$S=f\,(X_a, \ldots X_k, X_l \ldots, X_r, X_s \ldots X_z)$$

with

S = an indicator for schooling output, such as test scores.

$X_a, \ldots X_k$ = measuring the schooling environment, such as the amount and quality of resources available at school or the level of teacher education.

$X_l \ldots, X_r$ = measuring external influences that are unrelated to the schooling environment, accounting for the effect of the learner's home environment.

$X_s \ldots X_z$= measuring the learner's innate ability.

The typical production function arrives at estimates of average population parameters. But, because of heterogeneity across the learner population, the ideal policy intervention is expected to be different across discrete points of the learner distribution[6] (Glewwe & Muralidharan, 2016). Glewwe & Muralidharan (2016) highlight the now stylized dilemma of increased access yet only little acquisition of human capital in many developing countries. Learners simply do not learn what is specified in the curricula and then fall behind. Fortunately, research into estimating the structural parameters of the education production function has increased dramatically over the last decade. This movement is further accelerated by the increased application of randomised evaluations to measure interventions' impact, as well as an ever-improving access to survey and administrative datasets (Glewwe & Muralidharan, 2016). This has helped tremendously in unpacking the black box problem repeatedly encountered in the discourse on the production of education.

[5] See for instance Heckman et al. (2016).

[6] For the South African case, Spaull (2012) argues that there are essentially two different education systems that ultimately lead to two different data generating processes. As such, the production process that uses educational inputs is profoundly different for the two systems, and when aggregating the two systems, spurious results and misleading policy conclusions will be the outcome of the econometric analysis.

In summarizing the most recent evidence, Glewwe & Muralidharan (2016) categorize the technology components behind the production process as belonging to intervention mechanisms that pertain to: **demand** (with focus on the decision-making process in the household); **supply** (including teacher quantity/quality and the provision of school resources such as textbooks, food or healthcare); improved school **governance**; or **pedagogy** (which changes the technology of instruction). Though the focus of this dissertation is set on the impact of pedagogy in the education production process, the following section will briefly summarise the available evidence on demand- and supply-side programmes, as well as governance interventions.

Managing education demand: conditional cash grants can buy more education output

Policies that aim at boosting enrolment and attendance typically do so by trying to increase the household's immediate benefits of sending a child to school: households often make sub-optimal decisions when it comes to their children's education (Glewwe & Muralidharan, 2016). This has various reasons, including imperfect information on the returns to education, credit-constraints, high risk-aversion or discount rates, or simply not knowing about education's social externality effects (Glewwe & Muralidharan, 2016). By factoring in this sub-optimal decision making, demand-side interventions aim to decrease the cost of school attendance. Interventions are either information-based, cash-transfers or scholarships.

In general, consequences of demand-side programmes are consistent with the theory underpinning their design. This means that programmes that intend to increase the direct (perceived) benefit of being enrolled or attending classes have been very effective. Most notably, the receipt of conditional[7] cash grants has proven incredibly promising in raising attendance. The PROGRESA program in Mexico has received particularly much attention. Enrolment among those that receive money based on their attendance had increased by 3.4% among Grade 1 to Grade 8 learners, with the largest increase experienced by Grade 6 girls[8] (Schultz, 2004). Positive effects of conditional grants did not only materialize in the Mexican setting, but can actually be observed throughout the developing world. Large effects were found in Malawi (Baird et al., 2011), China (Mo et al., 2013a), Columbia (Barrera-Osorio et al., 2011), Nicaragua (Gitter & Barham, 2008), Brazil (Glewwe & Kassouf, 2012), Honduras (Galiani & McEwan, 2013), and Cambodia (Barrera-Osorio & Filmer, 2016). These effects vary in size. For example, in China dropout rates reduced by 8% (Mo et al., 2013a) versus, in Brazil where dropouts decreased between 3-9.6% depending on the income-area (Glewwe & Kassouf, 2012).

Results on cash grants' effectiveness have indicated mixed implications for test score performance. In the Malawian programme, Baird et al. (2011) found an increase in English and Mathematics test scores of 0.14

[7] The evidence is quite clear on the necessary prerequisite of conditionality; in turn, two unconditional cash transfers, evaluated by Baird et al. (2011) (in Malawi) and by Benhassine et al. (2015) (in Morocco) showed much smaller effects.
[8] The programme had been expanded and in following up with the programme's long-term effect, Behrman et al. (2011) confirmed benefits on not only schooling, but also a positive shift away from agricultural to non-agricultural work among beneficiaries post-schooling.

and 0.12 standard deviations, respectively. Boys that participated in the Nicaraguan programme scored 0.2 standard deviations higher than boys that did not receive the grant (Barham et al., 2013). However, test scores changes in Colombia were either insignificant (Mathematics) or negative and significant (English) (Baez & Camacho, 2011). Similarly, the studies by Mo et al. (2013) (China), Benhassine et al. (2015) (Morocco) and Barrera-Osorio & Filmer (2016) (Cambodia) showed no statistically significant effect on test scores.

In addition to income effects, high present bias may also affect education choices: learners may know about the returns to schooling yet still make sub-optimal choices. Thus, by making returns to effort more immediate, merit-based scholarships overcome this bias and reward students more instantaneously. Their effectiveness has been shown to work in practice with increases in test score performance ranging between 0.2 and 0.27 standard deviations depending on the setting of the study and the respective mechanics of the scholarship[9]. Kremer et al. (2009) measured the impact of a Kenyan scholarship programme that promised the top 15% of Grade 6 girls an aggregate amount of $19.20 (to be split between learner and their parents). The authors found that this method increased participation by 3.2% and test scores by 0.27 standard deviations. In a follow-up study, Friedman et al. (2011) measured the long term impact five years post-intervention and found an even larger impact on enrolment (8.6%). There was, however, no statistically significant effect on completed grades. Nonetheless, Friedman et al. (2011) were able to confirm a long-term effect on test scores: after 5 years, scores were still 0.2 standard deviations higher among programme participants.

Evidence on other demand-side interventions, including information-based campaigns, is less clear-cut. Here, the idea is that due to imperfect information in the market of education, individuals make sub-optimal decisions when it comes to opting for a level of education. Yet, improving this information in practice does not seem to effectively reach the most targeted beneficiaries and often leads to the undesirable results. An example of this can be found in a Dominican Republic programme evaluated by Jensen (2010), which provided information on returns to staying in school to Grade 8 boys from low-income background. The authors found that boys who received this information were 4.1% more likely to stay in school for an additional year. Even four years post-intervention, participants completed 0.2 more years at school. However, this trend was most strongly visible for the least poor and the poorest of the poor did not mirror these results (Jensen, 2010). In terms of undesired results, Loyalka et al. (2013) found that information on education and career counselling to Grade 7 learners in China had little to no effect on the dropout rates, and actually caused a decrease in the time spent at school. The authors suggest that this was because learners

[9] Blimpo (2014) evaluated a programme in Benin that offered scholarships either based on individual performance, group average performance or school performance (in a tournament against other schools). They found that all treatment arms performed similarly, increasing scores by around 0.24 standard deviations. Another scholarship programme in China was based on tournament only, evaluated by Li et al. (2014). In one treatment group low and high-achieving students were randomly paired, and scholarships were based on both students' improvements. The high-achiever was ought to peer-tutor the low-achiever. The other treatment group was based on individual performance among low-achieving students only. Li et al. (2014) showed that individually-based performance scholarships implied no changes in test scores; however, the matching of the low and high-achievers improved the latter's' test scores by 0.27 standard deviations.

did not know about the relatively hard entry requirements to secondary schooling. In a different Chinese intervention, Wang et al. (2014) found that the impact of school counselling to reduce anxiety among learners decreased dropout rates by 2% during the first six months, but this effect had entirely disappeared by the end of the school year.

Lastly, methods that indirectly increase demand for education, target explicit resource deficiencies at the household level. However, evidence remains ambiguous, potentially because the range of contexts and their external validity remains quite limited[10].

Managing education supply: indirect inputs are effective

On supply-side interventions, educational input programmes are wide-ranging. Evaluations include measuring the impact of increasing the number of schools, materials/facilities and teacher quantity. In summarizing cross-country evidence, programmes that increase school access by simply building more schools reduce not only indirect costs incurred from longer transportations and safety concerns, but also from foregone working hours. These programmes have large effects on attendance, which ultimately increases test scores. Duflo (2001) estimated the impact of an Indonesian school construction intervention carried out in the 1970s and found that additional schools increased boys' school attendance by 0.19 years. Similarly, in Mozambique, each additional school successfully increased learners' probability of being enrolled by 0.3% as estimated by (Handa, 2002). In Pakistan, Alderman et al. (2003) found that additional funding for more single-sex girl schools, increased urban girls enrolment by 25%, compared to 15% increases in girl enrolment in rural areas. Additional primary schools in rural Afghanistan as estimated by Burde & Linden (2013) increased girls' enrolment by 51.1% and test scores increased by 0.66 and 0.41 standard deviations for girls and boys, respectively. Further, a programme that built girl-friendly schools in Burkina Faso, increased their enrolment by 21.9% as estimated by Kazianga et al. (2013). Boys' enrolment increased by 16.3% and all learners' test scores improved on average by 0.41 standard deviations. Lastly, Bellei (2009) and Orkin (2013) find that in Chile and Ethiopia, simply keeping schools open for longer also increased learning outcomes.

The impact of other supply-side inputs is inconclusive both with respect to attendance and test score performance. For example, textbook provision to Kenyan Grade 8 learners increased their probability of completing school. However, the lower grades which were more at risk of drop-outs, and thereby the more important beneficiaries were unaffected (Glewwe et al., 2009). Further, Glewwe et al. (2009) found no effect on test scores for the average learner and demonstrated that positive effects occurred only for the high-achievers. Glewwe et al. (2009) argued that this is indicative of the ineffectiveness of the curriculum that is too hard for the low-performing learners, who then get left behind. Moreover, another textbook

[10] Oster & Thornton (2011) interrogated the impact of a female sanitary intervention conducted in Nepal, but found no effect on girls' attendance. Muralidharan & Prakash (2013) who measured the effect of a bicycle grant programme offered to families with girls carried out in India, estimated that it proved successful in raising enrolment by 5.2%.

programme evaluated by Sabarwal et al. (2014) did not increase attendance nor learning in Sierra Leone. The reason for this was not the textbook content; instead, textbooks never reached any of the students. As argued by Sabarwal et al. (2014), administrators of programme schools were uncertain about the continuous nature of textbook provision after the intervention had ended, and resorted to storing the books instead of distributing them. On uniforms, Hidalgo et al. (2013) found an unexpected 2% reduction in learner attendance after learners were given uniforms in an Ecuador programme. Glewwe et al. (2004) estimated the effect of flipcharts provided to programme schools in Kenya but found no conclusive effects on student learning. Lastly, the provision of school libraries in treatment schools in India evaluated by Borkum et al. (2012) had no effect on attendance nor learning.

Due to variations in education and training, evidence on the impact of teacher quality is lacking and most studies are merely able to account for a change in teacher quantity. The latter effectively changes pupil-teacher ratios, yet even here the evidence is not as straightforward as one might expect. Simply put, employing more teachers does not necessarily translate into higher test scores[11] (Chin, 2005; Urquiola, 2006; Duflo et al., 2015). More promisingly, interventions that target indirect inputs such as school meals prove to be very effective in most cases, both in terms of raising test scores and keeping programme costs low[12] (Tan et al. 1999; Adrogue & Orlicki, 2013; Kazianga et al. 2013).

Regarding healthcare inputs, some medical programmes evaluated provide deworming tablets, which have shown to be extremely promising in raising attendance and performance; and even in later schooling and labour market outcomes (Glewwe & Muralidharan, 2016). In much of the developing world, the incidences of worm infections are large. Thus, given the detrimental effect of worm infection on attentiveness and concentration, the provision of deworming tablets has been found to be a valuable and cost-effective way to increase enrolment/attendance and test scores (Glewwe & Muralidharan, 2016). For example, Miguel & Kremer (2004) assessed the impact of a deworming programme implemented in Kenya and found that it reduced absenteeism by 7-8 %. In a follow-up study, Ozier (2016) specifically interrogated the spillover effects of the programme and found that those children that at programme inception were 1 year olds captured large test score gains, corresponding to between 0.5 and 0.8 years of schooling. To assess additional long-term effects, Baird et al. (2015) looked at the impact of the Kenyan deworming programme ten years post-intervention and found that women who had received treatment were 25% more likely to be

[11] Chin (2005) evaluated an Indian programme that provided additional teachers and materials and showed that it successfully improved learners' school completion by 1 – 2 %. Yet, it is not possible to crystalize which input (teacher or material) actually caused this. Moreover, the impact of additional teachers on test scores is also ambiguous: in Bolivia, Urquiola (2006) found that decreased ratios have a negative effect on language test scores; in Chile, Urquiola & Verhoogen (2009) found negative effects both on language and math tests. Lastly, a programme that assigned additional contract teachers in Kenya was evaluated by Duflo et al. (2012; 2015). Regarding pupil-teacher ratios, the authors found no significant differences in test scores among learners who were taught by civil-service teachers in the treatment (which accounted for reduced class sizes) versus those in the control groups.

[12] Adrogue & Orlicki (2013) found that in an Argentinean school meal programme, treatment learners did statistically significantly better in language (0.17 standard deviations). In the Philippines, Tan et al. (1999) evaluated a school meal programme that proved successful in increasing Grade 1 learners' test scores by 0.25 and 0.16 standard deviations for math and language, respectively. In addition, Tan et al. (1999) found that a programme that combined school meal provision and a parent-teacher partnership did increase learning outcomes; however, it is impossible to disentangle which component of the programme had this effect. Further, in Burkina Faso, providing school meals increased math scores by 0.1 standard deviations, as estimated by Kazianga *et al.* (2013). In contrast, Adrogue & Orlicki (2013) found that take-home meals had no impact on student enrolment nor learning.

enrolled at secondary school. Further, Baird et al. (2015) estimated large gains in the labour market and projected that the return on investment was over 32%. Other than the deworming example, the evidence on other medical inputs is less clear[13].

Another group of supply-side programmes provide either large funds[14] or allow learners to attend so-called elite schools[15], which are equipped with better materials or have more qualified teachers. Evidence on this is very mixed and does not results into a clear directive.

Education governance: making salaries subject to test score performance

As a third decisive component of the education production function is the way in which the education system as whole is organized and monitored. Measuring education governance (or the lack thereof) entails identifying the corruption and capturing of funds and factoring in the realities of high teacher absenteeism (Glewwe & Muralidharan, 2016). For example, Reinikka & Svensson (2004) found that in Uganda 87% of public funds tagged as school grants never reached any of the beneficiaries. Further, Chaudhury et al. (2007) found that 25% of teachers in India and 27% of teachers in Uganda were absent during unannounced school visitations. In India, the cost of these levels of absenteeism is estimated at an annual $1.5 billion (Muralidharan et al., 2017).

Improving governance in the education setting is not an easy endeavour. As it comes to the scaling of programmes that tackle issues of mismanagement, there is little political will to tap into the benefits of decentralized decision-making[16]. Some approaches, like top-down teacher monitoring, are not too promising: Duflo et al. (2012) and Banerjee et al. (2010) found insignificant impact on students' time in school[17]. In terms of bottom-up monitoring, the results are not very assuring either: Banerjee et al. (2010), who evaluated a programme that measured community-level involvement on increasing teacher attendance

[13] A Chinese health insurance programme evaluated by Chen & Jin (2012) had no impact on enrolment rates, and the effect of the provision of iron supplements in the same country estimated by Luo et al. (2012) and by Sylvia et al. (2013) show no clear picture on the impact thereof on test scores. Promisingly, however, providing eyeglasses in the rural areas of China increased test scores of primary school learners by 0.16 standard deviations, as estimated by Glewwe et al. (2016).

[14] A programme that provided large amounts of money to Bolivian treatment schools had no impact on dropout rates as found by Newman et al. (2002). In contrast, an intervention that offered students that dropped out remedial education and mobilized their networks was evaluated by Pridmore & Jere (2011) who found that it did reduce dropout rates. However, which component of the programme actually prevented dropouts is unclear. Another way to increase provided resources is to give schools grants to be spent on educational inputs. One such programme was evaluated in India by Das et al. (2013) who found that in the first year of the grants (approximately $3 per student) test scores did improve by 0.09 standard deviations; however, in the second year, as households expected the grant the effect completely diminished. A programme that directly provided village councils in Indonesia with grants, where the continuation thereof was either conditional or unconditional on health and education outcomes had no impact on learning outcomes (Olken et al., 2014).

[15] In Romania, Pop-Eleches & Urquiola (2013) find that learners who attend an elite school did perform better in their graduation exams, however, which of the many aspects of a "better" school actually caused this tendency is very unclear. Adversely, a programme in Kenya that allowed some learners to attend elite schools had no effect on learning outcomes (Lucas & Mbiti, 2014).

[16] Glewwe & Muralidharan (2016) note how this somewhat puzzling results may stem from asymmetric power relations between schools and their respective communities, as well as collective action problems in terms of coordinating monitoring activities, particularly in areas of disadvantaged communities.

[17] Similarly, effects of such monitoring on student educational outcomes are not clear-cut either. In schools in India, Duflo et al. (2012) used cameras to monitor teachers and paid teachers according to the time they were actually in school; this improved test scores by 0.17 standard deviations and decreased absenteeism by half. Yet, it is not clear whether the incentives linked to being present, or the actual monitoring caused this effect. A low-stake version of monitoring including feedback sessions had no impact in schools evaluated by Muralidharan & Sundararaman (2010), who argued that consequences of being absent need to be high-stake in order for teachers to comply.

in rural India, found no significant impact[18]. Nonetheless, programmes that aim to improve teacher effort and their accountability have shown to improve learning outcomes (Glewwe & Muralidharan, 2016).

Generally, as the well-known dilemma of principal-agent relations predicts, productivity and effort are difficult to measure. Nonetheless, performance indicators such as test scores can be credible measures of effort (Glewwe & Muralidharan, 2016). In the same vein, the credible threat of losing employment contingent on low performance can create sound commitment to exert high effort. The evidence on making teacher salaries subject to test performance is pretty clear: the payment formula that sets salaries based on objective indicators such as learner attendance or their test scores increases effort and thereby educational outcomes. Test score improvements range from 0.17 (Muralidharan & Sundararaman, 2011) to 0.54 (Muralidharan, 2012), standard deviations subject to the setting of the intervention. For example, Muralidharan & Sundararaman (2011) evaluated an Indian programme that provided additional payments based on the performance of the teacher's class and found that learners in treatment schools scored 0.27 and 0.17 standard deviations higher for Mathematics and Language, respectively. Students in treatment classes also did much better in Science and Social Studies, which points to positive externalities of the programme in other subjects. Interestingly, Muralidharan & Sundararaman (2011) were able to distinguish between the effects of either resource inputs or performance-based payments. The authors found that the latter had a stronger positive effect than resources that were of the same cost as teacher salaries. Muralidharan (2012) conducted a follow-up study of this programme and found that five years post-intervention treatment students still had higher test scores (0.54 and 0.35 standard deviations for Mathematics and Language, respectively).

Contreras & Rau (2012) found that a Chilean nation-wide programme that delivered bonuses to teachers whose students did well improved learners' performance by 0.29 standard deviations in Mathematics. Lastly, a Kenyan tournament-type programme where high-performing schools were given prizes based on highest average scores was evaluated by Glewwe et al. (2010). The authors found that it increased teacher preparation effort but had no impact on teacher absenteeism and only improved learner performance in high-stake tests. Actually, this effect diminished post-intervention, suggesting that those types of tournaments had no impact on long-run outcomes. Glewwe et al. (2010) argue that due to the school-level nature of this tournament, many teachers may have been free-riding and thus complicate collective action.

The positive impact of additional contract-teachers may stem from either reduced pupil-teacher ratios or the contract-teachers' performance based on contingency rewards. As such, Duflo et al. (2015) were able to account for either effects in a setting in Kenya, where additional contract-teachers were randomly allocated to schools, which decreased the number of pupils per teacher. The latter mechanism did not

[18] Similarly, an intensive information campaign that informed parents about participating in bottom-up monitoring in several Indian states improved teacher attendance, yet had insignificant impact on educational outcomes (Pandey et al., 2009).

explain higher test scores (0.29 standard deviations); however, being taught specifically by a contract-teacher was statistically significant in explaining this effect; a trend that is evident in other settings, as well[19].

To wrap up the review on education governance, evidence on private schools shows that they may be more *efficient* at delivering inputs to learners than public schools (in terms of time and costs), yet they do not necessarily do better in improving test scores (Glewwe & Muralidharan, 2016). Given the inability of most low-income families in the developing world to afford these schools, proponents argue for school vouchers. Effects of voucher programmes are statistically insignificant when it comes to improving enrolment or time spent in school; still, there are positive effects documented with test score improvements ranging between 0.16 (Angrist et al., 2002) and 0.2 standard deviations (Angrist et al., 2006). In Colombia, recipients of private school vouchers showed a test score improvement of 0.16 standard deviations three years post-intervention (Angrist et al., 2002). A long-run follow-up of this programme conducted by Angrist et al. (2006) found that seven years post-treatment, voucher recipients were 5.6% more likely to graduate from high-school. In addition, participants scored 0.2 standard deviations higher in college entrance-exams. Lastly, Lara et al. (2011) and Hsieh & Urquiola (2006) estimated the impact of private schooling on test score performance in Chile. Neither of the studies found a statistically significant effect on Language or Mathematics tests in secondary private schools for Grade 10 (Lara et al., 2011), or Grade 4 and 8 (Hsieh & Urquiola, 2006) learners.

Pedagogy: teaching at the right level with trained volunteers

The quest behind finding the "right" pedagogy stems from the puzzling result that mere input provision does not effectively increase attendance or performance (as seen, for instance, in the textbook programme evaluated by Glewwe et al., (2009). In other words, understanding the best way of how to use those inputs (that is, the *technology of instruction*) needs to be emphasized. Given largely heterogeneous student bodies in developing countries, it is indeed surprising that instructional methods do not account for different levels of abilities and instead, teachers are often found "teaching to the top" (Glewwe & Muralidharan, 2016: 706). This status quo stems from systems that are designed to teach to the top end of the distribution instead of increasing the human capital of all learners, of which many are first-generation learners (Banerjee et al., 2007). In changing these archaic dynamics, several interventions have been set in place to find the "right" pedagogical approach. These interventions include the use of Information and Communication Technology in the classroom, and the application of more in-depth, remedial instruction. The latter consists of classroom tracking, "teaching at the right level" and – with specific relevance for the focus of this dissertation – reading literacy interventions. In general, studies have shown that those interventions that ensure the acquisition of basic skills as well as "teaching at the right level" approaches seem to be particularly promising (Glewwe & Muralidharan, 2016).

[19] Muralidharan & Sundararaman (2013) found that two years after randomly allocating additional contract-based teachers in India, treatment schools' performance in Mathematics and Language improved by 0.16 and 0.15 standard deviations, respectively. In addition, the authors discussed how absenteeism was a lot lower for contract-teachers who were absent 16% of the time, compared to 27% for civil-service teachers.

Technology optimists have long argued for the benefits of Information and Communication Technology that may also impart in the context of education (Gulek & Demirtas, 2005; Cristia et al., 2014; Kong, 2014). Due to the unchanged nature of teaching styles in many developing country classrooms, many believe that different media of instruction may create so-called "disruptive innovation" (Glewwe & Muralidharan, 2016: 708). This disruption may capitalize on the ability of technology to be tailored to individual instruction in line with the individual learner's ability and preparation; or, due to its interactive nature, technology can more rapidly transmit feedback to learners (Kong, 2014; Glewwe & Muralidharan, 2016). Moreover, using technology in the classroom is of particular interest when considering its potential substitution or complementary effects (Linden, 2008; Beuermann et al., 2013). As such, digital media can be used either to decrease pupil-teacher ratios by taking some children out of the classroom and let them work independently on computers, or it could be used together with conventional teaching and go into the lesson's content in more depth (Beuermann et al., 2013; Cristia et al., 2014; Glewwe & Muralidharan, 2016).

Whilst there is comparatively much evidence assessing the impact of technology in the classroom, null results often occur, or effects disappear quickly entirely. On null results, Cristia et al. (2014) found that increasing computer and internet access had no effect on dropout rates in Peru. Similarly, no impact was found in studies that looked at programmes in Latin America. For instance, Barrera-Osorio & Linden (2009) estimated that a computer programme included in Columbian treatment schools had no effect on learning outcomes. Similarly, a Peruvian home computer programme evaluated by Beuermann et al. (2013) had no impact in terms of test scores. On rapidly diminishing effects, Banerjee et al. (2007) found that a two-hour weekly Mathematics computer programme increased Mathematics test scores by 0.48 standard deviations in India. However, this effect did not persist one year after the programme had ended, where test scores were only 0.1 standard deviations higher in the treatment cohorts. Particularly, it appears that technology alone may not be beneficial if it is not effectively integrated with the conventional method of instruction and curriculum, and that this failure to integrate may in fact lead to negative impacts[20]. Positive test score results include increases in standard deviations from as low as 0.14 found by Lai et al. (2015), up to 0.48 found by Banerjee et al. (2007)[21].

[20] Linden (2008) who found positive effect for the out-of-school version of a computer programme, confirmed a negative impact for the in-school version thereof which decreased test scores by 0.55 standard deviations in treatment groups. Linden (2008) went on and argued that this may not be exclusively due to the technology nature of the programme but instead with teachers being reluctant and/or unable to effectively include technology instruction in their conventional ways of using the curriculum. Similarly, a Romanian programme that gave middle-school learners vouchers to purchase computers had a negative effect on their GPA as estimated by Malamud & Pop-Eleches (2011), who argued that by merely supplying computers, learning outcomes may actually decrease.

[21] An outside of school programme that provided instruction based on computer programmes in India was evaluated by Linden (2008) who estimated positive impact of 0.29 standard deviations for treatment learners one year after being part of the intervention. In China, several studies have found positive impact, varying in size. A programme that provided 40-minute long sessions of remedial math classes for grade 3 migrant children was effective in raising the math scores by 0.14 standard deviations compared to control learners (Lai et al., 2015). Similarly, Yang et al. (2013) find small yet statistically significant effects of computer programmes for learners in various Chinese provinces as it raised scores by 0.12 standard deviations. Lai et al. (2015)went on and evaluated the programme they had assessed for migrant learners in more rural areas, and found that it had increased language and math test performance by 0.2 and 0.22 standard deviations, respectively. Lastly, Mo et al. (2014) measured the impact of a technology-instruction programme in several other rural areas in China and estimated that it led to increases in test scores both for grade 3 and grade 6 learners, by 0.25 and 0.26, respectively.

Due to the heterogeneous nature of classrooms often due to differing levels of preschool preparation or family environments, teachers face difficulty in streamlining their lessons. To reduce this variation, remedial instruction may prove helpful in ensuring that every learner progresses on some learning trajectory, irrespective of initial skill levels (Banerjee et al., 2007; Lai et al., 2015). Remedial instruction programmes often make use of volunteer-based tutors, or teachers with only little formal qualification or training. Improvements in test scores can range from as little as 0.14 (Banerjee et al., 2007) and up to 0.74 (Lakshminarayana et al., 2013) standard deviations. One of the most famous of these promising programmes is the *Balsakhis* programme organized by Pratham: learners who did not acquire basic skills in Mathematics and Reading were taken out of the classroom and schooled instead by a *Balsakhi,* a trained tutor with secondary education (Banerjee et al., 2007). This programme was extensively evaluated by Banerjee et al. (2007) who found that those learners that were trained by a *Balsakhi* benefitted the most. This means, the reduced pupil-teacher ratio did neither positively nor negatively affect the learners who remained in the usual classroom[22].

Another positive finding on low performers came out of a different study by Banerjee et al. (2010). The authors looked at the effect of providing after-school reading camps with minimally trained volunteers. Impact on learners who started from a very low reading ability were large: the proportion that was able to identify letters at the end of the programme increased by 60%, and for the average learners it increased by 7.9% (Banerjee et al., 2010). Similarly, another remedial education programme based on community volunteers in India proved also very promising. Lakshminarayana et al. (2013) who evaluated this programme found that learners' test score improvements remained 0.74 standards deviations higher even two years after the completion of the programme. Lastly, an out-of-school "teaching at the right level" approach used in a targeted instruction programme in India resulted into increases of up to 0.7 standard deviations in Language test scores (Banerjee et al., 2016).

Literacy programmes: adjusting teaching techniques and involving families

Certainly, reading literacy programs differ a great deal in terms of logistics, inputs and intervention dimensions. To gain more insights and build a model specification based on the available evidence, the following discusses a selection of reviewed studies that evaluate reading literacy programmes. To warrant a wider assessment of the relevant publications, the review includes studies from both developing and developed countries. Some of the dimensions considered can be observed in Table 1, which also provides some remarks on key findings made.

[22] Though this programme did not have no effect on time spent at school, it did increase every learners' test performance by 0.14 standard deviations one year and by 0.28 standard deviations two years' post-intervention on average (Banerjee et al., 2007).

Table 1: Amalgamation of literacy reading evaluations

Author(s)	*Date*	*Location*	*n*	*Age group*	*Method of intervention*	*Findings*
Brooks et al.	2006	U.K.	155	11 – 12	ICT	ICT not effective as a method of delivering spelling. ICT significantly decreased reading (~ 0.5 SDs)
Justice et al	2009	US	106	3 – 5	Read-alouds	Use of print referencing effective (~0.5 SDs) for print concept and alphabet knowledge, but not for name-writing abilities.
Kim & Guryan	2010	US	370	10	Family literacy event + books handout	Reported reading has no effect on comprehension/vocabulary. Strongest predictor is vocabulary score and English proficiency number of books read predicted post-test scores in comprehension but not in vocabulary Learners lack decoding and fluency skills; do not benefit from programme, neither do their parents.
Piasta et al.	2012	US	550	4	Shared-reading and increased print referencing	Significant impact on children's early literacy skills (reading, spelling, comprehension) for two years post-intervention; 0.26 – 0.31
Abebere et al.	2014	Philippines	5510	10	Reading marathon	students' reading increased by 0.13 SDs encourages students to read more at home No evidence of spill over to other subjects effect size: 0.13 SDs (0.06 SDs after three months)
Banerjee et al.	2007	India		9 - 10	*Balsakhis* programme	0.28 SDs improvement bottom third benefits the most
He et .	2009	India		3 - 5	The *Shishuvachan* classes, and the child library	Most effective in the originally designed out-of-school preschool classes for which it was originally designed, particularly for low-performing students; Most robust gains when run as a complement vs. a stand-alone, 0.12 - 0.7 SDs
Steensel et al.	2011	International		3 - 10	Meta-Analysis	Family literacy events effective: d=0.18 (general literacy ability), d=0.22 (comprehension) and d=0.17 (code)

Evidence from developed countries: digital media cannot replace shared reading time

The first study considered is Brooks et al. (2006) who evaluated the effect of using Information and Communication Technology (ICT) on literacy achievement in a secondary school (Grade 7) in the UK. Brooks et al. (2006) assessed whether an ICT intervention proves effective in a conventional classroom environment. Further, the evaluation looked specifically at how the use of computer software impacts learners' progress in spelling and reading. The only controls used were the learners' baseline test scores, age and gender. As there were not sufficient laptops available for the entire cohort of learners, six learners at a time were able to use the computer programme. Eventually, all learners were exposed to the conventional teaching programme that did include the use of the regular ICT methods already installed at the school and English lessons, as well as the additional ICT treatment. Those that were randomised to be on the waiting list, received the usual literacy teaching, whereas the treatment group was exposed to the ICT programme that was designed to increase their phonological awareness[23]. The authors took assessment data three times:

[23] Phonological awareness is the ability to detect and process components of spoken language in auditory form; as well as the skill of distinguishing between syllables, phonemes or whole spoken words (The National Institute for Literacy, 2008).

to account for balancing concerns after randomizing assignment, and to measure whether there are any medium and/or long-term effects.

As an interesting aspect, Brooks et al. (2006) highlight that during the duration of the treatment, learners take part in place of their regular classes; which implies that some time is given up from their English class to take part in the intervention. Effectively, the software programme intervention was thus evaluated as a substitute to normal English classes, not as a complement. Brooks et al. (2006) conclude with the finding that there were no significant benefits of using ICT when it comes to spelling and reading. In fact, they found that the use of software had a statistically significantly negative effect on reading– there was no effect in either direction for spelling. Even though the study's main limitation is that only one environment and only one particular literacy software was considered, Brooks et al. (2006) highlight the need of using randomised methods to evaluate costly educational interventions before scaling them up.

With regards to different teaching styles that could be easily applied in a reading class, Justice et al. (2009) measured the impact of using print referencing[24] methods in preschools in the US. In the study by Justice et al. (2009), pre-schoolers were allocated to a treatment group where the classroom teacher applied print referencing during read-alouds. Study subjects were children who were at risk in terms of learning development due to the poverty exposure at home. Justice et al. (2009) assessed the pre-schoolers' general print knowledge and awareness which included their understanding of print, word segments and the alphabet, as well as the ability to write their own name. The impact of the programme on learner language growth[25] was also evaluated. The *only* aspect that differed between control and treatment groups was the teacher's intentional inclusion of verbal and non-verbal behaviours related to print referencing during the reading lessons[26]. The analysis showed that treatment participants developed higher knowledge scores in print and alphabet, but no statistically significant effects in their name-writing ability. Similarly, language growth did not differ significantly between the treatment and control pre-schoolers. Justice et al. (2009) effect sizes are in the medium range: print and alphabet knowledge increased by 0.5 and 0.56 standard deviations, respectively, and name-writing abilities increased by 0.42 standard deviations. As the performance in these areas is typically linked to later school outcomes[27] in reading and spelling, Justice et al. (2009) conclude that the mere adaptation of print referencing by teachers proved to be an effective value-added in boosting pre-schoolers' early literacy.

Another US study specifically looking at the impact of using print referencing during shared reading was carried out by Piasta et al. (2012). In their longitudinal study, Piasta et al. (2012) observed the differences

[24] Print referencing implies using techniques that specifically heighten the learner's focus on print during read-alouds or paired reading; this can include asking direct questions regarding print following text whilst conducting read-alouds.
[25] This was assessed by their ability to structure sentences and words; have expressive vocabulary; and core language skills (Justice et al., 2009).
[26] In both groups, teachers were requested to read an assigned book to their class at the beginning of a week, and then re-read that particular book in the remaining three sessions in that week.
[27] A thorough analysis of the predictive validity of early literacy on later school outcomes can be found in Chapter 2 of the Report of the National Early Literacy Panel (The National Institute for Literacy, 2008).

between a high-dose print referencing group and a low-dose group. Over three years, Piasta et al. (2012) measured phonological awareness and alphabet knowledge outcomes; and took performance scores in reading, spelling and comprehension one and two years post-intervention. The authors found that print referencing had a positive effect on phonological awareness and alphabet knowledge immediately post-intervention, but also that later skills in reading, spelling and comprehension were enhanced one and two years after the implementation of the programme. Effect sizes varied between 0.26 and 0.31, subject to the skill assessed. The causality here is between print referencing and its effect on increasing attention and time allotted to different types and purposes of books: the learner is allowed more time to process information about encoding various forms of print and instantaneous learning is taking place (Piasta et al., 2012).

As is highlighted in the work by Heckman (2006) parents and the general home environment play an important– if not *the* most important - role in the development of children's education. Consequently, many literacy programmes also include a parent-child interaction. For example, Kim & Guryan (2010) evaluated a reading intervention programme that took place over the summer school break in the US, specifically targeting language minority learners in Grade 4. Using the summer break as the time for a reading intervention is because that is when those learners tend to fall behind: the long break can create a gap of up to three months in reading test scores between middle and low-income background learners (Kim & Guryan, 2010). Some of the reasons behind this trend is that low-income families simply own fewer books, spend less time discussing books or engaging in reading activities (Kim & Guryan, 2010). As early language skills have to continuously be practiced, a summer break reading programme is meant to prevent reading skills to diminish. Two treatment versions of the programme were assessed. In one group, learners were given ten books to read during vacation; in the other group, learners were given books and their parents were invited to three literacy workshops which informed them on the practice and benefits of paired reading (Kim & Guryan, 2010). As such, silent reading had been shown to be rather ineffective, and the programme aimed to capitalize on the evidence that paired reading between child and parent has been found to increase reading fluency and comprehension (Kim & Guryan, 2010).

Even though learners in both treatments reported to have read more books during vacation, Kim & Guryan (2010) found that the number of books read had some prediction on comprehension, but not in vocabulary – implying that giving the opportunity to read more books may increase actual reading activity, but not necessarily develop reading skill. In fact, post-intervention reading scores were most strongly predicted by the individual learners' English proficiency. Kim & Guryan (2010) conclude that language minority learners often lack decoding skills and fluency, and thus cannot reap programme benefits. Thus, given the mismatch between learner and assigned book (that is, their reading ability did not match with the readability of the book), the programme was ineffective and additional scaffolding techniques provided by parents became pointless.

Evidence from developing countries: trained reading volunteers are effective

In a developing country context, Abeberese et al. (2014) measured the impact of the Philippines *Sa Aklat Sisikat* programme. The treatment included a 31-day reading marathon to see whether an intense short-run reading programme with trained facilitators had any impact on learners' reading skills. The marathon's participants were encouraged to read as many books as they could - this was tracked by using a publicly viewable chart[28]. Abeberese et al. (2014) found that this short-run reading marathon improved learners reading skills by 0.13 standard deviations immediately after the marathon, and by 0.06 standard deviations after 7 months, at the end of the academic term. The number of books read increased on average by 7.2 and there was a differential impact by baseline score: the effect of the programme was an increasing function of the pre-intervention test score (Abeberese et al., 2014). This implies that stronger performing students were simply better enabled to read independently at home, and thereby read more. Moreover, Abeberese et al. (2014) did not find any reading externality effects on other subjects, such as Mathematics, but did mention that the programme is costly.

Comparatively inexpensive is the Indian *Balsakhi* programme evaluated by Banerjee et al. (2007)[29]. Observing test scores in basic reading literacy, as well as attendance and dropout rates, Banerjee et al. (2007) found that the programme increased test performance by 0.28 standard deviations. By controlling for the pre-intervention test score distribution, Banerjee et al. (2007) observed that the impact is twice as large for the bottom third performers than it is for the top third, which shows that the programme is most beneficial for those students for which it was specifically designed for. Banerjee et al. (2007) highlighted that due to the low cost of running the *Balsakhis* programme, this programme is one of the cheapest with a cost of about 0.67 USD per standard deviation increase.

Another impact evaluation of an Indian programme, the *Shishuvachan* classes, was conducted by He et al. (2009), in a longitudinal RCT. Learners in *Shishuvachan* classes were taught foundational skills for their later enrolment in primary school, where the programme specifically focused on comprehension techniques. These were concentrated on story-telling and literacy games (He et al., 2009). By randomizing at either the school or the community level, the authors were able to measure the effect of various versions of this programme, as well as in-school and out-of-school dynamics. To observe how test scores change, He et al. (2009) utilised the frequently used Pratham reading assessment that looked at the learners' knowledge of the alphabet, the ability to recognise words, reading of paragraphs and short stories. Their parents had to take the same test. He et al. (2009) found that the learners' reading ability was strongly linked with their parents reading skills: those whose parents could read a whole paragraph scored 0.36 standard deviations higher than those whose parents could not (He et al., 2009). In addition, the programme proved to be most effective for those learners who were attending the out-of-school version *Shishuvachan* classes, especially for

[28] This may have also induced a social recognition dynamic (Abeberese et al., 2014).

[29] This is the same study mentioned in the Pedagogy section, but here specifically the reading literacy results of the programme are discussed.

low performers (He et al., 2009). Moreover, the most robust gains were apparent when the programme was run as complement rather than as a substitute to pre-existing classes. That is, when differentiating between the in-school and out-of-school versions of the programme, He et al. (2009) found that the latter increased test score performance by 0.24 standard deviations, over and above the 0.26 standard deviations increases caused by the former.

That the ecosystem of the family may in some respects be more impactful than the schooling environment holds both for developed and developing countries. As a result, literacy programmes that place importance on the function of the family environment have gained momentum everywhere (van Steensel et al., 2011). In evaluating the effectiveness of various family literacy programmes on kindergarten and preschool children, van Steensel et al. (2011) conducted a meta-analysis of thirty international studies. The authors selected studies that focused specifically on the respective programmes' impact on comprehension- and code-related[30] skills. The important observations included the dwindling number of programmes that only focus on decoding skills. This illustrates the move away from merely emphasizing reading readiness and instead paying attention towards a more holistic approach in emergent literacy, which includes vocabulary learning as a pre-requisite for literacy. Additionally, van Steensel et al. (2011) noted how there are several aspects of at-home factors that impact evaluations cannot take account for, given the various inter-dependent dynamics that take place in the learner's home environment. That is, some programmes take parents' skills for granted or are ignorant of the sensitive and emotional relationship between learner and parent, which gives rise to pressure in an at-home learning setting. In aggregating the reviewed studies' results, van Steensel et al. (2011) found small but significant effect sizes: d=0.18 (for general literacy ability), d=0.22 (comprehension) and d=0.17 (decoding). This renders programmes that focus on family support during emerging literacy to be an important component of literacy skill development.

[30] Decoding skills fall under the so-called conventional literacy skills which also include reading, comprehension, spelling and writing (The National Institute for Literacy, 2008).

Chapter 2: The Shining Stars of Shine Literacy

Shine literacy is a South African non-profit organisation (NPO) that provides reading assistance to learners who are identified to be at literacy risk at the end of Grade 1 (Shine Literacy, 2016). The first pilot centre opened in 2000 at the Observatory Junior School with residents of the suburb acting as "reading partners" for Grade 2 or Grade 3 learners who were identified by their teachers as requiring reading assistance (The DG Murray Trust, 2012). After positive results[31] confirmed by the Western Cape Education Department (WCED), Shine Literacy officially registered as an NPO, opened more centres at various schools across Cape Town during the late 2000s, and formalised its programme (The DG Murray Trust, 2012). In essence, the programme's strategy is to provide primary schools with trained volunteers that assist learners during normal school hours. This assistance takes place in the form of one-on-one sessions, for two hours between volunteers and Grade 2 learners twice every week (The DG Murray Trust, 2012).

Regarding its content, the programme offers access to storybooks and other reading material, both in English as well as the learners' home language (Shine Literacy, 2016). During the bi-weekly sessions, a Shine-trained volunteer works together with a learner to improve their phonetic understanding, word recognition and decoding, as well as the paired reading of whole paragraphs (Shine Literacy, 2016). Moreover, Shine invites learners' caregivers to two workshops during Grade 1, before the commencement of the programme (Shine Literacy, 2016). This provides caregivers with the necessary resources and information on how they can scaffold their children's reading progress at home (Shine Literacy, 2016). Shine's goal is to offer an important high-quality input into resource-poor schools and address educational shortfalls. Moreover, the programme aims to ensure that Grade 2 and 3 learners are equipped with their developmental age/grade adequate literacy (The DG Murray Trust, 2012; Shine Literacy, 2016). This is emphasised by the recognition that these grades are crucial in guiding literacy acquisition and that by Grade 4, learning to read will become increasingly difficult (The DG Murray Trust, 2012).

At Shine Literacy, the literacy ability of all children in the classroom is assessed at the end of Grade 1 with the same test instrument across all participating schools (The DG Murray Trust, 2012). This test diagnostic[32] contains a battery of various literacy assessment items and assesses learners' ability to identify alphabet sounds, write down three letter words after being given picture hints, and write a sentence that is dictated to them. Shine has identified these skills as essential in learning how to read for meaning (Shine Literacy, 2016). After taking the test, learners are sorted into a literacy proficiency ranking, categorising them into either being "At risk", "Poor", "Satisfactory" or "Good" with regards to their abilities. The

[31] The WCED Grade 3 Systemics had shown a continuous increase in literacy scores at Observatory Junior School, starting from 50% in 2002, up to 71.1% by 2006 (The DG Murray Trust, 2012).

[32] The diagnostic was designed by three South African educationists in continuous collaboration with centre staff and external Shine partners such as the South African NPO Wordworks. The diagnostic itself has not been psychometrically validated, due to funding constraints. However, Shine has emphasised that the diagnostic itself has to be seen as a formative assessment, as opposed to a summative assessment (Volbrecht, personal communication 2018, January 30). More on the different functions and uses of these two forms of assessments can be found in the Guide to Assessment in Early Childhood: Infancy to Age Eight by Slentz et al., (2008).

following test scores are associated with the rankings: 0 – 30% ("At risk"); 31% – 59% ("Poor"); 60% – 70% ("Satisfactory"); and 71% – 100% (Good) (Volbrecht, personal communication 2018, January 30). In practice, programme assignment does not clearly follow the underlying assumption behind this ranking. Because of varying degrees of availability among volunteers at the centres, some learners that are outside the "At risk" or "Poor" category are often assigned to the programme. That is, whenever there are more volunteers than needed, rather than sending them away, Shine assigns a learner to them. In the same vein, often too few volunteers are available, and some "At risk" or "Poor" learners do not get assigned a volunteer (Dornbrack, personal communication 2017, January 20). Mostly, the concern is that there are too few, rather than too many, volunteers on the school roster. Nevertheless, the Programme Manager for Shine Centres reports that once the roster is established for the year, volunteer numbers seem to stay constant over the school terms with only minor fluctuations. The number of volunteers across all Shine schools is currently at around 770 – 800 (Volbrecht, personal communication 2018, February 6).

By the middle of Grade 2, a second test diagnostic is taken by all learners, both non-participants and participants. This is to measure whether the latter have made sufficient progress to leave the programme, or whether they should stay in the programme for another year, during Grade 3. The programme does not currently extend beyond Grade 3 and all Shine participants go back to the general learner population (The DG Murray Trust, 2012). Generally, the progress of all Shine and non-Shine learners during the whole of Grade 2 is monitored, and feedback is given to the teacher, parents, Heads of Department and schools (The DG Murray Trust, 2012; Shine Literacy, 2016).

A previous evaluation of Shine was conducted by Schkolne (2014) who used data from eight Shine participating locations in the Western Cape for the years 2011 to 2013. Schkolne (2014) used a sharp cut-off discontinuity regression to show that the test score targets[34] set by Shine were met or in many cases exceeded. This includes that after half a year, most Shine participants' results increased to the 77.77% grade level average and was shown to be a statistically significant result of attending Shine even after 12 months at 10% significance. However, Schkolne (2014) found that this effect diminished after 18 months and learner improvement was not a direct result of having attended Shine[35].

Data and an introduction to the sample

To evaluate the impact of Shine on literacy, this analysis made use of three datasets: Shine base- and endline test score data; Shine attendance data; and Systemic test scores and demographic data from participating schools obtained from the WCED for 2013 to 2016. By combining the Shine and WCED data using the unique Central Education Management Information System (CEMIS) identifier, it was possible to fill in

[34] The specific target mentioned in Schkolne (2014) refers to Shine's first medium-term goal of improving each Shine participant's score by 20% (Grigg et al., 2016).
[35] Interpreting Schkolne's (2014) results needs to be done with caution, since Shine does not strictly adhere to the cut-off points for selection. Therefore, the cut-off used in her work may result into – at best - a fuzzy discontinuity.

missing information in the Shine datasets from the WCED data. This included information on gender, birthdates and home language of the learners. The instruction (i.e. the way it is assessed) and content (i.e. what is assessed) of Shine's test instrument had remained constant between 2013 and 2016, which is the time period underlying this analysis[36]. In that way, learner progress can be compared across and within cohorts, given the constant nature of the test content. Pooling the available Shine test score data and dropping observations with missing test scores resulted in a dataset of n=1543 unique observations. This dataset contains information on Shine participants and non-participants across ten schools over four years. Programme participation was recorded by Shine with a categorical variable indicated as "Never attended[37]", "Participating", "Graduated", "Waiting" or "Special Needs". In the data cleaning process, learners that were listed as "Graduated" were re-coded as participants. Similarly, learners that were indicated as "Waiting" (n=12) were also re-coded as participants, given the intention to treat those individuals. Lastly, learners that were identified to require special education (n=5) were dropped from the sample. This resulted in a sample of 836 non-participants, and 707 participants.

In the Western Cape, ten primary schools have hosted[38] an independent Shine Centre by 2018. All schools are located in the metro area of the City of Cape Town. The following table gives a breakdown of the participating schools, the total number of Grade 2 pupils, the fraction enrolled at Shine, data availability by year from each school, as well as Shine attendance rates.

Table 2: Participating Schools

School	**Total Grade 2 Pupils (n)**	**Shine Enrolment**	**Years Data Availability**				**Attendance Rates**
			2013	2014	2015	2016	
Claremont Primary	123	52.8%	x	x	x	x	77.9%
Good Hope Seminary Junior	177	41.8%	x	x	x	x	55.9%
Observatory Junior	68	77.9%	x	x			79.3%
Prestwich Street Primary	264	42.4%	x	x	x	x	71.2%
Rosmead Central Primary	159	38.9%			x	x	71.5%
St. Agnes's Primary	174	43.6%		x	x	x	54.3%
St. Paul's Primary (Wynb)	252	40%	x	x	x	x	55%
Walmer Estate Primary	96	48.9%		x	x	x	59.4%
Zonnebloem Boys Primary	128	61.7%	x	x	x	x	53.4%
Zonnebloem Girls Prac. School	102	35.2%	x	x	x		64.2%
TOTAL	1543	45.8%					63.5%

The largest school in the sample was Prestwich Street Primary with 264 Grade 2 learners, followed by St. Paul's Primary with 252 learners. In contrast, much smaller was Observatory Junior with 68 pupils, and Walmer Estate Primary with 96. Across all schools, Shine enrolment was 45.6% on average. The highest fraction of Shine participants was found in Observatory Junior[39] where 77.9% of Grade 2 pupils were

[36] Note, Shine has adopted a new testing diagnostic, which has been piloted since 2017.
[37] To avoid confusion: this is how Shine labelled those learners that were not assigned to the programme, i.e. non-participants.
[38] On the process of school selection, Shine identified their hosting schools based on the following school/learner characteristics: little to no formal pre-school education for learners; little available remedial support; teacher-learner ratios; little access to books or libraries; and mostly second language learners from low-income households (The DG Murray Trust, 2012; Shine Literacy, 2016).
[39] Note, first pilot Shine centre opened in 2000 at the Observatory Junior.

enrolled in the programme. This is followed by 61.7% of learners enrolled at Zonnebloem Boys Primary and 52.8% at Claremont Primary. In contrast, at Zonnebloem Girls only 35.2% of Grade 2 learners were enrolled. All ten schools were identified to fall into the NQ5 Quintile[40]. In terms of the Language of Teaching and Learning, schools exclusively teach in English.

There are attendance data available for all 707 Shine participants. Learner attendance at Shine sessions are generally recorded on a monthly basis. This was aggregated for each learner over the whole school year. The result is a list of information on how many sessions there were available for each learner, and out of that how many sessions the learner actually attended. The resulting intensity index measures the fraction of available sessions attended by the learner. On average, Shine learners attended 63.5% of their sessions. This varies by school: the highest average Shine attendance was recorded at Observatory Junior, where Shine learners attended almost 80% of their sessions. In contrast, Shine participants at Zonnebloem Boys Primary attended 53.4% of their sessions.

To get a deeper understanding of the available data, and to assess the internal validity of the analysis, Table 3 and Table 4 take a more detailed look at the sample characteristics. Table 3 displays the differences in demographics between participants and non-participants to examine the balance in the two sample populations[41]. The overall sample was almost equally split between boys and girls, but boys were significantly more likely to be in the participant population. Across languages, both samples seem to be similar: there are no statistically significant differences with the exception of IsiXhosa speakers who were more likely to be participating in the programme. Regarding learner age, no differences exist. On average, a learner was 8.5 years old. The majority (29.8%) of the sample spoke IsiXhosa, which was followed by English (12.7%).

Table 3: Sample Demographics (%)

		%	Participant	Non-Participant
All (n, %)		1543 (100)	707 (45.8)	836 (54.2)
Gender	*Female*	50.8	38.7	61.3***
Language[42]	*IsiXhosa*	29.8	51.1	48.9***
	English	12.7	47.7	52.3
	French	0.8	25	75
	Afrikaans	0.5	57.1	42.9
	Other[43]	3.2	56	44
Age (mean, SD)		1543	8.51 (0.5358)	8.48 (0.4309)

*** p<0.01, ** p<0.05, * p<0.1

[40] Three schools have ex-department indicators: Good Hope was administered by the "House of Representatives"; Observatory Junior and St. Agnes by the "Cape Education Department". From 1979 until 1994, the House of Representatives was the department that administered schools for Coloureds; and the Cape Education Department administered schools for Whites (Plüddemann *et al.*, 2004).
[41] For a breakdown of demographics by school, refer to Appendix A.
[42] Given that language data were imputed from the Systemic dataset and not all learners were able to be matched with Systemic information, there were a total of 821 missing data on home languages.
[43] This category included: IsiZulu, IsiNdebele IsiZulu, Pedi, Sesotho, Tswana, Venda, and Xitsonga.

To gain an idea of how test scores varied between participants and non-participants, Table 4 shows pooled test score averages for the Shine and Systemic tests, by participation. At baseline, Shine pupils scored an average of 44.4% on the Shine test diagnostic, compared to 70.5% scored by non-participants. This difference is statistically significant. At endline, Shine pupils' test score average increased to 66.4%, but their scores remained statistically lower than non-participants who averaged a score of 74.8 %. When looking specifically at the literacy ranking assigned at baseline, the vast majority of Shine participants scored either in the "At risk" (27%) or "Poor" (49.4%) categories. 19.2% of those that were assigned to the treatment scored "Satisfactory" and 4.4% "Good". In comparison, the vast majority (65.6%) of non-participants were ranked as "Good". This was followed by "Satisfactory" learners who made up just over 30% of the non-participant population. Only 2.5% and 1.8% of the non-participants scored in the "At risk" and "Poor" categories, respectively.

Table 4: Shine and Systemic Test Scores

			n	Average all	Participants	Non-Participants
Shine Test Score	*Baseline (Grade 1)*		1543	58.5%	44.4%	70.5%***
	Literacy Categories, %	*At risk*	212		27%	2.5%
		Poor	364		49.4%	1.8%
		Satisfactory	388		19.2%	30.1%
		Good	579		4.4%	65.6
	Endline (Grade 2)		1543	70.9%	66.4%	74.8%***
	Literacy Categories, %	*At risk*	46		3.5%	2.2%
		Poor	130		17.3%	1%
		Satisfactory	269		29.4%	7.3%
		Good	1094		49.8%	89.6%
Systemic Test Score	*Language*		727	52.3%	44.1%	60.6%***
	Maths		727	56.8%	48.1%	65.6%***

*** p<0.01, ** p<0.05, * p<0.1

At endline, the majority (49.8%) of Shine participants ranked in the "Good" literacy category. This is followed by 29.4% that scored "Satisfactory". Only 3.5% of Shine learners scored in the "At risk" category and 17.3% scored in "Poor". In the case of non-participants, 89.6% of learners were ranked as "Good", whereas 7.3% fell in the "Satisfactory" range. Lastly, 1% of learners who were not assigned to the programme scored in "Poor", and 2.2% were considered "At risk".

In the WCED's Systemics (n=727), which is written during Grade 3 when the programme is completed, Shine participants scored 44.1% for Language, compared to non-participants who averaged 60.6%. This is a statistically significant difference. Scores for Mathematics were generally slightly higher than Language Systemics, with an average across the learner population of 56.8%. Shine learners scored 48.1% and non-participants got 65.6%. Again, this difference by treatment is statistically significant.

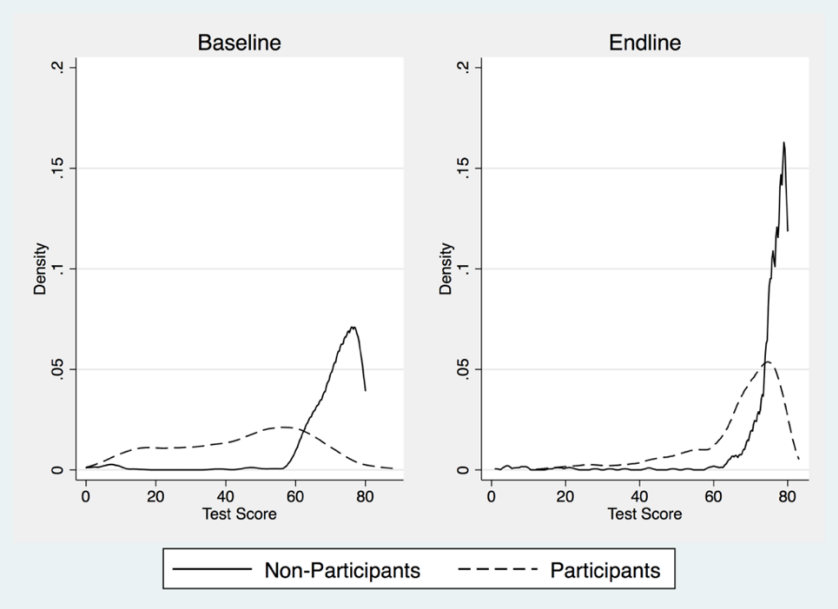
Undeniably, Shine participants' score increase of 22% between baseline and endline is quite drastic. Conversely, non-participants' scores increased by just under 5%. Most importantly, the proportion of Shine

learners that were scoring "At risk" or "Poor" decreased from a combined 76.4% at baseline, down to 20.8% by endline. The remaining proportion scored either "Satisfactory" or "Good". At the same time, non-participants do better in the Systemic tests during Grade 3, which begs the question of what would have happened to the Shine learners had they not been assigned to the programme in Grade 2.

Shine's diagnostic: how do participants compare to non-participants?

As highlighted by Ho & Yu (2015), normality in test score distributions rarely every holds in practice. A quick Skewness and Kurtosis[44] hypothesis test for normality on Shine scores showed that neither the baseline scores nor the endline scores are distributed normally. Instead, the distributions are heavily skewed and have much thinner tails[45] than a normal distribution. This is illustrated in Figure 2, which maps out the distribution of Shine test scores at baseline and endline by treatment.

Figure 2: Shine test score distribution over time



Note: Baseline non-participants' kurtosis and skewness: 22.02 and –3.04, participants' kurtosis and skewness: 2.12 and -0.3; Endline non-participants' kurtosis and skewness: 32.76 and -5.21; Participants' kurtosis and skewness: 5.82 and -1.72.

At baseline, non-participants' scores are much higher and concentrated around 70%, compared to a more flattened and spread out score distribution among participants. By endline, however, both samples' tails

[44] With the formulae for skewness and kurtosis, respectively:

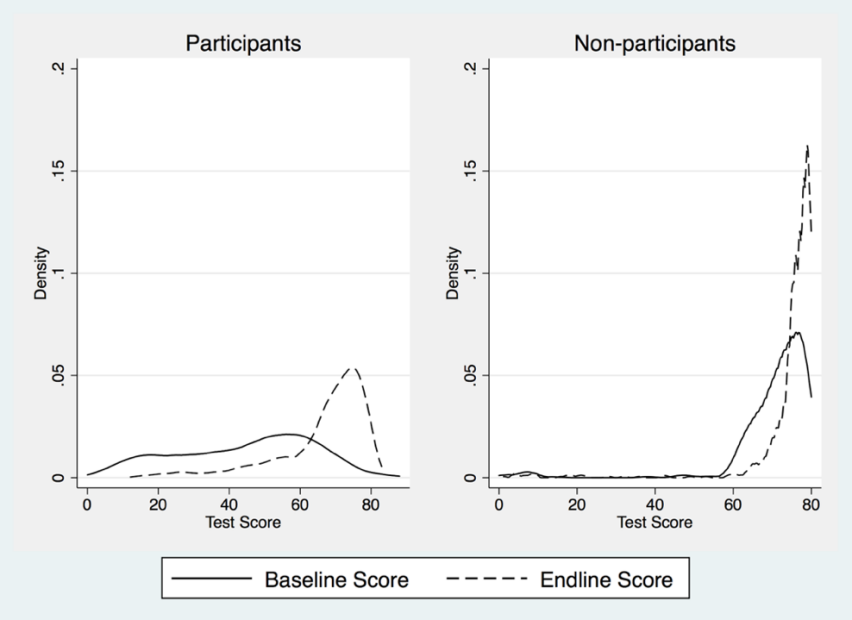
$$s = \frac{\sqrt{n(n-1)}}{n-2} \frac{\frac{1}{n}\sum_i (x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_i (x_i-\bar{x})^2\right)^{3/2}}$$

$$k = \frac{n(n+1)(n-1)}{(n-2)(n-3)} \frac{\sum_i (x_i-\bar{x})^4}{\left(\sum_i (x_i-\bar{x})^2\right)^2}$$

[45] Hypothesis p-values: Skewness Baseline: p=0.000, Kurtosis Baseline: p=0.01312; Skewness Endline p=0.000, Kurtosis Endline: p=0.000.

become tighter, meaning that not only do the score means increase, the distribution becomes less spread out. This decreases the variance in learner performance and essentially makes instruction for the teacher much easier. To really observe the magnitude of these shifts in the two distributions by participation, Figure 3 shows specifically how participants' scores change from baseline to endline, compared to how non-participants scores shift.

Figure 3: Shine test score distribution by participation



Participants' score distribution moved from its spread-out nature during baseline to a much tighter distribution at endline. A much higher fraction of Shine learners was found above the "Satisfactory" cut-off line. Recall that just under 30% scored in "Satisfactory" and around 50% in "Good". As such, mean test scores were still lower than non-participant learners' scores. However, the magnitude of this mean increase among Shine learners is striking and is already pointing towards large difference-in-differences.

Difference-in-Differences: "At risk" learners and isiXhosa speakers take it home

The Shine test score data lend itself to the use of difference-in-differences techniques, which has become a widespread method in quasi-experimental studies where assignment to treatment is not randomised (Khandker et al., 2010; Gertler et al., 2012; Lance et al., 2014). The most straightforward set-up is to observe the same test diagnostic for both non-participants and participants before and after the treatment. Non-participants were not assigned to the treatment, but participants were. To remove any bias in the endline comparison between the participants and non-participants that are either due to permanent differences between the groups or the result of time-variant[46] trends, the difference-in-differences method subtracts

[46] Time-variant trends would be the tendencies of test scores to improve generally as learners proceed through the grade, in line with age developmental progress.

the average difference in the non-participant group from the average difference in the participant group. Given the Shine data, the performance of all Shine participants is then compared to all non-participating learners, where both populations were spread across all four literacy ranking categories. This made difference-in-differences estimation an attractive method to apply.

First, consider the difference-in-differences equation if the data were treated as repeated cross sections. Shine treatment status was indicated by S = 0, 1 where 0 indicated no treatment, i.e. non-participants or the comparison group, and S = 1 indicated a Shine learner. Further, all learners were observed over two time periods, t =0, 1 where 0 was the time period at baseline, and 1 was the period at endline. Every learner was indexed by the letter i =1, …N and learners have two observations. Then, $\overline{\text{testscore}}_0^{S}$ and $\overline{\text{testscore}}_1^{S}$ are the average outcomes in test scores for Shine learners at baseline and endline, respectively. For non-participants the average outcomes were $\overline{\text{testscore}}_0^{NS}$ and $\overline{\text{testscore}}_1^{NS}$. The difference-in-differences are then:

$$\zeta=(\overline{\text{testscore}}_1^{S}-\overline{\text{testscore}}_0^{S})-(\overline{\text{testscore}}_1^{NS}-\overline{\text{testscore}}_0^{NS}) \quad \textbf{(1)}$$

Manually taking these differences across the available independent variables, resulted in a list of difference-in-differences. This is shown in Table 5, which displays difference-in-differences across gender, home language, schools, as well as the baseline literacy categories.

Table 5: Test Score difference-in-differences, %

		Participants			Non-Participants			ζ
		Baseline	Endline	Difference (x)	Baseline	Endline	Difference (y)	x - y
All		44.4	66.4	22	70.5	74.8	4.3	*17.7****
Gender	*female*	47.7	68.7	21	71.7	75.8	4	*16.9****
	Male	41.4	64.2	22.8	68.5	72.8	4.3	*18.4****
Language	*isiXhosa*	41.1	66.2	25.1	70.9	74.7	3.8	*21.5****
	English	52.3	67.9	15.6	72.2	77.1	4.9	*10.6****
	French	50.3	70.4	20.1	74.5	78.3	3.8	*16.3****
	Afrikaans	41.5	62.2	29.7	74.7	77.7	3	*17.7*
	Other	47.9	67.9	20	71.9	76	4.1	*15.9****
School	*Claremont Primary*	58.6	75.7	17.1	75.9	78.1	2.2	*14.9****
	Good Hope Seminary Junior	45.4	68.6	23.3	73.4	75.9	3.5	*19.8****
	Observatory Junior	35.1	68.7	33.6	52.3	57.6	5.3	*28.3****
	Prestwich Street Prim.	45.7	67.7	22.5	69.9	73.7	3.8	*18.7****
	Rosmead Central Prim.	57.1	72.6	15.5	75.8	73.7	2.3	*13.2****
	St. Agnes's Prim.	51.2	62.9	11.7	70.8	76.6	5.8	*5.9****
	St. Paul's Prim. (Wynb)	32	54.2	22.2	64.1	69.6	5.5	*16.7****
	Walmer Estate Prim.	36.9	63.5	27.6	68.7	73.4	4.7	*22.9****
	Zonnebloem Boys Prim.	36.9	65.3	28.3	68.1	74.1	6	*22.3****
	Zonnebloem Girls Prac. Sch.	49.7	70	20.4	73.6	76.8	3.2	*17.1****
Baseline Literacy	*At risk*	18.6	53.2	34.6	5.8	15.7	9.9	*24.7****
	Poor	47.4	69.7	22.3	51.1	69.9	18.8	*3.5*
	Satisfactory	65.5	73.1	8.6	66.1	73.8	7.7	*0.9*
	Good	76.5	75.1	-1.9	75.6	77.5	1.9	*-3.3****

*** p<0.01, ** p<0.05, * p<0.1

The difference-in-differences for all categories were relatively large throughout, and with the exception of the Afrikaans, "Poor", and "Satisfactory" categories, all are statistically significant. The difference-in-differences for Shine participants on test scores indicated an increase by about 17.7% due to attending Shine. The largest effect across languages among Shine participants was found for isiXhosa speakers with a difference-in-differences of 21.5%. When examining the schools, the highest test score difference-in-differences was recorded in Observatory Junior with 28.3%; followed by Walmer Estate Primary with 22.9%. The lowest difference-in-differences was found for St. Agnes's Primary with 5.9%. Lastly, across literacy rankings, "At risk" Shine learners saw a difference-in-differences in their test scores of 24.7%. In comparison, "Good" participants' scores actually saw a negative difference-in-differences of 3.3%.

Making use of the two-time period structure of the Shine test score dataset, it was possible to estimate the following model and arrive at a consistent estimator of the average treatment effect of Shine:

$$\Delta \text{test score}_{it} = \alpha + \beta S_{it} + \gamma\ \text{testscore}_0 + \delta(S_{it} * \text{testscore}_0) + \varphi x_{it} + \varepsilon_{it} \quad \textbf{(2)}$$

where

β = measures the Shine-specific average treatment effect.

γ = captures the impact of baseline test scores.

δ = measures the effect of the interaction between Shine treatment and baseline test scores.

φ= captures the effect of a vector of independent characteristics, including school dummies.

ε= as an idiosyncratic, unobserved error term.

To see how robust the coefficient on the treatment indicator was, this model included covariates of gender, learner's age and age squared (to account for a possible quadratic effect), learner's home language, schools and learner's attendance, including a squared attendance term. Robust standard errors were used and clustered at the school level. The results of this estimation are displayed in Table 6. In addition to the estimation of the average effect on the whole learner sample, various interactions were also included to observe a potential avenue for differential effects. The effect of those interactions is added as additional columns in the table.

Table 6: Change in test scores as a function of Shine

Dependent Variable: **Change in Shine Test Score**	(1)	(2)	(3)	(4)	(5)	(6)
Shine	26.94***	26.85***	24.94***	4.592*	3.942	-3.717
	(2.381)	(2.482)	(2.552)	(2.070)	(2.397)	(3.676)
Shine Baseline Score	-0.240***	-0.241***	-0.260***	-0.490***	-0.483***	
	(0.0574)	(0.0576)	(0.0597)	(0.0904)	(0.0943)	
female				1.688***		
				(0.469)		
Shine X female				0.120		
				(1.320)		
Learner's age		13.31	10.29	13.23	13.83	6.788
		(15.51)	(16.29)	(17.70)	(17.41)	(17.25)
Learner's age squared		-0.683	-0.534	-0.702	-0.742	-0.327
		(0.888)	(0.926)	(1.000)	(0.988)	(0.985)
Shine X Baseline Score	-0.347***	-0.345***	-0.326***			
	(0.0405)	(0.0392)	(0.0371)			
isiXhosa					-0.841	
					(0.701)	
Shine X isiXhosa					2.382	
					(1.428)	
At Risk Category						9.980***
						(1.537)
Poor Category						16.65***
						(1.701)
Satisfactory Category						6.510***
						(0.759)
Shine X Poor						7.055*
						(3.752)
Shine X At Risk						27.01***
						(4.337)
Shine X Satisfactory						2.844
						(2.605)
Good Hope Seminary Junior			-0.572	0.0104	-0.0652	-1.463***
			(0.334)	(0.663)	(0.705)	(0.311)
Observatory Junior			3.634**	4.082**	3.694**	3.134***
			(1.311)	(1.752)	(1.464)	(0.769)
Prestwich Street Prim.			-1.335**	-1.151	-1.315	-2.640***
			(0.514)	(0.918)	(0.803)	(0.474)
Rosmead Central Prim.			-1.528***	-0.966**	-1.071***	-1.495***
			(0.219)	(0.311)	(0.257)	(0.378)
St. Agnes's Prim.			-3.278***	-3.269***	-3.276***	-5.288***
			(0.302)	(0.616)	(0.808)	(0.378)
St. Paul's Prim. (Wynb)			-4.067***	-4.048**	-4.200**	-4.653***
			(0.908)	(1.643)	(1.770)	(0.421)
Walmer Estate Prim.			0.0599	0.572	0.457	-1.391*
			(0.851)	(1.414)	(1.321)	(0.631)
Zonnebloem Boys Prim.			1.002	2.412	1.352	-0.327
			(0.956)	(1.762)	(1.316)	(0.477)
Zonnebloem Girls Prac. Sch.			-0.952**	-1.205**	-0.456	-2.068***
			(0.315)	(0.438)	(0.460)	(0.509)
Constant	21.15***	-42.41	-24.63	-22.43	-23.89	-29.74
	(4.132)	(66.12)	(69.72)	(75.52)	(74.00)	(74.88)
Observations	1,543	1,543	1,543	1,543	1,543	1,543
R-squared	0.699	0.702	0.719	0.695	0.693	0.691

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Estimating equation (2) resulted in an average treatment effect ranging between 24.9% and 26.8%, depending on the specification. These coefficients are statistically significant at 1% and robust throughout. In practical terms, this implies that the effect of being assigned to the programme increased test scores between 1.7 to 1.9 standard deviations. Compared to other education programmes discussed in Chapter 1,

this is already an enormous test score increase. An impact of this magnitude falls in the range of large Cohen's *d* effect sizes, which is implied by anything greater than 0.8 standard deviations (Cohen, 1992). Educationally speaking, programmes that increase scores by over 0.2 standard deviations are already viewed as meaningful (Cohen, 1992; Wasik, 1998).

Although girls across the learner population generally did around 1.7% better at endline, the coefficient on the interaction of Shine treatment and gender is small and statistically not different from zero. This implies that Shine impacts all learners the same, regardless of their gender. In terms of age, though there was the expected quadratic effect (impact of age decreases as age increases), in none of the specifications was the learner's age statistically significant. Unsurprisingly, the lower the learner's baseline score, the lower they scored at endline. This holds for both the general population and the Shine learners. Including an isiXhosa dummy carried no effect – presumably because of fixed school effects[47]. From column 6, it can be seen that large positive gains are experienced by "At risk" learners with a test score improvement of 27%, which is statistically highly significant at 1%.

To gain more insight on the potentially heterogeneous effects that Shine has on its learners, and to see who benefitted the most from participating, the following regression procedure allowed for the estimation of the differential impact on various sub-samples. The results of this estimation are displayed in Table 7, where in each column the same specification as per (2) above was run but restricting the sample on the relevant population.

[47] This is probably indicative of different demographics represented at each school, and in the case here, the reality that some schools are more heavily skewed toward isiXhosa speakers. For more insights on how school differ by demographics the reader is referred to Appendix A.

Table 7: Differential impact of Shine by baseline literacy profile

Dependent Variable: **Change in Shine Test Score**	(1) Sample: Girl Children only	(2) Sample: isiXhosa Children only	(3) At Risk learners	(4) Poor learners	(5) At Risk & Poor
Shine	20.09*	24.14***	20.79**	12.03	34.31***
	(9.231)	(6.789)	(6.501)	(7.608)	(4.721)
Shine Baseline Score	-0.381**	-0.296***	-0.592	-0.486***	0.124**
	(0.142)	(0.0889)	(0.648)	(0.137)	(0.0464)
Shine * Baseline Score	-0.261*	-0.296**	0.349	-0.221	-0.619***
	(0.127)	(0.0967)	(0.761)	(0.143)	(0.112)
Good Hope Seminary Junior	-0.300	-1.156*	-6.191***	-3.116**	
	(0.495)	(0.549)	(0.114)	(1.218)	
Observatory Junior	1.659	4.045***	0.948	-0.522	
	(0.923)	(0.952)	(0.747)	(1.155)	
Prestwich Street Prim.	-1.427**	-1.537***	-6.303***	-4.687***	
	(0.609)	(0.303)	(0.574)	(1.194)	
Rosmead Central Prim.	-1.099**	-2.482***	2.531***	-3.984**	
	(0.439)	(0.129)	(0.555)	(1.682)	
St. Agnes's Prim.	-2.809***	-4.569***	-14.71***	-9.818***	
	(0.587)	(0.219)	(0.102)	(1.799)	
St. Paul's Prim. (Wynb)	-2.015	-1.011	-14.94***	-9.804***	
	(1.237)	(0.873)	(1.097)	(1.525)	
Walmer Estate Prim.	0.756	-1.977**	-3.893***	-3.836***	
	(0.789)	(0.666)	(0.314)	(1.210)	
Zonnebloem Boys Prim.		-2.089**	-0.821	-4.466***	
		(0.706)	(0.539)	(1.118)	
Zonnebloem Girls Prac. Sch.	-0.893	-1.146***	-0.510**	-2.657**	
	(0.493)	(0.244)	(0.204)	(1.240)	
Constant	32.64**	26.45***	25.74***	48.77***	10.99***
	(10.60)	(6.640)	(3.384)	(7.334)	(1.389)
School FE					YES
Observations	781	460	212	364	576
R-squared	0.748	0.788	0.340	0.508	0.362

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In the first column of Table 7, it was investigated how girl learners perform by participation: that is, to see how Shine girls fare against non-Shine girls. While there was a 20.1% effect increase, this is only significant at the 10% level. In contrast, restricting the learner population to isiXhosa speakers only, it became clear that they benefitted greatly from the programme: test scores increased by 24.1% when assigned to the programme – an effect that is statistically significant at 1%. Furthermore, across baseline literacy profiles, the "At risk" learners were the strongest beneficiaries. Test scores among those who struggled with literacy the most increased by 20.8% due to participating in the programme. This effect increase is significant at the 5% level. It is, however, important to note that the sample for this estimation decreased down to 212 learners. Of these learners, the majority were Shine participants, which led to the problem of small sample comparison. To mitigate this problem, the specification was run on a pooled sample of "At risk" and "Poor" learners and its results reported in column 5. The treatment effect here was strong and statistically significant, even with school fixed effects. In terms of schools generally, Good Hope, Prestwich Street, St. Agnes's, St. Paul's, Walmer Estate, Zonnebloem Boys and Zonnebloem Girls all did statistically worse than Claremont Primary, which is the omitted school across the included school dummies. The interaction between Shine and baseline scores is negative throughout almost all specifications, and significant at varying levels of statistical significance.

Transitions to better literacy

To get an understanding of how learners improved between baseline and endline, Table 8 displays how the general learner population transitioned along the literacy rankings from baseline to endline. The way this transition varied by Shine participation is shown in Table 9.

Table 8: Literacy Categories all, %

All learners		Grade 2			
		At risk	Poor	Satisfactory	Good
Grade 1	At risk (n=212)	19.3	42.9	22.1	16.5
	Poor (n=364)	0.3	9.1	37.1	53.6
	Satisfactory (n=388)	0.3	1	19.9	78.9
	Good (n=579)	-	0.4	2.1	97.6

In general, fewer than 20% of baseline "At risk" literacy learners stayed in this category by endline, 42.9% moved up to "Poor", 22.1% to "Satisfactory" and 16.5% to "Good". "Poor" learners at baseline were generally not seen again in this category: under 10% remained there by endline. In contrast, 37.1% moved to "Satisfactory" and more than half scored in the "Good" range. Out of those learners that were categorised as "Satisfactory" at baseline, just under 20% stayed in "Satisfactory" and the majority (78.9%) scored "Good". Lastly, the overwhelming majority (97.6%) of "Good" literacy learners stayed in "Good". Only 2.1% slid down to "Satisfactory", and a small fraction of 0.4% down to "Poor".

Table 9: Literacy Categories by participation, %

Participants		Grade 2				**Non-Participants**		Grade 2			
		At risk	Poor	Satisfactory	Good			At risk	Poor	Satisfactory	Good
Grade 1	At risk (n=191)	12.6	45.6	23.7	18.3	Grade 1	At risk (n=21)	81	19	-	-
	Poor (n=349)	0.3	8.9	37.5	53.3		Poor (n=15)	-	13.3	16.7	60
	Satisfactory (n=136)	-	2.2	21.3	76.5		Satisfactory (n=252)	0.4	0.4	19.1	80.2
	Good (n=31)	-	3.2	9.7	87.1		Good (n=548)	-	0.2	1.6	98.2

When looking at the transition by participation displayed in Table 9, only 12.6% of those learners who were categorised as "At risk" and assigned to the programme remained there. In contrast, 45.6% moved to "Poor", 23.7% to "Satisfactory" and 18.3% moved up to "Good". Similarly, around 9% of learners who scored "Poor" at baseline stayed there. In fact, 37.5% moved up one category to "Satisfactory" and the majority (53.3%) increased by two categories to "Good". "Satisfactory" learners who were assigned to Shine

at baseline stayed in that category 21.3% of time, while the overwhelming majority of 76.5% moved up to "Good".

On the other hand, non-participant learners that were considered "At risk" (n=21) remained in that category 81% of the time. The remaining "At risk" fraction moved up to "Poor", however, none moved to 'Satisfactory" or "Good". This essentially implies that those "At risk" learners that were not assigned to the programme, stayed stuck. They did not transition to better literacy by endline. However, it needs to be highlighted that the sample size of those "At risk" learners is very small to begin with. Another small, yet striking insight into how learners transition is found in the "Good" category: a higher proportion (98.2%) of the non-participants who were considered "Good" remained in "Good" when compared to participants (87.1%); however, this only pertained to n=31 participant cases. In the instances of "Good" participants, roughly 10% actually moved down to "Satisfactory". This compares to only 1.6% of "Good" non-participants (n=548) who moved down to "Satisfactory". Admittedly, the drastically smaller sample of "Good" learners that participated against the much bigger sample of "Good" non-participants, does not give rise to any serious concern about this movement[48].

Attendance: it is (quite) important to show up

Shine learners who were identified in the attendance dataset (n=707) went to their one-on-one lessons on average 63.5% of the time. In 84 cases, there were Shine participants who never attended any session, and thus received an attendance index of zero. For the purpose of the following analysis, non-participants got assigned a zero for their attendance intensity indicator. Given that scoring 60% in the Shine test diagnostic was the cut-off score to move to a "Satisfactory" literacy ranking, the dependent variable in the following estimation was a binary variable indicating a score of at least 60% at endline. Provided that the majority of Shine's treatment takes place when volunteer and learner interact, the question that remains is how important it is for the learner to show up to improve their literacy significantly and eventually leave the programme. In Table 10 the average marginal probabilities of reaching 60% and greater at endline are reported. This estimation considers attendance intensities and a range of controls. Further, given two of the sub-samples on which Shine has already shown heterogeneous impact ("At risk" and "Poor" learners), the results of a restricted sample analysis are also shown.

[48] Nevertheless, specifically with the difference-in-differences of -3% of "Good" participants' test scores from Table 5 in mind, this movement should be observed more in-depth over time with more available data. What it could suggest, for instance, is that the programme is not well catered to those who are already doing well with regards to their literacy. It could be that the intervention is not fully fleshed out to accommodate those learners who are already at a good standing, and who would benefit more from regular class interaction than practicing literacy with a volunteer. Though it would imply sending volunteers away – which will rarely ever take place in practise – it could be that "Good" learners might not benefit, and instead do worse. Thus, it may be worthwhile to explore this possibility and consider asking "extra" volunteers to attend to a different Shine Center.

Table 10: Average Marginal effects after Probit

Dependent Variable: **Endline Score > 60%**	(1) All	(2) All	(3) All	(4) Sample: At Risk	(5) Sample: Poor
Shine Attendance	-0.122	-0.257**	-0.143	-1.210***	0.501**
	(0.0995)	(0.122)	(0.0916)	(0.332)	(0.219)
Attendance squared	0.182*	0.235**	0.193**	1.322***	-0.370
	(0.110)	(0.118)	(0.0921)	(0.388)	(0.252)
Learner's age	0.310	0.295*	0.316*	1.642**	-2.037**[49]
	(0.197)	(0.178)	(0.181)	(0.837)	(0.922)
Learner's age squared	-0.0163	-0.0157	-0.0166	-0.0894*	0.120**
	(0.0110)	(0.00998)	(0.0103)	(0.0474)	(0.0537)
female	0.0256**	0.0281***	0.0260**	0.0684	0.121***
	(0.0120)	(0.00979)	(0.0121)	(0.0661)	(0.0373)
isiXhosa	-0.0211	-0.0198	0.00683	-0.0517	-0.0444
	(0.0204)	(0.0196)	(0.0130)	(0.0632)	(0.0382)
Shine Baseline Score	0.00545***	0.00466***		0.0189***	0.00913***
	(0.000115)	(0.000455)		(0.00317)	(0.00215)
Shine Attendance X Baseline Score		0.00240**			
		(0.00120)			
Baseline Category: At risk			-0.283***		
			(0.0211)		
Attendance X At Risk			-0.000829		
			(0.0454)		
Baseline Category: Poor			-0.118***		
			(0.0306)		
Attendance X Poor			0.0152		
			(0.0523)		
Good Hope Seminary Junior	-0.00193	-0.00786		-0.191	
	(0.00682)	(0.00946)		(0.229)	
Observatory Junior	0.00551	0.0107**		-0.0421	
	(0.00462)	(0.00453)		(0.222)	
Prestwich Street Prim.	-0.0335***	-0.0340***		-0.268	
	(0.00427)	(0.00493)		(0.216)	
Rosmead Central Prim.	0.0230**	0.0157		0.00538	
	(0.00996)	(0.0134)		(0.281)	
St. Agnes's Prim.	-0.137***	-0.133***		-0.445**	
	(0.0149)	(0.0141)		(0.219)	
St. Paul's Prim. (Wynb)	-0.0598***	-0.0582***		-0.389*	
	(0.0140)	(0.0136)		(0.221)	
Walmer Estate Prim.	-0.00958	-0.0125		-0.261	
	(0.00782)	(0.00887)		(0.222)	
Zonnebloem Boys Prim.	0.0207***	0.0219***		-0.0982	
	(0.00744)	(0.00689)		(0.219)	
Zonnebloem Girls Prac. Sch.	0.0205**	0.0181**		0.198	
	(0.00839)	(0.00776)		(0.265)	
Observations	1,543	1,543	1,543	212	295

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Unsurprisingly, the higher the baseline score, the higher the likelihood of reaching 60%: an increase of 1% in baseline scores increased this likelihood by around 0.5%. Moreover, isiXhosa speakers were less likely to score in the "Satisfactory" or above category by endline. However, this was statistically not different from zero. By including the attendance index and a squared attendance term, the estimation results show how the impact of attendance increased as learners showed up to more sessions. This can be seen in all sub-

[49] The "Poor" scorers are only ones where the coefficient on age is negative and significant, and the effect of age increases as learners get older given the positive coefficient on the squared age term. This may suggest something along developmental lines, by summing age by baseline literacy categories, it was shown that "Poor" learners are generally a little younger than the remaining samples.

samples, except for when the sample is restricted to "Poor" learners[50]. From the second column, it is shown how a higher baseline score together with higher attendance increased the likelihood of reaching "Satisfactory" literacy, increasing baseline scores and attendance each by 1% increased this likelihood by 0.2%. Lastly, with respect to the "At risk" literacy learners, their probability of reaching 60% at endline increased to 81% if they would marginally increase their attendance by 10% (from the average of 52.2% to 57.4%). To guarantee a jump to the "Satisfactory" category, "At risk" learners would have to attend roughly 70% of the sessions that are offered to them, which is a finding that is statistically significant at 1%.

Propensity score matching: finding the right Shine match

The Shine data are also fitting for the use of another quasi-experimental method: propensity score matching. Matching can be used regardless of the programme's assignment strategy, given that at least one group had not been assigned to the programme and that there is other observable data on the whole learner population (Gertler et al., 2011). These characteristics are used to construct a comparison learner group. Using this method in the Shine context assumes that there are no unobservable characteristics of participants and non-participants that would affect test scores. Admittedly, this may be a strong assumption to make, but these estimates are provided as robustness checks on the treatment effects reported above.

Here, the goal with matching is to find non-participant learners that were very similar to Shine learners given the observable characteristics of baseline scores and age[51]. In so doing, the probability of being part of the Shine programme given these characteristics is then depicted as:

$$PR\,(S=1 \mid \text{Shine baseline score, age})$$

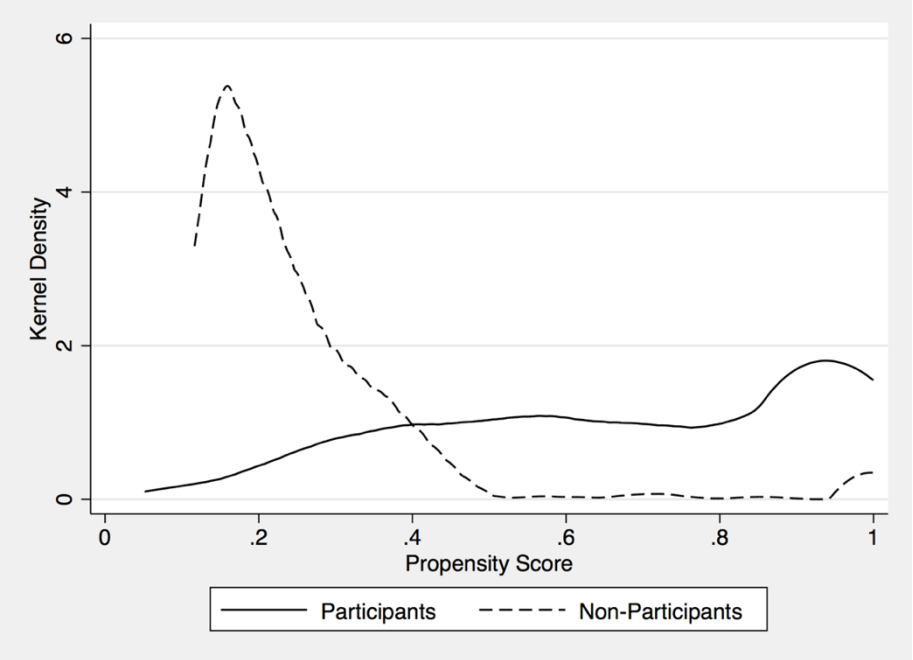
The result of this estimation, which was a probit regression of Shine treatment on baseline test score and age, is displayed in Appendix B. The estimation resulted in a predicted value for each learner's probability of attending Shine given these baseline characteristics, which is their propensity score. Once constructed, the new comparison group, which is comparatively similar to Shine learners in terms of age and baseline performance, mimicked the counterfactual of what would have happened to Shine learners, had they not been enrolled in the programme.

Naturally, given the inclusion of baseline scores as a matching variable, the distribution of propensity scores was much more likely to be closer to 1 for participants than for non-participants. However, given that Shine's assignment rule is not completely stringent on taking on learners solely based on sharp cut-offs, but also on volunteer availability, this method did not rule out the pre-requisite of common support. That is, because there were "At risk" and "Poor" learners in the non-participant population, but also "Satisfactory"

[50] This could be driven by the fact that "Poor" learners on average attended more sessions (61.5% compared to 52.2% attended by "At risk" learners, for instance). Also, it could be that they generally benefited, regardless of how many times they showed up.
[51] Trying to balance on any other combination of baseline characteristics proved unsuccessful in satisfying the balancing property and a good overlap of propensity scores. These are pre-requisites in propensity score matching.

and "Good" learners in the participant population, there was at least some part of the sample that had relatively good overlap of propensity scores between non-participants and participants (Gertler et al., 2011). After constructing the propensity, the sample used for the estimation of the average treatment effect was then restricted to those units that had common support[52], the overlay of which was computed in Figure 4.

Figure 4: Propensity scores common support, by participation



By subtracting the average endline score outcome of the constructed comparison group from the average endline score outcome of Shine learners, this method arrived at a measure of the average treatment effect as per:

ATE[53]=E(Endline Test Score Participant -Endline Test Score Constructed Comparison | Propensity Score, Shine=1)
=E (Endline Test Score Participant | Propensity Score, Shine=1)
- E (Endline Test Score Constructed Comparison| Propensity Score, Shine=0) **(3)**

To see how robust the estimated treatment effects were, the results of the default – and quite restrictive – method of Nearest Neighbour matching (where each treated Shine learner was matched with the constructed comparison learner that had the closest propensity score) were compared to another algorithm: Kernel matching. With Kernel, all comparison learner observations were used and weighed by how close they were to a participant observation based on their estimated propensity (Khandker et al., 2010). This avoids the risk of using only those non-participants who satisfied the strict criteria of matching as a nearest neighbour, which decreased the sample size for the comparison group quite drastically, as seen below (Khandker et al., 2010).

[52] The region of common support was calculated in the range [0.05246704, 0.99897727].
[53] This stands for "Average Treatment Effect"

Given the relatively strong assumption of no unobservable time-invariant differences under propensity score matching, this method can also be combined with difference-in-differences, which results in a matched difference-in-differences outcome[54] (Khandker et al., 2010; Gertler et al., 2011). The results of the treatment effects under the conventional propensity score matching method, which uses only endline test scores, are listed alongside the matched difference-in-differences estimation output in Table 11. To further investigate whether "At risk" and isiXhosa learners' large effect sizes as estimated under difference-in-differences could also be found with this method, the computation of the average treatment effect was also done with isiXhosa and "At risk" learners as sub-samples.

Table 11: Average treatment effects of Shine with propensity score matching

All	**Endline Test Score**	n= Shine	n= Constructed Comparison	Average Treatment Effect
	Nearest Neighbour Matching	707	148	10.2 (4.5)**
	Kernel Matching	707	835	13.2 (2.2)***
	Differenced Test Score			
	Nearest Neighbour Matching	707	148	9.1 (2)***
	Kernel Matching	707	835	9.2 (2.5)***
"At risk" learners	**Endline Test Score**	n= Shine	n= Constructed Comparison	Average Treatment Effect
	Nearest Neighbour Matching	191	14	38.5 (3.9)***
	Kernel Matching	191	20	37.5 (2.4)***
	Differenced Test Score			
	Nearest Neighbour Matching	191	14	33.2 (9.9)***
	Kernel Matching	191	20	25.5 (3.491)***
isiXhosa learners	**Endline Test Score**			
	Nearest Neighbour Matching	235	35	9.7 (7.646)
	Kernel Matching	235	182	16.4(2.520)***
	Differenced Test Score			
	Nearest Neighbour Matching	235	35	5 (3.525)
	Kernel Matching	235	182	11.5 (5.118)**

Note: for Nearest Neighbour the numbers of treated and controls refer to actual nearest neighbour matches; Kernel Matching: Bandwidth: 0.6, 5 replications bootstrapped standard errors; *** p<0.01, ** p<0.05, * p<0.1

As expected, given its restrictive requirements, using Nearest Neighbour as opposed to Kernel matching decreased the sample for the constructed comparison groups substantially. This was true for the estimation of the treatment effect on both endline and differences test scores, as well as for the two sub-samples investigated. When looking at the general learner sample and only at endline test scores, the average treatment effect under Nearest Neighbour was estimated as 10.2% and statistically significant at the 5% level. In contrast, when comparing this to Kernel matching, not only did the comparison group increase from 148 matched cases to 835[55], the treatment effect increased to 13.2%, and became even more statistically significant. However, once investigating the effect on differenced test, the estimation results were somewhat more conservative: under Nearest Neighbour, the impact Shine had was computed as 9.1%, compared to 9.2% under Kernel matching. Both effects were statistically different from zero at the 1% level of significance.

[54] Propensity score matching does not take into consideration those unobserved characteristics that may not only explain participation but also explain the outcome of interest. But, by combining it with a difference-in-differences outcome, one can at least account of unobserved characteristics that are time-invariant in both non-participant and participant groups (Gertler et al., 2011).
[55] Note, only one observation did not fall in the region of common support.

When restricting the sample to "At risk" learners, the sample size for the constructed comparison decreased quite drastically: with Nearest Neighbour to 14, with Kernel to 20 learners. Of course, this gives room for concerns regarding sample size comparison. Under the former method, the effect using endline test scores only amounted to 38.5%, and with differenced test scores around 33%, both statistically significant at 1%. Similarly, estimated impact with Kernel was also highly statistically significant: when using endline test scores the effect computed was 37.5% and with differenced test scores 25.5%. The latter effect size strongly resonates with the difference-in-differences calculated for "At risk" learners displayed in Table 5, with 24.7% and Table 6, with 27%. However, given the large decrease in the comparison group's sample size under both algorithms used, these effect sizes need to be interpreted with caution.

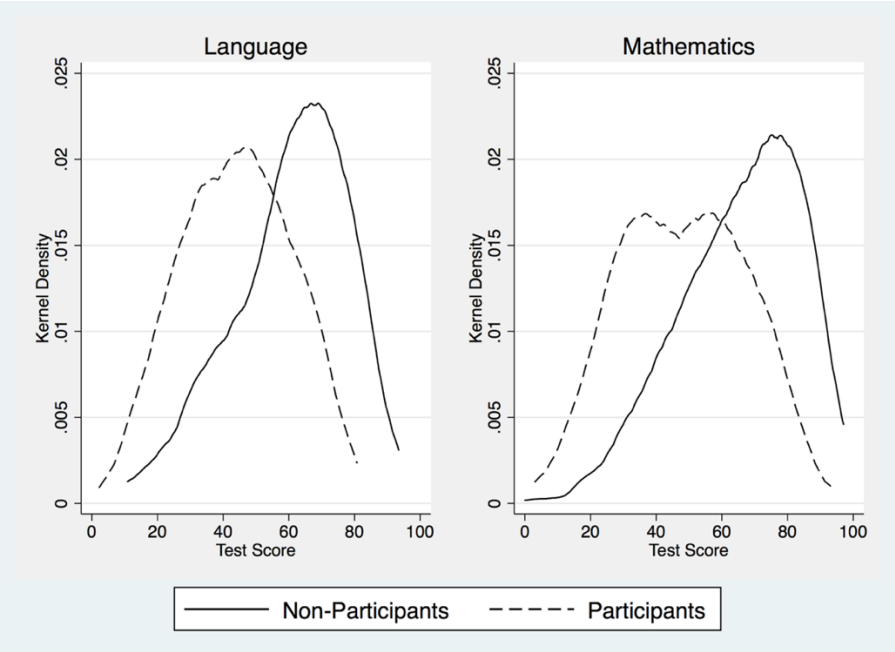
Lastly, limiting the sample to isiXhosa learners, the estimated treatment effects with Nearest Neighbour matching were small: using endline and differenced test scores as the outcomes, effects calculated were 9.7% and 5%, respectively. However, none of those were statistically different from zero. On the other hand, when using Kernel matching the estimated impact computed on endline test scores was 16.4%, which was statistically significant at 1%. A smaller treatment effect of 11.5% was estimated when using differenced scores as the outcome variable. This estimate was significant at the 5% level of significance. In comparison, treatment effects estimated under difference-in-differences using this sub-sample of learners were much higher and amounted to 21.5% and 24.1% as per Table 5 and Table 7, respectively.

In general, the treatment effects of Shine estimated with propensity score matching as a robustness check were more conservative in size when compared to the difference-in-differences computation. Nevertheless, it did provide some certainty that the impact of Shine Literacy on test scores in general can be expected to start from around 9%. This, entails an increase in standard deviations of at least 0.6 and thereby is an effect size that falls under the medium range as per Cohen's *d* (Cohen, 1992). Furthermore, though the construction of comparison groups for the "At risk" learners was not ideal, it could be confirmed that isiXhosa learners appeared to be the principal beneficiaries of the programme.

Western Cape Systemics: when good literacy predicts later mathematics

By merging the Shine test score and attendance data with the WCED Systemic test score datasets, it was possible to account for a potentially more long-term effect of Shine Literacy on Language or Mathematics tests, which are held in Grade 3. Again, the data was compiled with the knowledge in mind that the way in which the Systemics were assessed and the content that it assessed had remained constant over the relevant time period. Circa one year after the Shine programme started in each cohort, all learners in participating schools were also writing the Systemic tests. Using the CEMIS identifier, a total of 727 learners from the Shine data were identified in the Systemic datasets. Those learners were almost equally split between participants and non-participants: 365 vs 362. To get a first impression, Figure 5 depicts how Shine participants performed against non-participants in the Language and Mathematics Systemics

Figure 5: Systemic test score distribution by participation



Undeniably, non-participants test score distribution was centred around a higher mean than participants', and non-participants' distribution of test scores was slightly more skewed[56]. Non-participants scored an average of 60.6% in Language, compared to 43.7% scored by Shine participants. In Mathematics, scores were slightly higher, generally: non-participants scored 65.6% vs. 48.1% scored by participants. Both these differences are highly statistically significant. Additional test score differences in the Language and Mathematics Systemic test scores by participation and by the available demographic information are displayed in Table 12.

[56] Language Non-Participant: Skewness: -0.5757118; Kurtosis: 2.806082, Participants: Skewness: -0.0331736, Kurtosis: 2.28982; Mathematics Non-Participant: Skewness: -0.5989651; Kurtosis: 2.854932, Participants: Skewness: 0.0002325, Kurtosis: 2.239663.

Table 12: Systemic Score differences by treatment

		Language			Mathematics		
		Participants (y)	Non-participants (x)	Difference (x-y)	Participants (y)	Non-participants (x)	Difference (x-y)
All		44.1	60.6	16.5***	48.1	65.6	17.5***
Gender	female	46.7	63.2	16.5***	49	66.6	17.5***
	male	42.1	56.5	14.5***	47.4	64.1	16.7***
Language	IsiXhosa	41.2	58.7	17.5***	46	64.4	18.3***
	English	49.6	64.4	14.8***	52.6	68.2	15.6***
	French	51.1	63.6	12.5	52	62.1	10.1
	Afrikaans	43.7	34.3	-9.4	37.1	62.2	25.1
	Other	49.2	64.1	14.9***	51.5	68.4	16.9***
School	Claremont Primary	52.5	66.8	14.3***	59.8	75.6	15.7***
	Good Hope Seminary Junior	52.6	70.6	18***	59	79.3	20.3***
	Observatory Junior	42.1	52	9.9	45.1	58.8	13.6**
	Prestwich Street Prim.	43.4	60	16.6***	44.1	67.1	22.9***
	Rosmead Central Prim.	44.3	59.7	15.4***	53.3	63.5	10.2**
	St. Agnes's Prim.	47.2	62	14.8***	50	66.2	16.2***
	St. Paul's Prim. (Wynb)	47.5	70.3	22.8***	49.1	69.1	19.9***
	Walmer Estate Prim.	33	48.4	15.4***	38.2	55.4	17.2***
	Zonnebloem Boys Prim.	35.7	49	13.3***	40.3	52.1	11.7***
	Zonnebloem Girls Prac. Sch.	44.7	60.8	16.1***	50.9	62.9	12***
Baseline Literacy Cat	At risk	37.8	29.3	-8.6	39.9	18.4	-21.5**
	Poor	44.8	59.5	14.7**	49.5	59	10
	Satisfactory	47.4	54,8	7.4***	53.5	58.9	5.4**
	Good	57.2	63.8	6.5	58.4	69.7	11.3***

*** p<0.01, ** p<0.05, * p<0.1

As per Table 12, girl learners did slightly better in Language and Mathematics than boys, both as participants and non-participants. The differences in test scores by participation for both genders are highly statistically significant at 1%. In terms of home language, the largest difference between non-participants and participants for the Language Systemics test was seen for isiXhosa speakers, who scored 17.5% higher as non-participants, which is also statistically significant at 1%. Similarly, isiXhosa non-participants scored 18.3% higher than their participant counterparts in Mathematics. Learners who spoke English at home, scored 17.5% higher as non-participants in Language, and 18.3% higher in Mathematics; this difference is highly statistically significant, as well. Speakers of other (African) languages also scored higher as non-participants: for Language and Mathematics by 14.9% and 16.9%, respectively. By the same token, French non-participants scored higher, yet, neither the differences for Language or Mathematics are statistically different from zero. Lastly, though Afrikaans non-participants scored 25.1% higher in Mathematics, but Afrikaans participants scored 9.4% in Language, neither difference is statistically significant at any of the conventional levels.

Before being able to comment on the test score differences by literacy category assigned at baseline, a necessary caveat is the sample sizes across learner's literacy ranking[57]. But, a potentially interesting dynamic showed up: "At risk" Shine participants scored 21.5% higher in Mathematics than their non-participant counterparts. This difference is statistically significant at 5%. In addition, this same group also scored higher in Language: a difference of 8.6%. However, this is statistically not different from zero. Moreover, "Poor" non-participants got statistically higher marks in Language (14.7%), yet not in Mathematics. In the case of "Satisfactory" learners, learners who were not assigned to Shine did better both in Language and Mathematics scoring 7.4% ($p<0.01$) and 5.4% ($p<0.05$), respectively. Lastly, while "Good" non-participants did better (11.3%) in Mathematics, the score difference of 6.5% for Language Systemics is not statistically different from zero. Thus, again, the problem of the counterfactual arises: how would Shine learners have done were they not part of the programme, and most importantly, did Shine make sure that they did not do worse? Regardless of the sample size caveat, these results are certainly encouraging, considering how far Shine learners lagged behind in terms of Shine's baseline diagnostic one year earlier.

Systemics: estimating the counterfactual with OLS and propensity scores

While Shine learners did not do as well in the Systemic tests as their non-participant counterparts, the real objective of the following analysis was to model the counterfactual of Shine learners, i.e. to see how they would fare if they were not part of Shine in the first place. To get a first impression of how this would have unravelled in practise, the following model was estimated via OLS:

$$\text{Systemic test score}_i=\alpha+\beta S_i+\gamma\ \text{Shine Endline Score}_i+\delta(S_i*\ \text{Shine Endline Score}_i)+\mu\ \text{Shine Attendance}_i+\varphi x_i+\varepsilon_i \quad \textbf{(4)}$$

where

β = measures the Shine treatment effect.

γ = captures the impact of Shine endline test scores.

δ = measures the effect of the interaction between Shine treatment and Shine endline test scores.

μ= measures the effect of Shine attendance.

φ= captures the effect of a vector of independent characteristics, including school dummies.

ε= as an idiosyncratic, unobserved error term.

By including Shine endline scores and the interaction between Shine treatment and endline scores, it was possible to approximate the total effect Shine had on participant learners' Systemic test scores. Given the above insights on performance by baseline literacy ranking, the analysis also differentiated by learner sub-samples to see whether there was a differential impact based on literacy category assigned at baseline. In addition, attendance indicators were included, too. The results displayed in Table 13 and 14, which inspect Language and Mathematics separately, differentiate between the general learner population, "At risk" and

[57] "At risk" non-participants: 4, participants: 95; "Poor" non-participants: 6, participants: 185; "Satisfactory" non-participants: 112, participants: 67; "Good" non-participants: 240, participants: 18.

"Poor", as well as "Satisfactory" and "Good" learners, as indicated by the column headings. Both analyses included the additional controls of age and robust standard errors were clustered at the schools.

Table 13: Impact of Shine on Language Systemic Test scores

	(1) All	(2) At Risk & Poor	(3) Sat & Good	(4) All	(5) At Risk & Poor	(6) Sat & Good	(7) All	(8) At Risk & Poor	(9) Sat & Good	(10) At Risk & Poor	(11) Sat & Good
Shine	2.183	-7.028	34.82	2.183	-7.028	34.82					
	(9.528)	(8.666)	(31.09)	(6.957)	(6.987)	(27.75)					
Shine Endline Score	0.73***	0.62***	1.11***	0.73***	0.622***	1.11***	0.75***	0.73***	0.97***	0.705***	0.924***
	(0.106)	(0.132)	(0.263)	(0.126)	(0.119)	(0.274)	(0.129)	(0.122)	(0.240)	(0.109)	(0.202)
Shine X Endline Score	-0.182	-0.101	-0.578	-0.182*	-0.101	-0.578	-0.1***	-0.169	-0.0677	-0.202**	-0.0741
	(0.129)	(0.154)	(0.416)	(0.0990)	(0.158)	(0.356)	(0.033)	(0.123)	(0.0740)	(0.0891)	(0.0470)
Attendance							70.36	23.36	477.7***		
							(69.14)	(58.21)	(139.2)		
Attendance squared							-83.16	-24.72	-610***		
							(85.71)	(69.70)	(179.8)		
Attendance X Endline Score							-1.195	-0.470	-6.72***		
							(1.043)	(0.945)	(2.053)		
Attendance squared X Endline Score							1.349	0.449	8.472***		
							(1.263)	(1.097)	(2.525)		
Learner's age	-2.042	-19.76	24.47	-2.042	-19.76	24.47				-18.51	25.44
	(16.76)	(19.08)	(33.37)	(23.20)	(32.03)	(27.77)				(19.06)	(35.47)
Learner's age squared	-0.155	0.899	-1.739	-0.155	0.899	-1.739				0.834	-1.802
	(0.947)	(1.052)	(1.942)	(1.276)	(1.771)	(1.630)				(1.053)	(2.056)
Attendance >60										-0.458	-3.728
										(2.073)	(3.806)
Constant	33.60	118.2	-106.1	33.60	118.2	-106.1	3.660	7.997	-13.55	106.8	-95.46
	(75.04)	(87.65)	(143.9)	(107.5)	(147.2)	(107.3)	(10.16)	(9.347)	(19.02)	(85.84)	(154.0)
School cluster				YES	YES	YES	YES	YES	YES	YES	YES
Observations	727	290	437	727	290	437	727	290	437	290	437
R-squared	0.292	0.174	0.145	0.292	0.174	0.145	0.279	0.158	0.130	0.173	0.142

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In the first six columns reported in Table 13, neither of the treatment coefficients is statistically different from zero. Merely Shine endline scores appeared to be statistically significant ($p<0.01$) and explained between 0.6 – 1.1% of Language Systemics. This is as expected, since better endline literacy should explain better Language test results. Further, learners' age showed the expected quadratic form (with the exception of "Satisfactory" and "Good" learners) but was never statistically significant. In the case of all learners as reported in the fourth column, the coefficient on Shine treatment and endline score interaction is -0.182% and statistically significant at 10%. Including the effects reported as per the treatment coefficient results in a predicted score increase of approximately 2%. As reported in column 7 to 9, attendance had diminishing marginal returns generally, however, remained statistically not different from zero for the "At risk" and "Poor" sub-samples. For "Satisfactory" and "Good" learners, attendance interacted with endline scores was statistically significant at 1% and explained roughly 10% score increases. Moreover, the interaction between treatment and Shine endline score was statistically significant ($p<0.1$) for "At Risk" and "Poor"

learners and accounted for a decrease of roughly 0.2%. Including a dummy for attending at least 60% of Shine sessions as per column 10 and 11, left Shine endline scores statistically significant, however, the coefficient on this minimum attendance dummy was statistically not different from zero. With only this list of results, Shine appeared to have minimal (negative) to no effect on Language Systemics.

The potential predictive validity of improved literacy on numeracy as assessed in the Grade 3 WCED Systemic Mathematics tests can be observed in Table 14. In the same way as Table 13 above, the general learner sample, as well as baseline literacy ranking restrictions are imposed on the sample of interest, as per column headings.

Table 14: Impact of Shine on Mathematics Systemic Test score

	(1) All	(2) At Risk & Poor	(3) Sat & Good	(4) All	(5) At Risk & Poor	(6) Sat & Good	(7) All	(8) At Risk & Poor	(9) Sat & Good	(10) At Risk & Poor	(11) Sat & Good
Shine	12.45	3.412	54.45*	12.45	3.412	54.45*					
	(9.934)	(8.829)	(31.91)	(10.39)	(11.99)	(26.9)					
Shine Endline Score	0.934***	0.771***	1.105***	0.934***	0.771***	1.1**	0.82***	0.67***	1.04***	0.7***	0.78***
	(0.109)	(0.151)	(0.332)	(0.0701)	(0.134)	(0.341)	(0.0627)	(0.163)	(0.285)	(0.127)	(0.218)
Shine X Endline Score	-0.33**	-0.165	-0.85**	-0.326*	-0.165	-0.8**	-0.1***	-0.0513	-0.0376	-0.096	-0.104*
	(0.135)	(0.175)	(0.423)	(0.155)	(0.150)	(0.368)	(0.0257)	(0.0836)	(0.0853)	(0.103)	(0.0561)
Attendance							50.80	25.89	138.2		
							(75.08)	(87.12)	(189.2)		
Attendance squared							-63.10	-41.98	-25.79		
							(91.18)	(104.0)	(228.5)		
Attendance X Endline Score							-1.040	-0.649	-2.402		
							(1.109)	(1.336)	(2.746)		
Attendance squared X Endline Score							1.187	0.850	0.892		
							(1.339)	(1.588)	(3.213)		
Learner's age	-14.93	-51.08*	26.20	-14.93	-51.08	26.20				-49.2*	22.62
	(23.77)	(27.67)	(35.23)	(20.36)	(28.21)	(28.85)				(28.62)	(36.17)
Learner's age squared	0.701	2.673*	-1.556	0.701	2.673	-1.556				2.564	-1.373
	(1.365)	(1.575)	(2.053)	(1.126)	(1.589)	(1.642)				(1.632)	(2.101)
Attendance >60										-1.739	-1.612
										(2.230)	(4.663)
Constant	70.94	243.4**	-128.2	70.94	243.4*	-128.2	3.392	9.250	-13.54	238.6*	-86.77
	(103.4)	(121.9)	(152.2)	(91.64)	(120.4)	(108.1)	(5.446)	(8.386)	(22.42)	(124.8)	(156.9)
School cluster				YES	YES	YES	YES	YES	YES	YES	YES
Observations	727	290	437	727	290	437	727	290	437	290	437
R-squared	0.288	0.184	0.114	0.288	0.184	0.114	0.285	0.168	0.117	0.185	0.106

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

As reported in the first column, which considers the entire learner population identified in the Systemics dataset, the coefficient on the Shine treatment is not statistically significant at any conventional levels, but the combined effect of Shine and Shine endline scores is significant at 5%. This resulted in a combined estimated effect of roughly 12%. In the case of "At risk" and "Poor" learners, age is significant at 10% as shown in column 5. When considering specifically "Satisfactory" and "Good" learners as per column 3,

those who were assigned to the programme did better in their Mathematics Systemics by around 53%. This large effect stayed the same with clustered standard errors at the school level, and the coefficient remained statistically significant at 10%. Regarding attendance, the expected increasing returns with increasing attendance took place, however, none of the estimated effects were statistically significant at any of the conventional levels.

Given that learners were not randomly assigned to the Shine programme, the above OLS procedures wrongly assumed internal validity after random assignment. Hence, the results need to be considered as "naïve". On the other hand, a difference-in-differences estimator would rely on the data to be set up in a panel structure, where learners would write the same Systemics before and after the programme. This, of course, was not available. Another option to explore, then, was to see whether Shine had a treatment effect on Grade 3 Systemics via propensity score matching. Given that based on this score the sample was reasonably comparable, it was possible to mimic a counterfactual situation. Using the same propensity score as predicted by baseline scores and age earlier, resulted in average treatment effects of Shine as reported in Table 15. Again, the rather strict results of the algorithm used under Nearest Neighbour were compared to Kernel matching, both for Language and Mathematics.

Table 15: Average treatment effects of Shine on Systemics with propensity score matching

Language Systemics	n= Shine	n= Constructed Comparison	Average Treatment Effect
Nearest Neighbour Matching	365	69	- 14.7 (6.3)**
Kernel Matching	365	361	- 4.9 (2.9)*
Mathematics			
Nearest Neighbour Matching	365	69	5.2 (7.5)
Kernel Matching	365	361	6.3 (2.2)***

Note: for Nearest Neighbour the numbers of treated and controls refer to actual nearest neighbour matches; Kernel Matching: Bandwidth: 0.6, 5 replications bootstrapped standard errors; *** p<0.01, ** p<0.05, * p<0.1

Again, the differences between the Language and Mathematics results proved interesting: if any effect, Shine seemed to have a stronger (and positive) impact on Mathematics than on Language. Though the positive effect as estimated by Nearest Neighbour matching was statistically not different from zero, the estimation of the average treatment effect under Kernel matching proved insightful. Firstly, it increased the size of the comparison group drastically from only 69 learners to 361, and its effect is statistically significant at 1%. Under this scenario, Shine learners' test score in Mathematics were 6.3% than non-participant learners.

On the other hand, the treatment effects of Shine on Language Systemics is negative. However, these should be viewed with caution. The method of Nearest Neighbour restricts the available comparison group sample quite immensely down to 69 learners. Further, the treatment effect of -4.9% as estimated under Kernel is significant at 10% only. Nevertheless, this does give some room for concern on how learners' performance in Language Systemics is influenced by attending Shine Literacy. It would be reasonable to assume that Language were positively affected by better Literacy, but it might not be the case here.

Shine Literacy as the South African counterpart to Shishuvachan classes

Volunteering in education programmes has increased momentum over the last two decades[58] and it does not only seem to be causing "warm glow" among volunteers, but also holds educational promises for the learners. Shine Literacy's method that is largely based on interactions between volunteers and learners serves as a great example – and it seems to be working well. Across educational interventions reviewed in Chapter 1, the impact Shine Literacy has on its learners' performance is much greater in magnitude. Complex education programmes such as conditional cash grants as evaluated by Baird et al. (2011), hiring additional contract teachers as analysed by Duflo et al. (2015) or providing school vouchers as examined in Angrist et al. (2006) resulted in small standard deviation improvements of between 0.12 to 0.29. If looking more specifically at literacy programmes, the large standard deviation upwards of 1.7 standard deviations observed for the Shine Literacy case is unparalleled. Teachers using print referencing techniques as evaluated by Justice et al. (2009) or applying ICT methods in the classroom as examined in Brooks et al. (2006) caused medium sized effect changes of around 0.5 standard deviations in participant learners' literacy. In contrast, even if using one of the more conservative estimates as estimated under propensity score matching, Shine Literacy can almost compete with the famous *Shishuvachan* classes as evaluated by He et al. (2009). Children assigned to the *Shishuvachan* sessions improved their scores by up to 0.7 standard deviations – Shine, at the very least, increases learners' literacy by 0.6 standard deviations. Could it be that Shine is the South African success equivalent to the Indian *Shishuvachan* classes?

Similar to the other Indian-based results in trained volunteer interventions as evaluated in Banerjee et al. (2007) and Lakshminarayana et al. (2013), Shine Literacy can attest to the potential that holds in recruiting an army of volunteers. Similarly to Banerjee et al. (2007), this analysis has shown that the lowest performers have benefitted the greatest from attending Shine. One of its most laudable results is that Shine Literacy aids those at the bottom the most: only 12.6% of those who were considered "At risk" at baseline stayed in that category by endline. In fact, 23.7% moved up to "Satisfactory" and 18.3% to "Good" literacy. In the difference-in-differences results, a programme impact of 20.8% for "At risk" learners was detected. In the same vein, only 9% of those who scored "Poor" before the programme were still considered "Poor" afterwards. The remaining fractions moved either up to "Satisfactory" (37.5%) and the large majority of 53.3% increased by two literacy rankings to "Good". In contrast, those who were considered "At risk" but were not assigned to Shine stayed there 81% of the time – they stayed stuck.

In effect, this means that Shine is very efficient in its focus on the learners that need attention the most. As opposed to education programmes that only serve high-achievers such as information campaigns (Jensen, 2010) or the provision of textbooks (Glewwe et al., 2009), Shine Literacy sets the focus on pulling those learners that struggle most out of a potentially never ending literacy deficiency. The underlying goal is to

[58] see for instance, Wasik (1998); Elliott et al. (2000) or Porter & Johnson (2004).

put them on good footing for later school success. Most importantly, enabling learners' literacy to now be on par with what is expected from them later by Grade 3 and Grade 4 ensures that learners can proceed on their learning journey, and do not get left behind early on. In effect, learners do not need to exhaust their working memory becoming frustrated in trying to learn how to read, but can now read to learn.

Moreover, Shine Literacy does not only ensure that bottom-end learners increase their mean performance, but also that the distribution of test scores across the learner population becomes much tighter. The shifts in test score distribution before and after the intervention are immense, moving the combined mean of participants from 44.4% at baseline, to 66.4% at endline. Though non-Shine learners do better in literacy on average, those learners that were far away from such literacy levels before they attended the programme are coming closer to similar literacy proficiency afterwards. In practise, this means that instruction in the classroom can be streamlined more easily, and thereby making the lives of already overrun and tired teachers much easier. In that way, teachers are enabled to focus on the content that they are teaching, instead of trying to make sure that every learner is understanding written text or instructions. This becomes a particularly meaningful insight when considering that just employing more teachers and thereby changing pupil-teacher ratios may not actually have any effect on learners' performance as indicated by the null (or negative) results found in Chin (2005), Urquiola (2006) and Duflo et al. (2015). In contrast, taking learners that seem to struggle with their reading literacy out of the classroom and giving them the opportunity to work on their individual shortfalls one-on-one with a dedicated volunteer proves to be quite effective in really improving performance.

Those interested in examining gender equity effects in schooling outcomes will be unsurprised that boy learners were more likely to be programme participants due to their lower literacy performance. However, they will also be excited to see that it seemed that boy children benefitted just as much as girl children when part of the programme. That is, Shine takes a step towards closing the achievement gap[59] between girls and boys when it comes to literacy. Moreover, bearing in mind that all schools examined in this analysis use exclusively English as the Language of Teaching and Learning, the immense literacy gains undergone by isiXhosa speakers are also indicative of Shine's success. Those learners who switch between speaking isiXhosa at home to speaking English at school experienced literacy improvements of 24.1%. This also implies that Shine successfully supports learners that do not speak English as their first language and also improves their language proficiency – a large, and important stepping stone to better reading ability (Pretorius & Spaull, 2016).

When looking at the results of the Western Cape Systemic results, thought-provoking dynamics unravelled. As expected, non-participant fared better generally in the tests than their Shine counterparts. However, the

[59] The reader interested in an assessment on the gender achievement gaps for both mathematics and reading is referred to Robinson & Lubienski (2011). Here, it is explored to what extend the reading gaps between boys and girls exist in the US context, and how this gap widens specifically among those at the bottom end of the performance distribution.

impact Shine may have had on Mathematics performance points to interesting spillover effects. As opposed to Abeberese et al. (2014), who found no effects of the *Sa Aklat Sisikat* reading marathon in the Philippines on other subjects including math skills, Shine may have made a difference in learners' numeracy. Improved literacy due to attending Shine increased Systemic Mathematics by 6.3%, an effect size that was highly statistically significant under Kernel matching.

Reading skills hold the promise of being applicable across disciplines, and being able to read and write equips learners with skills that are not only associated with being able to read written text (Jerman & Mirman, 1974; Caponera et al., 2016; Neufeld, 2005). It has been shown that better readers are also more likely to be better at effectively filtering out relevant information and managing their time during tests better (Caponera et al., 2016). The strong positive influence that improved literacy can have on math performance was shown for the Italian case in Caponera et al. (2016) using PIRLS data. Using Shine data, it was shown that improved literacy implied predictive validity on numerical skills as assessed in the Grade 3 Systemics. In contrast, when not assigned to Shine but considered "At risk" during the baseline assessments, puts learners on bad footing when it comes to Mathematics: Shine "At risk" learners performed better by 21.5% that their non-participant counterparts.

Keep shining bright: future research avenues and limitations

As with any quasi-experimental impact evaluation of an education programme, the current analysis aimed to determine the empirical association between participation in the Shine Literacy programme and learners' educational success. The effectiveness of Shine's support via trained volunteers is documented in this dissertation. However, it is important to remember that assignment to Shine Literacy is not randomised, forcing us to rely on other econometric techniques to try and assess impact. As with many educational programmes that are evaluated in quasi-experimental studies, there are various factors, observed and unobserved, that unravel in the real world that econometric methods simply cannot account for. For example, not much is known about the learner population's innate ability, the factors they are exposed to at home, or even an indicator of how vested their parents are in their children's literacy. From Heckman (2006) and Heckman & Masterov (2004), it is known how important parental efficacy is in learners' academic development. Therefore, accounting for parental investment would be an interesting avenue to explore. Nevertheless, both the difference-in-differences and propensity score matching methods pointed to a strong and positive effect of Shine Literacy on learner outcomes. In the context of potential scale up, these effect sizes are educationally speaking very meaningful.

The available data context pertains to ten schools in the Western Cape. As such, external validity is quite limited. Regarding Systemic test score results, the fact that only 727 out of 1542 learners were identified, limited the sample size drastically. This points to issues with many learners not having been identifiable with the CEMIS indicator and should be rectified.

Several future avenues for evaluative work on Shine Literacy can be imagined. Especially with the similarity in success to the *Shishuvachan* classes in mind, it could be interesting to see how Shine Centres compare to their skeleton model counterparts: the Shine Chapters. These are set up as independently operated community franchises, that deliver the Shine method. It is Shine's own way of scaling up their model and making it more accessible and inexpensive to more parts of South Africa (Shine Literacy, 2016). Currently, there are chapters in the Western Cape, the Eastern Cape and KwaZulu-Natal. It would be interesting to investigate if the literacy gains are similar to the Shine Centres' success story. In so doing, it could then also be compared in terms of cost-effectiveness. Being able to account for how much a standard deviation improvement actually costs, will helpful when it comes to decision-making on the scale up of Shine across South Africa and receive the dedicated support. Anecdotally, one would assume that Shine does not need a lot more other than a specified learning space, some books and a dedicated volunteer.

There are additional interesting aspects to explore in future work. By itself, children's interaction with the enthusiastic adult volunteers has been identified as a great driver of learners' confidence and self-esteem. As put by Lufefe:

> *"I was too shy to speak. My Shine volunteer helped me to open up. It was scary at first, but then it was good. It's nice to know there are people who care about you. Leigh-Anne was my volunteer. She was wonderful."*
>
> (Shine Literacy, 2016: 12)

Based on a myriad of anecdotal evidence recounted by the Shine staff and drawn on in their annual reports, participant learners have benefitted immensely by having dedicated, focussed attention been given to them. To turn these testimonials into hard evidence, a behavioural experiment could be designed where learners would be given a small game to play before or after interacting with their volunteer. Increased confidence could be measured by asking children to bet on their future success in that game, similar to the method used in Hoff & Pandey (2014) who figured out a way to put an index on learner confidence. Of course, given that in Hoff & Pandey (2014) slightly older learners were participating, this experiment would have to be adapted to the early primary school case.

Similarly, given that Shine relies largely on its roster of volunteers, it would be interesting to explore what drives individuals to volunteer for Shine. By identifying which drivers function as motivators, it may become easier to target potential volunteers and grow Shine's volunteer roster. Lastly, it would be interesting to see whether the impact on Mathematics persists beyond Grade 3.

Whilst these are all useful expansions, this dissertation has provided some sound evidence on the effectiveness of Shine in assisting learners that were deficient with their literacy. At baseline, Shine learners were a universe away from the rest of their cohort – at endline, this gap is greatly reduced. Perhaps, this is how *Words Can Change Worlds*.

References:

Abeberese, A.B., Kumler, T. & Linden, L. 2014. Improving reading skills by encouraging children to read: A randomized evaluation of the Sa Aklat Siskat reading program in the Phillippines. *Journal of Human Resources*. 49 (3): 611–633.

Adrogue, C. & Orlicki M.E. Do In-School Feeding Programs Have an Impact on Academic Performance? The Case of Public Schools in Argentina. *Education Eolicy Analysis Archive*. 21 (50): 1 - 23.

Alderman, H., Kim, J. & Orazem, P.F. 2003. Design, evaluation, and sustainability of private schools for the poor: The Pakistan urban and rural fellowship school experiments. *Economics of Education Review*. 22 (2003): 265–274.

Angrist, J., Bettinger, E., Bloom, E., King, E. & Kremer, M. 2002. Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. *The American Economic Review*. 92 (5): 1535–1558.

Angrist, J., Bettinger, E. & Kremer, M. 2006. Long-Term Educational Consequences of Secondary School Vouchers Evidence from Administrative Records in Colombia. *The American Economic Review*. 96 (3): 847–862.

Baez, J.E. & Camacho, A. 2011. *Assessing the long-term effects of conditional cash transfers on human capital: Evidence from Colombia*. Discussion Paper Series Forschungsinstitut zur Zukunft der Arbeit, Number. 5751. Bonn, Germany.

Baird, S., Mcintosh, C. & Ozler, B. 2011. Cash Or Condition ? Evidence From A Cash Transfer Experiment. *The Quarterly Journal of Economics*. 126(4):1709–1753.

Baird, S., Hicks, J.H. Kremer, M & Miguel, E. 2015. Worms at work: Long-run impacts of a child health investment. UC Berkeley. Berkeley, CA.

Banerjee, A. V, Cole, S., Duflo, E. & Linden, L.L. 2007. Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*. 122 (3): 1235–1264.

Banerjee, A. V, Banerji, R., Duflo, E., Glennerster, R. & Khemani, S. 2010. Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. *American Economic Journal: Economic Policy*. 2 (1): 1–30.

Banerjee, A. V, Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M. & Walton, M. 2016. *Mainstreaming an Effective Intervention: Evidence From Randomized Evaluations of "Teaching At the Right Level" in India*. National Bureau Of Economic Research Working Paper Series Number 22746. Cambridge, MA.

Barham, T., Macours, K. & Maluccio, J.A. 2013. *More Schooling and More Learning? Effects of a Three-Year Conditional Cash Transfer Program after 10 Years*. Inter-American Development Bank Working Paper Series, Number 432. Washington, D.C.

Barrera-Osorio, F. & Filmer, D. 2016. Incentivizing Schooling for Learning: Evidence on the Impact of Alternative Targeting Approaches. *Journal of Human Resources*. 51 (2): 461–499.

Barrera-Osorio, F. & Linden, L.L. 2009. *The use and misuse of computers in education: Evidence from a randomized experiment in Colombia*. The World Bank, Policy Research Working Paper Number 4836, Impact Evaluation Series No. 29 Washington, D.C.

Barrera-Osorio, F., Bertrand, M., Linden, L.L. & Perez-Calle, F. 2011. Improving the Design of Conditional Cash Transfer Programs: Evidence from a Randomized Evaluation in Colombia Organ donations. *American Economic Journal: Applied Economics*. 3 (April): 167–195.

Behrman, J.R., Parker, S.W., Todd, P.E., Behrman, J.R., Parker, S.W. & Todd, P.E. 2011. Do Conditional Cash Transfers for Schooling Generate Lasting Benefits? A Five-Year Followup of PROGRESA/Oportunidades. *The Journal of Human Resources*. 46 (1): 93–122.

Bellei, C. 2009. Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile. *Economics of Education Review*. 28 (5): 629-640.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P. & Pouliquen, V. 2015. Turning a Shove into a Nudge?

A "Labeled Cash Transfer" for Education. *American Economic Journal: Economic Policy*. 7 (3): 86–125.
Beuermann, D., Cristia, J., Cruz Aguayo, Y., Cueto, S. & Malamud, O. 2013. *Home computers and child outcomes: short term impacts from a randomized experiment in Perú*. NATIONAL BUREAU OF ECONOMIC RESEARCH Working Paper Series Number 18818. Cambridge, MA.
Blimpo, M.P. 2014. Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin. *American Economic Journal: Applied Economics*. 6 (4): 90-109.
Borkum, E., He, F. & Linden, L.L. 2012. *School Libraries and Language Skills in Indian Primary Schools: A Randomized Evaluation of the Akshara Library Program*. JPAL Working Paper April 2012. MIT. Boston, MA.
Bowles, S. 1970. Education and Production Functions. In *Education, Income, and Human Capital*. V. I. W.L. Hansen, Ed. National Bureau of Economic Research. 11–70.
Brooks, G., Miles, J., Torgerson, C.J. & Torgerson, D.J. 2006. Is an intervention using computer software effective in literacy learning? A randomised controlled trial. *Educational Studies*. 32 (2): 133–143.
Burde, D. & Linden, L.L. 2013. Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. *American Economic Journal: Applied Economics*. 5 (3): 27-40.
Caponera, E., Sestito, P. & Russo, P.M. 2016. The influence of reading literacy on mathematics and science achievement. *Journal of Educational Research*. 109 (2): 197–204.
Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K. & Rogers, F.H. 2007. Missing in Action: Teacher and Health Worker Absence in Developing Countries. *The Journal of Economic Perspectives*. 20 (1): 91–116.
Chen, Y. & Jin, G.Z. 2012. Does health insurance coverage lead to better health and educational outcomes? Evidence from rural China. *Journal of Health Economics*. 31 (1): 1–14.
Chin, A. 2005. Can redistributing teachers across schools raise educational attainment? Evidence from Operation Blackboard in India. *Journal of Development Economics*. 78 (2): 384–405.
Cohen, J. 1992. A Power Primer. *Psychological Bulletin*. 112 (1): 155–159.
Contreras, D. & Rau, T. 2012. Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile. *Economic Development and Cultural Change*. 61 (1): 219–246.
Cristia, J., Czerwonko, A. & Garofalo, P. 2014. *Does technology in schools affect repetition, dropout and enrollment? Evidence from Peru*. Inter-American Development Bank Working Paper Series, Number 477. Washington, D.C.
Das, B.J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K. & Sundararaman, V. 2013. School Inputs, Household Substitution, and Test Scores. *American Economic Journal: Applied Economics*. 5 (2): 29–57.
Department of Basic Education. 2016. *Education Statistics in South Africa*. Pretoria.
Dhaliwal, I., Duflo, E., Glennerster, R. & Tulloch, C. 2012. Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries. *A General Framework with Applications for Education*. JPAL Working Paper December 2012. MIT. Boston, MA.
Duflo, E. 2001. Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *The American Economic Review*. 91 (4): 795–813.
Duflo, E., Hanna, R. & Ryan, S.P. 2012. Incentives Work: Getting Teachers to Come to School. *The American Economic Review*. 102 (4): 1241–1278.
Duflo, E., Dupas, P. & Kremer, M. 2012. *Teacher management, pupil-teacher ratios, and education quality: Experimental evidence from Kenyan primary schools*. National Bureau of Economic Research Working Paper 17939. Cambridge, MA.
Duflo, E., Dupas, P. & Kremer, M. 2015. School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*. 123 (1): 92–110.
Elliott, J., Arthurs, J. & Williams, R. 2000. Volunteer Support in the Primary Classroom: the long-term impact of one initiative upon children's reading performance. *British Educational Research Journal*. 26 (2): 227–244.

Friedman, W., Kremer, M., Miguel, E. & Thornton, R. 2011. *Education as Liberation*. National Bureau of Economic Research Working Paper Series Education Number 16939. Cambridge, MA.
Galiani, S. & McEwan, P.J. 2013. The heterogeneous impact of conditional cash transfers. *Journal of Public Economics*. 103 (1): 85–96.
Gertler, P., Martinez, S., Premand, P., Rawlings, L.B. & Vermeersch, C.M. 2011. *Impact Evaluation in Practice*. The International Bank for Reconstruction and Development, The World Bank, Washington D.C.
Gertler, P.J., Patrinos, H.A. & Rubio-Codina, M. 2012. Empowering parents to improve education: Evidence from rural Mexico. *Journal of Development Economics*. 99(1):68–79.
Gitter, S.R. & Barham, B.L. 2008. Women's power, conditional cash transfers, and schooling in Nicaragua. *World Bank Economic Review*. 22 (2): 271–290.
Glewwe, P. & Kassouf, A.L. 2012. The impact of the Bolsa Escola/Familia conditional cash transfer program on enrollment, dropout rates and grade promotion in Brazil. *Journal of Development Economics*. 97 (2): 505–517.
Glewwe, P. & Muralidharan, K. 2016. Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications. In *Handbook of the Economics of Education*. 1st ed. V. 5. Elsevier B.V. 653–743.
Glewwe, P., Kremer, M., Moulin, S. & Zitzewitz, E. 2004. Retrospective vs. prospective analyses of school inputs: The case of flip charts in Kenya. *Journal of Development Economics*. 74 (1): 251–268.
Glewwe, P., Kremer, M. & Moulin, S. 2009. Many Children Left Behind ? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*. 1 (1): 112–135.
Glewwe, P., Ilias, N. & Kremer, M. 2010. Teacher Incentives. *American Economic Journal: Applied Economics*. 2 (3): 205–227.
Glewwe, P., Park, A. & Zhao, M. 2016. A better vision for development: Eyeglasses and academic performance in rural primary schools in China. *Journal of Development Economics*. 122 (1): 170–182.
Grigg, D., Joffe, J., Okeyo, A., Schkolne, D., van der Merwe, N., Zuma, M., Mulenga, C., Boodhoo, A. 2016. The role of monitoring and evaluation in six South African reading programmes. *Southern African Linguistics and Applied Language Studies*. 34 (4): 359–370.
Gulek, J.C. & Demirtas, H. 2005. Learning with Technology: The impact of laptop use on student achievement. *The Journal of Technology, Learning and Assessment*. 3 (2): 3-38.
Handa, S. 2002. Raising Primary School Enrolment in Developing Countries. *Journal of Development Economics*. 69 (2002): 103–128.
He, F., Linden, L.L. & MacLeod, M. 2009. *A better way to teach children to read? Evidence from a randomized controlled trial*. JPAL Working Paper: May 2009. MIT. Boston, MA.
Heckman, J.J. 2006. Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*. 312 (June):1900–1902.
Heckman, J.J. 2011. The economics of inequality: The value of early childhood education. American Educator: 31–36.
Heckman, J.J. & Masterov, D. V. 2004. *The Productivity Argument for Investing in Young Children*. Working Paper 5, Invest in Kids Working Group Committee for Economic Development.
Heckman, J., Humphries, J.E. & Veramendi, G. 2016. Returns to Education: The Causal Effects of Education on Earnings, Health and Smoking. National Bureau of Economic Research Working Paper Series Number 22291. Cambridge, MA.
Hidalgo, D., Onofa, M., Oosterbeek, H. & Ponce, J. 2013. Can provision of free school uniforms harm attendance? Evidence from Ecuador. *Journal of Development Economics*. 103 (1): 43–51.
Ho, A.D. & Yu, C.C. 2015. Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects. *Educational and Psychological Measurement*. 75 (3): 365–388.
Hoff, K. & Pandey, P. 2014. Making up people - The effect of identity on performance in a modernizing society. *Journal of Development Economics*. 106 (1): 118–131.

Howie, S., Venter, E. & van Staden, S. 2008. The effect of multilingual policies on performance and progression in reading literacy in South African primary schools. *Educational Research and Evaluation*. 14 (6): 551–560.
Hsieh, C.T. & Urquiola, M. 2006. The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program. *Journal of Public Economics*. 90 (8–9): 1477–1503.
Jensen, R. 2010. The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics*. 125 (2): 515–548.
Jerman, M.E. & Mirman, S. 1974. Linguistic and computational variables in problem solving in elementary mathematics. *Educational Studies in Mathematics*. 5 (3): 317–362.
Justice, L.M., Kaderavek, J.N., Fan, X., Sofka, A. & Hunt, A. 2009. Accelerating Preschoolers' Early Literacy Development Through Classroom-Based Teacher-Child Storybook Reading and Explicit Print Referencing. *Language, Speech, and Hearing Services in Schools*. 40 (1): 67–86.
Kazianga, H., Levy, D., Linden, L.L. & Sloan, M. 2013. The Effects of " Girl-Friendly " Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. *American Economic Journal: Applied Economics*. 5 (3): 41 - 62.
Khandker, S., B. Koolwal, G. & Samad, H. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. 1st ed. The World Bank. Washington, D.C.
Kim, J.S. & Guryan, J. 2010. The Efficacy of a Voluntary Summer Book Reading Intervention for Low-Income Latino Children from Language Minority Families. *Journal of Educational Psychology*. 102 (1): 20–31.
Kong, S. 2014. Developing information literacy and critical thinking skills through domain knowledge learning in digital classrooms: An experience of practicing flipped classroom strategy. *Computers and Education*. 78 (1): 160-173.
Kremer, M. 2003. Randomized Evaluations of Educational Programs in Developing Countries : Some Lessons. *The American Economic Review*. 93 (2): 102–106.
Kremer, M., Miguel, E. & Thornton, R. 2009. Incentives to Learn. *The Review of Economics and Statistics*. 91 (3): 437–456.
Lai, F., Luo, R., Zhang, L., Huang, X. & Rozelle, S. 2015. Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education Review*. 47 (2015): 34–48.
Lakshminarayana, R., Eble, A., Bhakta, P., Frost, C., Boone, P., Elbourne, D. & Mann, V. (in press). The Support to Rural India's Public Education System (STRIPES) Trial: A Cluster Randomised Controlled Trial of Supplementary Teaching, Learning Material and Material Support. *PLoS ONE*. 8 (7): 1 - 17.
Lance, P., Guilkey, D., Hattori, A. & Angeles, G. 2014. *How Do We Know If a Program Made a Difference? A Guide to Statistical Methods for Program Impact Evaluation*. 1st ed. Chapel Hill, North Carolina: MEASURE Evaluation, University of North Carolina
Lara, B., Mizala, A. & Repetto, A. 2011. The Effectiveness of Private Voucher Education: Evidence From Structural School Switches. *Educational Evaluation and Policy Analysis*. 33 (2): 119–137.
Li, T., Han, L., Zhang, L. & Rozelle, S. 2014. Encouraging classroom peer interactions: Evidence from Chinese migrant schools. *Journal of Public Economics*. 111 (2014): 29–45.
Linden, L.L. 2008. *Complement or Substitute? The Effect of Technology on Student Achievement in India*. JPAL Working Paper June 2008. MIT. Boston, MA.
Linnakylä, P., Malin, A. & Taube, K. 2004. Factors behind low reading literacy achievement. *Scandinavian Journal of Educational Research*. 48 (3): 231–249.
Loyalka, P., Liu, C., Song, Y., Yi, H., Huang, X., Wei, J., Zhang, L., Shi, Y. 2013. Can information and counseling help students from poor rural areas go to high school? Evidence from China. *Journal of Comparative Economics*. 41 (4): 1012–1025.
Lucas, A.M.. & Mbiti, I.M. 2014. Effects of School Quality on Student Achievement: Discontinuity Evidence from Kenya. *American Economic Journal: Applied Economics*. 6 (3): 234–263.
Luo, R., Shi, Y., Zhang, L., Liu, C., Sharbono, B., Yue, A., Zhao, Q. & Martorell, R. 2012. Nutrition and

Educational Performance in Rural China 's Elementary Schools: Results of a Randomized Control Trial in Shaanxi Province. *Economic Development and Cultural Change*. 60 (4): 735–772.
Malamud, O. & Pop-Eleches, C. 2011. Home computer use and the development of human capital. *Quarterly Journal of Economics*. 126 (2): 987–1027.
McCarthy, J. & Oliphant, R. 2013. *Mathematics Outcomes in South African Schools. What are the facts? What should be done?* The Centre for Development and Enterprise. Johannesburg, South Africa.
Miguel, E. & Kremer, M. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*. 72 (1): 159-217.
Mo, D., Zhang, L., Yi, H., Luo, R., Rozelle, S. & Brinton, C. 2013. School Dropouts and Conditional Cash Transfers: Evidence from a Randomised Controlled Trial in Rural China's Junior High Schools. *Journal of Development Studies*. 49 (2): 190–207.
Mo, D., Zhang, L., Wang, J., Huang, W., Shi, Y., Boswell, M. & Rozelle, S. 2014. *The Persistence of Gains in Learning from Computer Assisted Learning (CAL): Evidence from a Randomized Experiment in Rural Schools in Shaanxi Province in China*. REAP China Working Paper 268. Stanford University: San Francisco.
Moloi, M.Q. & Chetty, M. 2011. *Progress in Gender Equality in Education: South Africa*. SACMEQ Policy Brief Number 6 (September).
Mullis, I.V.S., Martin, M.O., Foy, P. & Hooper, M. 2017. *Progress in International Reading Literacy Study 2016: International Results in Reading*. Boston, MA.
Muralidharan, K. 2012. *Long-term effects of teacher performance pay: Experimental evidence from India*. National Bureau Of Economic Research Working Paper Series Number 15323. Cambridge, MA.
Muralidharan, K. & Prakash, N. 2013. *Cycling to School: Increasing Secondary School Enrollment for Girls in India*. Forschungsinstitut zur Zukunft der Arbeit Discussion Paper Number 7585. Bonn, Germany.
Muralidharan, K. & Sundararaman, V. 2010. The Impact Of Diagnostic Feedback To Teachers On Student Learning: Experimental Evidence From India. *The Economic Journal*. 120 (546): F187–F203.
Muralidharan, K. & Sundararaman, V. 2011. Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*. 119 (1): 39–77.
Muralidharan, K. & Sundararaman, V. 2013. *Contract Teachers: Experimental Evidence From India*. National Bureau Of Economic Research Working Paper Series Number 19440. Cambridge, MA.
Muralidharan, K., Das, J., Holla, A. & Mohpal, A. 2017. The fiscal cost of weak governance: Evidence from teacher absence in India. *Journal of Public Economics*. 145 (1): 116–135.
Neufeld, P. 2005. Comprehension Instruction in Content Area Classes. *The Reading Teacher*. 59 (4): 302–312.
Newman, J., Pradhan, M., Rawlings, L.B., Ridder, G., Coa, R. & Evia, J.L. 2002. An impact evaluation of education, health, and water supply investments by the Bolivian Social Investment Fund. *The World Bank Economic Review*. 16 (2): 241–274.
Olken, B.B.A., Onishi, J. & Wong, S. 2014. Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia. *American Economic Journal: Applied Economics*. 6 (4): 1–34.
Orkin, K. (2013). The Effect of Lengthening the School Day on Children's Achievement in Ethiopia. Oxford Department of International Development, University of Oxford. Working Paper 119.
Oster, E. & Thornton, R. 2011. Evaluation Menstruation, Sanitary Products, and School Attendance: Evidence from a Randomized Evaluation. *American Economic Journal: Applied Economics*. 3 (1): 91–100.
Overett, J. & Donald, D. 1998. Paired reading: effects of a parent involvement programme in a disadvantaged community in South Africa. *British Journal of Educational Psychology*. 68(3):347–356.
Ozier, O. 2016. *Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming*. The World Bank Policy Research Working Paper 7052. Washington, D.C.
Pandey, P., Goyal, S. & Sundararaman, V. 2009. Community participation in public schools: impact of information campaigns in three Indian states. *Education Economics*. 17 (3): 355–375.
Piasta, S.B., Justice, L.M., McGinty, A.S. & Kaderavek, J.N. 2012. Increasing Young Children's Contact

With Print During Shared Reading: Longitudinal Effects on Literacy Achievement. *Child Development*. 83 (3): 810–820.
Du Plessis, E., Naude, H. & Viljoen, J. 2003. A conceptual framework for accelerating emergent literacy skills of disadvantaged pre-schoolers in South Africa. *Educare*. 32 (1): 20–35.
Plüddemann, P., Braam, D., October, M. & Wababa, Z. Dual-medium and parallel- medium schooling in the Western Cape: from default to design. PRAESA Occasional Papers Number 17. University of Cape Town. Cape Town, South Africa.
Pop-Eleches, C. & Urquiola, M. 2013. Going to a Better School: Effects and Behavioral Responses. *The American Economic Review*. 103 (4): 1289–1324.
Porter, C.L. & Johnson, J.E. 2004. Parents as Classroom Volunteers and Kindergarten Students' Emergent Reading Skills. *The Journal of Educational Research*. 97 (5): 235–247.
Pretorius, E.J. & Spaull, N. 2016. Exploring relationships between oral reading fluency and reading comprehension amongst English second language readers in South Africa. *Reading and Writing*. 29 (7): 1449–1471.
Pridmore, P. & Jere, C. 2011. Disrupting patterns of educational inequality and disadvantage in Malawi. *Compare: A Journal of Comparative and International Education*. 41 (4): 513–531.
Reinikka, R. & Svensson, J. 2004. Local Capture: Evidence from a Central Government Transfer Program in Uganda. *The Quarterly Journal of Economics*. 119 (2): 679–705.
Robinson, J.P. & Lubienski, S.T. 2011. The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings. *American Educational Research Journal*. 48 (2): 268–302.
Sabarwal, S., Evans, D.K. & Marshak, A. 2014. *The permanent input hypothesis: the case of textbooks and (no) student learning in Sierra Leone*. The World Bank Education Global Practice Group & Africa Region, Policy Research Working Paper 7021.
Salisbury, T. 2016. Education and inequality in South Africa: Returns to schooling in the post-apartheid era. *International Journal of Educational Development*. 46 (1): 43–52.
Schkolne, D.S. 2014. An Outcome Evaluation of the Shine Centre's Literacy Hour Programme. Masters Thesis. School of Management Studies, University of Cape Town. Cape Town, South Africa
Schultz, T.P. 2004. School Subsidies For The Poor: Evaluating The Mexican PROGRESA Poverty Program. *Journal of Development Economics*. 74:199–250.
Shine Literacy. 2016. *Shine Literacy Annual Report 2016*. Wynberg, South Africa.
Slentz, K.L., Early, D.M. & McKenna, M. 2008. A Guide to Assessment in Early Childhood: Infancy to Age Eight. Washington State Office of Superintendent of Public Instruction. Washington, D.C.
Spaull, N. 2011. *Primary school performance in Botswana, Mozambique, Namibia, and South Africa*. SACMEQ Working Paper Number 8.
Spaull, N. 2012. *Poverty & Privilege : Primary School Inequality in South*. (13/12). Maitland, South Africa.
van Steensel, R., McElvany, N., Kurvers, J. & Herppich, S. 2011. How Effective Are Family Literacy Programs? Results of a Meta-Analysis. *Review of Educational Research*. 81 (1): 69–96.
Sylvia, S., Luo, R., Zhang, L., Shi, Y., Medina, A. & Rozelle, S. 2013. Do you get what you pay for with school-based health programs? Evidence from a child nutrition experiment in rural China. *Economics of Education Review*. 37 (2013): 1–12.
Tan, J., Lane, J. & Lassibille, G. 1999. Student Outcomes in Philippine Elementary Schools: An Evaluation of Four Experiments. *The World Bank Economic Review*. 13 (3): 493–508.
The DG Murray Trust. 2012. The Shine Centre: Words can Change Worlds. Learning from our implementing partners. Learning Brief 11. Claremont, South Africa.
The National Institute for Literacy. 2008. *Developing Early Literacy: A Scientific Synthesis of Early Literacy Development and Implications for Intervention*. Jessup, MD.
Urquiola, M. 2006. Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia. *The Review of Economics and Statistics*. 88 (1): 171–177.

Urquiola, M. & Verhoogen, E. 2009. Class-Size Caps, Sorting, and the Regression-Discontinuity Design. *The American Economic Review*. 99 (1): 179–215.
Wang, H., Chu, J., Loyalka, P., Tao, X., Shi, Y., Qu, Q., Yang, C. & Rozelle, S. 2014. *Can School Counseling Reduce School Dropout in Developing Countries?* REAP China Working Paper 275. Stanford University: San Francisco.
Wasik, B. A. 1998. Volunteer Tutoring Programs in Reading: A Review. *Reading Research Quarterly*. 33 (3): 266–291.
The World Bank (2016). *World Development Indicators*. Washington, DC: Oxford University Press. Available: http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators# [2016, November 24].
The World Bank (2018). *World Development Indicators*. Washington, DC: Oxford University Press. Available: http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators# [2018, January 13].
Yang, Y., Zhang, L., Zeng, J., Pang, X., Lai, F. & Rozelle, S. 2013. Computers and the academic performance of elementary school-aged girls in China's poor communities. *Computers and Education*. 60 (1): 335–346.

APPENDIX A: Demographics by School % (n)

School	female	Language					Average age (SD)	Baseline Literacy				Average baseline Score (SD)	Average Endline Score (SD)	Average Language Systemic Score (SD)	Average Mathematics Systemic Score (SD)
		isiXhosa	English	Afrikaans	French	Other		At risk	Poor	Satisfactory	Good				
Claremont Primary	59.4% (73)	30.9% (38)	27.6% (34)	0% (0)	0% (0)	2.4% (3)	8.6 (0.4)	2.4% (3)	19.5% (24)	23.6% (29)	54.5% (67)	66.8% (14.5)	76.8% (5.3)	59.2% (14.5)	67.2% (18.8)
Good Hope Seminary Junior	55.9% (99)	23.2% (41)	4.5% (8)	0% (0)	0% (0)	4% (7)	8.4 (0.5)	10.2% (18)	22.6% (40)	19.8% (35)	47.5% (84)	61.7% (18.8)	73.4% (9)	61.3% (18.2)	68.8% (20.3)
Observatory Junior	50% (34)	61.8% (42)	8.8% (6)	0% (0)	0% (0)	1.5% (1)	8.6 (0.5)	35.3% (24)	35.3% (24)	20.6% (14)	8.8% (6)	42.2% (22.5)	69.3% (12.7)	44.1% (17.5)	47.9% (19.5)
Prestwich Street Prim.	51.1 (135)	44.7% (118)	6.4% (17)	1.5% (4)	0% (0)	5% (12)	8.5 (0.4)	11.7% (31)	25% (66)	25.8% (68)	37.5% (99)	59.4% (18.7)	71.1% (11.5)	52% (17.6)	56.1% (21.1)
Rosmead Central Prim.	(51.6% (82)	25.2% (40)	2.5% (4)	0.6% (1)	7.6% (12)	1.3% (2)	8.6 (0.4)	4.4% (7)	8.8% (14)	25.2% (40)	61.6% (98)	68.5% (14.3)	76% (5.8)	55.3% (18.2)	60.6% (16.9)
St. Agnes's Prim.	54.6% (95)	3.5% (6)	35.1% (61)	0.6% (1)	0% (0)	8.1% (14)	8.5 (0.5)	6.9 (12)	24.3% (44)	33.3% (58)	34.5% (60)	62.3% (16.5)	70.6% (11.2)	54% (20.7)	57.3% (21.6)
St. Paul's Prim. (Wynb)	45.6 %(115)	13.5% (34)	4.8% (12)	0% (0)	0% (0)	1.2% (3)	8.3 (0.5)	24.2% (61)	23.4% (59)	24.2% (61)	28.2% (71)	51.3% (23.9)	63.6% (18.7)	59.1% (17.7)	59.3% (20.8)
Walmer Estate Prim.	47.9 % (46)	36.5% (35)	12.5% (12)	1% (1)	0% (0)	3.1% (3)	8.7 (0.5)	20.8% (20)	28.1% (27)	29.1% (28)	21.9% (21)	52.7% (20.8)	69.9% (11.9)	40.9% (18.5)	47% (19.1)
Zonnebloem Boys Prim.	0% (0)	44.5% (57)	14.1% (18)	0% (0)	0% (0)	2.3% (3)	8.5 (0.6)	24.2% (31)	35.2% (45)	24.2% (31)	16.4% (21)	48.9% (21.7)	69.2% (10.3)	40.5% (18.3)	44.6% (19.4)
Zonnebloem Girls Prac. Sch.	100% (102)	48% (49)	24.5% (25)	0% (0)	0% (0)	2% (2)	8.6 (0.5)	4.9% (5)	20.6% (21)	23.5% (24)	51% (52)	65.2% (15.8)	74.6% (6.9)	55.7% (17.9)	59.1% (17)

APPENDIX B: Probit – Estimating the propensity score of Shine Participation

Dependent variable:	**Shine Treatment**
Baseline Score	-0.0534***
	(0.00234)
Age	0.00172
	(0.0801)
Constant	3.068***
	(0.705)
Observations	1,543

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1