Art. # 933, 12 pages, http://www.sajournalofeducation.co.za

# Large-scale instructional reform in the Global South: insights from the mid-point evaluation of the Gauteng Primary Language and Mathematics Strategy

**Brahm Fleisch**
Division of Education Leadership and Policy Studies, Wits School of Education, University of the Witwatersrand
Brahm.Fleisch@wits.ac.za
**Volker Schöer**
AMERU / School of Economic and Business Sciences, University of the Witwatersrand

This paper reports on a mid-point evaluation of the Gauteng Primary Language and Mathematics Strategy (GPLMS), an innovative large-scale reform designed to improve learning outcomes. Using data from universal testing of all learners in 2008 on a provincial systemic evaluation, and data from the 2011 and 2012 Annual National Assessment (ANA) test, this paper addresses the key research question, namely whether the GPLMS is effective in closing the gap between performing and underperforming schools. Given the evidence we have presented of an instrument effect, namely that various versions of the ANA may not be strictly comparable, no definitive conclusions can be drawn about the effectiveness of the GPLMS.

**Keywords:** large-scale education reform, literacy, regression discontinuity design

## Introduction

There is an emerging consensus that the primarily challenge of Post 2015 Education for All is the improvement of learning outcomes for all, particularly in primary school language and mathematics (King, 2013). Education systems in the global south need to develop and share new approaches to, and evidence of, effective large-scale education improvement, particularly innovations that fundamentally transform classroom instruction (see for example Rincón-Gallardo & Elmore , 2012 in Mexico and Banerji & Mukherjee, 2008 in India). Although South Africa's experience with large-scale reform aimed at improving instruction is at an early stage, insights from provincial initiatives have the potential to contribute to the new knowledge base emanating from the global south. The context for the South African large-scale instructional innovation is the mounting evidence from cross-national studies and government studies of the continued crisis in academic performance in South Africa (Fleisch, 2008). This research shows that schoolchildren in South Africa are underachieving in reading and mathematics, with an estimated 70% of learners not meeting the minimum curriculum policy standard (Spaull, 2013). South African learners are the poorest performers in international studies (Howie, Venter, Van Staden, Zimmerman, Long, Du Toit, Scherman & Archer, 2008). Research shows the profound inequality in achievement, with the overwhelming majority of children from historically "black" schools performing very poorly compared to children (both black and white) from formerly advantaged schools (Howie et al., 2008).

As a result of this research, the education crisis, particularly the underachievement in reading and mathematics in primary schools, has become a major theme in government planning. In 2010, South African President Jacob Zuma identified primary school achievement as a national priority and set targets for the Department of Basic Education to ensure that 60% of learners performed at grade level by 2014 on the ANAs for Grades 3 and 6. The National Planning Commission (2013) has similarly identified education as a national planning priority and has recommended a range of interventions. Within the Department of Basic Education (DBE) (2011), the Education Plan 2014 has identified indicators and targets for achievement in primary schools. In addition, it has begun to put in place key policies, such as the ANAs, the DBE workbooks, and training strategies such as the professional learning communities.

While there is growing national momentum to develop policies and programmes to address the problem of underachievement in reading and mathematics in primary schools, two provinces have developed their own unique approaches to large-scale improvement. Since the early 2000s, the Western Cape has been implementing programmes to improve primary school achievement. This province's achievement was highlighted in the 2010 McKinsey report titled *How the world's most improved school systems keep getting better*.

More recently, the Gauteng Department of Education initiated its own intervention strategy, the Gauteng Primary Language and Mathematics Strategy (GLPMS). It built on ideas developed in an earlier generation of policies, for example, the provision of lesson plans in the Foundations for Learning Campaign (Meier, 2012). Working initially in 792 underperforming schools (which constitute about 65% of all public schools), the dual aim of the strategy is to raise the overall performance of the province in reading and mathematics and to close the gap between the historically advantaged schools and the historically disadvantaged schools, assuming that performance is associated with this categorisation. A growing body of research suggests that the intervention is successful (Fleisch, 2013; Hellman, 2012), but as yet no published studies have used large-scale learner assessment data to evaluate the relative performance of this intervention. As the design of the strategy is based on current 'change knowledge', specifically adapted to the situation in a middle income country, the insight from the study has wider relevance.

This article is the first study to begin to address this gap. Using data from universal testing of all learners in 2008 on a provincial systemic evaluation, and data from the 2011 and 2012 ANA tests, this article addresses the key research question, namely whether the GPLMS approach is effective in closing the gap between performing and underperforming schools.

## Literature Review

Why have education policies failed, and what are the alternatives? There is a growing body of "change knowledge" that suggests promising insights. One of the major shifts that have taken place in the field of education policy/educational change is the realisation that conventional input/output models for policy development and interventions are problematic (Cohen, Raudenbush & Ball, 2003; Raudenbush, 2005). It is not resources themselves that create achievement (output), but, rather, the way that resources (inputs) are utilised in the instructional process, and it is this instructional process, itself, that impacts learning. While this is a rather abstract idea, it has significant implications for theorising the centrality of instructional practices. It suggests the need to have an in-depth understanding of existing instructional practices, and the forces or factors that reproduce or potentially re-engineer them. Raudenbush (2005) has advanced the argument that to change the instructional core, we need to introduce new instructional regimes. In his view, these new teaching and learning programmes will be evidence-based, such as the highly prescriptive programme *Success for All*. The assumption is that if these instructional regimes are implemented with fidelity, stable and predictable learning outcomes could be expected.

There is recognition of the centrality of alignment and coherence of policies designed around instructional practice (Cohen, 2011). Cohen and Spillane (1993) distinguish between educational policies in general and education policies that have the potential to influence, and ultimately change, instruction in classrooms. The latter they refer to as policies that offer instructional guidance. They identify five categories of policy that have genuine potential to change instruction: curriculum frameworks; external assessment of learner performance; provision of instructional materials; monitoring of classroom instruction; and policy requirements for teacher education, and licensure. While governments formulate education policies to regulate other aspects of education, Cohen and Spillane (1993) argue that it is only the above-mentioned categories of policies that have the potential to contribute to change in instructional practice. The relative success of policies that promote instructional guidance depends on the extent to which the policies are consistent, that is, the degree to which the various instructional policies not only speak to each other, but are consistent and aligned. Cohen (2011) also suggests that the degree of specificity or prescriptiveness is key. Instructional policies can be deliberately designed to be vague to allow for a wide variation of interpretation and adaptation, or they can be clear and detailed, specifying the what, when, and how of teaching.

One of the major contributions to "change knowledge" has been the experiences of England's National Literacy and Numeracy Strategy (NLNS) and Michael Barber's insights about it. Fullan (2010) suggests that England was the first government in the world to use an explicit theory of large-scale change as the basis for bringing about system reform. Barber (2007) describes the thinking behind the NLNS as a high-challenge, high-support approach with five key components: ambitious standards, good data and clear targets, prescribed lesson plans, quality professional development, and accountability and intervention in direct proportion to success.

One of the most recent contributions to "change knowledge" has come from an international consulting firm, McKinsey and Company. Mourshed, Chijioke and Barber's (2010) study of how the world's most improved school systems keep getting better begins with the assumption that education systems are at different points in the change journey, and that depending on the stage in the journey, certain policy approaches are preferred. For example, for systems moving from "poor" to "fair", for example the Western Cape in South Africa, Minas Gerais in Brazil, and Madhya Pradesh in India, a cluster of interventions that include highly prescriptive mandated lessons, the monitoring of compliance by having regular class visits, and the setting of performance targets based on universal external assessments would work best. Mourshed et al. (2010) note that systems which currently have poor achievement levels often have mistakenly experimented with policies that favoured high levels of teacher autonomy and unstructured peer learning.

## Description of the Strategy, and the Theory of Change

The GPLMS was developed in 2010, based on the current change knowledge outlined above. Five basic principles were central in the development of the strategy. From a management perspective, the GPLMS was to be feasible, affordable, and within the capacity of the province to manage. Second, the strategy required a strong commitment to partnerships, particularly partnership with education non-governmental organisations (NGOs) in the province. The third principle required ongoing, dynamic internal monitoring and external evaluation to guide the evolution of the project over time. Fourth, the initiative was strongly committed to fostering alignment and coherence, both between components of the initiatives and, possibly more importantly, with the various external policy and programmatic initiatives, both in the province and from national government. Finally, the strategy recognised that the long-term sustainability of the initiative was ultimately dependent on teacher learning. As such, all aspects of the strategy were to be geared to facilitate and consolidate teachers learning new teaching practices.

In terms of the theory of literacy and learning, the strategy provided an upfront statement of its pedagogy, which was referred to as a "simple reading approach". The founding document described it as such:

> [It is] premised on the assumption of the importance of both 'decoding' and 'comprehension', - word recognition processes and language cognition processes. This is sometimes referred to as a balanced approach, combining phonics and whole language. There is recognition in the Simple Literacy Approach that primary school children move from 'learning to read' to 'reading to learn' and 'reading for a purpose' and 'reading for pleasure'. Teaching primary school learners to be fluent readers requires that they have extensive and continuous access to books and other reading materials that are age- and language-appropriate and enjoyable (GDE, 2010:15).

Unlike some variations under the umbrella of "balanced approaches", the GPLMS puts a strong emphasis on the phonics component, as a discrete set of activities to be taught daily. The strategy required that the phonics programmes were to be selected based on research evidence, taking cognisance of the unique language contexts of both teachers and learners in the

schools. The phonic component was to be reinforced by a range of other literacy activities, both oral and textual, both reading and writing, with increasing emphasis on the use of a range of texts of increasing complexity. Comprehensive development of vocabulary, sentence structure, and exposure to both oral and written text was seen as essential to underpin reading and writing skills and language development in order to develop comprehension skills.

While much of the focus of the "simple reading approach" was on the Foundation Phase, the strategy recognised the unique challenges faced by learners at the Intermediate Phase. Studies have shown a widespread levelling off of reading skills after the Foundation Phase, either because of lack of access to interesting texts or because of poor reading skills (Snow & Biancarosa, 2003).

The strategy makes use of multiple-overlapping, mutually reinforcing components, what Cohen (2011) refers to as instructional infrastructure, all of which are tightly aligned, both in terms of their emphasis on classroom practice and in terms of the sequence and timing of their roll-out. While the use of standardised learner test results (ANAs) was seen as an important pillar of the strategy, the core component of the interventions designed were daily lesson plans, high-quality learning and teaching materials, and instructional coaching.

The scripted daily lesson plans provided systematic, paced, and easily accessible lessons for the teachers to follow through the year. The underlying purpose of the lesson plans was to introduce and gradually institutionalise a repertoire of practices that will improve teachers' time on task and establish new daily and weekly routines. In addition the programme to includes pre-designed assessment tasks, together with model answers, marksheets, and aligned homework activities. The lesson plans also integrate the use of the official DBE workbooks that have been provided to all schools. The lesson plans reduce teachers' planning and administrative workloads and allow them to concentrate on actual teaching, thus shifting the focus from interpreting the national curriculum to delivery of the curriculum.

The designers of the strategy recognised that the lesson plans alone would not, and, in fact, could not, transform classroom practices. The provision of whole class sets of quality learning materials, that is, phonics programmes, workbooks, and class sets of graded readers, were seen as a necessary condition for change. One of the features of these resource packages is that they make use of a systematic and planned approach to instruction. The strategy recognised that, together, the scripted lesson plans and the learner materials were the bedrock of the new practice, but that an additional component would be needed to translate the lessons and the materials into a new practice. For this one-on-one instructional coaching on a continual basis was provided to teachers. The coaches model the new teaching practice, support and encourage work on the new practice, and help establish new learner expectations. In addition, the coaches also provided "just-in-time" training, working through the scripted lesson plans with groups of teachers at the beginning of each term.

## Research design

### Test instruments

During the planning phase of the GPLMS intervention, the Gauteng Department of Education decided to make use of the 2008 Systemic Evaluation (SE) and ANA in the programme evaluation. The 2008 SE dataset was used as a baseline and determined inclusion into the GPLMS intervention. The statistical picture of change over time was to be drawn from the ANAs. The ANA in 2011 was conducted in February and was designed to evaluate performance in 2010, while the ANA in 2012 was administrated later in the year, in September. The DBE provided the national ANA 2011 and 2012 datasets electronically. The datasets include the name of the school and the national Education Management Information System (EMIS) number, as well as the number of marks per grade, including the grade average and the percentage of learners that achieved a mark above 50% for each grade.

Although the ANA 2011 and 2012 provide information for both language and Mathematics for Grades 1 to 6 and 9, for the purposes of this evaluation, emphasis is placed on the Grade 3 Language findings. There are three reasons for this. Firstly, while the schools administered Grade 1 and Grade 2 tests, only the Grade 3, 6 and 9 tests were rigorously piloted to ensure the validity and fairness of test items (DBE, 2012). Secondly, the splitting of the Intermediate Phase tests into Home Language (HL) and First Additional Language (FAL) makes the analysis more complicated, particularly as regards a comparison of GPLMS schools and non-GPLMS schools. Finally, the Grade 3 learners in GPLMS schools would have had almost 19 months of intervention by September 2012, providing a good picture of the actual effects of the intervention, compared to a mere seven months in the Intersen Language component, and even less time in the case of the Foundation Phase and Intermediate Phase Mathematics components, as is illustrated in Table 1.

**Table 1** GPLMS Roll-out 2011 and 2012

| | 2011 | 2012 |
|---|---|---|
| Grade 1 | Language | Language & Mathematics |
| Grade 2 | Language | Language & Mathematics |
| Grade 3 | Language | Language & Mathematics |
| Grade 4 | | Language & Mathematics |
| Grade 5 | | Language & Mathematics |
| Grade 6 | | Language & Mathematics |

## Results

Has the GPLMS impacted the overall performance in the province of Gauteng, and is it effective in improving the Language and Mathematics achievement of learners in underperforming schools in the province? To answer these questions, we begin by exploring the overall provincial achievement in Grade 3 Language of primary schools in Gauteng, as compared to other provinces. This is followed by an analysis of the relative performance of GPLMS schools and non-GPLMS schools in Gauteng, specifically exploring the performance gaps between these two groups of schools over time.

### Provincial comparison

In 2012, Gauteng's average percentage mark in Grade 3 Language was 54.8%, with 61.7% of learners achieving a mark of 50% or above. While Gauteng ranked third on this Grade 3 Language test, the province ranked first in the Grade 1 Language test and second in the Grade 2 Language test, ranking first in Grades 4-6 on the FAL Language tests. Although the DBE report notes an improvement from 2011 to 2012, given that a significant proportion of school marks were not captured in 2011, and that the test was administered at a different time of the year, we would advise caution in claims about change over time based on the 2011 and 2012 data alone. Nevertheless, we

can look at the ranking of the different provinces across the two years.

Although Gauteng moved up two places in the rankings between 2011 and 2012, it is imprudent to draw conclusions from changes in the ranking, particularly as the veracity of some of the provincial aggregate scores was unfairly influenced by the fact that the marks of some of the learners had not been captured.

### School comparison

A more robust approach to assessing the efficacy of the GPLMS was to compare the aggregate performance of schools included, and schools not included in the strategy. Assignment to the GPLMS initiative was based on test scores achieved in the literacy section of the 2008 SE in Gauteng. Specifically, all primary schools that obtained, on average, 40% or below in the Grade 3 literacy section of the 2008 SE were considered under

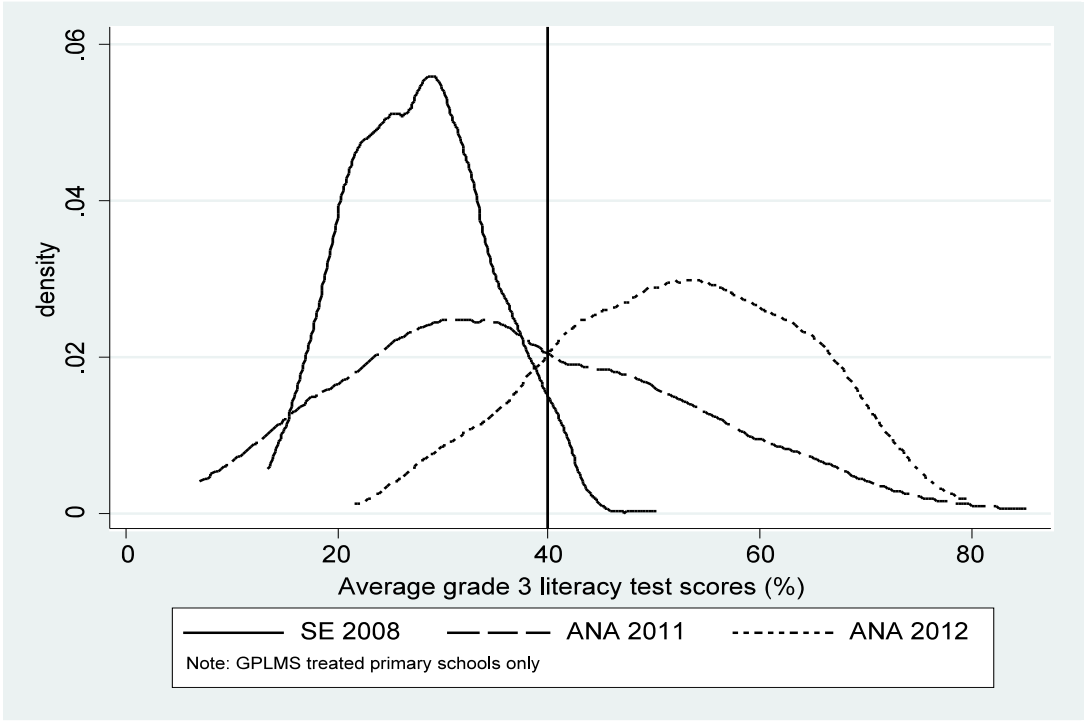


Figure 1 Distribution of Grade 3 literacy scores in SE 2008, ANA 2011 and ANA 2012 (GPLMS schools)

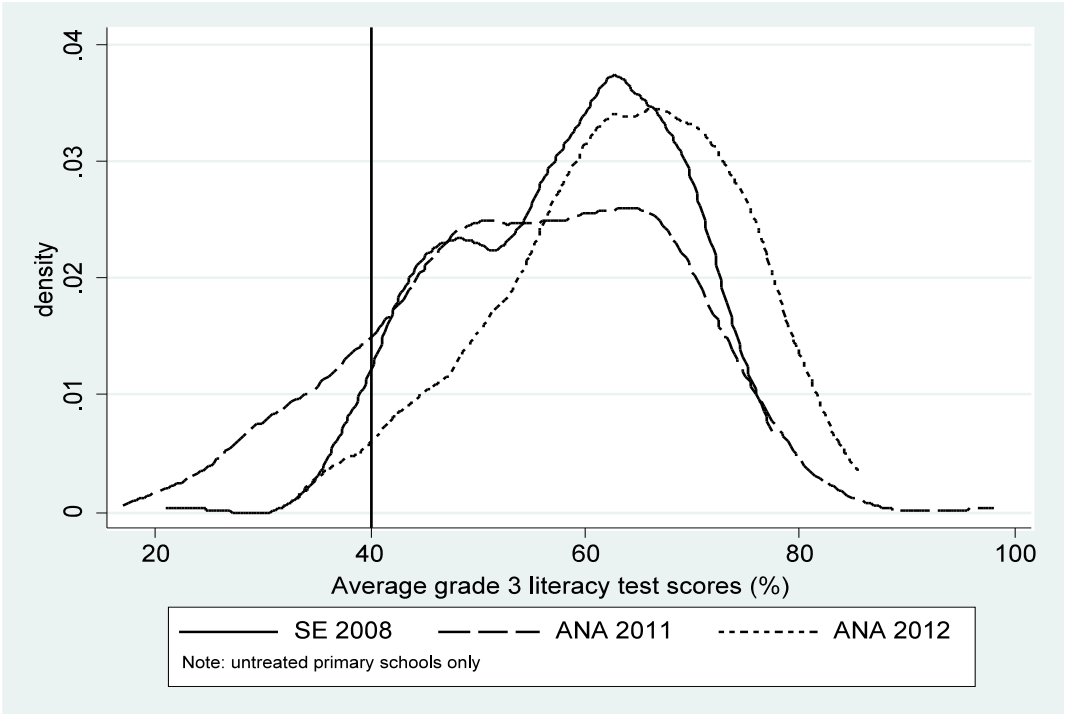


Figure 2 Distribution of literacy scores in SE 2008, ANA 2011 and ANA 2012 (Non-GPLMS schools)

performing, and were thus assigned to the GPLMS initiative. The initial number of primary schools that participated in the 2008 SE was 1,303. However, only 915 schools can be matched across all three test instruments used in this study, namely the 2008 SE, the 2011 ANA, and the 2012 ANA.[i] For the purposes of this study, we restricted our sample to primary schools that were either exposed to the GPLMS from the initial implementation of the initiative in 2011 or that were never exposed to the initiative. Thus, we exclude primary schools that, for whatever reason, were exposed to the GPLMS initiative only from a later stage onwards or were initially exposed but dropped out. This leads to the exclusion of an additional 17 schools from the sample[ii] and we use the remaining sample of 898 primary schools. Of these 898 schools, 609 were actually exposed to the GPLMS initiative from the start, while the remaining 289 schools were never exposed to it.

**Table 2** Achievement in Grade 3 Language for GPLMS and non-GPLMS schools for different test instruments (%)

| | SE 2008 | ANA 2011 | ANA 2012 | No. of schools |
|---|---|---|---|---|
| GPLMS Schools | 27.6 | 37.3 | 52.0 | 609 |
| Non GPLMS Schools | 58.2 | 54.7 | 63.3 | 289 |
| Difference (percentage points) | -30.6*** | -17.4*** | -11.3*** | 898 |

Note: *** differences in means significant at 0.01 level

Figures 1 and 2 show the distribution of GPLMS schools and non-GPLMS schools by initial performance in the 2008 SE, as well as the subsequent performances of these respective schools in the two ANAs. While the assignment threshold was set at 40% in the literacy section of the 2008 SE, a small number of primary schools that actually performed below the cut-off were never treated, while a small number of schools that actually performed above the cut-off were exposed to the initiative.[iii]

The frequency distribution in Figure 1 shows the initial performance of schools that were assigned to the GPLMS in the provincial SE in 2008 (solid line), and how the majority of the GPLMS schools still underperformed in the 2011 ANA, with a mark below the 40% threshold (long-dashed line). However, by 2012, the same sample of previously underperforming schools was able to improve its performance in the 2012 ANA dramatically, with a significant proportion of schools achieving around 60% (short-dashed line). This represents high levels of achievement, and only a relatively small cluster of schools fall below the 40% cut-off. In comparison, the frequency distribution for non-GPLMS schools in Figure 2 reflects much less change between the different test instruments in 2008, 2011, and 2012. While the entire sample of non-GPLMS schools improved between 2008 and 2012, theimprovement is not of the same magnitude as that experienced by the GPLMS schools.

Hence, when comparing the performance of GPLMS schools and non-GPLMS schools over the three test instruments (Table 2), we observe a trend for the GPLMS-treated schools to show a greater improvement relative to the non-GPLMS schools. While the GPLMS schools in our sample performed, on average, 30.6 percentage points below the non-GPLMS schools in the literacy section of the 2008 provincial SE, the difference in performance in the same sample of underperforming schools narrowed to 17.4 percentage points in the ANA 2011 literacy section, and to 11.6 percentage points in the ANA 2012 literacy section. Thus, it is evident that the achievement gap between the GPLMS schools and the non-GPLMS schools narrowed over the three test instruments. As with the Grade 1 and 2 results, the achievement gap between the underperforming schools that were assigned to the GPLMS intervention and other Gauteng primary schools in our sample that were not treated with the GPLMS intervention appears to be narrowing, suggesting reduced levels of education inequality.

However, the changes in the distribution of test averages might be driven by a number of factors, of which assignment to the GPLMS programme is only one. For instance, the improvement of the GPLMS-treated schools in the two ANA tests in 2011 and 2012 might indicate that the design of the ANA literacy sections and/or the marking thereof allowed weaker learners to perform only relatively well, that is, it was easier to obtain a mark closer to 50% than to perform well above 50%. Thus, the shift in the distribution of marks could simply be a test instrument effect. Furthermore, it is more difficult for already high-performing schools to improve their performance, while low-performing schools that start from a very low base might experience larger improvements more easily. It is therefore possible that there was simply an overall trend in reading/literacy improvement across all schools – GPLMS-treated schools, as well as untreated schools – which simply had a relatively larger effect on initially low-performing schools. This seems to be confirmed when looking at Figure 3, which shows that across all performance ranges in the 2008 provincial SE, primary schools improved their mean scores for the ANA 2012 Grade 3 literacy section, including the initially high-performing schools. A similar trend can be seen in the difference in schools' test scores between the SE 2008 and the ANA 2011, that is, prior to the implementation of the GPLMS intervention, which supports the argument of a test instrument effect. However, the fact remains that GPLMS schools experienced a significantly greater improvement in the ANA 2012 after the implementation of the GPLMS intervention. This is reflected in Figure 3 by the steeper slope of the solid line, which shows the improvement of schools from the 2008 SE to the 2012 ANA, compared to the slope of the dashed line, which shows the improvement from the 2008 SE to the 2011 ANA.

## Analysis

Simply comparing changes in the means across all GPLMS schools and non-GPLMS schools in our sample might lead us to misinterpret the reduction in the performance gap as being a result of the GPLMS intervention, rather than considering an equally plausible factor, namely a test instrument effect. In order to evaluate the efficacy of the GPLMS intervention, as well as the change model on which it is based, most evaluators would opt to use a randomised control trial (RCT) design, which is widely regarded as the "gold standard" in development evaluation. In a RCT, the untreated, or control, schools could be used as a counterfactual, which shows what the performance would be for schools that have not been exposed to the intervention but have the same characteristics as schools that have been exposed to the intervention. The difference in performance between these two groups of schools could then be attributed to the intervention, which would allow for a causal interpretation. However, given that this intervention was explicitly not designed to be a pilot, but to achieve change at scale during the years of the initiative (2010-2014), this approach to evaluation was not feasible. In order to investigate the impact of

the GPLMS initiative, it is necessary to have GPLMS schools and non-GPLMS schools that are relatively similar in their characteristics.

We therefore combine two impact evaluation methods for our analysis: regression discontinuity design (RDD), and difference in differences (DID). RDD would assume that schools which obtained mean scores just above and just below the 40% assignment threshold (for instance, within the range of 35-45%) are likely to be very similar in their characteristics. Thus, schools that fell just above the 40% assignment threshold and were not exposed to the GPLMS intervention can be used as a comparison group for those schools that fell just below the threshold, and therefore were exposed to the intervention. Put another way, rather than just looking at the change experienced by the treated group, the DID method compares change not only in the treated group but also in the control group. By comparing the differences in outcomes between the two groups, DID can differentiate any unobserved heterogeneity that might have affected programme participation.

However, key assumptions of the DID approach are that these unobserved heterogeneities are time-invariant and that both groups experience a common time trend.

To illustrate the DID method, consider the information presented in Table 2. The treatment group in the year 2008 had an average of 27.6% which in 2012 went up to 52%, an 24.4 per point increase. A simple before-after analysis would conclude that the GPLMS initiative led to an improvement of treated schools by 24.4 percentage points from 2008 to 2012. However, at the same time, the control group, i.e. schools that did not receive the treatment, averaged 58.2% in 2008 but also experienced an increase up to 63.3% in 2012, a 5.1 the percentage point increase. The DID is calculated by subtracting the gains not attributed to the intervention, but evident in the control group from the gains made specifically in the treatment group, that is 24.4 – 5.1 = 19.3. Thus, the treatment effect only accounts for 19.3 percentage points in the total change of 24.4 percentage points experienced by the treated primary schools between 2008 and 2012.

We proceed by restricting our sample to schools that obtained between 35 and 45% in the literacy section of the 2008 SE. We investigate the changes in performance from the 2008 SE to both ANAs for schools that obtained between 35 and 40% in the 2008 SE, and therefore were exposed to the GPLMS intervention, compared to schools that obtained between 40 and 45% in the 2008 SE, and therefore were not exposed to the GPLMS. Finally, we investigate two implementation regimes. In the first regime, we impose the initial strategy, where all schools that obtained a score of or below the cut-off in the 2008 SE were assigned to the GPLMS initiative, while all schools above the cut-off were not exposed to the intervention. We restrict our sample to such schools and exclude all schools that should have been exposed to the GPLMS on account of their performance in the 2008 SE but were not exposed, as well as schools that performed above the cut-off but were still exposed. This reduces our sample to 871 schools. The variable indicating exposure to the GPLMS initiative for this reduced sample is called *GPLMS*. The second implementation regime includes all 898 schools, irrespective of whether they should or they should not have been treated, according to the initial GPLMS strategy. The variable indicating exposure to the GPLMS initiative for this extended sample is called *Treated*.

### Regression analysis

Of interest to this study is the effect of being exposed to the GPLMS initiative over time. This is shown in Table 3 by the treatment effect variable at different years (for instance, *Treatment effect in 2011* in Table 3). The treatment effect variable shows the change in performance in the treated group compared to the control group over and above the time trend. Thus, we estimate the same value that we calculated manually in the above example for Table 2, i.e., the treatment effect of 19.3 percentage points for the treated group between 2008 and 2012.

Table 4 in the appendix reports the full set of results for different specifications of the DID regression.[iv]

The first specification (column 2) includes the full sample of primary schools that, according to the initial GPLMS strategy were correctly assigned to the GPLMS intervention and that were exposed to the GPLMS on account of their performance in the 2008 SE. Thus, we test the first implementation regime (GPLMS) as outlined above. The results confirm the findings of the descriptive statistics section and show that GPLMS-exposed schools had already experienced a relatively larger improvement in the 2011 ANA relative to their initial performance in the 2008 SE, despite the fact that the GPLMS initiative had not yet been initiated (see *Treatment effect in 2011*). Thus, we need to find a sample of schools where there is no difference in the change in performance between the exposed schools and the non-exposed schools between 2008 and 2011, prior to the GPLMS intervention.

The results for the second specification (column 3) indicate that our limited sample of GPLMS schools (which obtained scores between 35 and 45% in the 2008 SE), when we impose the first implementation regime, did not experience a change in their performance differently to that of the non-GPLMS schools between 2008 and 2011. This serves to confirm that our reduced sample experienced the same trend between 2008 and 2011, and that these schools might have relatively similar characteristics. However, GPLMS schools managed to outperform non-GPLMS schools by, on average, 4-5 percentage points in the change between their 2008 and 2012 performances, as can be seen by the coefficient of the variable *Treatment effect in 2012*. Nevertheless, large standard errors reduce the statistical significance of these differences.

The third specification (column 4) increases our sample range from 35 and 45% in the 2008 SE to 34-46% in the 2008 SE. While the increase is only marginal, our results for the GPLMS intervention (GPLMS) seem quite sensitive to such changes. Similar to the narrower range of the previous 35-45% sample, the GPLMS schools in the 34-46% sample experienced a larger increase in performance compared to the non-GPLMS schools between 2008 and 2012 but at a magnitude of more than 5 percentage points. Thus, the inclusion of previously lower-performing schools and higher-performing schools on either side affected the coefficient on the treatment effect variable (*Treatment in 2012*), which indicates that the effect is larger for previously lower-performing schools. This, again, suggests a possible test instrument effect, which allowed lower-performing schools to experience a larger improvement relative to higher-performing schools.

We test this by including all primary schools within the range of the assignment threshold that either were treated or were not treated, irrespective of whether they should have been treated, given their performance in the 2008 SE. Thus, we test the second implementation regime (*Treated*). We look at all schools that were exposed to the GPLMS intervention, irrespective of whether the school performed above or below the 40% threshold. While specification 4 (column 5) shows the results for the full sample of all primary schools, specifications

**Table 3** Difference in Difference Regression Output for Average Test Scores of Primary Schools in Gauteng in 2008 SE, 2011 ANA and 2012 ANA Grade 3 Literacy Section (percentage points)

| | Implementation regime 1: GPLMS | | | Implementation regime 2: Treated | | | Pseudo 1 | Pseudo 2 |
|---|---|---|---|---|---|---|---|---|
| 2008 SE Grade 3 literacy scores Specification | 0 - 100% Full sample (1) | 35 - 45% restricted sample (2) | 34 - 46% restricted sample (3) | 0 - 100% Full sample (4) | 35 - 45% restricted sample (5) | 34 - 46% restricted sample (6) | 45 - 55% restricted sample (7) | 25 - 35% restricted sample (8) |
| Treatment effect in 2011 | 13.90*** (1.202) | -1.108 (3.366) | 0.854 (3.132) | 13.23*** (1.199) | -2.242 (3.024) | -0.534 (2.843) | 1.737 (2.915) | 3.936** (1.914) |
| Treatment effect in 2012 | 20.26*** (1.202) | 4.286 (3.366) | 5.199* (3.132) | 19.39*** (1.199) | 2.342 (3.024) | 3.220 (2.843) | 4.660$ (2.915) | 5.445*** (1.914) |
| Observations+ | 2,613 | 297 | 372 | 2,694 | 372 | 447 | 210 | 897 |
| R-squared | 0.55 | 0.25 | 0.26 | 0.53 | 0.26 | 0.26 | 0.21 | 0.40 |

Standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ $ $p < 0.15$
Note: + the number of observations referred number of schools three observed over three years.

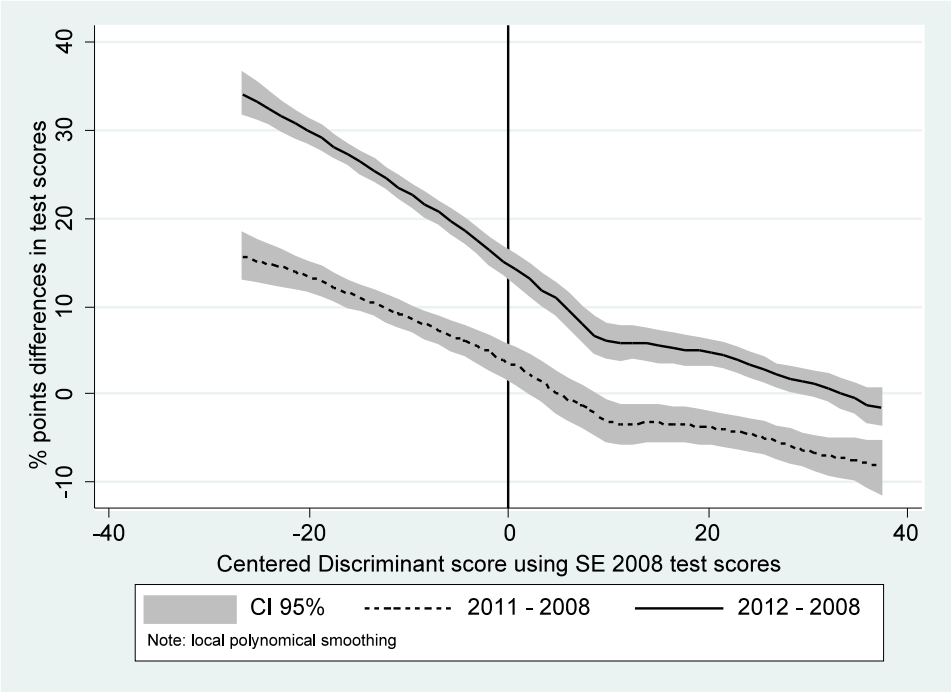


Figure 3 Percentage points differences between schools' performances in SE 2008 and both ANAs

5 and 6 (columns 6 & 7) report the findings for the narrower sample (35-45%) and the slightly larger range of primary schools that performed between 34 and 46 percentage points in the 2008 Provincial Systemic Evaluation. Including primary schools that were treated despite the fact that they actually performed above the initial cut-off, as well as primary schools that were not treated despite the fact that they performed below the initial cut-off, not only reduces the coefficient of the treatment variable (*Treatment in 2012*) but also provide results with no statistical significance. This suggests either that schools below the cut-off that were not treated still experienced a larger increase in performance from 2008 to 2012 compared to schools above the cut-off or that schools above the cut-off that were treated experienced less of an increase, or both of these explanations. It nevertheless confirms earlier findings that the ability to improve performance declines with higher performances in the 2008 SE. This could imply two possible and not mutually exclusive explanations: the test instrument used in the 2012 ANA does allow previously lower-performing schools to achieve relatively higher scores more easily, and/or the GPLMS intervention is most effective in schools that start from a very low performance base, while the efficacy of the GPLMS declines with increasing initial performance.

### Robustness checks

To test the robustness of our findings in the samples around the official assignment threshold of 40%, we created two pseudo samples which hypothetical assignment thresholds. The first sample, Pseudo 1, consisted of primary schools that initially performed above the assignment threshold (schools that obtained between 45% and 55% in the 2008 SE), and therefore were not exposed to the GPLMS initiative at all. We imposed a hypothetical assignment threshold at 50% and pretended that schools that obtained between 45% and 50% in the 2008 SE

actually did get exposed to the GPLMS initiative, while schools that performed above the 50% threshold now acted as our control group. The results of the DID regression for our Pseudo 1 sample are shown in specification 7 (column 8) in Table 3. Surprisingly, while there is no significant difference in performance in 2011, the hypothetically treated group (which obtained 45-50% in the 2008 SE) experienced, on average, a 5 percentage-point larger increase in performance between 2008 and 2012 compared to the hypothetically untreated group (which obtained 50-55% in the 2008 SE), despite the fact that neither group was actually exposed to the GPLMS intervention. Therefore, there seems to be a clear test instrument effect in the 2012 ANA literacy section, which allowed previously lower-performing schools to perform better than previously higher-performing schools.

We tested this finding with a second sample, Pseudo 2, which consisted of primary schools that initially performed below the assignment threshold (schools that obtained between 25% and 35% in the 2008 SE), and therefore were all exposed to the GPLMS initiative. Again, we imposed a hypothetical assignment threshold, but this time at the 30% cut-off, and pretended that only primary schools that obtained between 25% and 30% were actually assigned to the GPLMS initiative, while schools that obtained above 30% in the 2008 SE were not exposed to the GPLMS initiative. The results of the DID regression for our Pseudo 2 sample are shown in specification 8 (column 9) in Table 3. Similar to the first hypothetical sample (Pseudo 1), the treated group in Pseudo 2, that is, the group below the hypothetical cut-off of 30%, experienced a larger increase in performance between 2008 and 2012 compared to the group just above the hypothetical 30% cut-off. This increase, again, is, on average, around 5 percentage points, which is similar to the increase experienced by the sample around the true cut-off (40%), as well as the first hypothetical sample (Pseudo 1). This finding, however, could have two possible explanations: while it could confirm the test instrument effect, it could equally be the case that the GPLMS is most effective for schools that come from an initially very low performance level. Thus, given the set of different activities that constitute the entire GPLMS package, schools that experience very low performance levels might benefit most from these interventions. Unfortunately, with the available data, it is not possible for us to unpack these different explanations.

### Regression discontinuity at the 40% assignment threshold

While the DID analysis does not allow us to untangle a possible treatment effect from a test instrument effect, we use a further impact evaluation technique to illustrate the possible treatment effect of being exposed to the GPLMS intervention. As mentioned above, the idea of the RDD is that schools that obtained marks just above or just below the policy assignment threshold are likely to be very similar in their characteristics and their performance. Thus, schools around the assignment threshold actually had more or less the same probability of being assigned to the intervention or not being assigned. However, by some random coincidence, some schools did obtain an average score just below the cut-off, while other schools obtained a score just above the threshold. Therefore, the rationale of the RDD is that at the limit of the assignment threshold one should be able to observe a treatment effect as if these schools were randomly assigned to treatment similar to a RCT. In order to mimic a sharp RDD, we again restricted our sample to schools that were actually correctly assigned and consistently treated (that is, treated schools below the 40% cut-off), and we compared these schools to correctly untreated schools (that is, schools above the 40% cut-off that were never treated). Furthermore, we limited our analysis to a graphic analysis, and we investigated whether, at the limit, primary schools that were actually treated performed higher than schools that were not treated. The results are shown in Figures 4-7.

To control for different functional forms, we show the predicted values for first-, second-, third- and fourth-order polynomial regressions, as well as the bin means in each graph, respectively. The centred discriminant score refers to the assignment threshold and the distance of schools' performances in the 2008 SE from the 40% cut-off (indicated in each graph by 0). The Y-axis shows the difference in performance from the 2011 ANA to the 2012 ANA, that is, the performance before intervention and after intervention. All graphs show that, irrespective of the functional form, primary schools just below the cut-off experienced a larger increase in performance between 2011 and 2012 compared to schools just above the cut-off. Thus, at the limit of the assignment cut-off, the findings suggest that assignment to the GPLMS intervention did have a positive impact on the performance of schools. However, this argument can only be made for schools that performed close to the cut-off, and therefore only indicates a local treatment effect. To what extent this can be generalised across other performance ranges is not clear.

## Conclusions

Is the (GPLMS approach effective in closing the gap between performing schools and underperforming schools? Does a strategy built on international 'change knowledge' work in a middle income country? Or more broadly, what lessons can be learnt or insights gained that could contribute to the change knowledge for the global south?

Given the evidence we have presented of a possible test instrument effect, no definitive conclusions can be drawn about the effectiveness of the GPLMS at this stage. We have shown that the test instrument might allow previously lower-performing schools to improve on their low scores more easily than it allows previously higher-performing schools to improve on their already high scores. That said, the regression discontinuity analyses certainly indicate a local average treatment effect at the limit of the assignment cut-off.

Two concerns limit our ability to make strong claims about the relative effectiveness of the GPLMS. Firstly, although we establish a local average treatment effect for the sample of primary schools just below the assignment threshold, the same effect cannot be generalised across the whole population. Secondly, the actual mechanism of improvement, itself, is not self-evident. While assignment to the GPLMS intervention may be causally related to improvement, a variety of mechanisms could explain why the GPLMS schools performed better as the GPLMS intervention includes a package of different components. Our study cannot unpack which of the components, which combination of components, the relative dosage of which component or even 'gaming' effects, matters.

Notwithstanding these limitations, the GPLMS evaluation contributes important insights into large-scale instructional change in the global south. Although RCTs are the preferred method for establishing policy warrants, innovative alterative approaches, such as DID or RDD studies have the potential to provide genuine counterfactual evidence. This study also provides an important caution about the limitations of relying on national testing data in systems where system-wide testing is relatively new. Finally, although no definitive conclusions can

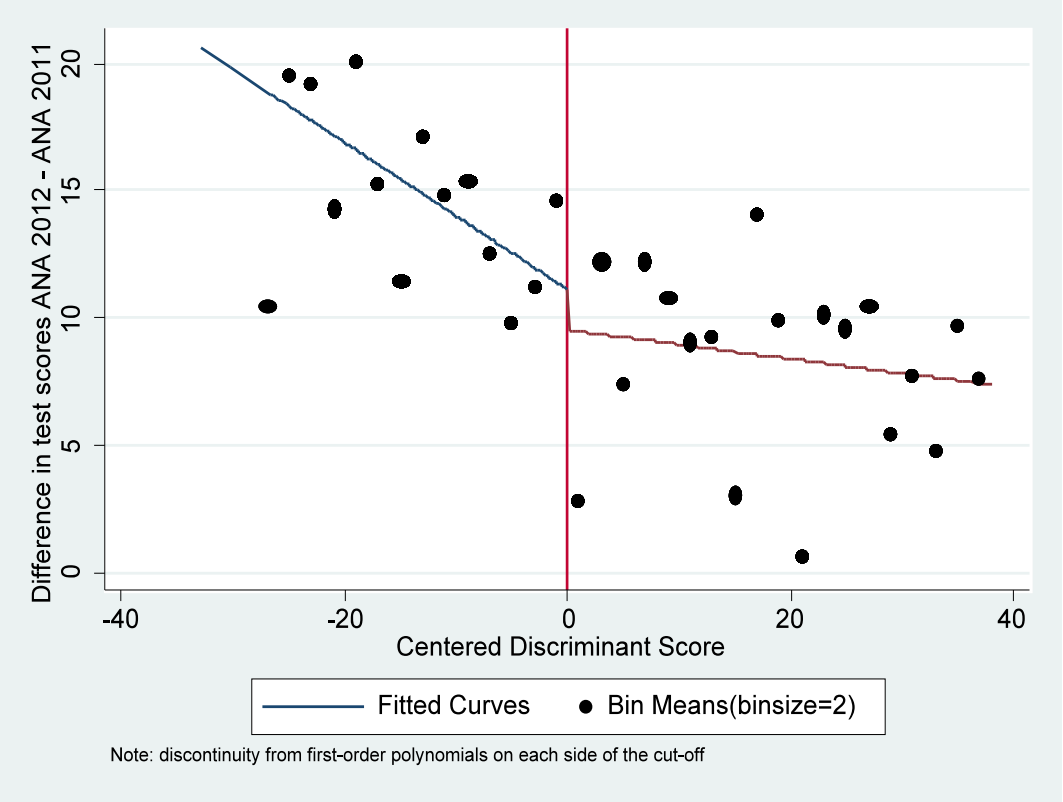


Figure 4 Regression discontinuity: difference in test scores ANA 2011 – ANA 2012, first order polynomial

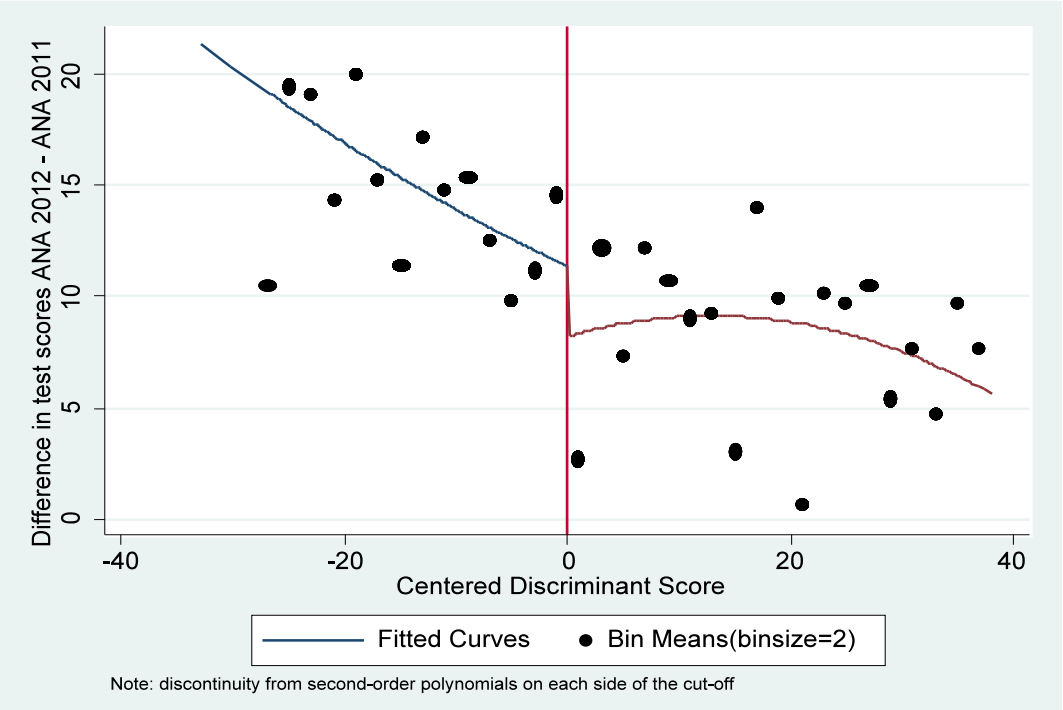


Figure 5 Regression discontinuity: difference in test scores ANA 2011 – ANA 2012, second order polynomial

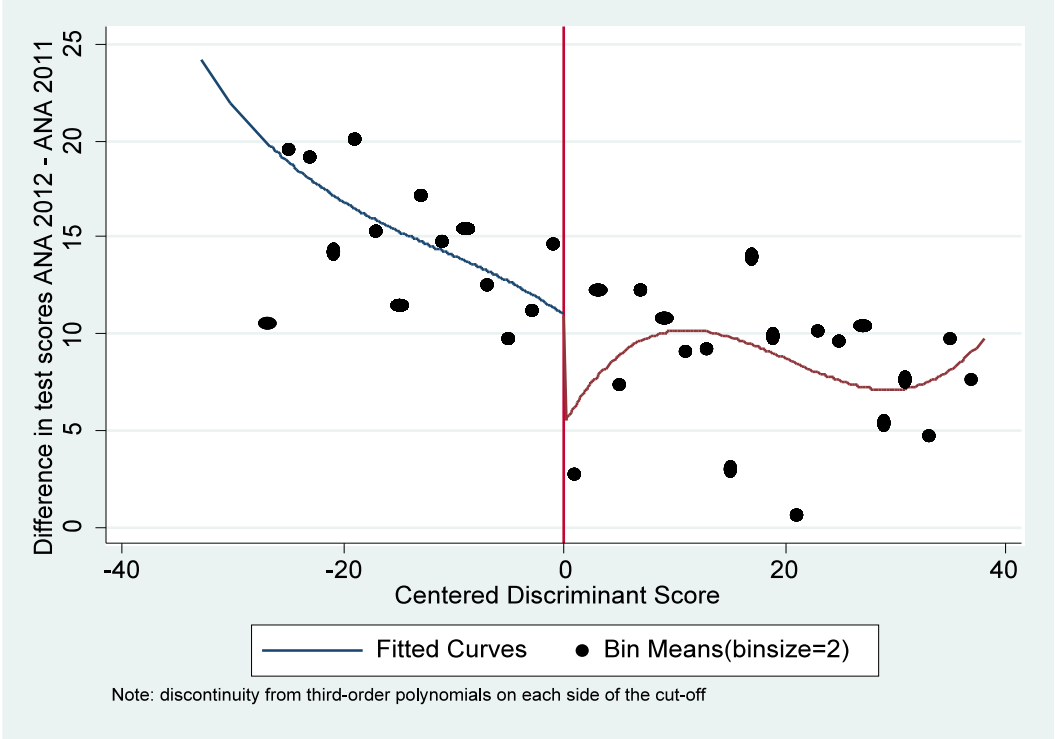


Figure 6 Regression discontinuity: difference in test scores ANA 2011 – Ana 2012, third order polynomial

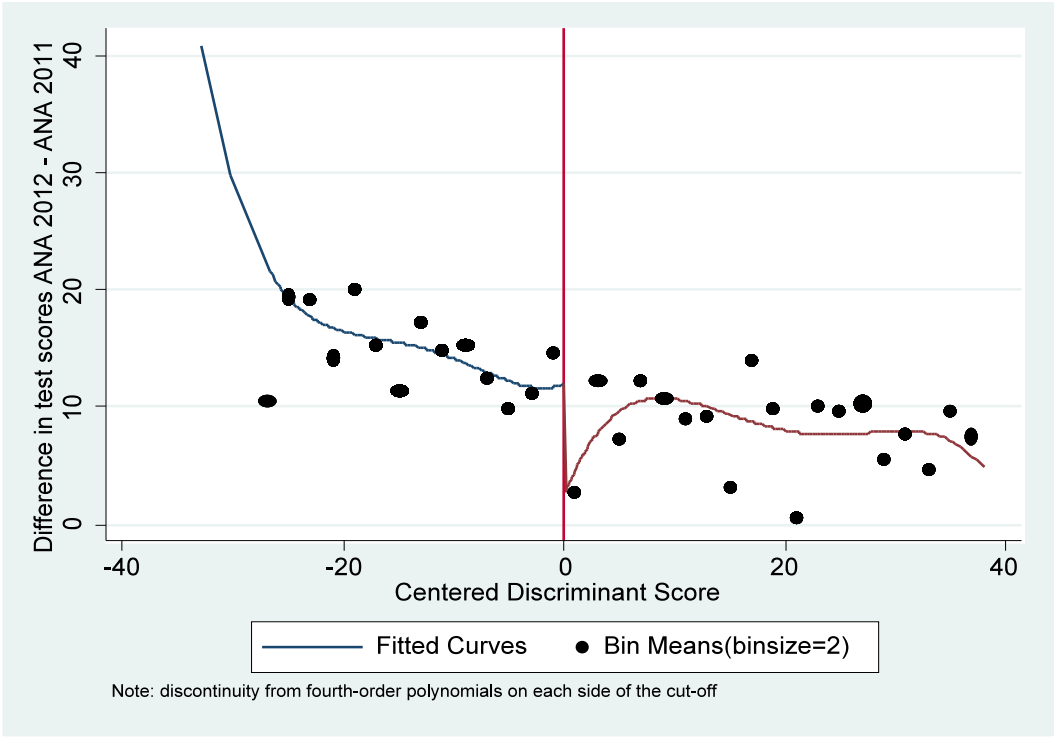


Figure 7 Regression discontinuity: difference in test scores ANA 2011 – 2012, fourth order polynomial

be drawn, the positive local average effect result provides a strong motivation for continued statistical analysis of the GP LMS programme and the model that it represents. Robust empirical studies of the kind described in this article are a sound basis for building powerful knowledge on large-scale reform of instruction as systems in the Global South begin to grapple with the Post 2015 quality challenge.

**Acknowledgements**

We thank Stephen Taylor, Martin Gustafsson, Nic Spaull and Gareth Roberts and the two anonymous reviewers for helpful comments. However, we take full responsibility for the content of this paper.

**Notes**

i To test the sensitivity of excluding these schools, separate analyses were performed where we include schools that can be matched only between 2008 SE and 2011 ANA as well as only between 2008 SE and 2012 ANA. While the results are slightly different in magnitude, the overall findings as presented in this paper remain the same.

ii As a robustness check, we included in our regression analysis all schools that were likely to have selected themselves into or out of the GPLMS programme. Their inclusion has a strong positive effect on the treatment coefficient. However, this effect is likely to be correlated with the decision of the schools to either participate or to exit the initiative, and therefore suffers from endogeneity. Furthermore, the low number of observations of such schools and limited access to school characteristics that would allow us to model their participation decision make it difficult to control for endogeneity.

iii This, according to the Gauteng Department of Education, was apparently less due to self-selection than it was due to an initial measurement/calculation error of the average scores.

iv For ease of reading the full regression output in Table 4: The constant indicates the average performance of the control group (primary schools that were not assigned to treatment) in the 2008 SE. The variable which indicates assignment to treatment (for instance, *GPLMS, Treatment*, and *Pseudo*) shows the average difference in performance in the 2008 Systemic Evaluation of the initially lower-performing group. As the samples tend to be over a 10 percentage-point range, the difference in the 2008 SE is around 5 percentage points, on average, by design. The two year dummies – 2011 and 2012 – reflect the average change in performance of the control group over the years compared to their performance in the year 2008. The actual treatment effect in each year is shown by the coefficients of the Treatment effect in 2011 and Treatment effect in 2012 variables.

**References**

Banerji R & Mukherjee AN 2008. Achieving universal elementary education in India: future strategies for ensuring access, quality and finance. *Margin: The Journal of Applied Economic Research*, 2(2):213-228. doi: 10.1177/097380100800200204

Barber M 2007. *Instruction to deliver*. London. Politico's Publishers.

Cohen DK 2011. *Teaching and its predicaments*. Cambridge, MA: Harvard University Press.

Cohen DK, Raudenbush SW & Ball DL 2003. Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, 25(2):119-142. doi: 10.3102/01623737025002119

Cohen DK & Spillane J 1993. Policy and practice: the relations between governance and instruction. In S Fuhrman (ed). *Designing coherent education policy: Improving the system.* San Francisco: Jossey-Bass.

Department of Basic Education (DBE) 2011. *Action Plan to 2014: Towards the Realisation of Schooling 2025*. Pretoria: DBE.

Department of Basic Education 2012. *Report on the Annual National Assessments 2012: Grades 1 to 6 & 9*. Pretoria: DBE. Available at http://www.education.gov.za/LinkClick.aspx?fileticket=YyzLTOk5IYU%3D&tabid=298. Accessed 14 January 2013.

Fleisch B 2008. *Primary education in crisis: Why South African schoolchildren underachieve in reading and mathematics*. Cape Town, SA: Juta.

Fleisch B 2013. *Change at the Instructional Core: Insights from the Intersen English Catch-Up Programme*. Paper Presented at the South African Education Research Association Meeting, January.

Fullan M 2010. *All systems go: The change imperative for whole system reform*. Thousand Oaks, CA: Corwin Press.

Hellman L 2012. *GPLMS Intersen Catch-up Programme: Analysis of Results*. Memo.

Howie S, Venter E, Van Staden S, Zimmerman L, Long C, Du Toit C, Scherman V & Archer E 2008. *PIRLS 2006 Summary report: South African children's reading literacy achievement*. Pretoria: Centre for Evaluation and Assessment. Available at https://web.up.ac.za/sitefiles/file/43/314/SA%20PIRLS%202006%20SUMMARY%20REPORT.pdf. Accessed 26 June 2014.

King K (ed.) 2013. Education and Development in the Post-2015 Landscapes. *NORRAG News 49*, October. Available at http://www.norrag.org/fileadmin/Full%20Versions/NN49.pdf. Accessed 26 June 2014.

Meier C 2011. The Foundations for Learning Campaign: helping hand or hurdle? *South African Journal of Education*, 31(4):549-560.

Mourshed M, Chijioke C & Barber M 2010. *How the world's most improved school systems keep getting better*. London: McKinsey & Company.

National Planning Commission 2013. *NDP 2030: Our future, make it work*. Pretoria: Sherino Printers. Available at http://www.npconline.co.za/MediaLib/Downloads/Downloads/NDP%202030%20-%20Our%20future%20-%20make%20it%20work.pdf. Accessed 26 June 2014.

Raudenbush SW 2005. Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, 34(5):25-31.

Rincón-Gallardo S & Elmore RF 2012. Transforming teaching and learning through social movement in Mexican public middle schools. *Harvard Educational Review*, 82(4):471-490.

Snow CE & Biancarosa G 2003. *Adolescent literacy and the achievement gap: What do we know and where do we go from here?* New York: Carnegie Corporation. Available at http://olms1.cte.jhu.edu/olms/data/resource/2029/class9_snow_biancarosa.pdf. Accessed 26 June 2014.

Spaull N 2013. Poverty & privilege: Primary school inequality in South Africa. *International Journal of Educational Development*, 33:436-447.

## Appendix

**Table 4** Difference in Difference Regression Output for Average Test Scores of Primary Schools in Gauteng in 2008 SE, 2011 ANA and 2012 ANA Grade 3 Literacy Section (percentage points)

| Grade 3 literacy scores | (1) Full sample | (2) 35-45% sample | (3) 34-46% sample | (4) Full sample | (5) 35-45% sample | (6) 34-46% sample | (7) 45-55% sample | (8) 25-35% sample |
|---|---|---|---|---|---|---|---|---|
| GPLMS | -31.81*** | -5.437** | -6.385*** | | | | | |
| | (0.850) | (2.380) | (2.215) | | | | | |
| Treatment effect in 2011 (Regime 1) | 13.90*** | -1.108 | 0.854 | | | | | |
| | (1.202) | (3.366) | (3.132) | | | | | |
| Treatment effect in 2012 (Regime 1) | 20.26*** | 4.286 | 5.199* | | | | | |
| | (1.202) | (3.366) | (3.132) | | | | | |
| Treated | | | | -30.63*** | -3.612* | -4.629** | | |
| | | | | (0.848) | (2.138) | (2.011) | | |
| Treatment effect in 2011 (Regime 2) | | | | 13.23*** | -2.242 | -0.534 | | |
| | | | | (1.199) | (3.024) | (2.843) | | |
| Treatment effect in 2012 (Regime 2) | | | | 19.39*** | 2.342 | 3.220 | | |
| | | | | (1.199) | (3.024) | (2.843) | | |
| Pseudo 1 (50% cut-off) | | | | | | | -5.187** | |
| | | | | | | | (2.061) | |
| Treatment effect in 2011 (Pseudo 1) | | | | | | | 1.737 | |
| | | | | | | | (2.915) | |
| Treatment effect in 2012 (Pseudo 1) | | | | | | | 4.660$^{\$}$ | |
| | | | | | | | (2.915) | |
| Pseudo 2 (30% cut-off) | | | | | | | | -4.606*** |
| | | | | | | | | (1.354) |
| Treatment effect in 2011 (Pseudo 2) | | | | | | | | 3.936** |
| | | | | | | | | (1.914) |
| Treatment effect in 2012 (Pseudo 2) | | | | | | | | 5.445*** |
| | | | | | | | | (1.914) |
| 2011 | -3.953*** | 4.433 | 3.310 | -3.529*** | 5.244** | 4.318* | -5.140** | 6.234*** |
| | (0.991) | (2.871) | (2.727) | (0.988) | (2.547) | (2.443) | (2.203) | (1.409) |
| 2012 | 4.492*** | 10.30*** | 10.05*** | 5.045*** | 12.49*** | 12.12*** | 2.881 | 19.83*** |
| | (0.991) | (2.871) | (2.727) | (0.988) | (2.547) | (2.443) | (2.203) | (1.409) |
| Constant | 59.01*** | 42.74*** | 43.02*** | 58.24*** | 41.66*** | 41.95*** | 52.40*** | 32.15*** |
| | (0.701) | (2.030) | (1.928) | (0.698) | (1.801) | (1.728) | (1.558) | (0.996) |
| Observations | 2,613 | 297 | 372 | 2,694 | 372 | 447 | 210 | 897 |
| R-squared | 0.55 | 0.25 | 0.26 | 0.53 | 0.26 | 0.26 | 0.21 | 0.40 |

Standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, $^{\$}$ $p < 0.15$