RESEP
Research on Socio-Economic Policy

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

forward together
sonke siya phambili
saam vorentoe

RESEP WORKING PAPER

Department of Economics, Stellenbosch University

Working Paper No.
02/25

DECEMBER 2024

This paper was produced as part of the MILAPS project, funded by Optima and utilises Data Driven Districts data

Linguistic interdependence?

Foundation Phase mastery in home language as a predictor of grade 4 repetition and EFAL marks

Author

Ros Clayton

OPTIMA

Linguistic interdependence? Foundation Phase mastery in home language as a predictor of grade 4 repetition and EFAL marks

Ros Clayton

December 2024

Abstract

For most South African learners, the destination language of instruction (typically English) is not their mother tongue; thus, proficiency in English is necessary for educational success. However, research has shown that development of linguistic skills is more effective in the mother tongue, and that these skills can be transferred to a second language (especially when high levels of proficiency are reached in the first language). Much of this research has been in the Global North, where the destination language is spoken by most of the population. However, this is not the case in South Africa, where English is spoken by fewer than 10% of the population; therefore, it is important to determine the nature of these relationships in the South African context. This study makes use of a large longitudinal dataset containing school-based assessment data for the Eastern Cape, Gauteng, KZN, Limpopo, Mpumalanga and North West to estimate the extent to which Grade 3 Home Language (HL3) mastery predicts Grade 4 repetition and Grade 4 English First Additional Language (EFAL) results. The results show that higher HL3 results are associated with lower repetition in Grade 4, and there is a pro-female bias in terms of lower predicted repetition which is largest in Quintile 1 schools, even after controlling for HL3 and other factors. Each one unit increase in HL3 was associated with a 0.4 to 0.5 unit increase in Grade 4 EFAL results, with females being advantaged by approximately 4-percentage points (controlling for other factors). This is suggestive of a growing gender gap between grades 3 and 4, especially in the poorest schools.

JEL Classification: I21, I28, C25

Keywords: Linguistic interdependence, South Africa, repetition, longitudinal analysis, primary school

Introduction

Many learners around the world, including in South Africa, cannot complete their education in their home language. A crucial component of these learners' education is therefore how well they master the "destination" language of instruction (LOI). Over the last four decades research has shown that development of linguistic skills in one's mother tongue is crucial to mastery of a second language. In South Africa, most learners do not have English as their mother tongue, yet this is the destination LOI and must be mastered for success at school. Cummins (1979) proposed the linguistic interdependence hypothesis, which states that linguistic competency in one language may be transferred to another, and greater levels of competency lead to greater levels of transfer. This theory emphasises the importance of mother tongue instruction, particularly at the start of the schooling journey.

Many studies have shown that higher levels of mother tongue (referred to as L1) literacy are associated with increased literacy in the second language (hereinafter referred to as L2) (see Melby-Lervåg and Lervåg (2011) for a review of the correlational evidence). However, many of these studies were conducted in the Global North, and often in contexts where the L2 is the predominant language in the country. In South Africa, however, English is the mother tongue of less than 10% of the population (Statististics South Africa, 2024); thus the applicability of these studies is unclear. More recently, some studies have demonstrated the validity of the linguistic interdependence hypothesis in the African context (de Galbert (2023), Kim and Piper (2019), and Humble (2024)), where the L2 is not the language of the majority. In South Africa, Taylor and von Fintel (2016) showed that learning in the L1 in the early grades caused higher results in L2 (English) in later primary grades (compared to using English (L2) as the LOI in the early grades). Mohohlwane et al. (2023) demonstrated the causal relationship of linguistic transfer from L1 to L2. The authors also showed that an intervention in the L2 (English) was detrimental to L1 literacy.

South African education policy states that learners should be taught in their mother tongue at least in the Foundation Phase (Grades 1-3). In practice this is largely the case, with most learners switching to either English (predominantly) or Afrikaans as the LOI in Grade 4. For linguistic transfer to take place in Grade 4 (when the LOI changes to English), learners would be expected to have sufficiently mastered their own home language. This study therefore seeks to contribute to the body of knowledge on linguistic transfer in South Africa by addressing two questions. Firstly, to what extent does home language mastery in the Foundation Phase predict the probability of repeating Grade 4? And to what extent does it predict EFAL performance in Grade 4?

The next section provides the background to the study, including a review of the existing literature on linguistic interdependence, as well as more detail on the context in South Africa. Section 3 provides a description of the dataset used for this study, while Section 4 provides details of the approaches used to address the research questions. Section 5 provides descriptive results,

while the estimation results are presented in Sections 6 and 7. Finally, a conclusion is offered in Section 8.

Background

Linguistic interdependence and the importance of mother tongue instruction

For many children in South Africa, and around the world, their mother tongue is not the predominant language of education. The destination language (English or Afrikaans, in South Africa) must be mastered for educational success, and there are many possible approaches for introducing the destination language in schooling. An influential theory in this field is the linguistic interdependence hypothesis (Cummins, 1979), which postulates that literacy skills involve an underlying proficiency that is independent of the specific language being studied. Therefore, linguistic skills in one language (such as phonological awareness and metalinguistic understanding about the function of text) can transfer to a second language. Greater proficiency in the first language is hypothesised to result in greater transfer of skills to the second language; as such, this view emphasises the importance of mother tongue instruction (as proficiency will be more easily achieved in a language known to the learner), especially in the early grades. When the learner's mother tongue is not the destination LOI, bilingual education is encouraged, and evidence shows that extended schooling in the mother tongue is not associated with later deficiencies in the destination language (Ball, 2011). Therefore, even when it is the case, as it is for many children across the world, that the destination LOI is not their mother tongue, it may still be beneficial for them to receive instruction in their L1, and to transition to the L2 as the LOI at a later stage.

The timing of the transition to L2 as the LOI is important. The most extreme approach is submersion, whereby learners are taught in L2 from the start of their schooling career. This approach has been shown to result in "subtractive bilingualism", where L2 is developed at the expense of L1 (Ball, 2011). Alternatively, learners may study in their mother tongue while simultaneously learning L2; this bilingual model has in fact been shown to have many advantages, provided L1 is prioritised and is the language of instruction (see reviews in Baker, 2001; Cummins, 2000; Dutcher, 1995). A transition after 1 to 3 years of L1 instruction is considered early exit, while later exit would be after 6 to 8 years, at the completion of primary education. Short cuts in bilingual education, such as early exit, have been shown to be detrimental to literacy acquisition due to inadequate development of literacy in L1, which impedes the transfer of linguistic competencies to L2 (Benson, 2002).

There is a large body of research supporting the linguistic interdependence hypothesis, and it is now taken as given that that early mother tongue education is imperative not only to ensure linguistic diversity and the preservation of culture, but also because it is optimal for acquisition of linguistic skills in both L1 and L2 (see Ball (2011) for a comprehensive discussion). Melby-Lervåg and Lervåg (2011) provide a meta-analysis of correlational studies investigating linguistic

transfer between L1 and L2, among 6- to 10-year-old learners. They found strong correlations in terms of phonology, and smaller correlations in the more complex domain of oral language, with large variations in the magnitude of the correlation across both domains. Larger correlations in terms of decoding were observed when learners received explicit instruction in L2 as a subject, indicating that including some instruction in L2 can be beneficial if it is done alongside L1. Usborne et al. (2009) used a longitudinal design to evaluate the relationship between L1 and L2 (English) amongst a small sample of Canadian Aboriginal people whose LOI was their mother tongue up to Grade 3 and who then transitioned to English in Grade 4. The authors used baseline (Grade 3) assessments in L1 to predict later English outcomes using a hierarchical linear model approach. They found that every 1 unit increase in the baseline L1 score was associated with a 0.45 unit increase in subsequent English scores (controlling for other factors). More recently, some randomised control trials have demonstrated the causal nature of linguistic interdependence, for example by showing that learners who received a comprehensive reading intervention in L1 (Spanish) also improved in L2 (English) (Vaughn et al., 2006).

It is notable that much of this research has been conducted in the Global North, where the language to which the learners must transition is often spoken by most of the population. However, in many African contexts the destination language is often an "international" language that is spoken by a minority of the population. This is certainly the case with English in South Africa and raises questions about the applicability of the aforementioned results to this context. Certainly Cummins (1998) argued that linguistic transfer is more likely to occur from a minority to a majority language; out-of-school language exposure matters.

Several recent studies provide support for linguistic interdependence in contexts where the destination language is not spoken by the majority of the population. Humble et al. (2024) provide correlational evidence of cross-linguistic transfer between L1 (Hausa) and L2 (English) in Nigeria for Grade 3 learners. In Uganda, de Galbert (2023) assessed literacy skills for a cross sectional sample of learners in their mother tongue (L1) and English (L2). The author found significant correlations between linguistic skills in L1 and L2 for learners whose LOI was their mother tongue; but no significant correlations when the LOI was English, thus providing support for the notion that L1 mastery is a prerequisite for linguistic transfer. In Kenya, Wawire and Kim (2018) used an RCT to provide causal evidence that a Grade 1 learner-focussed intervention in L1 caused improvements in both L1 and L2 (English), and vice versa. They also found that whether the learners' L1 subject was their mother tongue (or not) did not moderate the impact of the intervention (p. 456).

The South African context

South African language policy strongly encourages use of a learner's mother tongue (L1) as the instruction language and Home Language subject in the Foundation Phase (Grades 1 to 3), and this is de facto the case for a majority of learners in the country (van der Berg et al., 2020). While it is not required, most learners transition to the destination languages of either English (the

majority) or Afrikaans in Grade 4 as the LOI. From Grade 1 learners take an additional language subject, typically English. The transition to L2 (usually English) as the LOI by Grade 4 is considered an "early exit" from L1 as the instruction language and has been shown to be inferior to later exit, which allows learners to become "highly proficient" in L2 before switching to this as the LOI. This early-exit bilingual model that most South African learners follow has been shown to have many advantages (see reviews in Baker, 2001; Cummins, 2000; Dutcher, 1995), especially in comparison to "submersion", whereby learners are taught in L2 from the outset.

Repetition in South Africa

The first research question concerns repetition; thus a brief overview of the repetition literature in South Africa is presented here. van der Berg et al. (2019) used administrative and survey data to estimate and characterise repetition in South Africa. The authors estimated Grade 4 repetition to be, on average, 8.1% between 2014 and 2018 (using the GHS), and 12% between 2015 and 2016 (using Annual Schools Survey data). The authors also used EMIS data to estimate the proportion of overage learners in each grade, estimating 30-35% of learners to be overage by Grade 4 (for seven provinces, excluding FS and MP). The authors also found that females are less likely to repeat than males; this difference is present by Grade 1 (4 percentage point difference) and grows by Grade 4 (7 percentage point difference). More recently, the Department of Basic Education (2023) used data from the learner unit record system to estimate repetition rates between 2018 and 2020, estimating Grade 4 repetition to be 11% in 2018-2019, and dropping to 8% in 2020 (due to increased leniency during the Covid19 period). Wills (2023) found that early grade repetition was strongly tied to mastery of literacy skills (at least in the pre-Covid period).

Previously, Lam et al. (2011) showed that grade advancement for White and Coloured learners was more strongly determined by baseline characteristics (including literacy/numeracy skills and previous grade repetition) than for African learners, whose grade advancement through high school was found to have a larger stochastic component. However, an external evaluation (the Grade 12 matric examination) showed no difference in the predictive value of baseline characteristics by race group, leading the authors to attribute the difference in high school grade advancement predictability to poor measurement of ability of African learners in schools. This can be interpreted as the poor quality of school-based assessment (SBA) for these learners, who were concentrated in "African" secondary schools. While this study was done in the Western Cape, one of the two provinces not covered in the DDD dataset, it still raises important questions about the validity of SBA data in these "African" schools, which were typically Quintile 1-3 schools.

Promotion requirements in South Africa are laid out in a national policy statement (Department of Basic Education, 2011), which is updated via circulars. Typical Grade 4 pass requirements involve achieving a minimum threshold of 50% in the Home Language subject, 40% in a First Additional Language subject, 40% in Mathematics, and 40% in two other approved subjects. However, the Grade 4 promotion requirements for the time period under consideration for this

study were affected by Covid19 adjustments, which allowed learners to be promoted if they met all promotion requirements aside from the Mathematics requirement; additionally, teachers were formally allowed to make a mark adjustment of up to 5% in a maximum of three subjects if that would allow a learner to be promoted (Hoadley, 2023).

In addition to promotion requirements, repetition status may be affected by a learner's prior repetition in a schooling phase. Since 1998, the Department of Education has required that learners may repeat at most once per phase, and barred multiple repeats of one grade (Department of Basic Education, 2011). This policy would not impact Grade 4 repetition in this study, since Grade 4 is the first grade in the Intermediate Phase, but it would impact repetition in the Foundation Phase. Foundation Phase home language mastery would be expected to influence Grade 4 repetition directly through its impact on Grade 4 Home Language performance, and possibly through its impact on EFAL performance (if linguistic interdependence holds in this context).

Evidence for linguistic interdependence in South Africa

There are a small number of studies on linguistic interdependence in South Africa. Taylor and von Fintel (2016) provide convincing evidence that mother tongue instruction (as opposed to English instruction) in the Foundation Phase causes higher proficiency in English in Grades 4, 5, and 6. The authors leveraged data form a natural experiment that followed a change in language policy wherein schools transitioned from English to African languages as the LOI in the Foundation Phase. Different rates of transition allowed the authors to estimate the impact of having the learners' L1 as the LOI on later English results, compared to English (L2) as the LOI. Furthermore, the results tentatively suggest that receiving instruction in L1 is more beneficial than learning in another African language that is similar to one's mother tongue, although the impact is much smaller than the impact of learning in English.

Mohohlwane et al. (2023) evaluated the impact of a teacher professional development (TPD) program in either L1 or L2 (English), on learner outcomes in L1 and L2. The study was an RCT, although the two interventions are not perfectly comparable to each other due to differences in the learners' mother tongue languages. The authors found that a TPD in L1 caused improvements in both L1 and L2, whereas a TPD in L2 caused improvement in L2 but diminished performance in L1, especially for the bottom half of the distribution. This provides evidence that improved performance in the home language subject can be the cause of improvements in L2 (English, in this case) in the South African context.

This study seeks to contribute to this body of evidence by examining the extent to which Foundation Phase home language mastery impacts Grade 4 outcomes, specifically repetition and EFAL results. Learner SBA data for six provinces for the period 2017 to 2023 is used to create a longitudinal dataset containing individual learner outcomes in the Foundation Phase and Grade 4. The results show that higher home language results in Grade 3 are associated with

lower repetition rates, and higher EFAL results, in Grade 4. Both outcomes demonstrate a significant gender gap in favour of females.

Data

The data for this study are derived from SA-SAMS SBA data for learners in six provinces in South Africa (Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, and North West) between 2017 and 2023 (the Data Driven Districts, or DDD, dataset). The data is comprehensive, with the raw dataset including datapoints on a yearly average of 81% (Gauteng) to 98% (Eastern Cape, Limpopo, and Mpumalanga) of learners across provinces (see Table A 1). The data includes individual learner term 4 marks for Home Language, First Additional Language, and Mathematics, and covers Grades 0 to 7. Additionally, the dataset contains demographic data, including gender, age and the mother tongue language of the learner. The data contain unique identifiers, which were used to track the same learners over time to create a longitudinal dataset. The unique identifiers allow learners to be tracked even when they change schools, but while they also allow learners to be tracked when changing provinces, these learners are excluded from this sample due to the provincial analysis.

The population under consideration is Foundation Phase learners from Public Ordinary Schools whose mother tongue is not English or Afrikaans, and whose implemented home language subject is an African language, as these are the learners who face a language transition in Grade 4. van der Berg et al. (2020) showed that English is taken as a home language subject in the Foundation Phase at a significantly higher rate than the population of English speakers; this phenomenon was observed in this dataset, with approximately 14% of African mother tongue speakers in the six provinces under consideration taking English as a home language subject in Grade 3 in 2019 (with significant variation in this figure across provinces – see Table A 2). These learners were excluded from the sample as they do not face the language transition in the same way in Grade 4. Thus, both the potential sample and the final longitudinal sample exclude learners with English or Afrikaans as their Home Language subject. This significantly reduces the number of Quintile 4 and 5 schools in the sample, as these schools predominantly offer English and Afrikaans as the Home Language subject.

The number of observations excluded due to data errors (such as a single learner being recorded in multiple grades or schools in one year, or a learner regressing in grades across years) was small, affecting about 1 to 2% of observations in total. However, of greater significance is the exclusion of learners who were not observed in multiple years (to obtain a longitudinal dataset without any missing values on key variables). This resulted in the longitudinal sample containing between 73% (in Gauteng) and 88% (Limpopo) of the learners from the unbalanced DDD dataset[1] (see Table 1). Along with the completeness of the DDD dataset in relation to the actual

[1] This includes all learners in the complete DDD dataset whose Home Language is not English or Afrikaans, but these learners do not have records in all years.

learner counts (Table A 1), these figures can be used to approximate[2] the coverage of the longitudinal sample in relation to the actual learner population in the six provinces. This approach indicates that the longitudinal dataset contains approximately 64% of the target population of learners in Gauteng, up to approximately 86% in Limpopo.

Table 1. Completeness of the longitudinal dataset

	EC	GT	KZN	LP	MP	NW	Total
Number of learners in the DDD longitudinal dataset (African Home Language only)							
2017	107 488	55 380	93 669	111 802	62 714	51 192	482 245
2018	91 159	58 311	92 536	108 633	59 037	43 580	453 256
2019	80 891	61 109	107 866	107 128	57 601	43 265	457 860
Mean	**93 179**	**58 267**	**98 024**	**109 188**	**59 784**	**46 012**	**464 454**
Above counts as % of African Home Language learners in the DDD dataset							
2017	81	72	79	87	84	83	81
2018	83	72	71	89	86	83	81
2019	79	74	72	89	87	81	80
Mean	**81**	**73**	**74**	**88**	**86**	**83**	**81**

Source: DDD data (learners in public ordinary schools with an African Home Language subject).

The longitudinal sample includes learners from over 99% of schools in the complete dataset; that is, exclusion was not wholly determined by the school attended by the learner. However, learners in the longitudinal sample perform slightly better in Mathematics than learners in the complete dataset (see Figure A 1); excluded learners are weaker, on average, than learners who had sufficient data to be included in the longitudinal sample. Thus, the longitudinal sample is not representative of the population and the results may be biased towards stronger learners.

The longitudinal dataset includes learners who started Grade 1 in 2017, 2018, or 2019, and who were observed in Grade 4 in the dataset in subsequent years; these three groups are referred to as *cohorts* throughout the paper. Table 2 presents the number and proportion of learners from each cohort reaching Grade 4 in the given year. 71% of the 2017 cohort reached Grade 4 by 2020 (without repetition); 25% repeated once in the Foundation Phase, 3% repeated twice, and fewer than 1% repeated three times. The final year in the dataset is 2023; thus, learners from the 2019 cohort could have repeated at most once. For the 2017 cohort, all learners were assumed to be attempting Grade 1 for the first time, even though in fact this cohort would include learners who

[2] This can only be an approximation, since the School Realities Reports do not contain breakdowns by language groups. I therefore estimate population coverage as the product of the percent of learners retained in the longitudinal dataset (from Table 1), and the percent of coverage for all learners (from Table A 1). For example, approximate coverage in Gauteng is 73% of 88% = 64% of the population.

had already repeated Grade 1 (but it was not possible to determine which learners, since the dataset only started in 2017).

Table 2: Year reached Grade 4

	2020	2021	2022	2023	Total
Number of learners from each cohort (row) reaching Grade 4 in given year (column)					
2017	344 825	121 490	15 721	209	482 245
2018	0	329 883	109 769	13 604	453 256
2019	0	0	354 495	103 365	457 860
Above as a percentage of the cohort					
2017	72	25	3	0	100
2018	0	73	24	3	100
2019	0	0	77	23	100

Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017-2019 and who reached Grade 4 by 2023).

Repetition status was determined by either observing the learner in Grade 4 again (for repeating learners) or in Grade 5 (for passing learners) in the next year. For learners who were only observed in Grade 4, including the more than 115 000 learners who were in Grade 4 in 2023, repetition status was inferred by applying the CAPS promotion requirements (Department of Basic Education, 2011) to their observed results (incorporating the Covid19 adjustments for 2020-2022, as described in Hoadley (2023)). The actual Grade 4 repetition outcomes for the 2023 learners may differ from these inferred outcomes, as schools often choose to condone learners that do not meet the official pass requirements, and these decisions may also differ by province. However, excluding these learners would result in a sample that is (more) biased towards stronger learners[3], so the decision was made to retain these learners despite potential inaccuracies in inferring their actual repetition status.

Grade 3 absenteeism is used as a control variable in the analysis. However, due to Covid19 disruptions, it was only possible to obtain meaningful absenteeism data for those learners who were in Grade 3 in 2019: the learners from the 2017 cohort who did not repeat. Furthermore, learners without recorded absenteeism data were treated as missing data (rather than being absent for zero days), which further reduces the absenteeism sample. The proportion of learners who were recorded absent on at least one day varied significantly across provinces, from 63% of learners in Limpopo, to 94% of learners in Gauteng, with all other provinces below 78% (see Table A 3). This suggests significantly different patterns in absentee recording (or absenteeism outcomes) in Gauteng. Due to the reduced sample of learners with absentee data, a separate

[3] As learners who were in Grade 4 in 2023 are those who would have repeated at least once.

model will be run for the full longitudinal sample and for the absenteeism sample, and this will also be run by province to account for recording differences.

Estimation Approach and Strategy

Estimation (1): Foundation Phase Home Language mastery and Grade 4 repetition

Given the longitudinal dataset described above, Grade 3 Home Language subject results (hereinafter referred to as HL3) are the best metric for home language mastery in the Foundation Phase. The functional form of the relationship between HL3 and Grade 4 repetition was not assumed a priori but derived from the observed relationship between the two variables in the data. Gender and being overage are both important determinants of repetition in South Africa, and these are used as control variables. Additionally, Taylor and von Fintel (2016) found evidence that learning in an African language that differs from a learner's own (African) home language can be slightly detrimental to learning (specifically to later English performance). An indicator variable will therefore be used to control for whether Home Language subject differs from the learner's mother tongue. However, the impact of learning in a different home language subject may already be captured in HL3 results; this would attenuate the estimated impact of learning in a different home language in this model. There are several other variables in the dataset which were considered for inclusion in the model. These include Grade 1 Mathematics results (to control for general academic performance in a manner that is not too highly correlated with HL3); the difference in Mathematics results between Grade 3 and Grade 1 (the *mathematics delta*; to control for the learning trajectory); and Grade 1 repetition status.

In addition to these observed controls, there are many unobserved variables that are relevant to repetition. These unobserved characteristics occur at both the learner and school level. At the learner level, these unobserved variables include socio-economic status, family environment, learner motivation, and many others. It is not possible to adequately control for all these variables; however, it is likely that much of the impact of these factors is already captured in HL3. Thus, it is likely that the estimated impact of HL3 on repetition may be inflated, since higher HL3 may be positively associated with other unobserved factors that cause both the HL3 results and the Grade 4 repetition outcome.

At the school level, the quality of assessments, teaching, school management and other school-specific factors are not observed. It is expected that schools with better quality teachers and assessment might exhibit a more reliable relationship between Grade 3 results and Grade 4 outcomes (as suggested by Lam et al. (2011)). Higher quality schools would be expected to have higher SBA results, and lower repetition rates (although this would not always be the case, for example if some high-quality schools have higher assessment standards than schools with an equivalent learner base, they might have lower SBA results and possibly slightly higher repetition

rates). To avoid the biases that would occur due to these omitted school variables, the model is estimated using school fixed effects to control for time-invariant (and unobserved) school-level factors. In addition to school-level factors, it is possible that the relationships between the variables of interest differ systematically across the six provinces in this study. Therefore, the estimation results are also presented separately for each province, and by school quintile.

The repetition outcome is binary (1 if a learner repeated grade 4, and 0 if they passed grade 4), which lends itself to the use of a maximum likelihood estimator (such as the logit). While implementing a fixed effects approach with a maximum likelihood estimator is possible by using a dummy variable for each unit, this becomes computationally unfeasible when there are a large number of units (Greene, 2004), which is the case in this dataset. While "brute force" approaches to this method do exist (see Greene (2004)), they will not be attempted here.

I therefore estimate the impact of Grade 3 Home Language results, on the likelihood of repeating Grade 4 using OLS to estimate a linear probability model of the form:

$$Y_{is} = \rho_0 + \rho_1 H_{is} + \rho_2 H_{is}^2 + X_i\delta + M_i\alpha + \Upsilon_s + \varepsilon_{is} \quad [1]$$

where Y_{is} is the probability that learner i at school s repeats Grade 4. The key variable of interest is H_{is}, the Grade 3 Home Language mark for that learner (out of 100). The square of HL3 (H_{is}^2) is also included in the model to account for a possible diminishing impact of HL3 on repetition rates (as observed in the data). X_i is a vector of individual learner control variables that are expected to predict repetition outcomes (gender, overage status, and whether the Home Language subject differs from L1), and M_i is a vector containing measures of other Foundation Phase outcomes that may enhance the fit of the model. The model is developed in a stepwise manner, and variables are retained if they either enhance the fit of the model (without violating any important assumptions), or if they are important to report irrespective of their impact on the model. Time-invariant school fixed effects Υ_s are removed, and within-school effects are estimated. ε_{is} is an idiosyncratic error term which is clustered at the school level, allowing for correlation between the unobserved characteristics of learners within the same school. Robust standard errors are used to account for heteroskedasticity in the standard errors of a linear probability model (Wooldridge, 2010: 454).

Since this is a linear probability model, the R-squared does not have the standard interpretation. In the case of linear probability models, it is more useful to consider the percent correctly predicted (Wooldridge, 2013). Specifically, if the fitted value is less than a specific threshold (often 0.5) then the predicted value is set to 0, and 1 otherwise. The specific threshold value of 0.5 is essentially arbitrary, and may be replaced by, for example, the mean value of the outcome variable (Wooldridge, 2013). Both the 0.5 threshold and the mean value of repetition will be used, with the most accurate threshold presented. These predicted values are then compared to the actual repetition outcomes to determine the percent correctly predicted. However, the overall percent correctly predicted can be misleading for less likely outcomes (Wooldridge, 2010: 590); therefore the percent correctly predicted for each outcome is also presented.

Estimation (2): Foundation Phase Home Language Mastery and Grade 4 EFAL marks

This model uses similar control variables to the repetition model above, but the dependent variable is the learner's Grade 4 EFAL result (between 0 and 100). The independent variable of interest is once again HL3, and the functional form of the model is derived from the observed relationship between these two variables in the data. The following linear regression is estimated using OLS:

$$Y_{is} = \rho_0 + \rho_1 H_{is} + X_i\delta + M_i\alpha + \Upsilon_s + \varepsilon_{is} \quad [2]$$

where Y_{is} is the Grade 4 EFAL mark of learner i in school s. The rest of the model is the same as Model 1 above and is similarly developed in a stepwise manner to determine the best model.

Descriptive Results

Province and school quintile summaries

There is significant heterogeneity in the dataset by both province and school quintile. Table 3 presents mean values of some covariates for the longitudinal dataset by province.

Table 3. Mean values of select covariates, by province

	EC	GT	KZN	LP	MP	NW
Mean Grade 4 repetition rate (%)	9.4 [9.3 , 9.5]	7.5 [7.4 , 7.6]	8.2 [8.1 , 8.3]	10.0 [9.9 , 10.1]	7.0 [6.9 , 7.1]	13.0 [12.8 , 13.2]
Mean Grade 4 EFAL result (%)	57.3 [57.2 , 57.3]	61.7 [61.6 , 61.8]	59.6 [59.5 , 59.6]	60.3 [60.2 , 60.3]	61.8 [61.8 , 61.9]	58.9 [58.8 , 58.9]
Mean Grade 3 HL result (%)	66.2 [66.1 , 66.2]	67.9 [67.9 , 68.0]	68.4 [68.3 , 68.4]	70.3 [70.2 , 70.3]	69.6 [69.5 , 69.7]	69.1 [69.0 , 69.2]
Percentage of learners with condoned pass in Grade 3	5.4 [5.3 , 5.5]	8.3 [8.2 , 8.4]	8.6 [8.5 , 8.7]	5.3 [5.3 , 5.4]	8.1 [8.0 , 8.2]	7.8 [7.7 , 7.9]
Percentage of female learners	48.6 [48.4 , 48.8]	49.0 [48.7 , 49.2]	48.1 [48.0 , 48.3]	49.0 [48.8 , 49.1]	49.1 [48.8 , 49.3]	49.3 [49.0 , 49.6]
Percentage of overage learners (Grade 4)	45.7 [45.5 , 45.9]	36.6 [36.4 , 36.8]	36.4 [36.2 , 36.6]	25.4 [25.2 , 25.5]	34.4 [34.2 , 34.7]	39.1 [38.8 , 39.3]
Percentage of learners with HL subject = mother tongue	99.0 [99.0 , 99.1]	84.9 [84.7 , 85.0]	99.1 [99.1 , 99.1]	95.9 [95.8 , 95.9]	90.0 [89.9 , 90.2]	89.3 [89.2 , 89.5]
N	279 538	174 800	294 071	327 563	179 352	138 037

Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). 95% confidence intervals shown in brackets.

Grade 4 repetition rates in the sample vary significantly across provinces, from 7.0% in Mpumalanga to 13.0% in North West. There is also some variation in Grade 4 EFAL results and Grade 3 HL results, although the mean results are not obviously correlated with mean repetition rates: despite much higher Grade 4 repetition in North West, Grade 4 EFAL results are only second-lowest, and Grade 3 HL results are third highest. Limpopo has the lowest rates of

condoning failed learners through Grade 3 at 5.3%, while KZN has the highest rate at 8.6% (however, the relationship across provinces is affected by differences in mark inflation to pass levels across provinces – see Figure 2). In all provinces females are slightly under-represented, likely due to high levels of retention of males in the Foundation Phase. Overage learners are defined using the official age recommendations[4] as stipulated in the Education Laws Amendment Act of 2002 (Government of South Africa, 2002). The proportion of learners who are overage by Grade 4 varies significantly, from just 25.4% of learners in Limpopo, to 45.7% of learners in the Eastern Cape. Finally, 84.9% of learners in Gauteng have their mother tongue as the LOI, compared to 99% of learners in the Eastern Cape and KZN.

Table 4 presents the mean values of the same statistics, by school quintile. Unsurprisingly, Grade 4 repetition rates are highest in Quintile 1, and lowest in Quintile 5. Quintiles 4 and 5 have the highest rates of granting condoned passes, but this is in part due to lower rates of artificial mark increase to push learners through (see Figure A 2). Quintile 4 schools have the highest proportion of overage learners at 40.6%, while Quintile 2 schools have the lowest at 33.6%. The rather surprising result of more overage learners in high quintile schools may be because lower quintile schools tend to follow the official age guidelines more closely and admit learners to Grade 1 at a younger age than higher quintile schools (Böhmer, Forthcoming).

Table 4: Mean values of select covariates, by school quintile

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Mean Grade 4 repetition rate (%)	10.0 [9.9 , 10.1]	8.7 [8.6 , 8.8]	8.9 [8.8 , 9.0]	7.5 [7.3 , 7.7]	5.2 [4.8 , 5.6]
Mean Grade 4 EFAL result (%)	59.3 [59.2 , 59.3]	60.0 [60.0 , 60.1]	59.6 [59.6 , 59.7]	61.7 [61.6 , 61.9]	63.2 [62.9 , 63.5]
Mean Grade 3 HL result (%)	68.8 [68.7 , 68.8]	68.9 [68.9 , 69.0]	68.0 [68.0 , 68.1]	68.2 [68.1 , 68.4]	68.8 [68.6 , 69.1]
Percentage of learners with condoned pass in Grade 3	6.4 [6.3 , 6.5]	6.5 [6.4 , 6.5]	7.8 [7.7 , 7.9]	9.4 [9.2 , 9.6]	8.0 [7.5 , 8.5]
Percentage of female learners	48.8 [48.6 , 48.9]	48.8 [48.7 , 49.0]	48.7 [48.5 , 48.8]	48.7 [48.4 , 49.1]	49.9 [49.0 , 50.8]
Percentage of overage learners (Grade 4)	34.8 [34.6 , 34.9]	33.6 [33.4 , 33.7]	38.0 [37.8 , 38.1]	40.6 [40.2 , 40.9]	37.3 [36.5 , 38.2]
Percentage of learners with HL subject = mother tongue	94.5 [94.5 , 94.6]	94.6 [94.6 , 94.7]	95.0 [94.9 , 95.0]	90.9 [90.6 , 91.1]	84.3 [83.6 , 84.9]
N	461 222	421 707	424 719	71 867	12 393

Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). 95% confidence intervals shown in brackets.

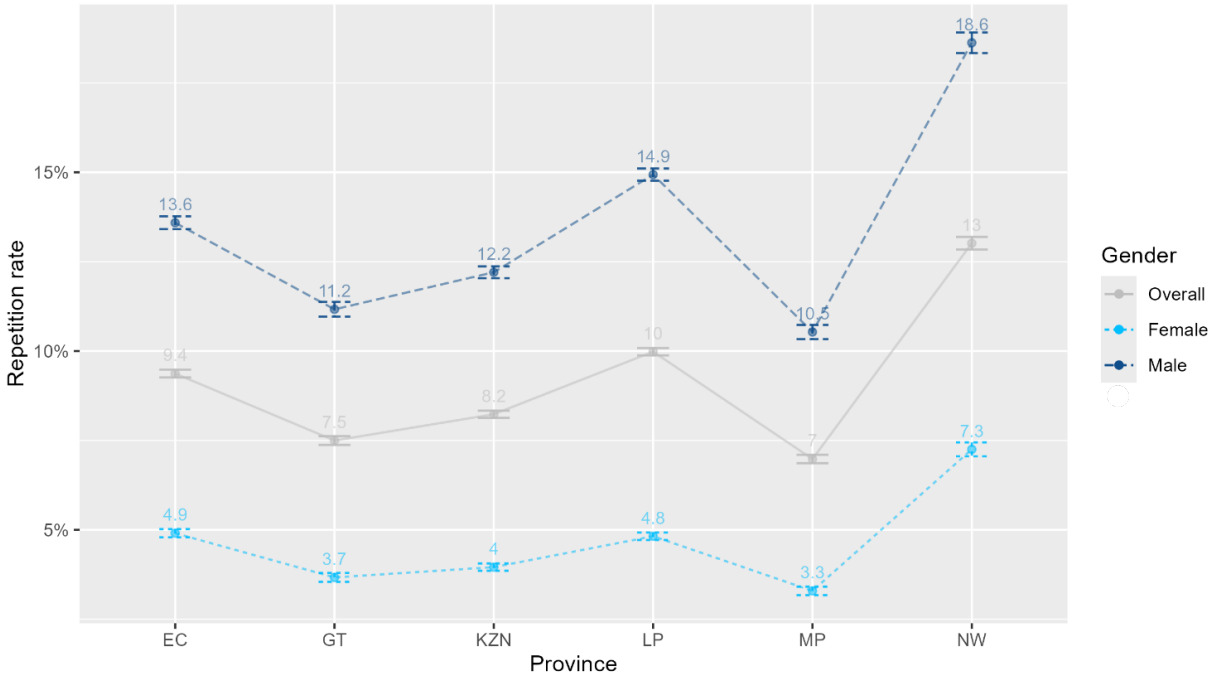
[4] The policy states that learners must enter Grade 1 if they turn 6 by 30 June of that year. Therefore, learners are classified as overage if they are older than their grade plus 5.5 years at the start of the year. This is a strict definition, and it appears that many schools may not have implemented this updated admission requirement, rather using the calendar year (enter Grade 1 if turning 7 in the year). It may be the case that some learners are classified as "overage", even though they did not repeat and would not have been considered overage while at school.

Grade 4 repetition

Grade 4 repetition by gender

Figure 1 shows Grade 4 repetition rates for each province, by gender. The gender gap in favour of females is fairly constant and large in provinces, at about 9 percentage points.

Figure 1. Mean Grade 4 repetition rates by province and gender



Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Error bars show a 95% confidence interval.

Learner characteristics (by repetition status)

Table 5 summarises the characteristics of the learners who pass Grade 4 on the first attempt, and those who repeat Grade 4 on the first attempt, as well as the percentage point difference between the means of the two groups. Higher HL3 results are strongly associated with passing Grade 4 on the first attempt, with passers achieving an average of 70.5%, and repeaters averaging 49.6% (a difference of 20.9 percentage points). In fact, only 3.7% of Grade 4 passers failed HL3 (and therefore received a condoned pass in Grade 3), compared to 33.6% of repeaters. Conversely, 40.9% of passers achieved at least 75% in HL3, compared to just 3.1% of Grade 4 repeaters. Grade 1 Mathematics results are similarly (but not as strongly) associated with repetition, with passers achieving 16.8 percentage points more for Grade 1 Mathematics than repeaters. Foundation Phase repetition is positively associated with Grade 4 repetition, with the most significant relationship with Grade 1 repetition: Grade 4 repeaters are 22 percentage points more likely to have also repeated Grade 1, while the difference in Grades 2 or 3 repetition is 8 and 6 percentage points respectively. Receiving a condoned pass in Grade 3 (progressing to Grade 4 despite not meeting the CAPS promotion requirements for Grade 3) is even more strongly associated with repetition. 35.4% of learners who repeated Grade 4 received a condoned pass in Grade 3, compared to 4.2% of learners who passed Grade 4; in fact, 92% of

learners who received a condoned pass failed their Home Language subject (figure not shown in table).

Table 5. Characteristics of learners that pass versus repeat Grade 4

	1. Passed Grade 4		2. Repeated Grade 4		Difference
	N	Mean / SE	N	Mean / SE	(1) - (2)
Grade 3 Home Language result	1 266 650	70.459 [0.0128]	126 711	49.557 [0.0411]	20.903***
Failed Grade 3 HL (<50%)	1 266 650	0.037 [0.0002]	126 711	0.336 [0.0013]	-0.299***
At least 75% in Grade 3 HL	1 266 650	0.409 [0.0004]	126 711	0.031 [0.0005]	0.377***
Grade 1 Mathematics result	1 265 860	70.576 [0.0153]	126 602	53.816 [0.0528]	16.760***
Repeated Grade 1	1 266 648	0.112 [0.0003]	126 711	0.333 [0.0013]	-0.221***
Repeated Grade 2	1 220 537	0.070 [0.0002]	121 780	0.146 [0.0010]	-0.076***
Repeated Grade 3	1 266 640	0.049 [0.0002]	126 706	0.105 [0.0009]	-0.056***
Grade 3 condoned pass	1 265 314	0.042 [0.0002]	126 598	0.354 [0.0013]	-0.312***
HL3 differs from actual HL	1 266 650	0.056 [0.0002]	126 711	0.056 [0.0006]	0.000
Number of days absent in Grade 3	256 656	5.548 [0.0125]	23 106	7.184 [0.0552]	-1.636***
Overage in Grade 4	1 266 650	0.326 [0.0004]	126 711	0.667 [0.0013]	-0.341***
Female	1 266 650	0.512 [0.0004]	126 711	0.244 [0.0012]	0.267***
Fee-paying school (Q4-5)	1 266 650	0.062 [0.0002]	126 711	0.047 [0.0006]	0.014***

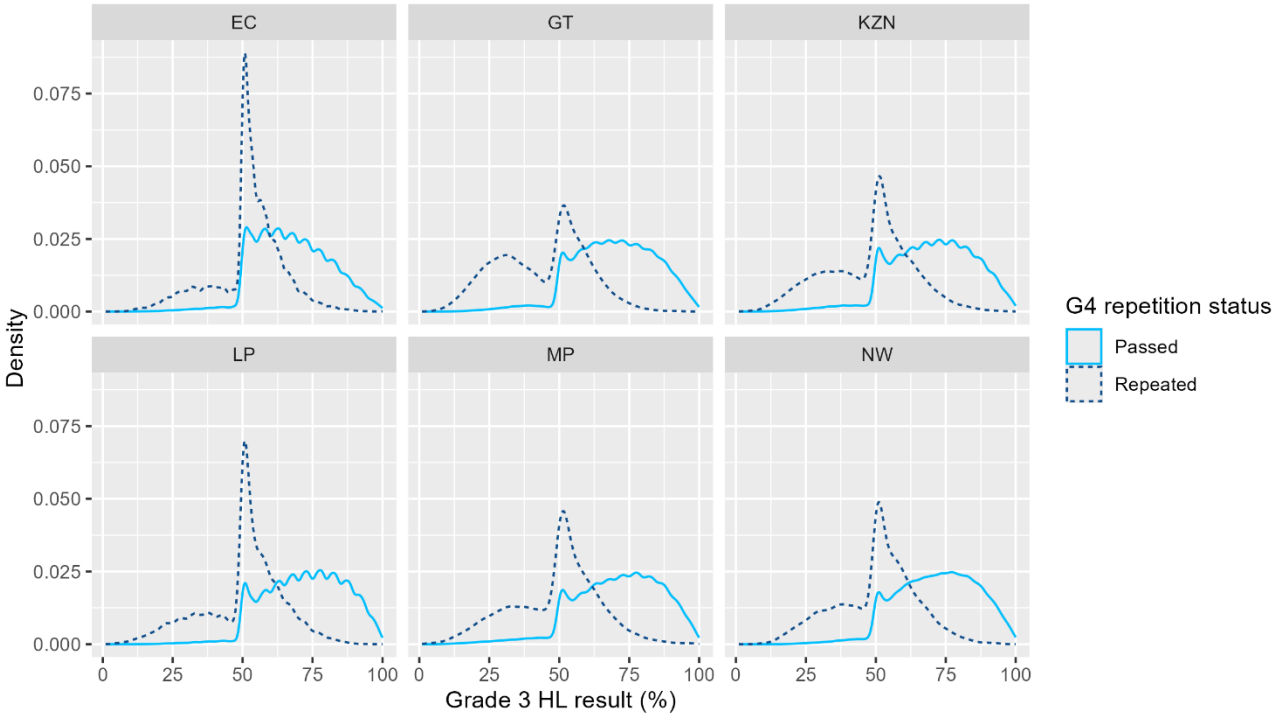
Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Notes: The value displayed for t-tests are the differences in the means across the groups. *** indicates significance at the 1 percent critical level. SE = standard error.

For both Grade 4 passers and repeaters, 5.6% took a home language subject which differed from their mother tongue, indicating no difference in the aggregate. Absenteeism is however associated with repetition: learners who passed Grade 4 were absent on 1.6 fewer days (in Grade 3) than repeaters. Being overage is also strongly associated with repetition, with two-thirds of Grade 4 repeaters being overage, compared to about one-third of non-repeaters. As expected, gender is also strongly predictive of repetition, with only 24.4% of repeaters being female. 6.2% of passing learners attended fee-paying schools, compared to 4.7% of repeaters; a difference of 1.4 percentage points.

Grade 4 repetition and Grade 3 Home Language results

Figure 2 presents the distributions of Grade 3 Home Language results, by Grade 4 repetition status.

Figure 2. Distribution of Grade 3 Home Language results, by Grade 4 repetition status



Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023).

All the graphs have a spike at 50%, suggesting that results just below 50% were artificially increased[5] to 50% (the minimum pass mark). The much larger spike at 50% in the repetition group indicates that a greater proportion of repeaters were pushed through to Grade 4 (indicating that the actual number of condoned passes among repeaters may be even higher than two-thirds). Finally, the practice of inflating marks to 50% appears to be most prevalent in the Eastern Cape and Limpopo, and least prevalent in Gauteng. However, there are also differences in the proportions of "officially" condoned passes (passes that are condoned despite not meeting the subject-specific requirements) across provinces (Table 3), with the Eastern Cape and Limpopo having the lowest rates of officially condoned passes.

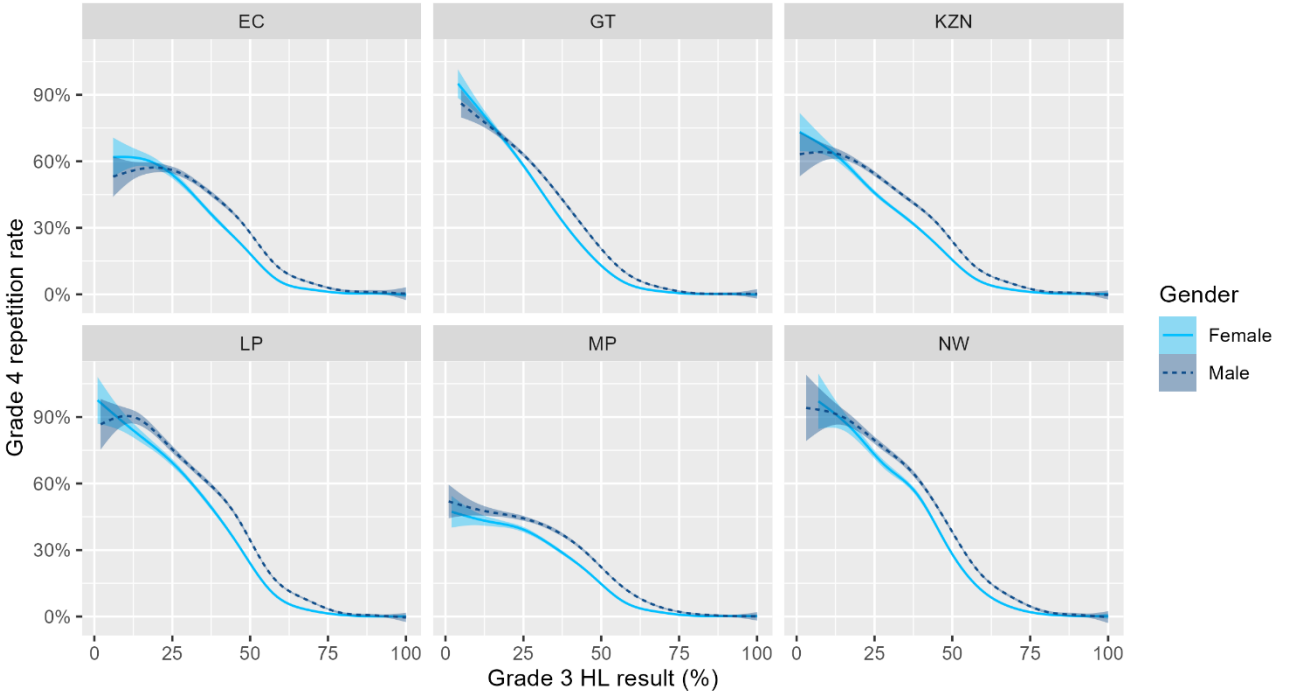
Figure 3 shows the relationship between Grade 4 repetition and HL3 by plotting loess curves with a 95% confidence interval on a random sample of up to 100 000[6] learners per province-gender group. The curves show that the relationship between HL3 and Grade 4 repetition rates is negative, and most significant (steepest) between 25-75%. It is slightly flatter at the bottom end

[5] While mark adjustments are an established practice, Covid19 adjustments to promotion requirements (between 2020-2022) formalised this practice by explicitly allowing for a 5-percentage point adjustment in up to three subjects (Hoadley, 2023).

[6] A random sample was used as the sample was too large to plot loess curves for some groupings (due to computational limitations).

of the HL3 distribution, and significantly flatter at the top end of the distribution, with almost all learners who scored 75% or higher, passing Grade 4. In fact, the repetition rate for the 28-42% of learners who achieved at least 75% for HL3 is less than 1.3% for all provinces (see Table A 5). For learners scoring between 25% and 75%, there is a gender gap of more than 5 percentage points in most provinces, indicating that gender is a significant predictor of repetition for many learners, even controlling for HL3. For learners at the bottom and top of the distribution, gender is less significant (controlling for HL3).

Figure 3. Grade 4 repetition and Grade 3 HL result, by gender

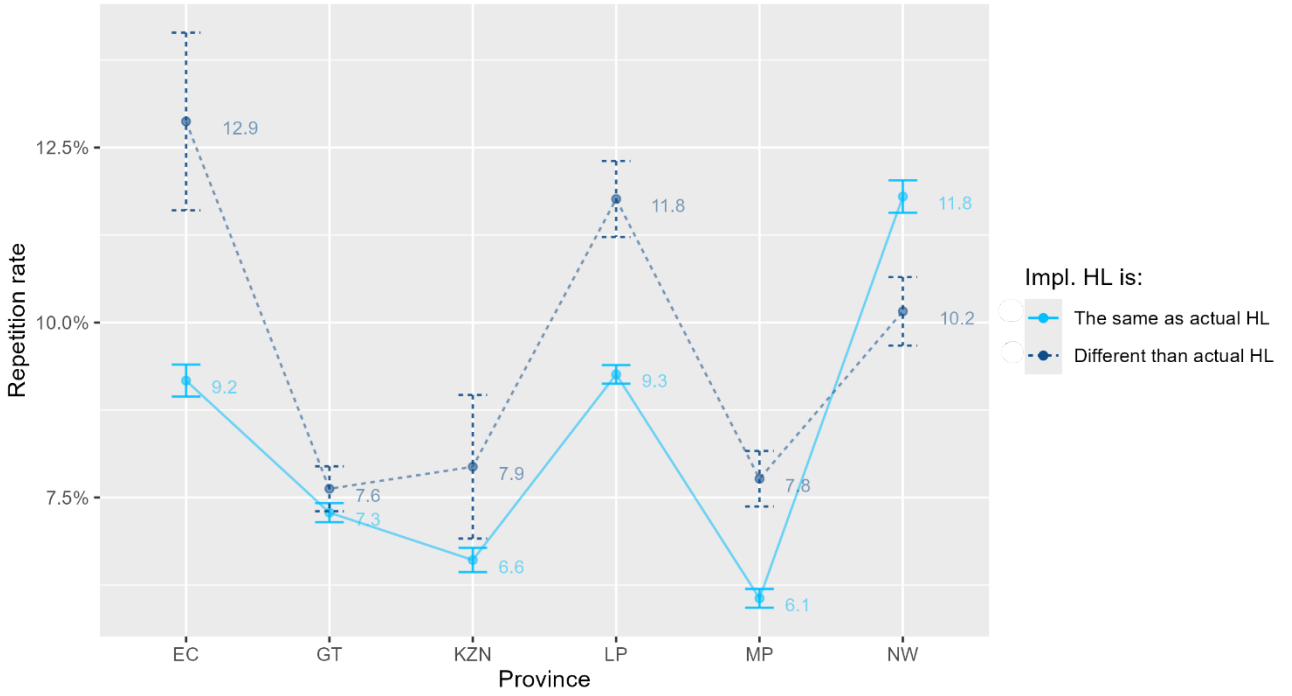


Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Loess curves on a random sample of up to 100 000 learners per province-gender group. Shading represents a 95% confidence interval.

Grade 4 repetition and mother tongue Home Language subject

Just 5.6% of the learners in the sample have a home language subject ("Implemented Home Language") that differs from the language that they speak at home ("Actual Home Language"). Figure 4 shows mean repetition rates by Home Language status, across province and by cohort. For comparability, the sample is restricted to those schools that have at least one learner whose implemented Home Language subject differs from their actual home language.

Figure 4. Grade 4 repetition rates by province and Home Language subject implementation status



Source: Implemented HL variation subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023, and whose grade includes learners whose mother tongue is not their Home Language subject). Error bars show a 95% confidence interval.

Learners whose Home Language subject differs from their mother tongue, repeat at significantly higher rates in the Eastern Cape, KZN, Limpopo, and Mpumalanga. In Gauteng there is no difference in mean repetition rates across the two groups, while in the North West learners whose implemented home language subject is not the same as their actual home language, repeat at a lower rate. This unexpected result in the North West occurs in the 2017 and 2018 cohorts, but not in the 2019 cohort (when the two groups repeat at equal rates). It is specific to Setswana Home Language subject, and is not explained by rural/urban, gender, or quintile. The results tentatively suggest that whether a learner's implemented Home Language subject is their mother tongue, or another African language, is not particularly important for Grade 4 repetition outcomes in Gauteng and KZN, although it is significantly associated with lower repetition rates in the Eastern Cape and Limpopo, and Mpumalanga to a lesser extent.

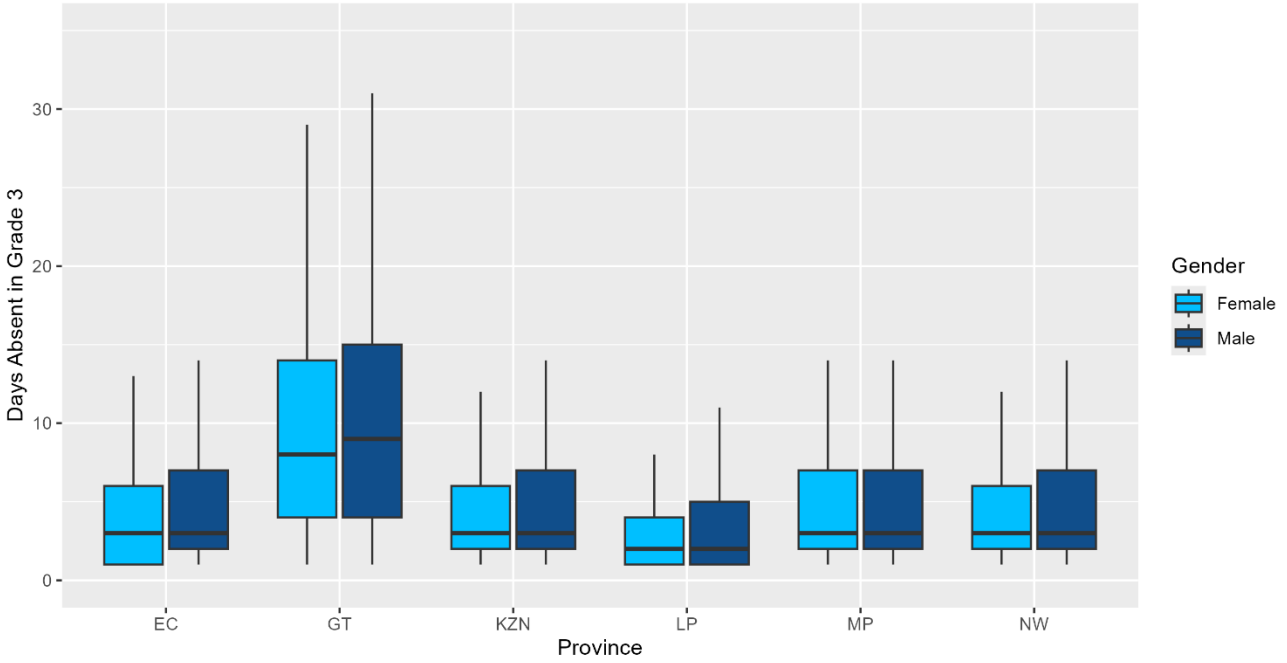
Grade 4 repetition and Grade 3 absenteeism

Figure 5 presents boxplots[7] of the distribution of days absent in Grade 3. Not only are a much higher proportion of Gauteng learners recorded as being absent at least once (see Table A 3), the overall distribution is much higher than other provinces, with the median of 8 days absent more than twice as large as the median in any other province. This reinforces the observation of significant differences in either absenteeism recording, or absenteeism itself, in Gauteng. The

[7] The bottom and top of each box represents the 25th and 75th percentile respectively, while the middle line represents the median. The whiskers give the largest or smallest value above or below the respective quartile plus (or minus) 1.5 multiplied by the interquartile range.

figure shows that absenteeism is roughly similar by gender for the bottom half of the distribution, but at 75th percentile boys tend to be absent slightly more often.

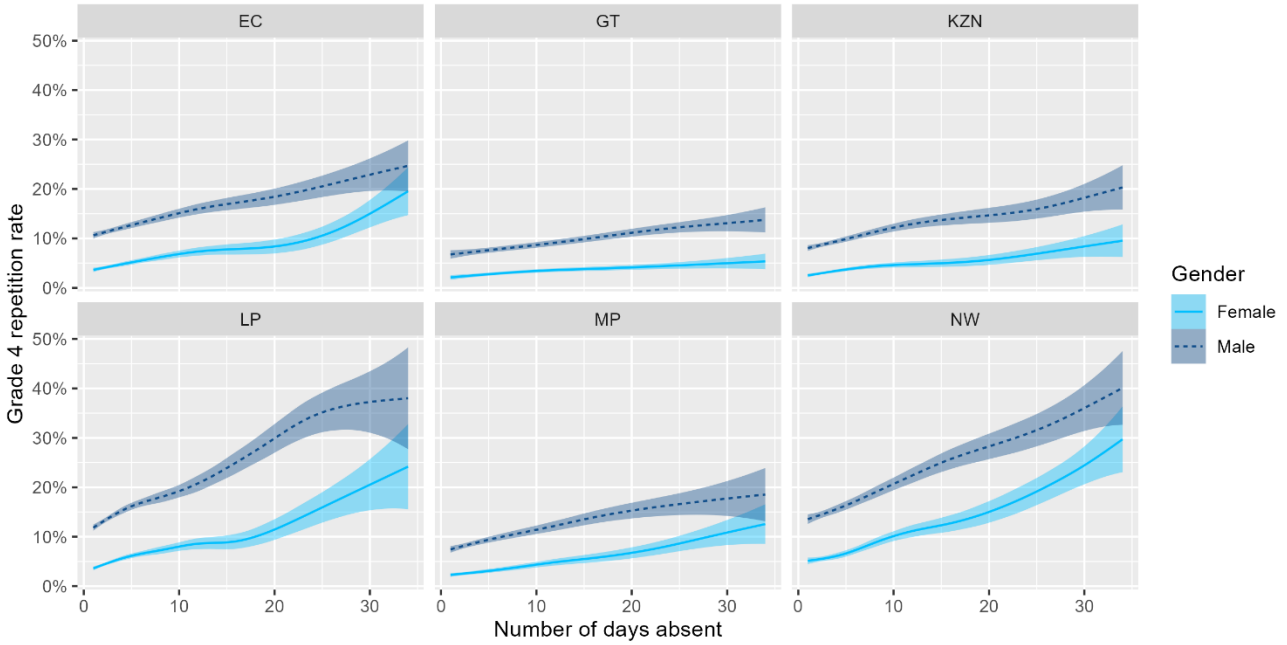
Figure 5. Distribution of days absent in Grade 3, by province and gender



Source: Absentee subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 in 2017 and who reached Grade 4 in 2020). Outliers not displayed.

Figure 6 shows a positive linear relationship between Grade 4 repetition and Grade 3 absenteeism, with a slope that is steeper in Limpopo and the North West, and flatter in the Eastern Cape, Mpumalanga, and especially in Gauteng.

Figure 6. Grade 4 repetition and absenteeism in Grade 3



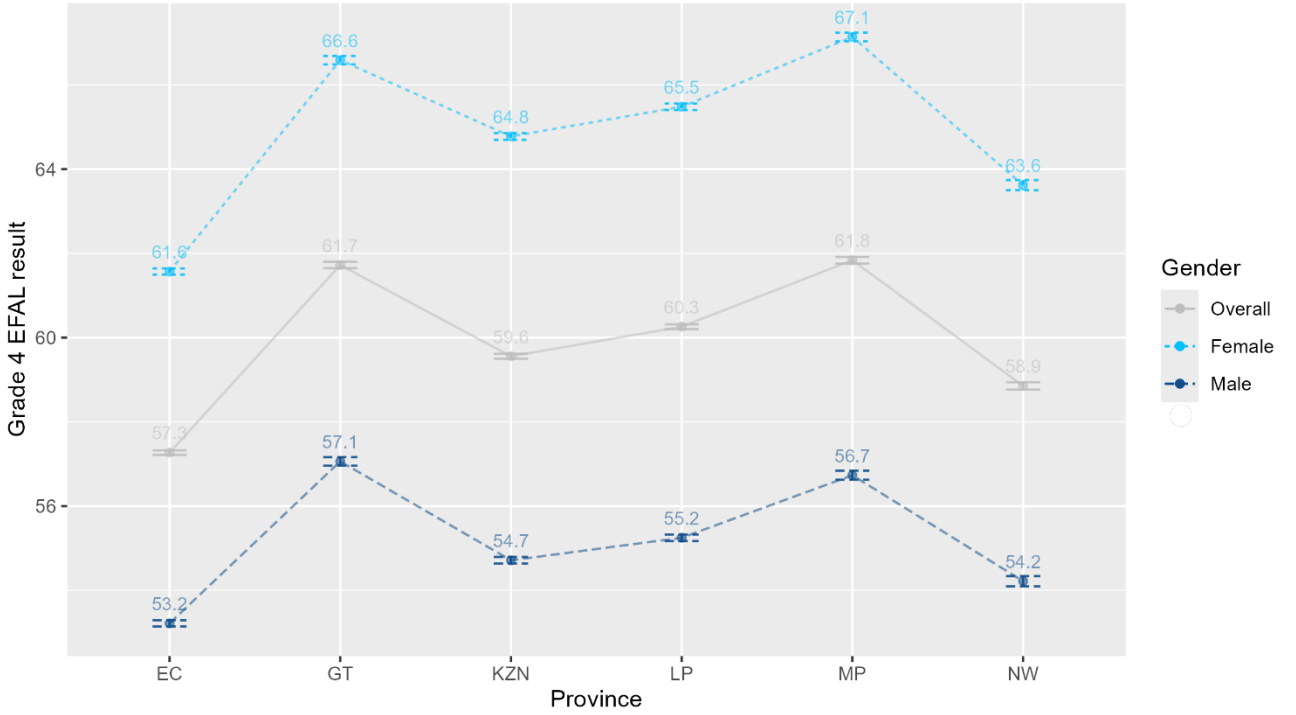
Source: Absentee subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 in 2017 and who reached Grade 4 in 2020). Loess curves on a random sample of up to 100 000 learners per province-gender group. Shading represents a 95% confidence interval.

Grade 4 EFAL results

Grade 4 EFAL results by province and gender

Figure 7 presents Grade 4 EFAL results, by province and gender. As with repetition, the gender gap in favour of females is large and relatively consistent across provinces.

Figure 7. Grade 4 EFAL results by province and gender

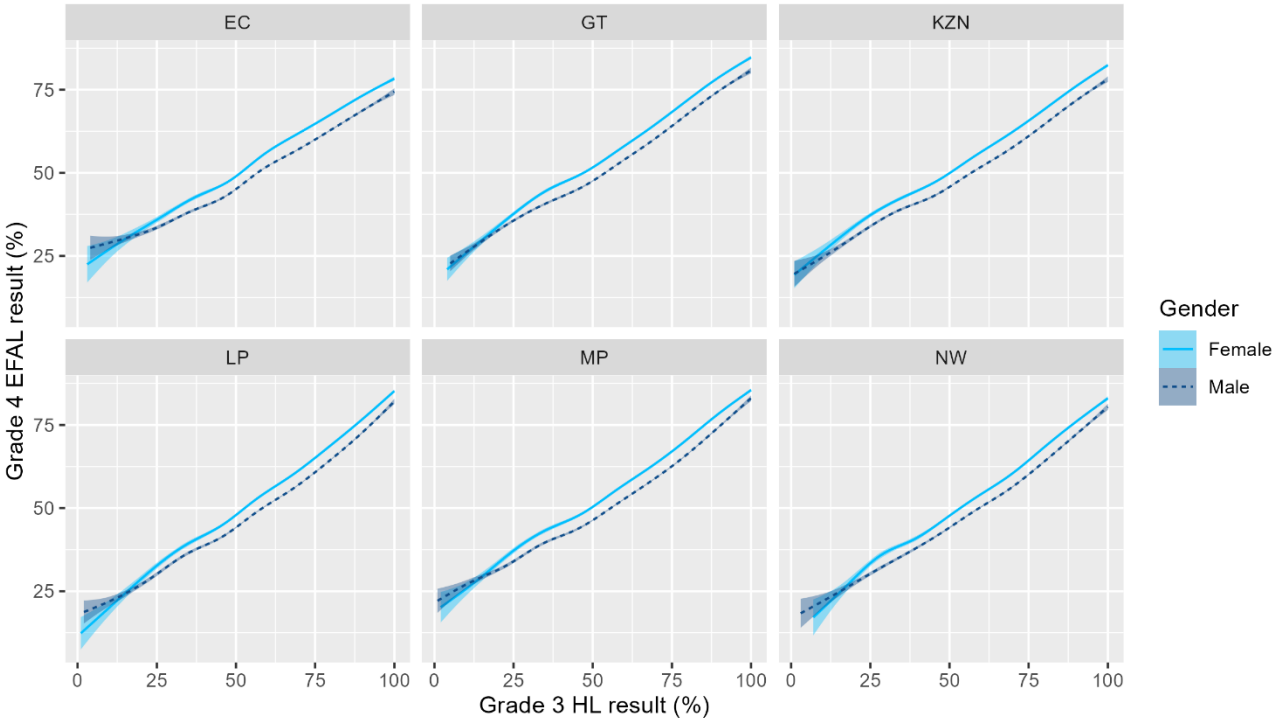


Source: EFAL subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who took EFAL, started Grade 1 between 2017 and 2019 and reached Grade 4 by 2023). Error bars show a 95% confidence interval.

Grade 4 EFAL results and Grade 3 Home Language results

The loess curves in Figure 8 plot the relationship between Grade 3 Home Language results, and Grade 4 EFAL results. There is a strong positive linear relationship between the two, which is relatively consistent across provinces. Barring those few learners at the bottom of the distribution (those who scored less than 25% for HL3), female learners score about 3-5 percentage points better, on average, than male learners at every level of HL3.

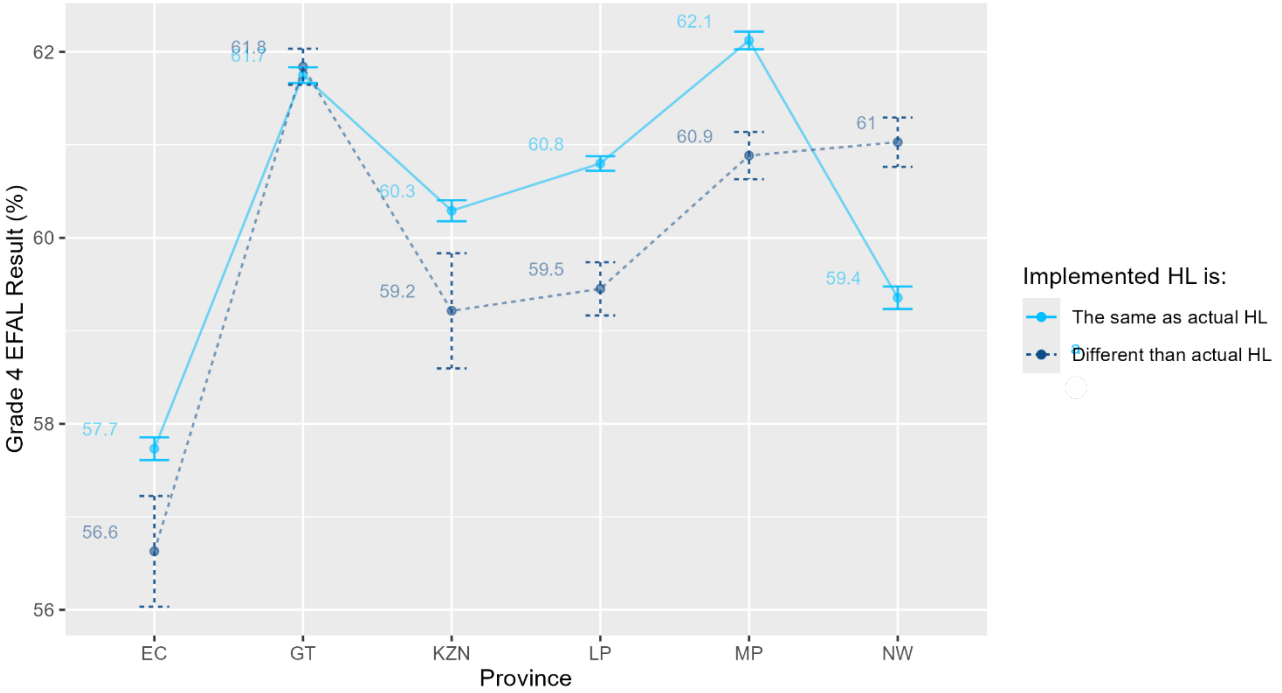
Figure 8. Grade 4 EFAL results and Grade 3 HL results



Source: EFAL subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who took EFAL, started Grade 1 between 2017 and 2019 and reached Grade 4 by 2023). Loess curves on a random sample of up to 100 000 learners per province-gender group. Shading represents a 95% confidence interval.

Grade 4 EFAL results and mother tongue Home Language subject

Figure 9. Grade 4 EFAL results and Grade 3 HL implementation status



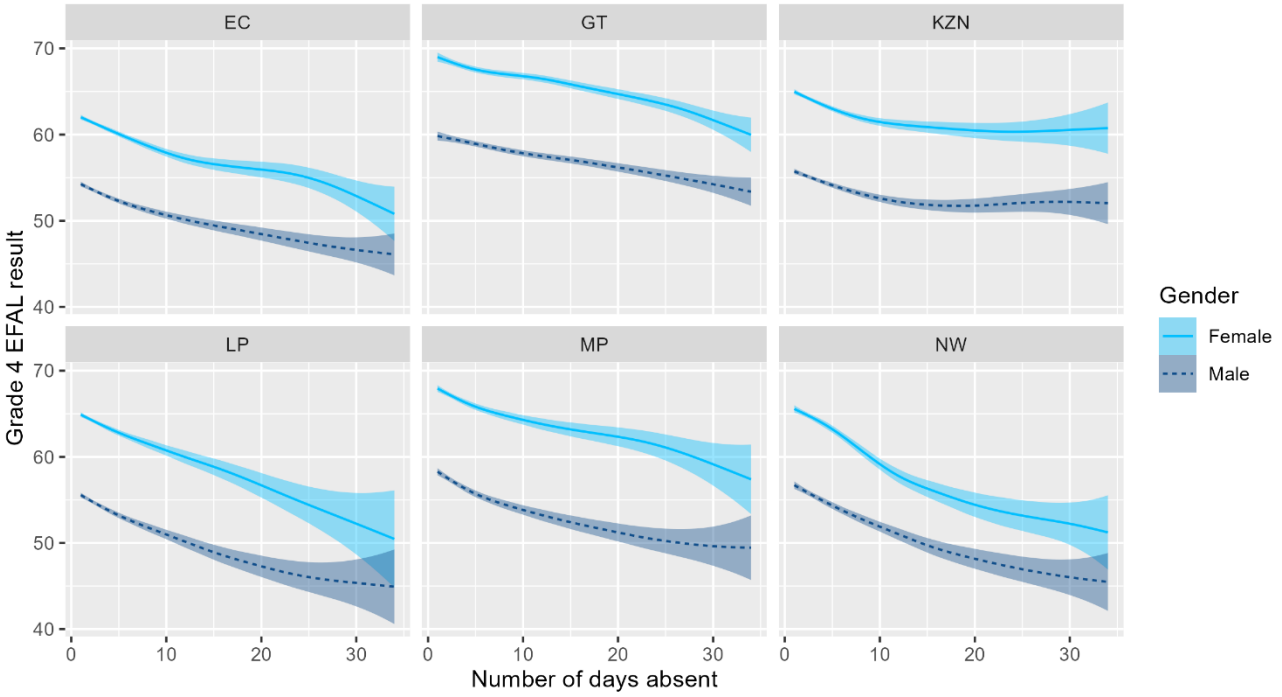
Source: EFAL subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who took EFAL, started Grade 1 between 2017 and 2019 and reached Grade 4 by 2023). Reduced sample of schools with variation in HL implementation status. Error bars show a 95% confidence interval.

Figure 9 presents the mean EFAL results for learners whose HL3 subject is the same as their mother tongue, and those whose HL3 subject is a different (African) language than their mother tongue. In the Eastern Cape, KZN, Limpopo and Mpumalanga, learners whose HL3 is their mother tongue achieve significantly higher EFAL results in Grade 4. In Gauteng there is no difference, and in the North West learners with a different HL3 subject perform better in Grade 4 EFAL. This unexpected result mirrors that shown for repetition in the North West.

Grade 4 EFAL results and absenteeism

Figure 10 presents the relationship between Grade 4 EFAL results and Grade 3 absenteeism (for learners that were in Grade 3 in 2019, since this was the only cohort with valid absenteeism data). In all provinces barring KZN, there is a negative linear relationship between number of days absent in Grade 3, and Grade 4 EFAL performance. In KZN the relationship becomes flat after approximately 15 days of absence.

Figure 10. Grade 4 EFAL results and Grade 3 absenteeism



Source: EFAL/absentee subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject and EFAL, who started Grade 1 in 2017 and who reached Grade 4 in 2020). Loess curves on a random sample of up to 100 000 learners per province-gender group. Shading represents a 95% confidence interval.

Estimation results (1): Impact of Grade 3 Home Language mastery on the probability of passing Grade 4

Estimating repetition: Model selection

Table 6 presents the results of various models considered, applied to Quintile 1 schools only. When interpreting the fit of the model, the adjusted R-squared is not particularly useful since this is a linear probability model (and values may lie below zero or above one). Overall

percentage correctly predicted is also not particularly insightful, since approximately 90% of learners do not repeat, and thus a model that predicts that every learner will pass will be correct 90% of the time. Therefore, when interpreting the fit of the model I will focus on percent correctly predicted of repeaters. The predicted outcome is based on a 0.5 threshold[8] to determine predicted outcomes; that is, a learner whose likelihood of repetition is greater than 0.5 will be classified as being predicted to repeat; otherwise, they will be predicted to pass.

Model 1, which includes only HL3 result, shows that HL3 is negatively associated with repetition: a learner who scores 10 percentage points higher on HL3 would be 8 percentage points less likely to repeat. However, this model only predicts repeaters correctly 1.4% of the time. The addition of the square of HL3 increases the accuracy of prediction of repeaters substantially, to 13.8%; a quadratic model is a far better fit for the data. The positive coefficient on HL3 indicates a convex quadratic; this is sensible insofar as the data would primarily fall on the decreasing portion of the quadratic, with the impact on repetition flattening out as HL3 increases (see Figure 3). The coefficients on these two variables of interest remain consistent in all the models. The inclusion of gender and overage status only slightly increases the percent of repeaters correctly predicted, to 14.1%. The inclusion of a dummy variable to indicate whether a learner's HL subject is not their mother tongue (Model 4) does not improve the fit of the model but is significant and the coefficient is negative; but this may be due to the types of schools where learners with differing HL subjects are located. Models 5 and 6 include potential control variables for learners' prior academic performance, namely Grade 1 Mathematics performance, as well as the difference between a learner's Grade 3 and Grade 1 Mathematics results (the *Math delta*, to capture their learning trajectory), and Grade 1 repetition status. The inclusion of these variables increases the fit of the model, but only Grade 1 repetition status is selected for the final model, for ease of interpretation.

In Model 7, the inclusion of province significantly increases the accuracy of the model in terms of predicting repetition, indicating significant differences in the relationships across provinces. Finally, Model 8 presents the best model with the inclusion of school fixed effects. This improves the accuracy of predicting repetition outcomes significantly, albeit only to 19.1%. Model 8 includes the dummy variable for HL3 implementation status, and while it is significant in the fixed effects model and positively associated with repetition, the effect is so small (a 0.4 percentage point increase in the predicted repetition rate if a learner's HL subject is not their mother tongue) as to have no impact on the accuracy of the mode. Prediction of passing is highly accurate throughout, with a correct prediction rate above 99% in all models.

[8] The mean repetition rate was also considered as a threshold for predicting outcomes (instead of the 0.5 threshold). While this increased the model accuracy in predicting repeaters, it substantially reduced the accuracy of predicting passers, thereby reducing the overall accuracy of the model.

Table 6: Model selection for estimating Grade 4 repetition (Quintile 1 sample)

Grade 4 Repetition	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Grade 3 HL result	-0.0076***	-0.0298***	-0.0288***	-0.0288***	-0.0272***	-0.0283***	-0.0280***	-0.0272***
Square of Grade 3 HL result		0.0002***	0.0002***	0.0002***	0.0002***	0.0002***	0.0002***	0.0002***
Female			-0.0384***	-0.0384***	-0.0401***	-0.0380***	-0.0371***	-0.0338***
Overage in Grade 4			0.0468***	0.0468***	0.0285***	0.0287***	0.0310***	0.0256***
HL3 differs from actual HL				-0.0110***	-0.0072***	-0.0098***	-0.0046***	0.0036*
Grade 1 Maths result					-0.0028***			
Math delta					-0.0016***			
Repeated Grade 1						0.0554***	0.0591***	0.0607***
GT							-0.0154***	
KZN							-0.0020*	
LP							0.0370***	
MP							-0.0022	
NW							0.0498***	
Constant	0.6233***	1.3146***	1.2550***	1.2559***	1.3353***	1.2353***	1.2119***	
Adjusted R-squared	0.15	0.18	0.19	0.19	0.20	0.20	0.20	0.26
Percentage correctly predicted (overall)	90.1	90.6	90.6	90.6	90.6	90.6	90.7	91.1
Percentage correctly predicted (repeaters)	1.4	13.8	14.1	14.1	14.6	14.6	15	19.1
Percentage correctly predicted (passers)	99.9	99.1	99.1	99.1	99.1	99.1	99.1	99.1
N (Learners)	461 222	461 222	461 222	461 222	460 529	461 222	461 222	461 222
N (Schools)	5 072	5 072	5 072	5 072	5 072	5 072	5 072	5 072
School fixed effect								Y

Source: Quintile 1 subset of DDD longitudinal dataset (all learners in public ordinary Quintile 1 schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Robust standard errors clustered at the school level. Significant at *10% level, **5% level, ***1% level.

Estimating repetition: By province

Table 7 presents the regression results by province (Model 8 only, with fixed effects). The coefficients on the key variable of interest, Grade 3 HL result and its square, are significant across provinces. The coefficients on HL3 and its square follow from the fit of the data on the decreasing portion of a convex quadratic, and the diminishing returns to the impact of HL3 results on repetition, especially above 75% (see Figure 8). The impact of gender differs slightly across provinces, ranging from a 2.1 percentage point reduction in predicted repetition for females in Gauteng, to a 3.8 percentage point reduction in North West. Being overage increases the probability of repetition in all provinces, from a 1.2 percentage point increase in predicted repetition for overage learners in Mpumalanga to a 3.9 percentage point increase in Limpopo. Learning in a language other than one's mother tongue has no impact in Gauteng, KZN, Limpopo and North West, but a 2.2 percentage point increase in predicted repetition in the Eastern Cape.

Finally, repetition of Grade 1 increases the likelihood of repetition in all provinces, and the effect ranges from a 4-percentage point increase in the Eastern Cape, to a 7.8 percentage point increase in the North West (controlling for other factors, including overage).

While the outcome of passing Grade 4 is predicted with over 98% accuracy in all provinces, predicting the repetition outcome is much more difficult and also more variable across provinces, ranging from just 8.4% of repeaters being correctly predicted in Mpumalanga (which also has the lowest repetition rate of the provinces), to 30.5% in the North West (which has the highest repetition rate – see Table 3).

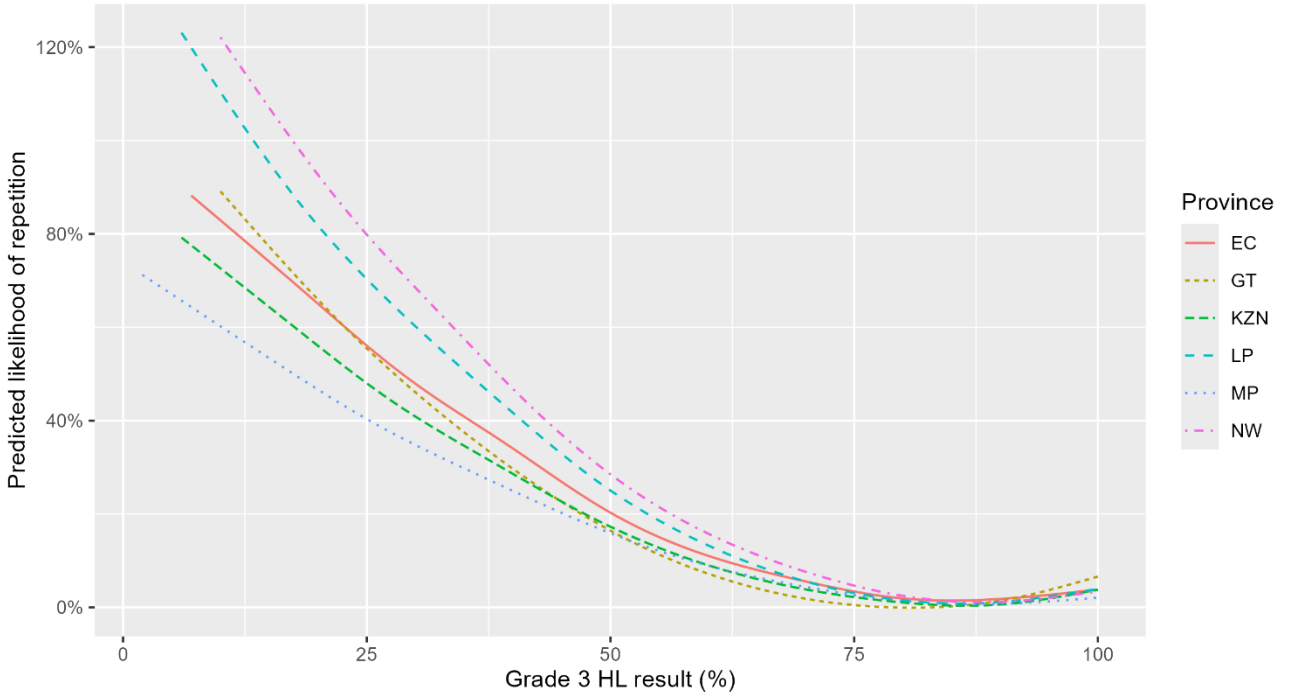
Table 7: Estimating Grade 4 repetition by province

Grade 4 Repetition	EC	GT	KZN	LP	MP	NW
Grade 3 HL result	-0.0250***	-0.0289***	-0.0224***	-0.0318***	-0.0179***	-0.0327***
Square of Grade 3 HL result	0.0001***	0.0002***	0.0001***	0.0002***	0.0001***	0.0002***
Female	-0.0362***	-0.0214***	-0.0255***	-0.0309***	-0.0246***	-0.0380***
Overage in Grade 4	0.0140***	0.0115***	0.0161***	0.0388***	0.0110***	0.0314***
HL3 differs from actual HL	0.0218***	0.0019	0.0017	0.0009	0.0053**	0.0008
Repeated Grade 1	0.0396***	0.0643***	0.0589***	0.0693***	0.0546***	0.0779***
Adjusted R-squared	0.21	0.30	0.25	0.30	0.20	0.33
Percentage correctly predicted (overall)	91.2	93.7	92.6	91.5	93.2	89.7
Percentage correctly predicted (repeaters)	11.8	26.9	17.6	22.7	8.4	30.5
Percentage correctly predicted (passers)	99.4	99.1	99.3	99.2	99.6	98.6
N (Learners)	279 539	174 800	294 070	327 562	179 352	138 037
N (Schools)	3 961	733	3 496	2 278	991	917
School fixed effect	Y	Y	Y	Y	Y	Y

Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Robust standard errors clustered at the school level. Significant at *10% level, **5% level, ***1% level.

Figure 11 presents the average predicted repetition by province. As expected with a linear probability model, some predicted values lie outside of the probability interval; however, this is only really the case for North West and Limpopo, with the predicted values for other provinces lying between zero and one (on average). Predicted repetition rates vary significantly across provinces across the HL3 distribution, with learners in Mpumalanga the least likely to repeat for a given HL3 result, and North West learners most likely to repeat. For example, a learner who increases their HL3 from 40% to 50% would be 9.3 percentage points less likely to repeat in Mpumalanga, and 16.8 percentage points less likely to repeat in North West. North West is also the province with the highest Grade 4 repetition rate (13% - see Table 3), and Mpumalanga the lowest (7%). This heterogeneity, especially in the face of similar mean HL3 results (Table 3), suggests that there is a larger gap between HL3 results and Grade 4 requirements in North West than in other provinces.

Figure 11: Provincial models: Average predicted Grade 4 repetition by HL3, for non-overage males with subject HL = mother tongue



Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Fitted values from the provincial models (with school Fixed Effects). Loess curves on fitted values. Fitted values above the horizontal line at 50% are predicted to repeat; below are predicted to pass.

The same model was applied to the absenteeism sample, with the addition of *days absent in Grade 3* as a control variable. Absenteeism is significantly associated with increased repetition, with each day absent in Grade 3 associated with between 0.1 (Gauteng) to 0.3 (North West) percentage point increase in predicted Grade 4 repetition, controlling for other factors. See Table A 6 for details.

Estimating repetition: By school quintile

Table 8 presents the estimation results by school quintile. In all quintiles, higher HL3 is associated with lower repetition, but the size of the effect differs across quintile. The gender gap is most notable in Quintile 1 (females 3.4 percentage points less likely to repeat), and decreases as the quintiles increase, to 1.4 percentage points in Quintile 5 schools. Similarly, there is more heterogeneity in repetition outcomes for overage learners in Quintile 1 (2.6 percentage points more likely to repeat) than in higher quintiles (no difference in Quintile 5 schools). Finally, having a Home Language subject that differs from one's mother tongue has a significant (but very small) impact only in Quintile 1 schools (with a 0.4 percentage point increase in predicted repetition). Repetition of Grade 1 is an important predictor of repetition across quintiles, with its impact ranging from a 5.2 percentage point increase on Grade 4 repetition in Quintile 3 schools, to an 8-percentage point increase in Quintile 5 schools. The accuracy of the model varies substantially across quintiles, from correctly estimating just 4% of Quintile 5 repeaters, to 22.7% of Quintile 4 repeaters. The very low accuracy in Quintile 5 may be due to provincial heterogeneity in this quintile, and also to the relatively low number of repeaters.

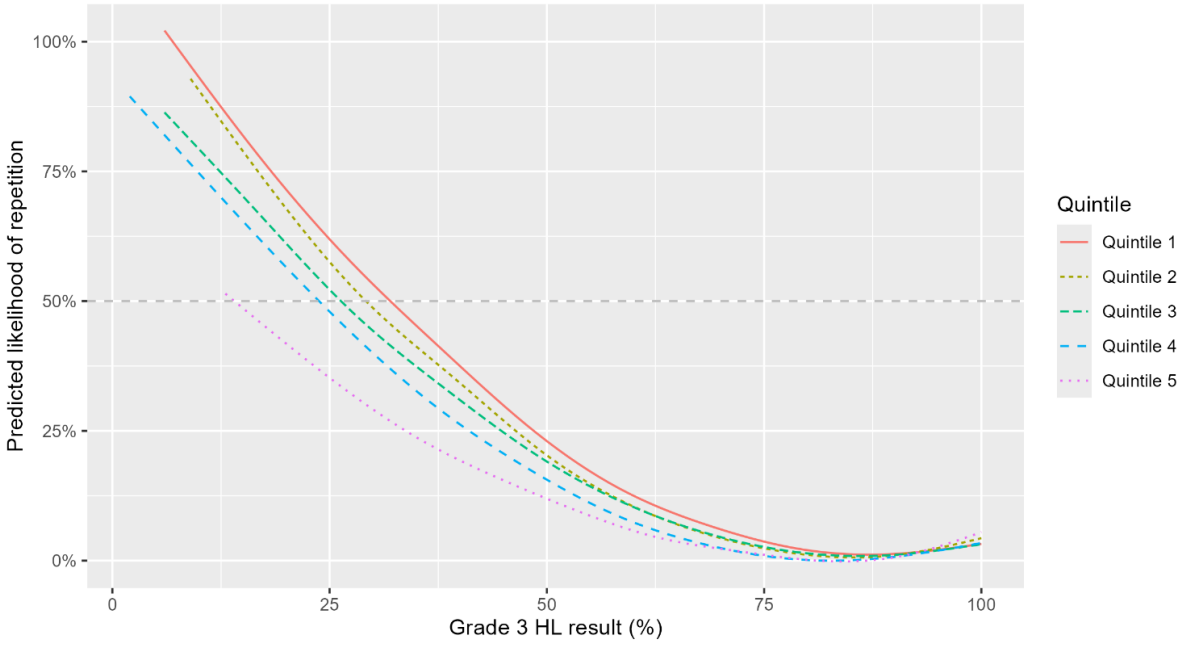
Table 8: Estimating Grade 4 repetition by school quintile

Grade 4 Repetition	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Grade 3 HL result	-0.0272***	-0.0271***	-0.0240***	-0.0241***	-0.0180***
Square of Grade 3 HL result	0.0002***	0.0002***	0.0001***	0.0001***	0.0001***
Female	-0.0338***	-0.0285***	-0.0273***	-0.0197***	-0.0139***
Overage in Grade 4	0.0256***	0.0226***	0.0152***	0.0150***	-0.0053
HL3 differs from actual HL	0.0036*	0.0017	0.0029	0.0055	0.0059
Repeated Grade 1	0.0607***	0.0577***	0.0522***	0.0606***	0.0809***
Adjusted R-squared	0.26	0.26	0.25	0.27	0.19
Percentage correctly predicted (overall)	91.1	92.3	92	93.4	94.9
Percentage correctly predicted (repeaters)	19.1	19.9	18.5	22.8	3.9
Percentage correctly predicted (passers)	99.1	99.2	99.2	99.1	99.8
N (Learners)	461 222	421 706	424 719	71 867	12 393
N (Schools)	5 072	4 021	2 890	329	50
School fixed effect	Y	Y	Y	Y	Y

Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Robust standard errors clustered at the school level. Significant at *10% level, **5% level, ***1% level.

Figure 12 illustrates that, for all values of HL3, predicted repetition is inversely related to quintile. A learner scoring 50% for HL3 would have a 12.5% chance of repeating Grade 4 if they attend a Quintile 5 school; but if they attend a Quintile 1 school this probability almost doubles, to just under 25%. This suggests that learners in Quintile 1 schools have more difficulty with the transition to Grade 4.

Figure 12: Quintile models: Predicted Grade 4 repetition by HL3 for non-overage males with subject HL = mother tongue



Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Fitted values from the quintile models. Loess curves on fitted values. Fitted values above the horizontal line at 50% are predicted to repeat; below are predicted to pass.

Estimation Results (2): Impact of Grade 3 Home Language mastery on Grade 4 EFAL results

Estimating EFAL results: model selection

Estimates of the relationship between Grade 4 EFAL results (EFAL4) and Grade 3 HL results (HL3) are shown in Table 9. Model 1 shows that a one percentage point increase in HL3 is associated with a 0.68 percentage point increase in EFAL4, and that this covariate alone explains 42% of the variation in Grade 4 EFAL results. Including a square of HL3 does not improve the accuracy of the model (no change in adjusted R-squared); this also follows from the observed linear relationship between HL3 and EFAL4 (see Figure 8). Therefore, HL3 squared is excluded from later models. Model 3 shows that there is a pro-female advantage of 4.1 percentage points on Grade 4 EFAL results, even controlling for Grade 3 Home Language results, while being overage in Grade 4 reduces the predicted value of EFAL4 by 2.9 percentage points. The inclusion of these learner characteristics increases explanatory power of the model by 2 percentage points. Model 4 indicates that having a HL subject that differs from one's mother tongue may have a small but positive impact on EFAL4, although the inclusion of this variable does not increase the explanatory power of the model. Model 5 demonstrates that controlling for general academic ability in the form of Grade 1 Mathematics performance improves the fit of the model by one percentage point, while Model 6 shows that the learning trajectory of the learner, as measured by the *Math delta* (the difference between Grade 3 and Grade 1 Mathematics results) is also important for predicting EFAL4. Model 7 shows that these two Mathematics covariates are more valuable for predicting EFAL4 than is Grade 1 repetition status. Therefore, the pair of mathematics controls will be selected for the final model, instead of Grade 1 repetition. Model 8 shows that there are significant differences in the relationships between the variables across provinces (with the Eastern Cape as the reference province), thus warranting presentation of the model by province. Model 8 presents the estimations of the preferred model with the addition of school fixed effects, which increases the R-squared to 60%. In this final model (with pooled provinces), the impact of HL3 subject differing from a learner's mother tongue becomes statistically insignificant.

Table 9: Estimating Grade 4 EFAL results (model selection) (Quintile 1 sample)

Grade 4 EFAL Result	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Grade 3 HL result	0.677***	0.360***	0.613***	0.613***	0.556***	0.411***	0.610***	0.414***	0.437***
Square of Grade 3 HL result		0.002***							
Female			4.108***	4.108***	4.106***	4.302***	4.100***	4.281***	4.032***
Overage in Grade 4			-2.852***	-2.855***	-1.138***	-1.259***	-2.416***	-1.319***	-1.027***
HL3 differs from actual HL				1.595***	1.334***	1.223***	1.567***	0.719***	-0.054
Grade 1 Maths result					0.126***	0.297***		0.296***	0.316***
Math delta						0.194***		0.192***	0.211***
Repeated Grade 1							-1.329***		
GT								1.457***	
KZN								-0.004	
LP								-0.883***	
MP								0.629***	
NW								-1.754***	
Constant	12.731***	22.615***	16.129***	16.048***	10.701***	9.197***	16.236***	9.406***	
Adjusted R-squared	0.42	0.42	0.44	0.44	0.46	0.47	0.44	0.47	0.60
N (Learners)	459 749	459 749	459 749	459 749	459 273	459 057	459 749	459 057	459 057
N (Schools)	5 072	5 072	5 072	5 072	5 072	5 072	5 072	5 072	5 072
School fixed effect									Y

Source: Quintile 1 subset of DDD longitudinal dataset (all learners in public ordinary Quintile 1 schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Robust standard errors clustered at the school level. Significant at *10% level, **5% level, ***1% level.

Estimating EFAL results: by province

Table 10 presents the estimation results of Grade 4 EFAL by province. The impact of HL3 on EFAL4 differs across provinces, from a predicted 0.40 unit increase in EFAL4 per one unit increase in HL3 in KZN, to 0.49 in North West. There is a pro-female gender gap in EFAL4 results in all provinces, with females predicted to score between 3.7 percentage points higher in North West, and 4.2 percentage points higher in KZN. Overage learners are expected to score between 0.78 (Limpopo) to 1.34 (KZN) percentage points lower than learners who are not overage. The impact of learning in a HL subject that is not a learner's mother tongue is only significant in the Eastern Cape, where it reduces the predicted EFAL4 result by 0.5 percentage points. The differing values of R-squared across provinces shows that the model predicts differing levels of the variation in the data across the provinces, from 54% in the Eastern Cape to 63% in Limpopo.

Table 10: Estimating Grade 4 EFAL results by province

Grade 4 EFAL Result	EC	GT	KZN	LP	MP	NW
Grade 3 HL result	0.405***	0.466***	0.398***	0.478***	0.463***	0.494***
Female	3.831***	3.875***	4.234***	3.804***	4.154***	3.719***
Overage in Grade 4	-1.089***	-0.861***	-1.345***	-0.784***	-1.176***	-0.982***
HL3 differs from actual HL	-0.517**	0.040	-0.335	0.184	0.069	0.193
Grade 1 Maths result	0.307***	0.282***	0.328***	0.330***	0.289***	0.305***
Math delta	0.215***	0.178***	0.224***	0.208***	0.191***	0.182***
Adjusted R-squared	0.54	0.60	0.59	0.63	0.61	0.62
N (Learners)	276 763	174 208	293 288	327 094	178 466	137 417
N (Schools)	3 958	733	3 497	2 278	991	917
School fixed effect	Y	Y	Y	Y	Y	Y

Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Robust standard errors clustered at the school level. Significant at *10% level, **5% level, ***1% level.

Estimating EFAL results: by school quintile

Table 11 presents the estimation results by school quintile. The impact of HL3 on EFAL4 increases across the quintiles, from 0.4 to 0.5 per unit increase. The pro-female impact differs slightly across provinces, from a boost of 3.7 percentage points in Quintile 5 schools, to a 4.0 percentage point advantage. Overage learners are most heavily penalised in Quintile 4 schools (EFAL4 result 1.4 percentage points lower), and the least in Quintile 5 schools (EFAL4 result 0.6 percentage points lower). When provinces are grouped in this manner, the impact of learning in a different HL is insignificant in all quintiles. Finally, a greater proportion of the variation is explained in Quintile 5 schools (64%) compared to the other four quintiles (59-61%). This is consistent with previous findings that there is a larger stochastic component in SBA data in poorer schools (Lam et al., 2011).

Table 11: Estimating Grade 4 EFAL results (by school quintile)

Grade 4 EFAL Result	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Grade 3 HL result	0.437***	0.448***	0.452***	0.453***	0.508***
Female	4.032***	3.892***	3.897***	3.958***	3.696***
Overage in Grade 4	-1.027***	-1.036***	-1.041***	-1.396***	-0.610**
HL3 differs from actual HL	-0.054	0.004	0.143	0.197	0.228
Grade 1 Maths result	0.316***	0.316***	0.307***	0.284***	0.289***
Math delta	0.211***	0.207***	0.198***	0.182***	0.183***
Adjusted R-squared	0.60	0.60	0.59	0.61	0.64
N (Learners)	459 057	420 025	422 699	71 665	12 337
N (Schools)	5 072	4 022	2 887	329	50
School fixed effect	Y	Y	Y	Y	Y

Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023). Robust standard errors clustered at the school level. Significant at *10% level, **5% level, ***1% level.

Discussion

Grade 3 Home Language results are predictive of Grade 4 repetition: learners with higher home language results in Grade 3 are less likely to repeat Grade 4. The relationship between HL3 results and repetition is roughly quadratic (on the decreasing portion of a convex quadratic), with diminishing returns at the top end of the HL3 results distribution. In fact, more than 98% of learners who achieve 75% or higher in their Home Language passed Grade 4 on the first attempt. This result supports established literature that home language mastery in the early years is important for learners' later outcomes. There was notable heterogeneity in outcomes along the lines of gender and age status, with females approximately 3 percentage points less likely to repeat, and overage learners 2-4 percentage points more likely to repeat. Learners in the North West are the most likely to repeat for any HL3, suggesting specific difficulty with the transition to Grade 4 in this province. Gender heterogeneity in predicted repetition is greatest the North West, and smallest in Gauteng. Quintile 5 learners are least likely to repeat at every given HL3 level, and Quintile 1 learners most likely. Thus, the difference in Grade 4 repetition rates across quintiles is not driven exclusively by differences in HL3 results, with learners in lower quintiles having greater difficulty with the transition to Grade 4. Gender has the largest impact on repetition in Quintile 1 schools, and the least impact in Quintile 5 schools.

Grade 4 EFAL results are positively related to Grade 3 Home Language results, with a consistent linear relationship throughout the distribution of Grade 3 Home Language results. Each one unit increase in HL3 results is associated with a 0.4-0.5 unit increase in EFAL4, controlling for other factors. This estimate is remarkably similar to the estimate of 0.45 found by Usborne et al. (2009) in a similarly designed study (albeit in a vastly different context). As with repetition, there was a pro-female boost of 3-4 percentage points, while overage learners were expected to score about one percentage point lower than their correct age-for-grade (or underage) peers (controlling for HL3 and other factors). The impact of HL3 on Grade 4 EFAL results varied across provinces, from a 0.40 (KZN) to 0.49 (North West) unit increase in EFAL4, per unit increase in HL3. Quintile 5 learners perform the best in EFAL4 across all levels of HL3: a learner who scores 50% HL3 would be expected to score 3.5 percentage points more in EFAL4 if they attend a Quintile 5 versus Quintile 1 school. Female learners are most advantaged in Quintile 1 schools, and least advantaged in Quintile 5 schools.

Of interest is the impact of learning in a language that is not a learner's mother tongue. The results suggest that learning in a different (African) home language subject does not, on average, have a significant impact on either repetition or EFAL results (controlling for HL3 and other factors). The exception to this was the Eastern Cape, where having a different HL subject (than mother tongue) was associated with slightly higher repetition and slightly lower EFAL results. The Eastern Cape also has the lowest proportion of learners whose HL subject differs from their mother tongue (at 1% of all learners in the sample), thus suggesting that the isolation of learners who are not learning in their mother tongue may be important for outcomes of these learners. However, this coefficient is likely to be biased downwards, as much of the impact of mother

tongue language status may already be captured in HL3. These results do not negate the finding that learning in a different African HL has a negative effect on outcomes (Taylor and von Fintel, 2016), but they do suggest that the effect may be small.

This correlational study contributes to existing literature on linguistic interdependence and the importance of learning in one's mother tongue by showing that home language mastery is an important predictor of Grade 4 repetition and EFAL results. The results hold across six provinces in South Africa (Eastern Cape, Gauteng, KZN, Limpopo, Mpumalanga and North West) and across school quintiles. The sample is slightly biased towards stronger learners; therefore, the estimates may not be representative of the very weakest learners. In addition, the results suffer from omitted variable bias insofar as many unobserved learner-level factors are likely to impact both Grade 3 Home Language results and Grade 4 repetition (and EFAL results). This would bias the estimated impact of HL3 upwards. Nonetheless, the results indicate that Grade 3 Home Language mastery is an important predictor of Grade 4 repetition and EFAL results and emphasise the importance of mother tongue literacy in the early grades.

References

Ball, J. (2011). Enhancing learning of children from diverse language backgrounds: Mother tongue-based bilingual or multilingual education in the early years. UNESCO.

Benson, C. J. (2002). Real and potential benefits of bilingual programmes in developing countries. International Journal of Bilingual Education and Bilingualism, 5 (6), 303-317.

Böhmer, B. (Forthcoming). Working paper. Stellenbosch, Research on Socio-Economic Policy, Stellenbosch University.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. Review of Educational Research, 49 (2), 222-251.

Cummins, J. (1998). Immersion education for the millennium: What have we learned from 30 years of research on second language immersion? *In:* Bostwick, M. R. C. R. M. (ed.) Learning through two languages: Research and practice. Second katoh gakuen international symposium on immersion and bilingual education. Japan, Katoh Gakuen.

De Galbert, P. G. (2023). Language transfer theory and its policy implications: Exploring interdependence between Luganda, Runyankole-Rukiga, and English in Uganda. Journal of Multilingual and Multicultural Development, 44 (1), 1-19.

Department of Basic Education (2011). National policy pertaining to the programme and promotion requirements of the national curriculum statement grades R - 12. Pretoria, Department of Basic Education.

Department of Basic Education (2023). Grade promotion, repetition and dropping out 2018 to 2021. Pretoria, Department of Basic Education.

Government of South Africa (2002). Education Laws Amendment Act, 2002. . Government of South Africa.

Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. The Econometrics Journal, 7 (1), 98-119.

Hoadley, U. (2023). Covid-19 and the South African curriculum policy response. Stellenbosch, Research on Socio-Economic Policy, Stellenbosch University.

Humble, S., Dixon, P., Gittins, L. & Counihan, C. (2024). An investigation of the cross-language transfer of reading skills: Evidence from a study in Nigerian government primary schools. Education Sciences, 14 (3), 274.

Kim, Y.-S. G. & Piper, B. (2019). Cross-language transfer of reading skills: An empirical investigation of bidirectionality and the influence of instructional environments. Reading and Writing, 32 (4), 839-871.

Lam, D., Ardington, C. & Leibbrandt, M. (2011). Schooling as a lottery: Racial differences in school advancement in urban South Africa. Journal of Development Economics, 95 (2), 121-136.

Melby-Lervåg, M. & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. Journal of research in reading, 34 (1), 114-135.

Mohohlwane, N., Taylor, S., Cilliers, J. & Fleisch, B. (2023). Reading skills transfer best from home language to a second language: Policy lessons from two field experiments in South Africa. Washington, Center for Global Development.

Statististics South Africa (2024). Census 2022 in brief. Pretoria, StatsSA.

Taylor, S. & Von Fintel, M. (2016). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach. Economics of Education Review, 50, 75-89.

Usborne, E., Caouette, J., Qumaaluk, Q. & Taylor, D. M. (2009). Bilingual education in an aboriginal context: Examining the transfer of language skills from inuktitut to English or French. International Journal of Bilingual Education and Bilingualism, 12 (6), 667-684.

Van Der Berg, S., Gustafsson, M. & Burger, C. (2020). School teacher supply and demand in South Africa in 2019 and beyond. DHET, Resep.

Van Der Berg, S., Wills, G., Selkirk, R., Adams, C. & Van Wyk, C. (2019). The cost of repetition in South Africa. Stellenbosch, Research on Socio-Economic Policy, Stellenbosch University.

Vaughn, S., Cirino, P. T., Linan-Thompson, S., Mathes, P. G., Carlson, C. D., Hagan, E. C., Pollard-Durodola, S. D., Fletcher, J. M. & Francis, D. J. (2006). Effectiveness of a Spanish intervention and an English intervention

for English-language learners at risk for reading problems. American Educational Research Journal, 43 (3), 449-487.
Wawire, B. & Kim, Y.-S. (2018). Cross-language transfer of phonological awareness and letter knowledge: Causal evidence and nature of transfer. Scientific Studies of Reading, 22 (6), 443-461.
Wills, G. (2023). Early grade repetition in South Africa: Implications for reading. Stellenbosch, Research on Socio-Economic Policy, Stellenbosch University.
Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. Cambridge, Massachusetts, The MIT Press.
Wooldridge, J. M. (2013). Introductory econometrics: A modern approach. Electronic copy, South-Western Cengage Learning.

Appendix

Table A 1: Completeness of the (unbalanced) DDD dataset

DDD dataset completeness							
	EC	**GT**	**KZN**	**LP**	**MP**	**NW**	**Total**
Number of learners in Grades 1-4 in Public Ordinary Schools in the DDD dataset after basic cleaning (unbalanced)							
2017	603 771	565 851	532 480	528 001	348 910	286 049	2 865 062
2018	613 338	706 125	675 663	547 342	354 212	276 121	3 172 801
2019	603 161	739 608	823 903	551 243	352 667	286 219	3 356 801
2020	583 945	676 572	741 722	542 449	343 634	271 535	3 159 857
2021	573 738	730 727	746 257	550 283	348 556	267 388	3 216 949
2022	552 193	727 916	771 324	534 378	345 323	273 701	3 204 835
2023	528 607	674 811	756 323	519 940	336 504	249 253	3 065 438
Mean	**579 822**	**688 801**	**721 096**	**539 091**	**347 115**	**272 895**	**3 148 820**
Above counts as % of comparable actual learner numbers (School Realities Reports)							
2017	98	67	56	93	94	96	78
2018	97	92	73	99	100	93	90
2019	98	96	94	99	98	97	97
2020	97	88	82	97	103	101	92
2021	98	95	84	99	98	95	94
2022	98	95	89	99	98	98	95
2023	98	88	89	99	97	91	93
Mean	**97**	**88**	**80**	**98**	**98**	**96**	**91**

Sources: DDD data (all learners in public ordinary schools) and DBE School Realities Reports.

Table A 2: Percentage of Grade 3 African mother tongue learners with English as Home Language subject in 2019

	EC	GT	KZN	LP	MP	NW	Total
Number of Grade 3 learners with African Home Language (2019)							
IsiNdebele	10	1 445	17	1 089	6 930	226	9 717
IsiXhosa	115 965	7 036	6 163	156	445	2 126	131 891
IsiZulu	1 474	30 643	153 908	815	19 266	708	206 814
SePedi	20	16 423	15	76 701	8 881	1 232	103 272
SeSotho	2 797	14 948	302	698	873	2 496	22 114
SeTswana	50	11 723	16	2 263	1 266	49 557	64 875
SiSwati	125	470	74	173	25 008	101	25 951
TshiVenda	3	1 509	13	20 745	67	207	22 544
XiTsonga	25	6 257	26	23 843	8 633	1 600	40 384
Total	**120 469**	**90 454**	**160 534**	**126 483**	**71 369**	**58 253**	**627 562**
% of the above with English as HL subject							
IsiNdebele	38	35	60	5	7	1	12
IsiXhosa	7	42	15	16	43	15	11
IsiZulu	5	35	14	8	19	11	18
SePedi	29	28	53	2	7	8	8
SeSotho	4	38	27	18	32	17	32
SeTswana	21	39	64	10	6	9	16
SiSwati	12	56	30	9	7	10	9
TshiVenda	57	62	96	4	50	12	14
XiTsonga	17	34	43	2	4	4	9
Mean	**7**	**36**	**14**	**3**	**11**	**9**	**14**

Source: DDD data (learners in public ordinary schools with an African language as their mother tongue).

Table A 3: Counts and proportions of learners with valid absenteeism data

	EC	GT	KZN	LP	MP	NW
Number of learners recorded as absent on at least 1 day in Grade 3 (2019)						
N	50 868	41 716	53 026	58 204	37 447	32 200
The above as a percent of the total learners in Grade 3 (2019)						
%	67	94	74	63	76	78

Source: Absentee subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 in 2017 and who reached Grade 4 in 2020).

Table A 4: Correlations between key variables

	G4 rep	G4 EFAL result	G3 HL result	Female	Overage (G4)	HL subj. ≠ L1	G1 rep	G2 rep	G3 rep	Condoned G3	G1 Math result	FP math delta	Days absent (G3)
G4 rep	1	.	.	.	.	.	.	.	.	.	.	.	.
G4 EFAL result	-.49	1	.	.	.	.	.	.	.	.	.	.	.
G3 HL result	-.39	.66	1	.	.	.	.	.	.	.	.	.	.
Female	-.15	.30	.27	1	.	.	.	.	.	.	.	.	.
Overage (G4)	.20	-.30	-.32	-.18	1	.	.	.	.	.	.	.	.
HL subj. ≠ L1	.00	.02	.00	.00	.00	1	.	.	.	.	.	.	.
G1 rep	.19	-.22	-.24	-.11	.49	-.01	1	.	.	.	.	.	.
G2 rep	.08	-.13	-.14	-.09	.37	-.01	-.04	1	.	.	.	.	.
G3 rep	.07	-.11	-.14	-.08	.31	.00	-.03	-.03	1	.	.	.	.
Condoned G3	.35	-.35	-.56	-.13	.27	.01	.22	.10	.11	1	.	.	.
G1 Math result	-.27	.45	.50	.17	-.47	.02	-.63	-.18	-.11	-.28	1	.	.
FP math delta	-.03	.07	.21	-.01	.24	-.02	.46	.09	.03	-.13	-.61	1	.
Days absent (G3)	.07	-.08	-.12	-.04	.11	.05	.07	.05	.05	.09	-.09	-.01	1

Table A 5: Grade 4 repetition and Grade 3 Home Language result bin

	EC	GT	KZN	LP	MP	NW
Grade 4 repetition rate (%)						
0-49%	42.4	47.5	41.8	60.6	35.4	62.5
50-74%	10.4	6.5	8.6	12.7	8.1	15.2
75-100%	1	0.3	0.7	0.8	0.6	1.2
Learner counts and proportions						
0-49%	13919 (5%)	13822 (7.9%)	23135 (7.9%)	15670 (4.8%)	13243 (7.4%)	9962 (7.2%)
50-74%	187472 (67.1%)	97046 (55.5%)	158838 (54%)	174642 (53.3%)	91123 (50.8%)	73083 (52.9%)
75-100%	78147 (28%)	63932 (36.6%)	112098 (38.1%)	137251 (41.9%)	74986 (41.8%)	54992 (39.8%)

Source: DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 between 2017 and 2019 and who reached Grade 4 by 2023).

Table A 6: Repetition model selection (absenteeism sample)

Grade 4 Repetition	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Grade 3 HL result	-0.0067***	-0.0272***	-0.0266***	-0.0266***	-0.0262***	-0.0286***	-0.0286***	-0.0268***
Square of Grade 3 HL result		0.0002***	0.0002***	0.0002***	0.0002***	0.0002***	0.0002***	0.0002***
Female			-0.0417***	-0.0417***	-0.0408***	-0.0355***	-0.0355***	-0.0305***
Overage in Grade 4			0.0276***	0.0276***	0.0148***	0.0158***	0.0158***	0.0125***
HL3 differs from actual HL				0.0411***	0.0420***	0.0527***	0.0527***	0.0252**
Repeated Grade 1					0.0421***	0.0866***	0.0866***	0.0788***
Number of days absent in Grade 3						0.0014***	0.0014***	0.0017***
Fee-paying school							0.1924*	
Constant	0.5398***	1.1811***	1.1473***	1.1467***	1.1300***	1.2045***	1.2046***	
Adjusted R-squared	0.10	0.13	0.14	0.14	0.14	0.17	0.17	0.26
Percentage correctly predicted (overall)	90.6	90.8	90.8	90.8	90.8	89.8	89.8	90.8
Percentage correctly predicted (repeaters)	0	7.4	7.6	7.7	7.6	13.4	13.4	24.3
Percentage correctly predicted (passers)	100	99.4	99.4	99.4	99.4	99.1	99.1	98.9
N (Learners)	279 539	279 539	279 539	279 539	279 539	65 070	65 070	65 070
N (Schools)	3 961	3 961	3 961	3 961	3 961	3 619	3 619	3 619
School fixed effect								Y

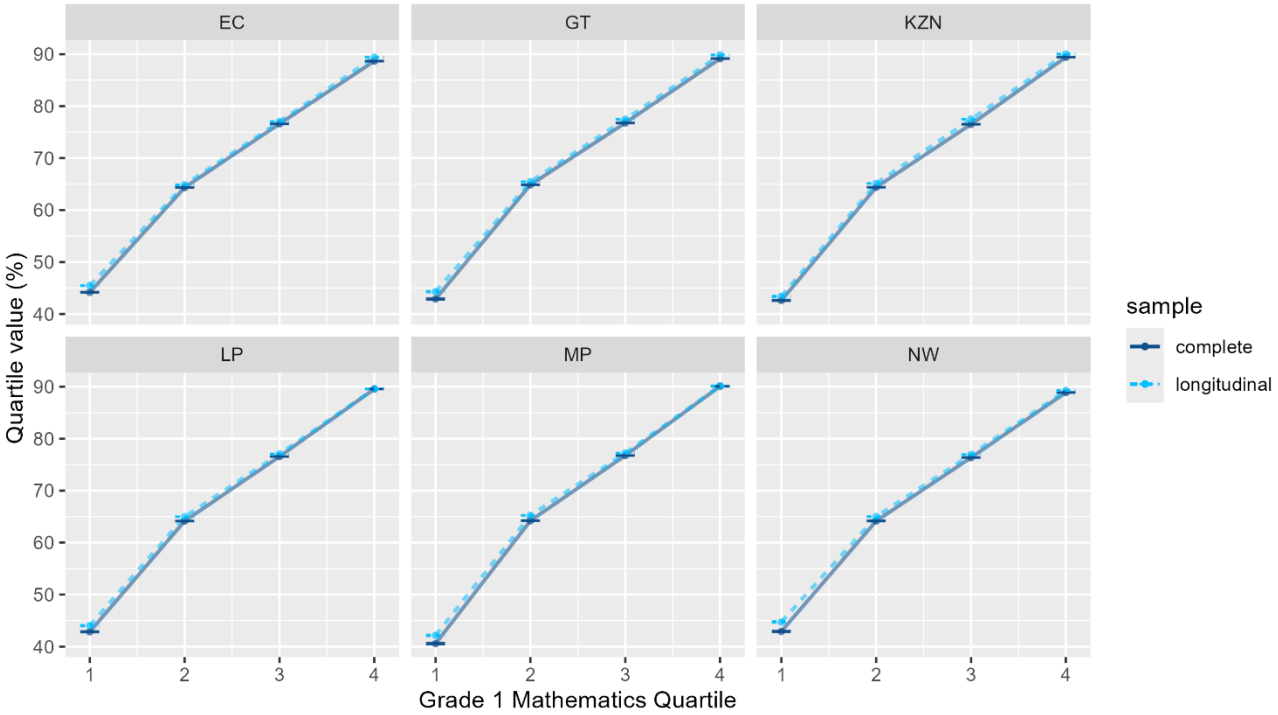
Source: Absentee subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 in 2017 and who reached Grade 4 in 2020). Robust standard errors clustered at the school level. Significant at *10% level, **5% level, ***1% level.

Table A 7: Estimating repetition in the absenteeism sample (by province)

Grade 4 Repetition	EC	GT	KZN	LP	MP	NW
Grade 3 HL result	-0.0268***	-0.0302***	-0.0255***	-0.0304***	-0.0223***	-0.0306***
Square of Grade 3 HL result	0.0002***	0.0002***	0.0001***	0.0002***	0.0001***	0.0002***
Female	-0.0305***	-0.0190***	-0.0251***	-0.0341***	-0.0254***	-0.0365***
Overage in Grade 4	0.0125***	0.0066**	-0.0004	0.0292***	0.0007	0.0281***
HL3 differs from actual HL	0.0252**	-0.0006	0.0052	0.0023	0.0087*	0.0005
Repeated Grade 1	0.0788***	0.1016***	0.1224***	0.1145***	0.1690***	0.1463***
Number of days absent in Grade 3	0.0017***	0.0011***	0.0017***	0.0022***	0.0018***	0.0030***
Adjusted R-squared	0.26	0.36	0.33	0.34	0.30	0.38
Percentage correctly predicted (overall)	90.8	92.6	91.5	90	92	88.8
Percentage correctly predicted (repeaters)	24.3	37.8	35.4	33.4	28.1	41.7
Percentage correctly predicted (passers)	98.9	98.6	98.6	98.6	98.9	97.9
N (Learners)	65 070	52 315	73 260	69 247	45 366	39 038
N (Schools)	3 619	725	3 096	2 232	983	906
School fixed effect	Y	Y	Y	Y	Y	Y

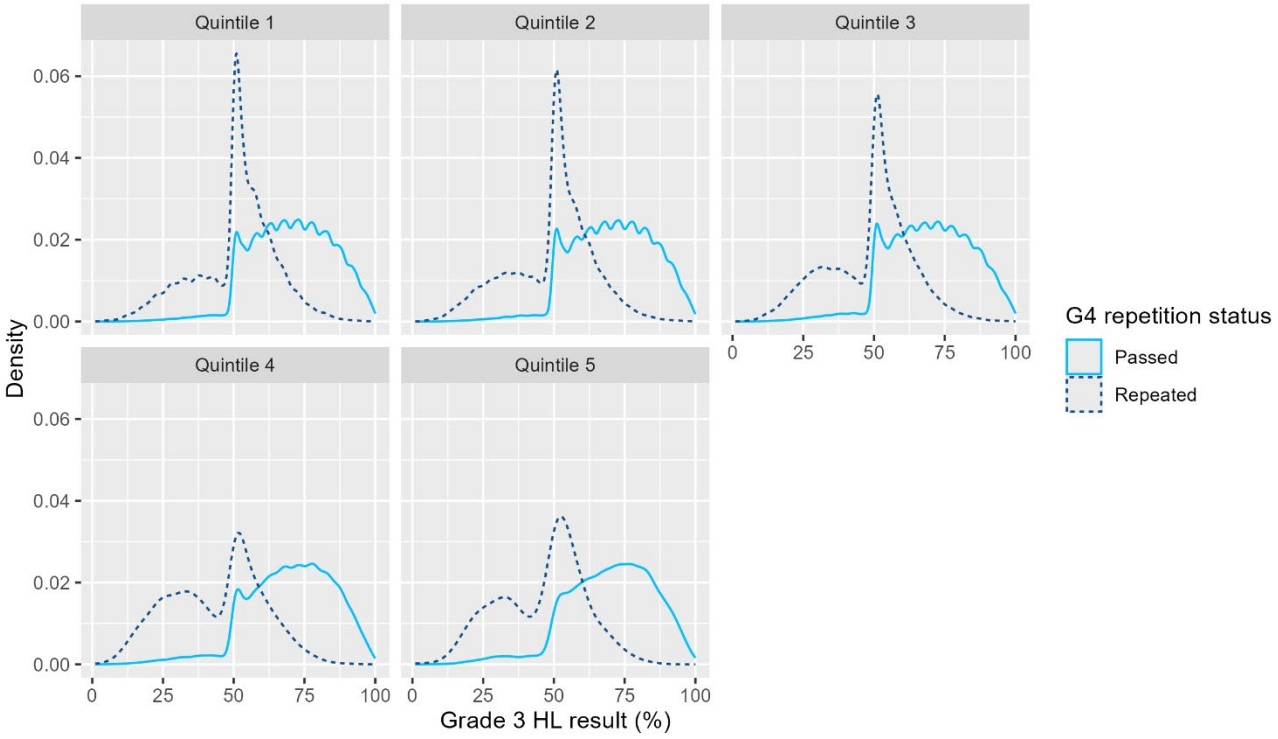
Source: Absentee subset of DDD longitudinal dataset (all learners in public ordinary schools with an African Home Language subject, who started Grade 1 in 2017 and who reached Grade 4 in 2020). Robust standard errors clustered at the school level. Significant at *10% level, **5% level, ***1% level.

Figure A 1: Grade 1 Mathematics Results in the African Home Language dataset, versus longitudinal sample



Source: DDD data (learners in public ordinary schools with an African Home Language subject). Quartiles by province and sample. Error bars show a 95% confidence interval.

Figure A 2: Grade 3 Home Language result distributions by Grade 4 repetition status and quintile



Source: DDD data.