Disentangling the language effect in South African schools: Measuring the impact of 'language of assessment' in grade 3 literacy and numeracy

NICHOLAS SPAULL

Stellenbosch Economic Working Papers: 19/16

KEYWORDS: LANGUAGE IN EDUCATION, ASSESSMENT, LITERACY, ENGLISH SECOND LANGUAGE
JEL: I24, I25, I28

NICHOLAS SPAULL
SARCHL CHAIR IN INTEGRATED STUDIES OF
LEARNING LANGUAGE, MATHEMATICS AND SCIENCE
IN PRIMARY SCHOOL
UNIVERSITY OF JOHANNESBURG
SOUTH AFRICA
E-MAIL: NICHOLASSPAULL@GMAIL.COM

UNIVERSITEIT STELLENBOSCH UNIVERSITY

BER BUREAU FOR ECONOMIC RESEARCH

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

Disentangling the language effect in South African schools: Measuring the impact of ‘language of assessment’ in grade 3 literacy and numeracy[*]

NICHOLAS SPAULL[†]

ABSTRACT

The aim of this paper is to exploit an unusual occurrence whereby a large group of South African grade 3 students were tested twice, 1 month apart, on the same test in different languages. Using a simplified difference-in-difference methodology, it becomes possible to identify the causal impact of writing a test in English when English is not a student’s home language for 3402 students. The article aims to address the extent to which language factors (relative to non-language factors) can explain the high levels of underperformance in reading and mathematics in South Africa. I find that the language of assessment effect is between 0.3 and 0.7 standard deviations in literacy and 0 and 0.3 standard deviations in numeracy. This is approximately 1–2 years worth of learning in literacy and 0–1 year worth of learning in numeracy. By contrast, the size of the composite effect of home background and school quality is roughly 4 years worth of learning for both numeracy (1.2 standard deviations) and literacy (1.15 standard deviations). These results clearly show that the ‘language effect’ should be seen within the broader context of a generally dysfunctional schooling system. They further stress the importance of the quality of instruction, not only the language of learning and assessment. The fact that the literacy and numeracy achievement of South African children is so low in grade 3 (prior to any language switch to English in grade 4) should give pause to those who argue that language is the most important factor in determining achievement, or lack thereof, in South Africa.

Keywords: Language in education, assessment, literacy, English Second Language
JEL codes: I24, I25, I28

[*] Competing interests: The authors declare that they have no financial or personal relationships which may have inappropriately influenced them in writing this article.

[†] The author would like to thank Professor Servaas van der Berg and Dr Dieter von Fintel as well as the rest of the Research on Socioeconomic Policy (RESEP) team at Stellenbosch University for useful comments on an earlier draft of this paper.

Introduction

The topic of language in education is a contentious one internationally, and this is particularly the case in the South African context. While many countries have suffered the subjugating effects of colonisation and linguistic imperialism – including South Africa under the British – South Africa was also subject to 46 years of legislated racial exclusivity and State-sponsored linguistic inequality under apartheid. The language policies introduced during apartheid held both symbolic and practical value for the ruling government and were consequently resented by the majority of black South Africans. This resentment reached its zenith in the Soweto Uprising on the 16 June 1976 when over 20 000 students protested in the streets in opposition to the introduction of Afrikaans as the medium of instruction (Ndlovu 2004). Tragically, the police massacred hundreds of the protesting students, creating one of the most infamous and influential moments of the anti-apartheid struggle in South Africa. For the purposes of the present discussion, it is worth including one excerpt from the minutes of the General Students' Council from 1976:

> The recent strikes by schools against the use of Afrikaans as a medium of instruction is a sign of demonstration against schools' systematised to producing 'good industrial boys' for the powers that be … We therefore resolve to totally reject the use of Afrikaans as a medium of instruction, to fully support the students who took the stand in the rejection of this dialect (and) also to condemn the racially separated education system. (Karis & Gerhart 1997:569 cited in Ndlovu 2004)

From this quote, one can see that the Soweto Uprising of 1976 was in resistance both to the Afrikaans language policy and also to the unequal quality of education offered in the separate education systems (see also Fiske & Ladd 2004; Mesthrie 2002). While it may seem strange to discuss the intricacies of the Soweto Uprising in an article dedicated to the causal impact of language on performance, this is done so as to highlight an important parallel between the two topics: the distinction between the language of instruction and the quality of instruction. More often than not, language scholars conflate these two issues of language and quality but then proceed to talk about only language, as if quality was somehow subsumed under the all-encompassing umbrella of language. As will become clear, it does not. Isolating the causal impact of either of these factors is particularly difficult in South Africa given that they are both highly correlated and also strongly associated with other factors that influence performance, factors such as parental education, teacher quality, resources, geographic location, school functionality and socio-economic status.

The aim of this article is to try and disentangle these two highly correlated impacts in order to provide some empirical evidence regarding the size of these effects and particularly the impact of language after accounting for quality and home background. To do so, I exploit two factors: (1) the fact that the vast majority of South African students are taught in their mother tongue for the first

3 years of schooling before switching to English[1] in grade 4 and (2) that it is possible to identify and match 3402 grade 3 students who were sampled and included in both the Systemic Evaluation of September 2007 and then also the National School Effectiveness Study (NSES) of October 2007. These two surveys used the same test instrument with the exception that the first test (Systemic Evaluation) was written in the language of learning and teaching (LOLT) of the school – typically an African language when the majority of the students are black – and the second test (NSES) written 1 month later was written in English. Furthermore, the NSES sample was a sub-sample of the Systemic Evaluation making it possible to match a significant number of students across the two surveys. Using these matched students and their performance in the two tests, one can identify what proportion of the score achieved by students in numeracy and literacy is attributable to writing in English and what proportion is attributable to other factors.

Literature review and background

Throughout the world, scholars have been at pains to stress the links between language and nationhood (Weber 1976), language and identity (Edwards 2012), language and culture (Kramsch 1993) and language and power (Fairclough 1989). Most of these scholars – and particularly those who deal with language and education – have argued that policy decisions about language in education must consider far more than simply communicative efficiency, test scores or functional literacy. Applying these insights to the South African context, Neville Alexander has argued persuasively that South Africa's colonial and apartheid history further cement these links between language, class, power and identity (see Alexander 2005 for an overview).

While it is true that that the issue of language in education cannot be reduced to a discussion of fluency, proficiency and literacy scores (in both home language and in English), it is also true that these are legitimate areas of enquiry when speaking about language in South Africa, or any other country. Given that this is the focus of the present study and that the broader issues have been discussed at length elsewhere (see Mesthrie 2002; Murray 2002 for overviews), the discussion turns to the relationship between language proficiency and academic achievement.

Fleisch (2008) and Hoadley (2012) usefully summarise the most prominent causal theories showing how these two outcomes (language and achievement) are inter-related. The five 'mutually reinforcing and interconnected causal mechanisms' (Fleisch 2008:105) that they identify are (1) transfer theory and the density of unfamiliar words, (2) emotions of second-language teaching, (3) code-switching, (4) English language infrastructure and (5) language and power. Table 1 summarises some of the literature from each of these areas and categorises each

[1] Technically, students can switch to either English or Afrikaans, but in reality almost all students who do switch language in grade 4 switch to English (Taylor & Von Fintel 2016). See also Figure 1. For the remainder of the article, I therefore speak about 'switching to English' rather than 'switching to English or Afrikaans'.

one according to the purposes of this study. These are (1) language factors, (2) non-language factors and (3) factors where there is an interaction between language and non-language factors. It further splits the literature by (1) learners/learning, households/parents (2) teachers/teaching and (3) assessment. The intention here is not to provide an exhaustive list of factors but rather a list that is indicative of the types of factors in each category.

TABLE 1: Factors related to Language of Learning and Teaching (LOLT) and student performance on assessments.

Factors related to LOLT and student performance on assessments	**Teachers/teaching**	**Learners/learning and households/parents**	**Assessment**
Language factors	(1) Teacher proficiency in LOLT (Cazabon, Nicoladis & Lambert 1997; Heugh 2012; Macdonald & Burroughs 1991), (2) teacher training in LOLT, (3) teacher confidence in LOLT, (4) lack of teacher support material in the LOLT (Welch 2011), (5) length of instruction in African language (Taylor & Von Fintel 2016)	(1) Density of unfamiliar words and the inability to 'move' to a new language (Heugh 2012; Macdonald & Burroughs 1991), (2) Emotions of learning in a second language (Probyn 2001), (3) Lack of exposure to English language infrastructure in the school, community and the home (especially for rural students) (Setati *et al*. 2002; Welch 2011)	(1) Lack of exposure to the test language (English) at home (Howie *et al*. 2007; Reddy 2006), (2) understanding of the language-content of the test, (3) the quality of the translation/versioning (Stubbe 2011)
Non-language factors	(1) Teacher content knowledge (N. Taylor & S. Taylor 2013; Venkat & Spaull 2015), (2) Pedagogical content knowledge (Ball, Hill & Bass 2005; Carnoy, Chisholm & Chilisa 2012), (3) curriculum coverage (Reeves, Carnoy & Addy 2013) (4) teacher absenteeism (Prinsloo & Reddy 2012), (5) teacher professionalism (NPC 2012; N. Taylor 2011), (6) school functionality (NEEDU 2013).	(1) Parental education and household socio-economic status (Timæus, Simelane & Letsoalo 2013), (2) exposure to quality preschool education (Heckman 2000), (3) nutrition, socio-emotional stimulation and child health (Shonkoff *et al*. 2012)	(1) Psychometric validity of the test, (2) difficulty level of the test, (3) length of the test (for overviews, see Greaney & Kellaghan 2008; Postlethwaite & Kellaghan 2008)
Interaction between language and non-language factors	(1) Teachers restrict classroom interactions to low-level cognitive tasks due to children's insufficient language proficiency (Heugh 2005a, 2005b; Macdonald 1990; Macdonald & Burroughs 1991), (2) teaching using code-switching and language translation takes additional time that the curriculum may not accommodate (Setati & Adler 2000).	(1) Students who cannot read (properly) in the LOLT cannot learn (properly) in the LOLT (Macdonald 1990; Mullis *et al*. 2011)	

The aim of the present article is not to discuss all the above literature in detail, but simply to show the main themes of existing language-related research. For a more comprehensive discussion, see Taylor and Taylor (2013b). One issue in Table 1 that is worth briefly discussing is the issue of transfer theory and the density of unfamiliar words (Fleisch 2008:105). Partially because this has received considerable scholarly attention (both locally and internationally) but also because it provides a good case study of the limitations of qualitative research and the inability or unwillingness of South African education researchers to adequately recognise and acknowledge these limitations.

Drawing on language acquisition theory and particularly the work of Cummins (1984, 2000) and Skutnabb-Kangas (1988, 2000), researchers have argued that students need to first master the decontextualised discourse of schooling before switching to a second language (Alidou *et al*. 2006; Heugh 1993, 2005a, 2005b, 2012). Macdonald (1990) identified that black grade 5 Setswana children had at most 700 words in English when the curriculum required at least 7000 (Hoadley 2012:189). This, together with their insufficient grasp of the linguistic structure of English seriously limited their ability to read (and particularly to read for meaning) in English. Following on from this, children who have not learnt to read cannot read to learn. One of the most prominent research projects looking at language and the transition from mother tongue to English was the Threshold Project carried out by Carol Macdonald and various colleagues in 1987. These case studies focused on the language learning difficulties of African children when they switch from their mother tongue to English in four schools. In their discussion of this project, Macdonald and Burroughs (1991:58) conclude as follows:

> In the DET[2] curriculum, the present policy means that not enough time is given to English in order to prepare the children for learning in English in Standard 3 [Grade 5]. In other words, English is merely taught as a subject in the lower primary, which is unsatisfactory if English is to become the language of instruction in Standard 3 [Grade 5]. Up to a third of the total teaching and learning time should be devoted to the learning of English.

The research emanating from the Threshold Project has been particularly influential as far as South African language policy and research is concerned. For example, despite being conducted in 1987, the above quote from 1991 essentially summarises the view that has subsequently found its way into the new curriculum (DBE 2011:9), which introduces a minimum time requirement for First Additional Language (English in most cases). It is also expressed in the National Development Plan, which states that, 'learners' home language should be used as medium of instruction for longer and English introduced much earlier in the foundation phase' (NPC 2012:304). The Threshold Project is still regularly referred to in the literature (Fleisch 2008; Heugh 2012; Hoadley 2012)

[2] Department of Education and Training (DET) referred to the education system reserved for Black South Africans under apartheid.

despite having been conducted in 1987. To be sure, the influence of these case studies is largely warranted given its in-depth, innovative and methodologically rigorous approach to the topic.

Notwithstanding the above, it is worth emphasising three points that call into question the external validity of the study: (1) the Threshold Project was essentially a case study of four schools (Lefofa, St Camillus, Selang and Seroto), which were all situated in one circuit (Moretele Circuit) in one homeland (Bophuthatswana) (Macdonald 1990:8), (2) because of the fact that homelands were linguistically zoned, all these students were Setswana speakers, which is 1 of the now 11 official South African languages, and (3) the majority of the research was conducted almost three decades ago in 1987 when there was a different curriculum, with different teacher training institutions and different levels of resources and when the language switch to English happened 1 year later (grade 5) than it does now (grade 4). It is unfortunate that the study has not been replicated in other contexts or in more recent years because these newer studies could point to context-specific factors (if there are any) or how things have changed since 1987.

In essence, the Threshold Project tells us a great deal about how the children in these four schools manage the transition from an African language to English in Grade 5. Many of these findings do seem to be generalisable to other African-language students who face similar constraints (linguistic and otherwise) when switching from an African language to English. This being said, we should be cautious about immediately generalising findings from any case study to all South African schools where students switch from an African language to English (i.e. the vast majority). The four schools that were included in the Threshold Project may have been more or less functional than the average school, may have had more or less resources than the average school, may have had more or less capable teachers than the average school, may have had students who were more or less linguistically homogenous than the average school. All these factors are likely to affect how students transition from their home language into English at school. While these four schools may have been relatively representative of primary schools in the Bophuthatswana homeland, one should be cautious of extending the generalisability to schools in other homelands, because Bophuthatswana may have been quite different to the other homelands. For example, Chisholm (2013) explains that by 1985, the vast majority of primary schools in Bophuthatswana (760/840 schools) had experienced the Primary Education Upgrade Programme (PEUP). In this regard, she explains that

> A decade after it was first introduced, the PEUP was described as having "infused primary education in Bophuthatswana with a new spirit and orientation" and for being responsible for its much better educational showing than other Bantustans. (Chisholm 2013:403; Taylor 1989).

The aim in highlighting these potential external validity concerns is not to call into question the findings of the Threshold Project – findings which seem to have been confirmed in other less in-depth studies (Setati *et al*. 2002; Taylor, Van der berg & Mabogoane 2013) – but rather to stress the paucity of rigorous research on language transition in South Africa post-apartheid. Thus, Hoadley (2012:193) is correct in stating that:

> The question of why, and by how much language and especially learning in an additional language, affects achievement remains open. Fleisch (2008) makes the important observation that it is very likely that the use of English as the language of instruction is likely to have different effects across different groups of learners, especially with regard to social class and those in rural and urban areas. In other words, a consideration of the social context in which any language is being taught needs to be considered.

This is in stark contrast to Heugh (2012) who summarises the 'large body of South African research on bilingual education and transitional bilingual programmes' and concludes that:

> There is no need for more research to identify the problem or how to remedy it. The answers to these questions have already been established through research conducted in South Africa. There is no reliance on international research in this regard. (Heugh 2012:14)

However, it is not entirely clear which large body of South African research Heugh is referring to. It is perhaps telling to look at the studies which Heugh (2012:13) presents as her selection of this large body. Apart from the work of Malherbe (1946), the remaining three references are two case studies and a policy document. The first case study (Ianco-Worrall 1972) observes 30 White Afrikaans-English bilinguals in Pretoria, the second (Macdonald 1990) looks at four schools in Bophuthatswana in 1987, as I have discussed above, and the policy document (LANGTAG 1996) is not even a research document and does not present research findings, it was meant to advise the Minister of Education on developing a National Language Plan for South Africa. For a similarly small, case study–type approach, Brock-Utne (2007) observes two classes of isiXhosa children and concludes that they learn better when being instructed in their home language. While case studies are especially important in this field, they cannot be generalised to large populations unless they are sampled in such a way that they are representative of that underlying population (which has never been done in South Africa) or are replicated in a number of different contexts. Case studies are indicative and can point to underlying problems and potential solutions, but before they can inform policy, they need to be replicated in multiple contexts or with a large sample of schools (both of which ensure the findings are not context-dependent). For a recent exception to this general paucity, see Taylor and Von Fintel (2016), who employ a quantitative approach using administrative and assessment data for 9180 schools in South Africa. They find that mother tongue instruction in the early grades significantly improves English acquisition, as measured in grades 4, 5 and 6. See also Pretorius and Spaull (2016), who use a large (1772) sample of grade 5 rural English Second Language students to estimate the relationship between oral reading fluency and comprehension.

Caveat and extension

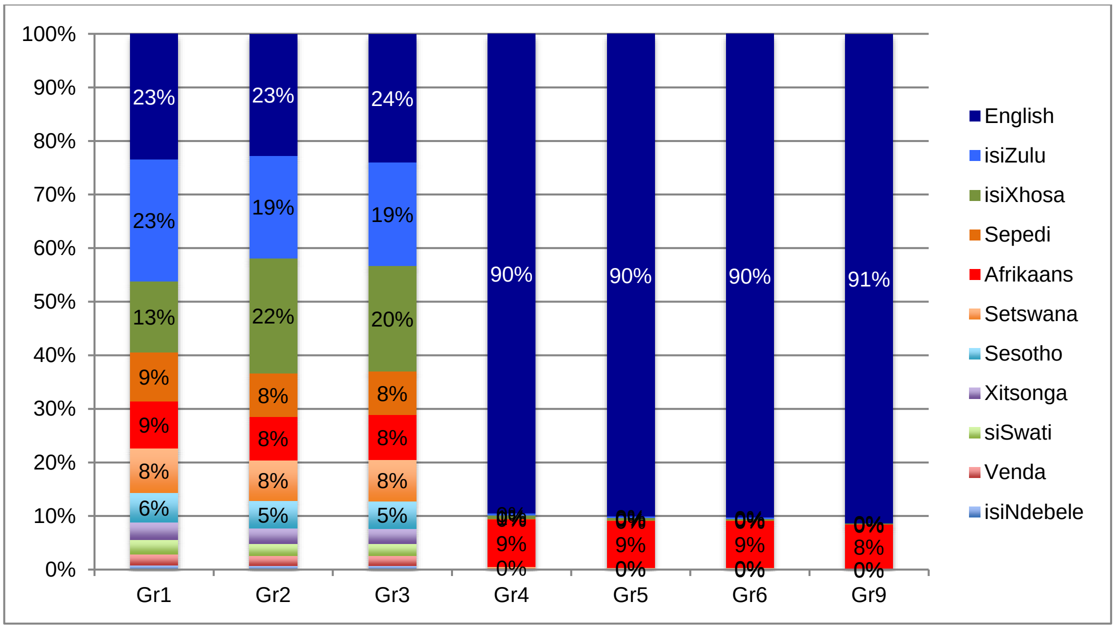
Where the present study differs from most previous quantitative work on language and achievement is that it focuses on grade 3, the period before students switch to English in grade 4. By observing students 'pre-switch', we are essentially controlling for all the 'language factors' in Table 1 and avoiding confounding influences inherent in any analysis of language post-switch. If one were to analyse students in grade 6, for example, it would be difficult to disaggregate what proportion of a student's performance was 'attributable' to language and what proportion to other factors like teacher quality, parental education or resources at home – all of which interact with language in complex ways. Given how highly correlated language and non-language factors are, if a non-English grade 6 student writes a test in English, it is unclear what proportion of their performance is attributable to language factors and what proportion to non-language factors. Even if we compared grade 6 students' performance on tests conducted in their home language and in English, it would not be clear what proportion of their achievement on tests conducted in their home language was because of language and what proportion was because of other language-related factors such as writing in a language (home language), which they are not currently taught in (English) or have been learning in since grade 4, or alternatively, the impact of a teacher who is not familiar with, or sufficiently proficient in teaching through, English as a medium of instruction. By looking at grade 3, these confounding factors fall away – students are assessed in the language they know best and in which they have been taught for 3 years, most teachers are teaching in their mother tongue (which is also the LOLT of the school) and students have not yet switched to English. Thus, there are few (if any) confounding language factors that could affect a child's numeracy or literacy performance at the end of grade 3. Put differently, one cannot talk about language-switching factors being a main cause of poor performance for non-English students at the end of grade 3, something which is probably not true of student performance in grade 4 or grade 6, for example.

By the end of grade 3, most non-English students have had very little (if any) exposure to English in or outside the classroom. English instruction was not timetabled in the National Curriculum Statement (NCS) for grade 3 – the prevailing curriculum in 2007, the period under analysis. Given that almost all non-English students switch to English as LOLT in grade 4, the difference in performance when students write a test in their home language relative to English is likely to be higher in grade 3 than in any subsequent grade. This is the reason why the estimates presented in this study cannot be generalised to higher grades. In higher grades, students' exposure to English should decrease the difference in performance between a test written in their home language and one written in English. Thus, one can think of the estimates presented here as the maximum possible language disadvantage attributable to writing a test in English for non-English students.

Language in education in South Africa

The language in education policy in South Africa supports children being taught in their home language for at least the first three grades of primary school and thereafter to switch to either English or Afrikaans. Figures from the 2011 Census show that only 23% of South African citizens speak either English or Afrikaans as their first language (StatsSA 2012:23), and consequently, it is the vast majority of students who experience a LOLT switch in grade 4. Figure 1 vividly illustrates this situation using data from the Annual National Assessments of 2013, which tested all students in grades 1–6 and 9 in languages and mathematics. From Figure 1, one can see that while 32% of students learn in English or Afrikaans in grades 1–3, this figure increases dramatically to 99% in grade 4. Almost all students who learn in an African language in grades 1–3 switch to English in grade 4.

Figure 1: Breakdown of language of learning and teaching (LOLT) by grade – Annual National Assessments 2013 (*n* = 7 630 240, own calculations using 'loa_lang').



Although the present study does not look at whether, when, why or how students should transition from an African language to English, this study is aimed at contributing some empirical evidence to the debate regarding how much language (as opposed to other factors) affects achievement.

Research questions

The aim of the present article is to isolate the causal impact of writing a test in English when English is not a student's home language. This broad research area can be broken down into the following research questions:

What is the 'cost' (in terms of marks forgone) when students are forced to write a numeracy test in English when English is not their home language?

How much worse do students do on high-language-content numeracy items versus no-language-content numeracy items when they are posed in English when English is not the student's home language?

What is the 'cost' (in terms of marks-forgone) when students are forced to write a literacy test in English when English is not their home language?

For students' whose home language is not English, does the 'cost' mentioned above differ between items testing the five different literacy processes of: (1) cloze items and items requiring students to match words to pictures, (2) items which require that students focus on and retrieve explicitly stated information, (3) items which require students to make straightforward inferences, (4) items which require students to interpret and integrate ideas and information and (5) items which require students to write sentences. If so, how large are these differences?

For students' whose home language is not English, does the 'cost' mentioned above differ when items are phrased in multiple-choice format or free-response format? If so, how large is the difference?

The major problem inherent in answering these questions in the South African context is that one cannot simply use a single test written in English and compare the outcomes of students whose home language is English with the outcomes of students for whom English is a second (or third) language. This is because English and non-English students differ in a number of observable and unobservable ways, which confound the comparison. This is a fact that is widely acknowledged in the South African literature:

> The extent to which language factors contribute to this low performance is not clear, given that language disadvantages are so strongly correlated with other confounding factors such as historical disadvantage, socio-economic status, geography, the quality of school management and the quality of teachers. (Taylor & Von Fintel 2016:75)

Data and identification strategy

To estimate the causal impact of test language on test performance in the South African context, one can employ one of two methods; either one can sample a large group of students and then randomly allocate half to writing the test in English and the other half to write it in their mother tongue. Provided that the group is sufficiently large, any observed or unobserved differences should be negligible across the two groups. Alternatively, one can test the same group of students twice in a relatively short space of time. The advantage of the second method is that one does not need as large a sample because factors that do not vary between the tests will be differenced out (things like teacher quality, home background, parental education, etc.). By using the same group of students across the two tests, one is effectively imposing *ceteris paribus* conditions with two exceptions: (1) Because students will have already seen the test, they may perform better on Test 2 than on Test 1 simply because they remember some of the items and (2) students may learn new skills or reinforce previous work in the period between the two tests, which would lead to better marks in the second test that are independent of language. Both these instances would lead to a positive bias in the second test. Given that our a priori is that students perform better on assessments when they are set in their home language, we would argue[3] that the best sequencing of the two tests would be to test students in their mother tongue first and in English second, rather than the other way around. This is the conservative method of estimating the difference because the positive biases mentioned above (if they exist) will decrease the difference between the two tests rather than increase the difference as would be the case if students were tested in English first.

Running a large experiment for the sole purpose of testing the causal impact of test language was not possible in the present instance; however, it was possible to exploit a unique situation in South Africa where a group of students happened to be sampled twice – for two different surveys – with tests written 1

[3] It is perhaps easiest to explain by example: if we assume that students score 25% when they write a test in English and 45% when they write the same test in their home-language the 'true' causal impact would be negative 20 percentage points. Let us further assume that the two biases mentioned above contribute to an additional 5 percentage points for the second test relative to the first test due to their 'learning effect'. Given that we do not know the size of this learning effect bias, if we tested students first in English and second in mother-tongue we would estimate the causal impact to be 25 percentage points (25% – (45% + 5%). If we tested students first in mother-tongue and second in English we would estimate the causal impact to be 15 percentage points (45% – (25% + 5%). Given that we would rather be conservative in our estimate we would argue that it is better to test students first in their mother-tongue and secondly in English and estimate a lower-bound causal impact of writing a test in English when English is not a student's mother-tongue. Furthermore, by including a *within*-test difference (in addition to the *between*-test difference), the present difference-in-difference analysis accounts for both of these biases as long as they affect all item categories equally – this is discussed in more detail later in the article where the difference-in-difference method is explained further.

month apart. In September 2007, the Systemic Evaluation tested a nationally representative sample of 54 298 grade 3 students from 2327 primary schools (DoE 2008:1). The aim was to measure the levels of achievement in literacy and numeracy relative to grade-appropriate curriculum outcomes. At the same time, the NSES was being planned and implemented by the Joint Education Trust, the same organisation who was providing technical support to government for the Systemic Evaluation (SE) Test. The NSES decided to test a sub-sample of grade 3 students from the Systemic Evaluation sample 1 month later (October) and tested approximately 16 000 students from 268 schools. The NSES used the same instrument as the Systemic Evaluation with one major exception: where the Systemic Evaluation tests (Test 1) were written in the LOLT[4] of the school at the grade 3 level, the NSES tests (Test 2) were written in English (Taylor *et al*. 2013:18). The implementers of the NSES explain their rationale as follows:

> While SE tests were written in the home language of the learners at Grade 3 level, the NSES tests were written in English. The reason behind this decision was that the NSES followed the same cohort of learners for 3 years, administering the same test annually. Because most schools for African learners change their medium of instruction in Grade 4 from mother tongue to English, we wanted to have comparable scores for the same learners for each of the three years. Thus while at Grade 4 level the learners would have been disadvantaged by writing in a language with which they are unfamiliar, this design enabled us to compare scores directly across the three years. Because the NSES schools were a subsample of the SE sample the design also provided a unique opportunity to compare scores by the same Grade 3 learners on the same test written first in their mother-tongue and second in English. (Taylor *et al*. 2013:18)

Matching students across tests

Given that South African students do not have unique identification numbers, it was not possible to match all students between the two tests. In addition, the selection procedures employed by the NSES were different from that of the Systemic Evaluation. Where the Systemic Evaluation randomly selected 25 students from a class, the NSES tested all students in the class (Taylor & Taylor 2013b:147). In their analysis of learner performance in the NSES, Taylor and Taylor (2013b) also compare the performance of students between the Systemic Evaluation and the NSES using a similar method to that employed here. To match individuals between the two samples, they used four matching criteria: (1)

[4] Although the Taylor *et al*. (2013) quote says 'in the home language of the learners', this is technically not true. To the extent that the home language of the learner corresponds to the LOLT of the school (which is not always the case), this is correct because the Systemic Evaluation was conducted in the LOLT of the school not in the home language of the learner.

the unique school administrative (EMIS[5]) number, (2) the first three letters of the child's surname, (3) the first letter of their first name and (4) the child's gender (Taylor & Taylor 2013b:147). Using this approach, they were able to match 2119 learners in both the NSES and the Systemic Evaluation data sets. The matching criteria employed by these authors is relatively stringent as the authors themselves identify:

> The matching process was conservatively done in the sense that errors of excluding learners who did in fact participate in both evaluations were far more likely than errors of false matches. (Taylor & Taylor 2013b:147)

Given that Taylor and Taylor (2013b) were only able to match 2119 students of the 16 000 that participated in NSES and that these 2119 may be quite different to the unmatched students, they provide a sensitivity analysis comparing performance on the NSES between the matched and unmatched sample – reproduced in Table 2.

Table 2: Taylor and Taylor's (2013b:150) comparison between the matched (SE and NSES) and unmatched (NSES only) samples (reproduced verbatim).

	NSES Literacy Score	**NSES Numeracy score**	**Number of learners**
Unmatched (NSES only)	17.34%	24.57%	14 384
Matched sample	23.08%	33.62%	2119

NSES, National School Effectiveness Study; SE, Systemic Evaluation.

Taylor and Taylor (2013) explain that the difference in performance between the matched and unmatched sample could be driven by two factors: (1) that weaker children were more likely to make mistakes writing their names than more literate children leading to more non-matches among weaker children and (2) because the selection of the 25 students in the Systemic Evaluation may not have been entirely random and instead teachers may have somehow ensured that better students were selected for the Systemic Evaluation (and thus effectively matched) (Taylor & Taylor 2013b:148).

For the purposes of the present comparison, we employed a different matching technique and were able to match significantly more students. To match students, we used two criteria: (1) the school's unique administrative (EMIS) code and (2) the student's birthday, birth month and birth year. Doing so allowed us to match 3402 unique students, which amounts to 61% more students than those matched by Taylor and Taylor (2013). The major problem with this matching strategy is that there is a relatively high probability that two children in a particular class will share a birthday. Using the formula below, one can see that in a class of 30 students the probability is 70.6% that two students share the same birthday.

[5] EMIS stands for the Education Management Information System. Schools' EMIS numbers uniquely identify all schools in South Africa.

$$p(n) = 1 - \frac{365!}{365^n(365-n)!}$$

While this may seem problematic at first, the reduction in sample size from dropping all students who share birthdays in a particular school is relatively small compared to more stringent matching criteria. Furthermore, we would argue that sharing a birthday with someone else in the class is completely random and therefore exogenous to student achievement or selection. Consequently, dropping these students from the analysis should not bias the results. However, given that we can only match students with non-missing birthday information, it is possible that in matching we select stronger students who are more numerate and therefore less likely to make mistakes. This is unavoidable but is also partially accounted for in the difference-in-difference analysis as discussed later. Table 3 shows the average numeracy and literacy scores for students in the Systemic Evaluation and the NSES for 'unique' students (i.e. no common birthdays) and duplicate students (common birthdays) as well as the total number of students. One possible reason why duplicates (or students missing date of birth information) perform worse is if weaker students are more likely to either forget their birthdays, make mistakes in writing them down, or forget to fill them in. If one compares the average numeracy and literacy scores for the total sample of students and those who do not share a birth date (i.e. unique observations after duplicates and missing data have been dropped), the average scores are not statistically significantly different. Throughout the present analysis, standard errors are calculated with clustering at the school level if average scores are being calculated and clustering at the individual level if the analysis is at the item level.

Table 3: Literacy and numeracy scores for grade 3 students in the Systemic Evaluation and NSES by uniquely identified individuals and duplicates.

	Test 2 – NSES Gr3 (October)			**Test 1 – Systemic evaluation Gr 3 (September)**		
	Total	**Unique**	**Duplicates and missing (on school and birth date)**	**Total**	**Unique**	**Duplicates and missing (on school and birth date)**
Mean literacy %	18.2%	19.2%	14.6%	32.4%	32.6%	30.2%
Standard error	0.75%	0.77%	0.87%	0.25%	0.25%	0.53%
Mean numeracy %	26.0%	27.5%	20.4%	33.8%	34.0%	31.7%
Standard error	1.18%	1.19%	1.50%	0.36%	0.36%	0.76%
Sample size	16 525	**13 033**	3492	54 298	**49 456**	4842

NSES, National School Effectiveness Study.

One further potential source of false matching is if students forget their birth dates and write something else down. This is unlikely to lead to false matches because it would require that two students both forget their birth date in one of the assessments and then both decide to pick the other student's birth date as their own in the other assessment. This is highly improbable.

Table 4 reports the average numeracy and literacy performance for the matched and unmatched samples of the NSES and the Systemic Evaluation. Summing the number of students between the unmatched Systemic Evaluation (46 054) and matched Systemic Evaluation and NSES (3402) provides the total unique observations in the Systemic Evaluation (49 456) in Table 3 and similarly for the NSES where the unmatched (9631) and matched (3402) samples sum to the total unique observations in the NSES (13 033) in Table 3.

Table 4: Average student performance in numeracy and literacy in the Systemic Evaluation and the NSES by matched and unmatched samples.

	Number of students	Numeracy		Literacy	
		SE	NSES	SE	NSES
Unmatched Systemic Evaluation Gr3 (Sept 2007)	46 054	34.0%		33.8%	
Standard error		0.10%		0.08%	
Unmatched NSES Gr3 (Oct 2007)	9631		25.7%		18.7%
Standard error			0.22%		0.15%
Matched NSES-SE sample	3402	33.4%	32.7%	34.4%	23.2%
Std. Err.		0.38%	0.40%	0.29%	0.26%

NSES, National School Effectiveness Study; SE, Systemic Evaluation.

From Table 4 one can see that matched students perform significantly better in the NSES than unmatched students in the NSES in both numeracy and literacy. However, for the Systemic Evaluation matched and unmatched students perform essentially the same.

Difference-in-difference analysis

For the present difference-in-difference analysis, the first difference is the difference between the student's score on a particular item in the Systemic Evaluation relative to that student's score on that item in the NSES, that is, a between-test difference. The second difference is the difference between item categories within a particular test, that is, a within-test difference. The between-test difference takes into account the difference in the language of the test and the within-test difference takes into account any student-specific or test-specific factors that may be different between the two tests but similar between item categories.

For the language test, the item categories follow the literacy-process categorisation of the items (match, retrieve, infer, interpret and write). For the numeracy test, the items are categorised according to the language content of the item (no language content, high language content and ambiguous language content). These categories are all discussed below.

Background information on the test instruments

Literacy test

The literacy test that was administered to grade 3 students in both the SE and the NSES was designed to reflect the reading and writing proficiency of grade 3 students in South Africa. Of the 40 items included in the test, most were set at the grade 3 level (30 items) but there were also questions set at earlier grade levels, specifically at the grade 2 (7 items) and grade 1 (3 items) levels. Taylor *et al*. (2013:31) have classified the 40 items that made up the literacy assessment according to the PIRLS[6] framework. PIRLS identifies four processes of comprehension: (1) focus on and retrieve explicitly stated information, (2) make straightforward inferences, (3) interpret and integrate ideas and information and (4) examine and evaluate content, language and contextual elements (Howie *et al*. 2007). Although PIRLS is a reading assessment, the literacy assessment used in the Systemic Evaluation and NSES covered both reading and writing. Consequently, Taylor and Taylor (2013) extend the PIRLS framework and include two additional categories: (1) cloze items and matching words to pictures and (2) writing tasks. The literacy test did not contain any items in the 'examine and evaluate content, language, and textual elements' category, and consequently, this category is dropped from the analysis in this article. Thus, Taylor and Taylor (2013) end up with five categories, which they refer as 'literacy processes'. Test items were also classified on whether they are multiple-choice items or free-response items. The distribution of test items by text type, literacy process and answering format can be seen in Table 5 (reproduced from Taylor & Taylor 2013b:33). For the present analysis, we use the same categorisation of items and collapse the categories of 'matching a word to a picture' and 'fill in a missing word (cloze)' primarily because the NCS, the prevailing curriculum at the time of testing, prescribes that these type of items should be mastered at the grade 1 level.

Table 5: Distribution of literacy test items in Test 1 and Test 2 according to text type and literacy process.

		Format	Purposes of reading (types of text)					Total no. Items
			Visual cue	Poster	Bar graph	Non-fiction descriptive	Fiction narrative	
Literacy processes	Matching word to picture	MC	1, 2					2
	Fill in missing word (cloze)	MC	3, 4, 5, 6, 7, 8, 9					7
	Retrieve	MC		10, 11	14, 15	19, 20, 21, 22, 23, 24	30, 31, 31	13
		FR			12, 13	25.26		4
	Infer	MC					33, 34	2

[6] PIRLS stands for the Progress in International Reading Literacy Study.

		FR				27, 28		2
	Interpret	MC					35, 36, 37	3
		FR				29	38, 39, 40	4
	Evaluate							0
	Write a sentence	FR	16, 17, 18					3
	Write a paragraph							0
Total number of items			12	2	4	11	11	40

MC, multiple choice; FR, free response.
Source: Taylor & Taylor 2013b:33.

Numeracy test

The numeracy test used in the Systemic Evaluation and the NSES consisted of 53 questions with items set at the grade 1 (2 items), grade 2 (14 items), grade 3 (30 items) and grade 4 level (7 items). Table 6 reports the breakdown of items by grade level and language content. The grade-level distinctions are sourced from Taylor and Taylor (2013:34). Given that the focus of the present analysis is the causal impact of writing a test in a second language, the 53 numeracy items were split into one of three categories based on the language content of the item. If a question consisted only of numbers and symbols (e.g. '24 ÷ 3 = ___'), it was classified as a 'No language content' item. If a question had some language content but could be solved by deductive reasoning without any understanding of the language, that item was classified as an 'Ambiguous item'. For example, question 4 is worded as follows: 'Count forward in 2s. Fill in the next number in the space provided; 74 76 78 ___'. An item was classified as a 'High language content item' if it was not possible to solve the problem without understanding the language content of the question. For example, question 22 asked, 'Mother is 77 years old. Father is 6 years older than her. How old is father? ____'. The aim in grouping items along a language-content dimension was to test the finding in the literature that students who write a test in a second language find word problems more difficult than those problems posed in symbolic format (for some examples, see Adetula 1990; Bernardo 1999; Ní Ríordáin & O'Donoghue 2008).

Table 6: Distribution of items in Test 1 and Test 2 grade 3 numeracy test by grade-level and language content.

		Language content			
		No language content	**Ambiguous items**	**High language content**	**Total**
Grade level	Grade 1	28		13	2
	Grade 2	35, 36	2, 3, 4, 16, 17	1, 10, 14, 19, 22, 29, 30,	14
	Grade 3	20, 21, 23, 24, 25, 37, 39, 42, 49	6, 7, 8, 18, 31, 32, 38, 45	9, 11, 12, 15, 33, 43, 44, 46, 47, 48, 51, 52, 53	30
	Grade 4	26, 34, 40, 41,	5	27, 50	7

	Total	16	14	23	53

Data structure

In order to perform the difference-in-difference analysis, the data need to be at the item level rather than the student level. That is to say that it should be transformed from a person-level database with N rows to an item-level database with $N \times K \times T$ rows, where N is the number of students, K is the number of items (40 in the case of literacy and 53 in the case of numeracy) and T is the number of tests (2). That is to say that the traditional data set of one row per student should be transformed, reshaping twice from wide to long to a data set of one row per item per test per student. In matrix-vector format, this transformation is represented as follows:

$$\boldsymbol{A}\text{:}\begin{bmatrix} \boldsymbol{q}_{11} & \boldsymbol{q}_{12} & \boldsymbol{q}_{13} & \cdots & \boldsymbol{q}_{1K} \\ \boldsymbol{q}_{21} & \boldsymbol{q}_{22} & \boldsymbol{q}_{23} & \cdots & \boldsymbol{q}_{2K} \\ \boldsymbol{q}_{31} & \boldsymbol{q}_{31} & \boldsymbol{q}_{31} & \cdots & \boldsymbol{q}_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{q}_{N1} & \boldsymbol{q}_{N2} & \boldsymbol{q}_{N3} & \cdots & \boldsymbol{q}_{NK} \end{bmatrix} \rightarrow \boldsymbol{B}\text{:}\begin{bmatrix} (\boldsymbol{q}_{11})' & (\boldsymbol{q}_{12})' & \cdots & (\boldsymbol{q}_{1K})' \\ (\boldsymbol{q}_{21})' & (\boldsymbol{q}_{22})' & \cdots & (\boldsymbol{q}_{2K})' \\ (\boldsymbol{q}_{31})' & (\boldsymbol{q}_{32})' & \cdots & (\boldsymbol{q}_{3K})' \\ \vdots & \vdots & \ddots & \vdots \\ (\boldsymbol{q}_{N1})' & (\boldsymbol{q}_{N2})' & \cdots & (\boldsymbol{q}_{NK})' \end{bmatrix} \rightarrow \boldsymbol{C}\text{:}\begin{bmatrix} (\boldsymbol{q}_{11})' \\ (\boldsymbol{q}_{12})' \\ (\boldsymbol{q}_{13})' \\ \vdots \\ (\boldsymbol{q}_{1K})' \\ (\boldsymbol{q}_{21})' \\ (\boldsymbol{q}_{22})' \\ (\boldsymbol{q}_{23})' \\ \vdots \\ (\boldsymbol{q}_{2K})' \\ \vdots \\ (\boldsymbol{q}_{N1})' \\ (\boldsymbol{q}_{N2})' \\ (\boldsymbol{q}_{N3})' \\ \vdots \\ (\boldsymbol{q}_{NK})' \end{bmatrix}$$

where $\boldsymbol{q}_{11} = [q_{1a} \quad q_{1b}]_{n=1}$, where a represents the NSES test and b represents the Systemic Evaluation.

It is not possible to use the weights provided in either the NSES or the Systemic Evaluation because the weights attached to students correspond to the original samples and not the smaller matched sample. Consequently, we do not weight the sample and do not claim that it is nationally representative. We do adjust the standard errors to account for clustering. When calculating mean scores, clustering is calibrated at the school level (student responses are clustered in schools), and when calculating mean scores for item categories, clustering is calibrated at the individual level (item responses are clustered within an individual).

Identifying home language

In order to estimate the causal impact of writing a test in English when English is not a student's home language, it is necessary to identify which students have English as their home language and which do not. This involves a second round of matching based on the question asking what a student's home language was. Table 7 shows the breakdown between matched and unmatched students by home language. From the table, one can see that 459 students from 158 schools could not be matched on the home-language variable across the two surveys, either because the variable was missing in one of the two surveys or because the listed home language was different between the two surveys. Note that the total number of matched schools (223) does not equal the sum of the total number of matched-schools-by-language. This is because it is possible to have students from multiple home languages in a single school. This is also the reason why the framing of the research question refers to students 'whose home-language is not English' rather than 'for whom English is a second language' because many of these students will only learn English as a third or fourth language.[7] The focus of most of this article is on the 2811 students who do not share a birthday with someone in their class (the first round of matching) and whose home language was consistently matched between the two tests (the second round of matching) and was also not English. The 132 successfully matched English home-language students will be used for robustness checks because these students wrote the same test twice in the same language 1 month apart and therefore create a useful reference category for test-specific differences.

Table 7: Total number of students matched consistently on home-language variable between Systemic Evaluation and National School Effectiveness Study.

Language groups matched consistently on home-language	**English home language**	**Non-English home-language**	**Total number of unique schools**	**Total**
Afrikaans		499	43	
English	132		24	
isiNdebele		49	10	
isiXhosa		498	58	
isiZulu		786	66	
Sepedi		286	37	
Sesotho		131	23	
Setswana		256	26	
SiSwati		109	13	
Tshivenda		66	7	
Xitsonga		131	17	
Total matched	132	2811	223	2943
Total unmatched	459		158	459
Total				3402

[7] For example, an isiXhosa student living in KwaZulu-Natal may be in an isiZulu school and therefore learning in isiZulu in grades 1–3 before switching to English (their third language) in grade 4.

Identification strategy

Estimating the difference-in-difference model for the language test can be accomplished in one of two ways. One could estimate the regression equation:

$$L_{nkt} = \lambda + \delta NSES_{nt} + \varphi_{1-4} Lit\_Cat_{ik} + \beta_{1-4}(NSES * Lit\_Cat)_{nkt} + \varepsilon_{nkt},$$

where L_{nkt} = is the average percentage correct in the literacy test for individual n on item category k in test t where $n \in (12\,811); k \in (1.5); t \in (0.1)$, where $t = 0$ for the Systemic Evaluation, $t = 1$ for the NSES, $k = 1$ for the 'cloze/word-matching' category of items, $k = 2$ for the 'retrieve' category of items, $k = 3$ for the 'infer' category of items, $k = 4$ for the 'interpret' category of items and $k = 5$ for the 'write a sentence' category of items. φ_{1-4} are the four coefficients corresponding to the four dummy variables of literacy categories (with 'cloze/word-matching' as reference group). This is typically the strategy employed where there are no data for the 'no treatment state' (Angrist & Pischke 2009:227). However, for the present analysis, we have data for all individuals on all items for both tests (i.e. for the treatment and control arms), and thus do not need to make additional assumptions about omitted variable bias and the required level of aggregation for differentiation, as one would typically need to do.

Given that we have data on all outcomes (treatment and non-treatment) for all students, using the regression equation to predict outcomes for sub-groups – the purpose of the present analysis – is mathematically equivalent to a table of means with t rows and k columns. Calculating the difference-in-difference from this table of means is equivalent to predicting the outcomes for each combination of literacy category (k) and specific test (t). Given that the regression coefficients are not directly interpretable (they must be summed across the combinations of dummy-variable categories and multiple interaction terms), we decided to rather use the table of means approach, which is more parsimonious and easier to interpret.

Findings – language results and literacy processes

Table 8, Figures 2 and 3 report the main findings from the literacy test analysis for students whose home language is English (*n* = 132) and those for whom it is not (*n* = 2811). As one would expect, students' whose home language is not English performed statistically significantly better when they wrote the test in the LOLT of the school (Test 1: average score 33%) than when they wrote it 1 month later in English (Test 2: average score 22%). Given that the standard deviation[8] for these students in the Systemic Evaluation literacy test (*n* = 2811)

[8] The standard deviations for the various groups are as follows: For all matched students (*n* = 3402), the standard deviation for the Systemic Evaluation (Test 1) literacy test was 16.4% and the standard deviation for the Systemic Evaluation numeracy test was 22.2%. For students whose home language was English (*n*

was 15.8%, one can say that students performed 0.69 (10.97/15.8) of a standard deviation worse in Test 2 (in English) than they did in Test 1 (in the LOLT of the school).

One could argue that the 0.69 estimate is a lower bound estimate because it is the net effect of the positive 'learning/familiarity' gain (from writing the same test twice, albeit in a different language) and the negative language cost (from writing Test 2 in English, a language with which they are unfamiliar). If we assume that the learning/familiarity gain among the English students between the two tests (2% points) is the same as the learning/familiarity gain among the non-English students between the two tests, then the language effect grows from 0.69 of a standard deviation to 0.82 of a standard deviation.

Observing the outcomes in the Systemic Evaluation (Test 1 – in the LOLT of the school), one can clearly see that students found the 'cloze/matching' items easiest (average score of 57%) and the 'interpret' questions most difficult (average score of 9%). Importantly, the average score for the whole test when written in the LOLT of the school was still only 33%. This is after students have been learning in their home-language for 3 years and before any switch to English in grade 4. This low level of performance 'pre-language-switch' provides some backing to the arguments made by Murray (2002) and reiterated by Hoadley (2012), who argue that there should be as much attention paid to the quality of instruction as there is to the language of instruction. This is one of the motifs that runs through much of the present analysis.

If one thinks that the three main factors affecting students' performance are (1) home background, (2) school quality and (3) language factors, it is possible to provide rough estimates for the size of the impact of (3) and a composite estimate of (1) and (2) combined. We have already seen that non-English students performed 0.69–0.82 of a standard deviation worse when writing in English relative to the LOLT of the school. This could be considered one estimate for the size of the 'language factor'. If one then only looks at the Systemic Evaluation and compares the performance of English home language students (average score 50%) and non-English home language students (average score 33%), the difference amounts to 1.08 (0.17/0.158) of a standard deviation.[9] This can be thought of as a composite estimate of (1) home background and (2) school quality. Disentangling (1) and (2) is far more difficult because one does not have

= 132), the figures for the Systemic Evaluation literacy test standard deviation were 18.6% and for the numeracy test 26.9%. If one looks only at students who do not speak English as a home language (*n* = 2811), the figures were 15.82% for literacy and 21.6% for numeracy.

[9] Using the standard deviation of non-English home language students in the Systemic Evaluation. One could argue for using a different standard deviation – perhaps the full Systemic Evaluation sample standard deviation; however, the differences can become confusing (and potentially misleading) because it is not only the difference that is changing but also the standard deviation that one is using to scale the difference. Furthermore, the difference in standard deviations between non-English Systemic Evaluation (15.8%) and total-matched Systemic Evaluation (16.4%) is not large. For this reason, I use the same standard deviation (Systemic Evaluation non-English sample) but have already reported alternate standard deviations in a previous footnote should anyone wish to use a different standard deviation.

exogenous variation in either (1) or (2) as we do for language with the two tests. Furthermore, separating out the effects of (1) and (2) is not the focus of this study.

Observing the outcomes in the NSES (Test 2 – in English), one can see that non-English-home-language students performed statistically significantly worse in three of the five categories (cloze/matching, retrieve, write a sentence), with roughly similar performance in the 'infer' and 'interpret' categories. By contrast, English-home-language students – who wrote the same test in English twice – performed better in Test 2 than in Test 1 for all literacy processes except the three 'write a sentence' items.

Table 8: Average performance (%) by literacy process in Test 1 (Systemic Evaluation) and Test 2 (National School Effectiveness Study) for students whose home language is and is not English [standard errors clustered at the individual level].

	Non-English-home-language students (*n* = 2811)					
	Cloze /matching-word-to-picture (9 items)	**Retrieve (17 items)**	**Infer (4 items)**	**Interpret (7 items)**	**Write a sentence (3 items)**	**Total (40 items)**
Test 2 (in English)	51.14	27.09	15.48	7.00	8.90	**22.07**
Standard error	0.41	0.36	0.31	0.16	0.34	**0.25**
Test 1 (in Home Language)	56.69	33.91	16.03	8.51	43.51	**33.04**
Standard error	0.45	0.40	0.32	0.18	0.58	**0.30**
Difference (Test 2 – Test 1)	−5.55	−6.82	−0.55	−1.51	−34.61	−**10.97**
Standard error	0.61	0.54	0.45	0.24	0.67	**0.39**

	English home-language students (*n* = 132)					
	Cloze /matching-word-to-picture (9 items)	**Retrieve (17 items)**	**Infer (4 items)**	**Interpret (7 items)**	**Write a sentence (3 items)**	**Total (40 items)**
Test 2 (in English)	81.57	64.87	39.85	22.04	42.23	52.06
Standard error	1.05	1.93	2.37	1.54	2.23	1.46
Test 1 (in Home Language)	75.42	57.58	36.21	14.60	57.32	50.04
Standard error	1.52	2.30	2.23	1.04	2.56	1.62
Difference (Test 2 – Test 1)	6.14	7.30	3.64	7.44	−15.09	2.02
Standard error	1.85	3.00	3.25	1.86	3.40	2.19

Figure 2: Average performance (%) in Test 1 (Systemic Evaluation) and Test 2 (National School Effectiveness Study) by literacy process for students whose home language is not English (n = 2811).

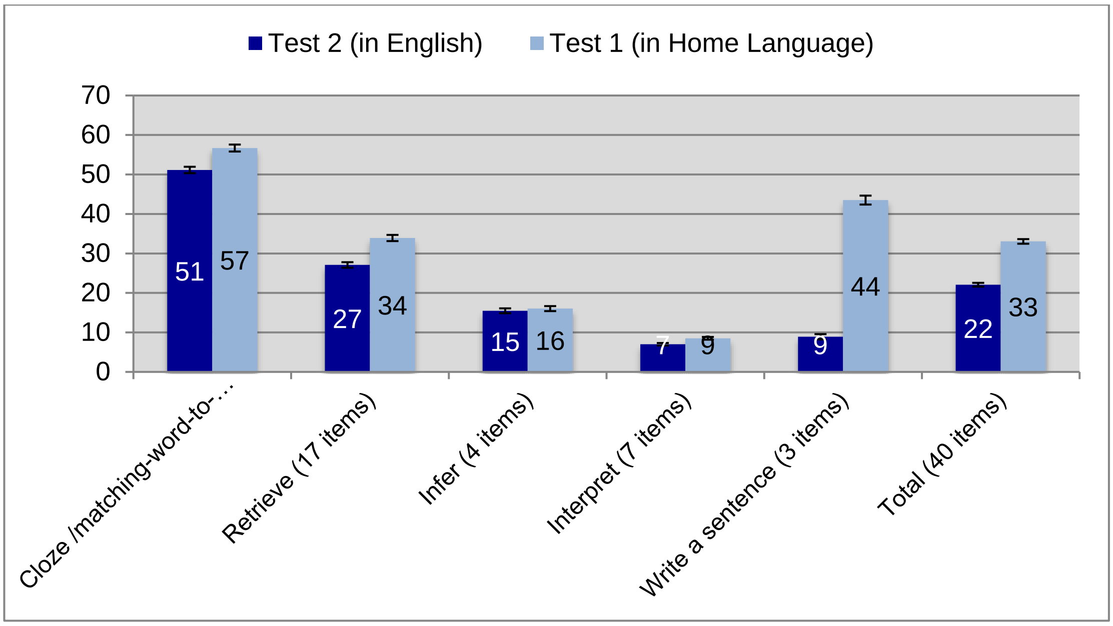
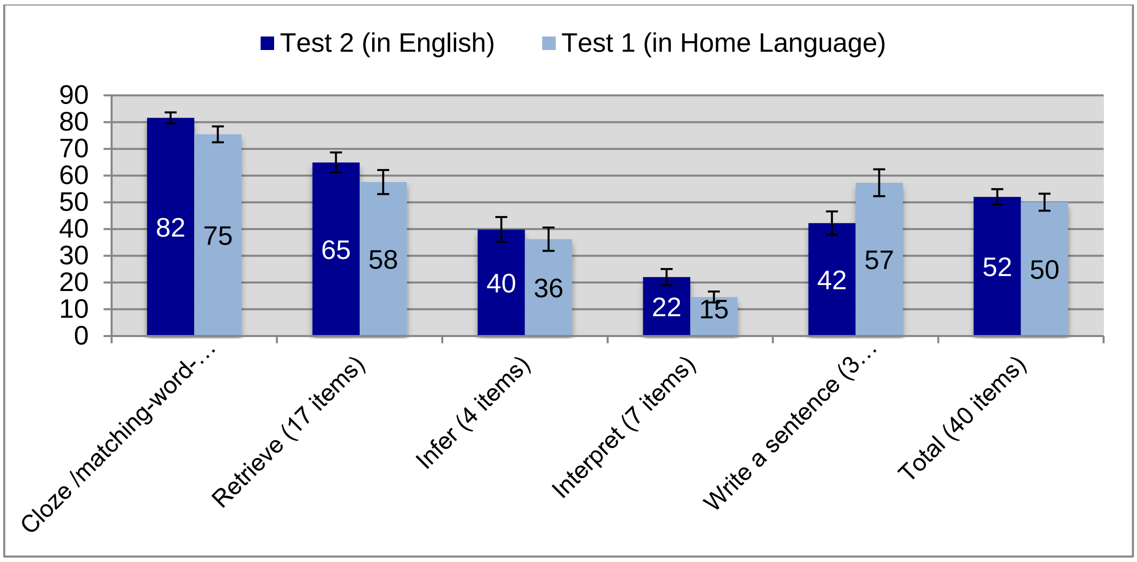


Figure 3: Average performance (%) in Test 1 (Systemic Evaluation) and Test 2 (National School Effectiveness Study) by literacy process for students whose home language is English (n = 132).



The most striking feature of the comparison between the two tests for non-English-home-language students is their performance on the three items that require students to write a sentence about a picture. On these three items,[10] students performed considerably better when they were able to write in the

[10] An example of one of these items is included in Appendix B.

LOLT of the school (average score 44%) than when they were forced to write in English (average score 9%). While this could reflect the fact these items were the most heavily influenced by the language of the test, it is also possible that the Test 2 markers marked these items more strictly than the Test 1 markers. Given that the 'write a sentence' items were out of four marks, there is more room for marker discretion than there is for the items in the other categories, which were mostly out of one mark. Given that the people marking the two tests were not the same people, it is possible that Test 2 markers marked more strictly than Test 1 markers.[11] This hypothesis is supported by the results of the 132 English-home-language students who performed worse in Test 2 only in the 'write a sentence' category. A priori, we would expect the English students to do the same, or better, on all items in Test 2 than in Test 1 given that they wrote the same test twice, both times in their home language. The fact that English students do worse in Test 2 on the 'write a sentence' questions is most likely because of differential marking practices on these items across the two tests. It is highly unlikely that their sentence-writing abilities have deteriorated substantially over the 1-month period.

One could look at English home-language students and use the difference between Test 1 and Test 2 on the 'write a sentence' items as a lower bound estimate of the cost of the harsher marking on the write-a-sentence items (i.e. negative 15.09% points). This is a lower-bound estimate because this is the effect of the positive learning bias, the positive test-familiarity bias, and the negative stringency bias from the harsher marking in Test 2. Using this estimate as a lower-bound estimate of the cost of harsher marking requires us to assume that Test 2 markers were equally strict when marking the scripts of English and non-English home-language students. If markers were not consistent across language groupings within an item category, it is not possible to benchmark across language groupings, as we do here. Comparing the differences across item categories and language groupings (English and non-English home language), it is clear that non-English-home-language students did considerably worse than English-home-language students and that this difference was largest for the 'write a sentence' items, even after accounting for harsher marking in Test 2.

Looking at the nine 'cloze/matching' items, students whose home-language is not English perform 9.8% worse (5.55% points) when they wrote the test in English as compared to writing the test in the LOLT of their school. Looking at the 17 'retrieve' items, these same students perform 20% worse (6.82% points) when they write the test in English as compared to writing the test in the LOLT of their school. There is a strong case to be made that both these estimates

[11] This was clarified through personal communication with Carla Perreira (2014), the Chief Operating Officer at JET Education Services (the technical adviser for the Systemic Evaluation, and the implementing agent for the NSES). The Systemic Evaluation markers were recruited and managed by the Department of Basic Education (DBE). Although JET provided training and supported the process of the Systemic Evaluation, the DBE was responsible for the marking and moderation processes. For the NSES, JET did the marking and moderation using the same marking memos and the same training procedures, albeit with different markers.

represent the causal impact of writing these kinds of items in English relative to the LOLT of the school, when a student's home language is not English.

Looking at the differences between the two tests for the four 'infer' items and the seven 'interpret' items, it is less clear that these differences represent the causal impacts of anything. It would seem that most students whose home language is not English found these items to be too difficult for them to provide meaningful information on the impact of language. When written in the LOLT of the school (Test 1), students scored an average of 16.03% on the 'infer' items and 8.51% on the 'interpret' items, dropping to 15.48% and 7%, respectively. Given that half of these questions were structured as multiple-choice questions with four choices (see Table 5) and that multiple-choice questions overestimate true ability because of random guessing, it is highly likely that the 'true score' here is essentially zero – that is if we corrected for guessing. In these instances it would seem that language is a second-order concern. If students already perform extremely poorly in their home language (as in the 'infer' and 'interpret' items) – perhaps because the cognitive demand was too high – then asking the same questions in English is unlikely to lead to a significant drop in average performance. On the other hand, if students are able to answer the questions in their home language but not in English, this suggests that the language content of the items is preventing them from understanding the questions rather than not having the ability, skill or understanding to answer the question (as in 'cloze/matching' and 'retrieve' items).

An alternative to grouping items by literacy process is to group items by item format, that is to say whether the item is a multiple-choice question or a free-response question. Table 9 reports the average literacy score by language groups and question format. From the table, one can see that both groups of students perform better[12] on the multiple-choice question items than on the free-response items, but that the difference is largest for students whose home language is not English. Given that there is no marking discretion for multiple-choice items, one can say that the causal impact of writing these 27 questions in English, when English is not a student's home language is –3.05% (–5.62 %points) without correcting for guessing.

For the 13 free-response items (which include the three 'write a sentence' items), the difference is much larger at –15.95% points for students whose home language is not English. Looking at the 132 English home language students, one can see that these students did slightly better in Test 2 than on Test 1 on free-response items and much better on the multiple-choice question items. These increases are, again, presumably a result of learning or test familiarity. The average impact of stricter marking on some free-response items in Test 2 (NSES) was clearly smaller than the learning/test-familiarity effect, yielding a net-positive result. Given that one cannot easily compare English home language

[12] As mentioned previously, one would expect students to do better on multiple-choice questions purely because of guessing.

and non-English home language students, it is not clear what proportion of the –15.95% point decline between Test 1 and Test 2 for students whose home language is not English is a result of stricter marking and what was because of writing in an unfamiliar language.

Table 9: Average literacy score (%) by question format between Test 1 (Systemic Evaluation) and Test 2 (National School Effectiveness Study) for students whose home language is not English and those for whom it is [standard errors clustered at the individual level].

	Non-English-home-language (*n* = 2811)			**English home-language (*n* = 132)**		
	Multiple choice questions (27 items)	**Free response items (13 items)**	**Total (40 items)**	**Multiple choice questions (27 items)**	**Free response items (13 items)**	**Total (40 items)**
Test 2 (in English)	37.46	7.75	22.07	70.48	34.90	52.06
Standard error	0.32	0.22	0.25	1.47	1.65	1.46
Test 1 (in Home Language)	43.08	23.70	33.04	62.99	37.98	50.04
Standard error	0.36	0.31	0.30	1.93	1.55	1.62
Difference (Test 2-Test 1)	−5.62	−15.95	−10.97	7.49	−3.08	2.02
Standard error	0.48	0.38	0.39	2.42	2.27	2.19

Numeracy results and language-content

In addition to comparing the literacy test results from Test 1 (written in the LOLT of the school) and Test 2 (written in English), one can also compare the numeracy test results between these two tests. One of the major advantages when looking at the numeracy test is that all items were either correct or incorrect (one mark questions) and therefore left little room for differential marking across the two tests, unlike the literacy test – as discussed above. Rather than compare numeracy processes across the two tests (see Taylor & Reddi 2013), the focus here is on the difference in performance on item groupings based on the language content of those items. The three groups are (1) high language items, (2) no language items and (3) ambiguous items (i.e. items that could not be classified as either 'high language' or 'no language' items).

Table 10 reports the numeracy results for Test 1 (in the LOLT of the school) and Test 2 (in English) for students whose home language is not English, and for those for whom it is. Looking first at students' whose home language is not English, it is interesting to note that the overall difference between Test 1 and Test 2 is not statistically significant – on average, students scored 33% on both tests. However, if one looks at the results disaggregated by language content, one can see that students did slightly better in Test 2 on the 'no-language' and 'ambiguous' items than they did in Test 1 and slightly worse in Test 2 on the 'high-language' items, as one might expect. On both tests, students found the 23 high-language items slightly easier than the 16 no-language items. From Table 10, one can see that students whose home language is not English scored 5.2% (1.87% points) worse when writing high-language content items in English compared to writing those same items in the LOLT of their school. This is arguably the causal impact of writing high-language content mathematics items in English when English is not a student's home language.

If students learned new skills or consolidated old skills in the month between the two tests, one would expect them to perform better on Test 2 than on Test 1. Similarly, if students became familiar with the test (remembered test items), we would also expect them to perform better in Test 2 than in Test 1. This is in fact what we see for the 'no-language' and 'ambiguous' items for students whose home language is not English. If we assume that these three effects are equal across the three item categories (something which may or may not be true), we can employ a second difference to difference out these biases. By comparing the difference between the two tests (first difference) and the difference between the item categories (second difference), one can estimate the causal impact of writing high-language content items relative to low-language content items for students whose home language is not English (accounting for all biases, assuming that these biases affect all three categories equally). Table 10 shows that this amounts to negative 8.1% (–2.91% points) for high-language content items. One can thus think of these two estimates (–5.2% and –8.1%) as a lower-bound and an upper-bound estimate of the causal impact of writing high-language content items in English relative to writing them in the LOLT of the school, when English is not a student's home language.

If one looks at students whose home language is English, one can see that they perform better in Test 2 (67%) than in Test 1 (60%) and that the gains are largest for the high-language content items. The difference-in-difference analysis shows that that the difference between high-language and no-language items was larger in Test 2 than in Test 1, that is to say that English students either (1) learned more content relating to the high-language items than the no-language items in the intervening month between the tests or (2) remembered the high-language items better than the no-language items between the two tests. While this is an interesting finding in and of itself, one could possibly use this information to inform the difference-in-difference analysis for students whose home language is not English. However, this would require that we assume that the same amount of learning takes place in schools that English-home-language students attend, and those that non-English-home-language students attend, something that is almost certainly untrue (Shepherd 2011; Spaull 2013; Taylor & Yu 2009). Furthermore, the sample of 132 English students is relatively small with concomitantly large standard errors.

Using a similar framework for the numeracy test as for the literacy test, one can identify what the difference in achievement that is attributable to (1) home background and (2) school quality (jointly); and (3) language is for non-English home language students. Table 10 shows that there is practically no difference between Test 1 and Test 2, suggesting that the language factor is only a very small part of the story in the underperformance of non-English students in mathematics. It would be prudent to ask whether the 'learning/familiarity gains' (between Test 1 and Test 2) and the 'language cost' (because of writing in English) are not simply cancelling each other out creating a net effect of zero. While this may be true, it is difficult to estimate the size of the 'learning/familiarity gain'. However, if we assume that non-English students learn as much in the intervening month as do their English peers, and remember as much of the test as their English peers (which is unlikely given that they have seen the test in two languages whereas the English students saw the exact same test twice), then we can use the gains seen in the English home language sample (7% points) as an upper bound estimate of the 'learning/familiarity gain' and thus as an upper-bound estimate of the 'language cost'. This amounts to 0.32 (0.07/0.22) of a standard deviation.

Using the standard deviation of 22.2% (from non-English students in the Systemic Evaluation Numeracy – Test 1), the difference between English home language students (average score 60%) and non-English home language students (average score 33%) amounts to 1.22 (0.27/0.222) standard deviations. This can be thought of as a composite estimate of the impact of (1) home background and (2) school quality.

Table 10: Average numeracy performance (%) by language-content in Test 1 (Systemic Evaluation) and test 2 (National School Effectiveness Study) for students whose home language is not English [standard errors clustered at the individual level]

	Non-English-home-language students (*n* = 2811)			
	High language items (23 items)	**Ambiguous items (14 items)**	**No language items (16 items)**	**Total (53 items)**
Test 2 (in English)	34.02	34.88	30.03	33.04
Standard error	0.41	0.49	0.48	0.41
Test 1 (in Home Language)	35.89	33.13	29.00	33.08
Standard error	0.39	0.49	0.47	0.41
Difference (Test 2 – Test 1)	−1.87	1.75	1.03	−0.04
Standard error	0.57	0.69	0.67	0.58
Difference-in-difference (relative to no language items)	−2.91	−2.78	-	−1.07
Standard error	0.88	0.97	-	0.89
	English-home-language students (n=132)			
	High language items (23 items)	**Ambiguous items (14 items)**	**No language items (16 items)**	**Total (53 items)**
Test 2 (in English)	69.47	68.99	62.26	67.17
Standard error	2.01	1.97	2.45	2.00
Test 1 (in Home Language)	59.22	64.94	56.82	60.01
Standard error	2.24	2.49	2.69	2.34
Difference (Test 2 – Test 1)	10.24	4.06	5.45	7.16
Standard error	3.01	3.18	3.64	3.08
Difference-in-difference (relative to no language items)	4.80	−1.39	-	1.72
Standard error	4.72	4.83	-	4.76

Summary of findings and robustness check

Table 11 presents the various 'effect sizes' discussed in this study. The composite effect of (1) home background and (2) school quality was calculated as the difference between the score of English students on the Systemic Evaluation and the score of non-English students on the Systemic Evaluation. Given that all students wrote the Systemic Evaluation in the LOLT of the school, we argue that this is the sum of all non-language factors (summarised as 'home background and school quality'). The effect of language was calculated as the difference between Test 2 (NSES written in English) and Test 1 (Systemic Evaluation written in LOLT of the school). The lower-bound estimate is the straight-forward difference between the two tests while the upper-bound estimate assumes that non-English students would have learnt as much in the intervening month as English students and would remember as much of the test as English students, and is thus calculated as the difference between Test 2 and Test 1 in addition to the learning/familiarity gain seen among the English students. All differences are expressed as a percentage of a standard deviation found among non-English students in the Systemic Evaluation test (15.6% for literacy and 22.2% for numeracy).

Table 11: Size of various 'effects' in standard deviations for students whose home language is not English.

	Literacy		**Numeracy**	
	Lower-bound	**Upper-bound**	**Lower-bound**	**Upper-bound**
(1) Home background	−1.08		−1.22	
(2) School quality				
(3) Language	−0.69	−0.82	0	−0.32
(1a) Home background	−1.13		−1.22	
(2a) School quality				
(3a) Language (excluding 3 write-a-sentence items)	−0.29	−0.71	0	−0.32

In addition to the effect sizes, Table 11 also reports the results of a sensitivity analysis for the literacy results. Given that NSES markers seemed to be more strict than the Systemic Evaluation markers on the 'write a sentence' questions (as discussed above), it was decided to re-do the analysis excluding the three 'write a sentence' items. These results are reported as (1a), (2a) and (3a) in Table 11**.** Given the data presented in Table 10 and the discussion about the different marking procedures employed, there is a strong case to be made that the results in (1a), (2a) and (3a) are more reliable than those in (1), (2) and (3).

Summary and conclusion

To summarise the main findings from this analysis:
Non-English grade 3 students performed between 0.29 and 0.71 standard deviations worse in **literacy** when writing the test in English compared to writing the test in the LOLT of the school. This impact can be regarded as causal.

Non-English grade 3 students performed 1.08 standard deviations worse in **literacy** than English grade 3 students when both groups of students wrote the test in the LOLT of their respective schools. This can be regarded as the size of the effect on literacy of home background and school quality factors combined.

Non-English grade 3 students performed between 0 and 0.32 standard deviations worse in **numeracy** when writing the test in English compared to writing the test in the LOLT of the school. This impact can be regarded as causal.

Non-English grade 3 students performed 1.22 standard deviations worse in **numeracy** than English grade 3 students when both groups of students wrote the test in the LOLT of their respective schools. This can be regarded as the size of the effect in numeracy of home background and school quality factors combined.

The analysis of the literacy tests showed that students whose home language is not English found it particularly difficult to write a sentence in English, even after accounting for the fact that the test markers for Test 2 (NSES in English) seemed to mark more strictly than those for Test 1 (Systemic Evaluation in the LOLT of the school). The results further showed that student performance on the 'infer' and interpret' items in Test 1 (written in the LOLT of the school) was so low to begin with that there was hardly any difference in performance when it was written in English in Test 2 (NSES).
The analysis of the literacy test confirmed findings in the international literature (Adetula 1990) that student's whose home language is different to the language of the test find free-response questions more difficult than multiple choice questions.

Analysis of the numeracy test for non-English students showed only slight differences in performance across the two tests, with a slightly larger 'cost' for high-language items relative to no-language items.

Where the present study differs from earlier research is that it focuses on the grade 3 level, which is before any LOLT switch to English. By taking this approach, we were able to isolate the impacts of language factors on the one hand and home background and school quality on the other. In essence, this study has extended the analysis of Taylor and Taylor (2013) in two important ways; firstly, by improving their matching algorithm (matching 61% more students), and secondly, by disaggregating students' numeracy and literacy performance by item category, language content and question format. This was done in an attempt to provide empirical estimates of the language cost associated with different literacy processes and question types for literacy; and for numeracy, the differences between high language and no language items.

Perhaps the most important finding emerging from the present analysis is that the size of the composite effect of home background and school quality is 1.6–3.9 times larger than the impact of language for literacy and at least 3.8 times larger for numeracy. To put this in terms of 'years worth of learning', if one uses 0.3 standard deviations as an approximation of 1 year of learning in South Africa (see Spaull & Kotze 2015), then the size of the 'language cost' is approximately 1 to 2 years worth of learning for literacy and a maximum of 1 year for numeracy. By contrast, the size of the composite effect of home background and school quality is roughly 4 years worth of learning for both numeracy (1.2 standard deviations) and literacy (1.15 standard deviations). This finding reiterates those expressed by other authors in the literature (Fleisch 2008; Hoadley 2012; Murray 2002); for example, Hoadley (2012) concludes that:

> Divided opinions over the language of instruction issue have masked the issue of poor literacy teaching per se, as is evident in the low home language literacy levels amongst learners…To a certain extent, in other words, debates around language deflect attention from the quality of instruction, irrespective of the language of instruction. (Hoadley 2012:192)

The intention of these authors is not to negate the importance of language, but rather to situate the language effect within the discussion of a generally dysfunctional schooling system. By doing so, these findings – including those presented in this article – aim to stress the importance of the quality of instruction, not only the language of learning. The fact that the literacy and numeracy achievement of South African children is so low prior to any language switch to English should give pause to those who argue that language is the most important factor in determining achievement, or lack thereof, in South Africa.

References

Adetula, L.O., 1990, 'Language factor: Does it affect children's performance on word problems?', *Educational Studies in Mathematics* 21(4), 351–365. http://dx.doi.org/10.1007/BF00304263.

Alexander, N., 2005, 'Language, class and power in post-apartheid South Africa', in *Harold Wolpe Memorial Trust Lecture,* Wolpe Trust, Cape Town.

Alidou, H., Boly, A., Brock-utne, B. & Satina, Y., 2006, *Optimizing learning and education in Africa – The language factor*, UNESCO pp. 1–186.

Angrist, J.D. & Pischke, J.-S., 2009, *Mostly harmless econometrics: An empiricist's companion. An empiricist's companion,* Princteon University Press, Princeton, NJ, p. 373.

Ball, D.L., Hill, H.C. & Bass, H., 2005, *Knowing mathematics for teaching*. American Educator, Fall 2005.

Bernardo, A.B.I., 1999, 'Overcoming obstacles to understanding and solving word problems in Mathematics', *Educational Psychology* 19(2), 149–163. http://dx.doi.org/10.1080/0144341990190203.

Brock-Utne, B., 2007, 'Language of instruction and student performance: New insights from research in Tanzania and South Africa', *International Review of Education* 53(5–6), 509–530. http://dx.doi.org/10.1007/s11159-007-9065-9.

Carnoy, M., Chisholm, L. & Chilisa, B., 2012, *The low achievement trap: Comparing schooling in Botswana and South Africa*, HSRC Press, Cape Town.

Cazabon, M.T., Nicoladis, E. & Lambert, W.E., 1998, *Becoming bilingual in the Amigos two-way immersion program*, Research Report 3.

Chisholm, L., 2013, 'Bantustan education history: The "progressivism" of Bophutatswana's Primary Education Upgrade Programme (PEUP), 1979–1988', *South African Historical Journal* 65(3), 403–420. http://dx.doi.org/10.1080/02582473.2013.787642.

Cummins, J., 1984, *Bilingualism and special education: Issues in assessment and pedagogy*, Multilingual Matters Limited, Clevedon, England. p. 306.

Cummins, J., 2000, *Language, power, and pedagogy: Bilingual children in the crossfire*, Multilingual Matters, Cleveland, England. p. 309.

DBE, 2011, *Curriculum and Assessment Policy Statement (CAPS): Foundation Pase Grades 1–3 First Additional Language*, Department of Basic Education. Pretoria, p. 94.

DoE, 2008, *Grade 3 systemic evaluation 2007 leaflet*, Department of Education Pretoria.

Edwards, J., 2012, 'Language and identity', in C. Chapelle (ed.), *The encyclopedia of applied linguistics,* pp411-420 Oxford, UK Wiley-Blackwell.

Fairclough, N., 1989, *Language and power*, Longman Group UK, London, p. 253.

Fiske, E.B. & Ladd, H.F., 2004, *Elusive equity: Education reform in post-apartheid South Africa*, Brookings Institution Press, Washington, DC.

Fleisch, B., 2008, *Primary education in Crisis: Why South African Schoolchildren underachieve in reading and mathematics*, Juta & Co, Cape Town, pp. 1–162.

Greaney, V. & Kellaghan, T., 2008, *Assessing national achievement levels in education (Vol. 1)*, World Bank, Washington DC Publications.

Heckman, J.J., 2000, 'Policies to foster human capital', *Research in Economics* 54, 3–56. http://dx.doi.org/10.1006/reec.1999.0225.

Heugh, K., 1993, Not so straight for English. Bua!, 8.2.

Heugh, K., 2005a, 'Mother tongue education is best', *HSRC Review* 3(3), 6–7.

Heugh, K., 2005b, 'The merits of mother tongue education', *SA Reconciliation Barometer* 3(33), 8–9.

Heugh, K., 2012, *The Case against Bilingual and Multilingual Education in South Africa* (No. 6), PRAESA Cape Town, pp. 1–42.

Hoadley, U., 2012, 'What do we know about teaching and learning in South African primary schools?', Education as Change 16(2), 37–41. http://dx.doi.org/10.1080/16823206.2012.745725

Howie, S., Venter, E., Van Staden, S., Zimmerman, S., Long, C. & Scherman, V., 2007, *PIRLS 2006 Summary Report: South African children's reading achievement*, Centre for Evaluation and Assessment Pretoria.

Ianco-Worrall, A., 1972, *Bilingualism and cognitive development*, Child Development. Vol 43(4) pp1390-1400

Karis, T. & Gerhart, G., 1997, *The 1976 Soweto Uprising. In from protest to challenge: A documentary history of African politics in South Africa, 1882–1990 Nadir and resurgence, 1964–1979*, UNISA Press, Pretoria, p. 569.

Kramsch, C., 1993, *Context and culture in language teaching*, Oxford University Press, Oxford, pp. 1–303.

LANGTAG, 1996, *Towards a national language plan for South Africa,* South African Department of Arts, Culture, Science and Technology, Pretoria.

Macdonald, C., 1990, *Crossing the threshold into standard three. Main report of the Threshold Project*, Human Sciences Research Council, Pretoria.

Macdonald, C. & Burroughs, E., 1991, *Eager to talk and learn and think: Bilingual primary education in South Africa*, Maskew Miller Longman, Cape Town, p. 87.

Malherbe, E., 1946, *The Bilingual School,* Longman, London.

Mesthrie, R., 2002, 'South Africa: A sociolinguistic overview', in R. Mesthrie (ed.), *Language in South Africa,* Cambridge University Press, Cambridge.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., Trong, K.L. & Sainsbury, M., 2011, *PIRLS 2011 assessment framework,* TIMESS & PIRLS International Study Center and IE, Boston, MA.

Murray, S., 2002, 'Language issues in South African education: An overview', in R. Mesthrie (ed.), *Language in South Africa,* Cambridge University Press, Cambridge.

Ndlovu, S., 2004, 'The Sowetho uprising', in *The road to democracy in South Africa, vol 2, 1970–1980*, UNISA Press.

NEEDU, 2013, *NEEDU National Report 2012,* Pretoria, p. 89.

Ní Ríordáin, M. & O'Donoghue, J., 2008, 'The relationship between performance on mathematical word problems and language proficiency for students learning through the medium of Irish', *Educational Studies in Mathematics* 71(1), 43–64. http://dx.doi.org/10.1007/s10649-008-9158-9

NPC, 2012, *National development plan 2030: Our future – Make it work*, Pretoria, p. 484.

Postlethwaite, T.N. & Kellaghan, T., 2008, *National assessments of educational achievement*, UNESCO, Paris, p. 31.

Pretorius, E. & Spaull, N., 2016, 'Exploring relationships between oral reading fluency and reading comprehension amongst English second language readers in South Africa', *Reading and Writing*. http://dx.doi.org/10.1007/s11145-016-9645-9.

Prinsloo, C. & Reddy, V., 2012, *Educator leave in the South African public schooling system,* HSRC Pretoria, p. 3.

Probyn, M., 2001, 'Teachers voices: Teachers reflections on learning and teaching through the medium of English as an additional language in South Africa', *International Journal of Bilingual Education and Bilingualism* 4(4), 249–266. http://dx.doi.org/10.1080/13670050108667731.

Reddy, V., 2006, *Mathematics and science achievement in South African Schools in TIMSS 2003,* Human Sciences Research Council, Cape Town.

Reeves, C., Carnoy, M. & Addy, N., 2013, 'Comparing opportunity to learn and student achievement gains in southern African primary schools: A new approach', *International Journal of Educational Development*. http://dx.doi.org/10.1016/j.ijedudev.2012.12.006

Setati, M. & Adler, J., 2000, 'Between languages and discourses: Language practices in primary multilingual mathematics classrooms in South Africa', *Educational Studies in Mathematics* 43(3), 243–269. http://dx.doi.org/10.1023/A:1011996002062.

Setati, M., Adler, J., Reed, Y. & Bapoo, A., 2002, 'Incomplete journeys: Code-switching and other language practices in Mathematics, Science and

English Language classrooms in South Africa', *Language and Education* 16(2), 128–149. http://dx.doi.org/10.1080/09500780208666824.

Shepherd, D.L., 2011, *Constraints to school effectiveness: What prevents poor schools from delivering results?* (No. 05/11), Stellenbosch, pp. 1–37.

Shonkoff, J.P., Richter, L., van der Gaag, J. & Bhutta, Z.A., 2012, 'An integrated scientific framework for child survival and early childhood development', *Pediatrics* 129(2), e460–e472. http://dx.doi.org/10.1542/peds.2011-0366.

Skutnabb-Kangas, T., 1988, *Minority education: From shame to struggle*, Multilingual Matters, p. 410.

Skutnabb-Kangas, T., 2000, *Linguistic genocide in education – Or Worldwide Diversity and Human Rights?*, Routledge, p. 824.

Spaull, N., 2013, 'Poverty & privilege: Primary school inequality in South Africa', *International Journal of Educational Development* 33, 436–447. http://dx.doi.org/10.1016/j.ijedudev.2012.09.009.

Spaull, N. & Kotze, J., 2015, 'Starting behind and staying behind in South Africa: The case of insurmountable learning deficits in mathematics', *International Journal of Educational Development,* 41, 12–24. http://dx.doi.org/10.1016/j.ijedudev.2015.01.002.

StatsSA, 2012, *Census 2011 in brief,* Statistics South Africa, Pretoria.

Stubbe, T.C., 2011, 'How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments', *Educational Research and Evaluation* 17(6), 465–481. http://dx.doi.org/10.1080/13803611.2011.630560.

Taylor, N., 1989, *Falling at the First Hurdle: Initial Encounters With the Formal System of African Education in South Africa*, Research Report No 1, Johannesburg: Education Policy Unit, University of Witwatersrand, p.38.

Taylor, N., 2011, *Priorities for addressing South Africa's Education and Training Crisis: A review commissioned by the National Planning Commission*, Johannesburg, p. 68.

Taylor, N. & Reddi, B., 2013, 'Writing and learning mathematics', in *Creating effective schools*, p. 281, Pearson, Cape Town.

Taylor, N. & Taylor, S., 2013a, 'Teacher knowledge and professional habitus', in N. Taylor, S. Van der berg & T. Mabogoane (eds.), *Creating effective schools,* Pearson, Cape Town.

Taylor, N., Van der berg, S. & Mabogoane, T., 2013, *What makes schools effective? Report of the National School effectiveness study,* Pearson, Cape Town.

Taylor, S. & Taylor, N., 2013b, 'Learner performance in the NSES', in *Creating effective schools,* Pearson, Cape Town.

Taylor, S. & Von Fintel, M., 2016, 'Estimating the impact of language instruction in South African primary schools: A fixed effects approach', *Economics of Education Review* 50, 75–89. http://dx.doi.org/10.1016/j.econedurev.2016.01.003.

Taylor, S. & Yu, D., 2009, *The importance of socioeconomic status in determining educational achievement in South Africa* (No. 1), Stellenbosch.

Timæus, I.M., Simelane, S. & Letsoalo, T., 2013, 'Poverty, race, and children's progress at school in South Africa', *Journal of Development Studies* 49(2), 270–284. http://dx.doi.org/10.1080/00220388.2012.693168.

Venkat, H. & Spaull, N., 2015, 'What do we know about primary teachers' mathematical content knowledge in South Africa? An analysis of SACMEQ 2007', *International Journal of Educational Development* 41, 121–130. http://dx.doi.org/10.1016/j.ijedudev.2015.02.002.

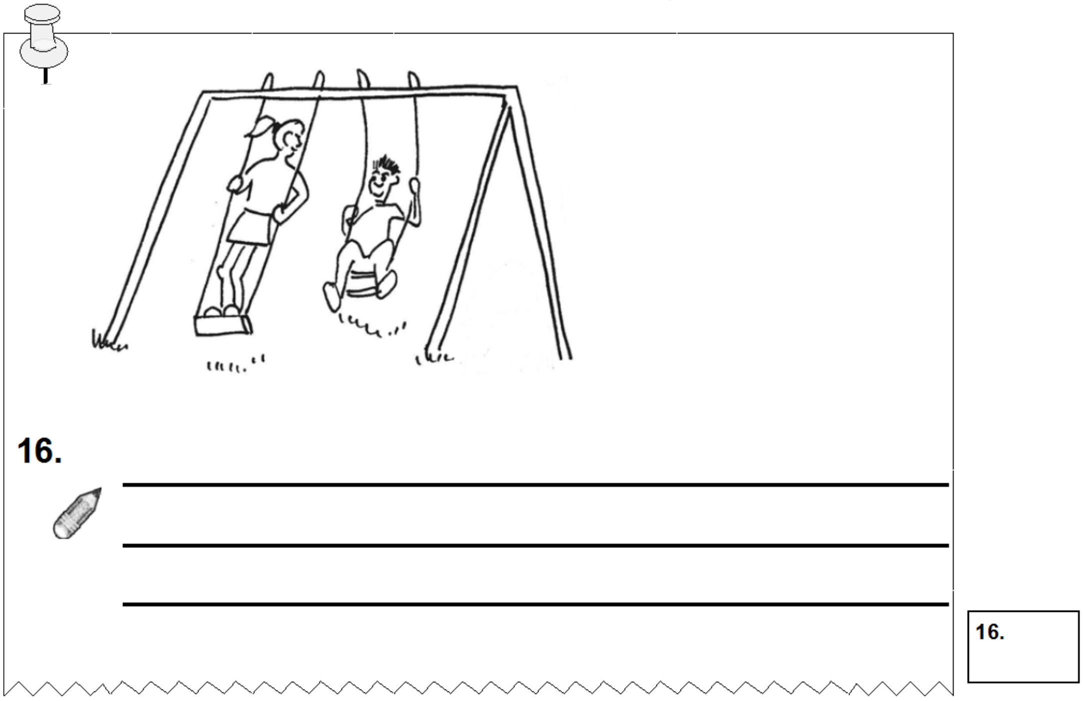
Weber, E., 1976, *Peasants into Frenchmen: The modernization of rural France, 1870–1914*, Stanford University Press, Palo Alto, pp. 1–632.

Welch, T., 2011, *Review of Foundation Phase Literacy Resource Packages for evaluation of Gauteng Primary Literacy Strategy*, Johannesburg, p. 218.

Appendix B

Examples of items:

'Write a sentence' item: The caption for the question read: 'Look at the pictures in questions 16 – 18. For each picture write a complete sentence about what the child or children are doing in the picture'.



16.

16.