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.
06/25

NOVEMBER 2025

The impact of early grade repetition on test scores: Evidence from a regression discontinuity design in South Africa

Keywords

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

JEL: I21, I25, C21, O15

Author

Ros Clayton

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

OPTIMA

The impact of early grade repetition on test scores: Evidence from a regression discontinuity design in South Africa

Ros Clayton

Abstract

This paper provides the first causal evidence on the effects of early grade repetition in South Africa. Using a large administrative dataset covering six provinces from 2017–2023, I implement a fuzzy regression discontinuity (RD) design exploiting published promotion thresholds to estimate the impact of repeating Grade 1 or Grade 4 on subsequent test scores. Grade 1 repetition raises achievement in Home Language, Mathematics, and English First Additional Language by over one standard deviation in the following grade, with effects diminishing but remaining sizeable until at least three grades after the repetition. Grade 4 repetition yields smaller initial gains, which fade less over time. The initial effects of Grade 1 repetition are larger than those reported in the most comparable RD studies, reflecting rapid cognitive development in early primary years, while the Grade 4 effects are large but consistent with international evidence. The findings indicate that repetition can be an effective remedial tool even in settings with limited structured support for repeaters, especially in the context of literacy deficits amongst English learners. The paper contributes to the literature on grade repetition, skill formation, and education policy in middle-income countries, providing evidence relevant for the design of promotion rules and remedial strategies.

JEL Classification: I21, I25, C21, O15

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

1 INTRODUCTION

Human capital is widely recognised as a key driver of economic growth, with evidence showing that improvements in educational outcomes can foster economic development (Hanushek & Woessmann, 2012). In South Africa, where inequality remains entrenched, economic growth sluggish (OECD, 2025), and the education system both weak and unequal (Department of Basic Education, 2024a; von Davier et al., 2024), enhancing learning outcomes is critical. In the face of high unemployment, returns to advanced education are high and rising (Köhler, 2024), and the challenge lies in ensuring that learners from all socioeconomic backgrounds are able to reach their full potential.

Heckman's (2006) model of skill formation emphasises that early learning begets later learning, underscoring the need to assess interventions and practices that shape early educational trajectories – including grade repetition. Although some studies suggest that early repetition may yield long-term benefits, such as higher future earnings (Eide & Showalter, 2001), the broader literature presents mixed evidence, with effects varying across contexts, grades, and research designs (Valbuena et al., 2021). In South Africa, where repetition rates are high relative to neighbouring countries (Wills, 2023), and where grade repetition is the primary remediation mechanism, understanding the consequences of this contentious practice is essential to advancing both educational quality and equity.

Earlier international research, particularly before the 2000s, generally associated repetition with adverse outcomes – such as lower academic achievement, diminished self-esteem, increased absenteeism, and higher dropout rates (Jimerson, 2001). However, these findings are often complicated by negative selection into repetition: pupils who repeat a grade tend to differ systematically from their promoted peers in unobserved attributes such as cognitive ability, motivation, and home support, making causal inference difficult. Recent quasi-experimental research applying regression discontinuity (RD) designs suggests that repetition often yields academic gains, especially when the repetition is undertaken in the early primary years and the analysis compares outcomes for repeaters when they reach the same grade as the non-repeaters (Greene & Winters, 2007; Winters & Greene, 2012; Mariano & Martorell, 2013; Schwerdt et al., 2017; Figlio & Özek, 2020; Hwang & Koedel, 2023; Quintero, 2025).

The purpose of this study is to determine the causal impact of early grade repetition on test scores in South Africa by addressing the following research questions:

1. Does repetition in Grade 1 or Grade 4 improve learner results in the following grade (Grade 2 or 5, respectively)?
2. Does the impact of repetition fade out in the three grades following the initial repetition?

3. Is earlier repetition (Grade 1) associated with better outcomes than later repetition (Grade 4)?

I use a comprehensive administrative dataset containing learner school marks across six provinces in South Africa between 2017 and 2023 and construct two longitudinal panels[1] that track individual learners over time. I exploit the existence of published test score thresholds for grade promotion (Department of Basic Education, 2011b) and implement a fuzzy RD design (Hahn et al., 2001) to estimate the local average treatment effect – specifically, the effect of repetition on repeaters whose marks fall just below the promotion thresholds.

The analysis indicates that Grade 1 repetition leads to large improvements in subsequent achievement, though these gains diminish over time. In Home Language, repeaters score 1.1 standard deviations higher in Grade 2, 0.6 standard deviations higher in Grade 3, and 0.3 standard deviations higher in Grade 4, with similar effects in Mathematics and English First Additional Language. Grade 4 repetition produces more modest initial benefits that are more sustained over time: marginal repeaters outperform their marginally promoted peers by 0.6 standard deviations in Grade 5, 0.5 standard deviations in Grade 6, and 0.4 standard deviations in Grade 7, again with comparable patterns across subjects.

The effects of repetition found in this study are much larger than those found in other effective remedial interventions in South Africa (Wills, 2025). However, the costs, both direct and indirect, may be higher. Policymakers should weigh the costs and benefits of repetition against alternative language remediation strategies, including structured pedagogy programs (Stern et al., 2024). Complementary language support for repeaters may also be worthwhile given evidence that such support enhances outcomes (Valbuena et al., 2021). These additional supports in literacy may be especially effective for English learners, as demonstrated by the large positive effects of repetition found in studies that focus on this subpopulation (Figlio & Özek, 2020; Quintero, 2025).

This study makes several novel contributions to the growing set of causal studies estimating the impact of grade repetition. It is, to my knowledge, the first to use an RD design to estimate the impact of Grade 1 repetition in any country, and the first to apply an RD design to estimate the impact of early grade repetition in a middle-income country context. It is also the first RD study which estimates the impact of early grade repetition in a context where no systematic support is provided to repeaters; an important contribution, given concerns that a large portion of the reported effects of repetition in the United States may be due to other factors besides the additional year of schooling (Berne et al., 2025).

[1] One panel for estimating the effect of Grade 1 repetition, and one for Grade 4 repetition.

The next section outlines the educational context in South Africa and reviews methodological considerations for RD designs, alongside international evidence on the effects of grade repetition on test scores. Section 3 describes the data used in this study, while Section 4 details the empirical approach, including strategies for addressing potential threats to the assumptions in an RD design. Section 5 presents descriptive statistics, and Section 6 reports the estimated treatment effects. Finally, Section 7 discusses the results, highlights the study's limitations, and offers recommendations for policymakers as well as directions for future research.

2 BACKGROUND

2.1 South African context

2.1.1 General education context

South Africa has seen substantial improvements in primary and secondary education outcomes in recent decades, with notable gains in international assessment performance between 2006 and 2016, albeit from a low base (van der Berg & Gustafsson, 2019; van Staden & Gustafsson, 2022). More young people than ever are attaining a matric qualification (Wills et al., 2024). However, despite these improvements, South Africa's performance prior to the COVID-19 pandemic remained below that of comparable middle-income countries (Nyamunda, 2024), and the pandemic itself caused substantial short-term learning losses (Ardington et al., 2021; van der Berg et al., 2022). These losses have persisted in Grades 4 and 5 in both language and mathematics, although Grade 9 mathematics performance has since recovered and now exceeds pre-pandemic levels in the Trends in International Mathematics and Science Study (TIMSS) (Department of Basic Education, 2023; Department of Basic Education, 2024b).

South African public schools are categorised into five (unequally sized) quintiles, which determine the extent of government per-learner funding (South Africa, 2006). Quintile 1 schools (the poorest) receive the most funding, while Quintile 5 schools (the wealthiest) receive the least; however, this does little to mitigate the entrenched inequalities in infrastructure that accumulated over decades of unequal apartheid-era spending (Adams, 2020). Quintile 4 and 5 schools supplement government funding by charging fees. Overall, education inequality remains high, although roughly in line with expectations for a country at South Africa's income level (van der Berg & Gustafsson, 2019). Of particular concern is the persistence of inequality along racial lines (van der Berg & Gustafsson, 2019).

In recent years, the Department of Basic Education has increased its focus on foundational literacy and numeracy (Department of Basic Education, 2025), and there have been numerous governmental and non-governmental initiatives aimed at supporting learners in achieving

essential early learning outcomes, particularly in literacy (Kika et al., 2022; Wills, 2025). Effect sizes for these interventions can be substantial. For instance, the Early Grade Reading Study, a two-year language intervention, improved literacy by up to 0.24 standard deviations in the short term, with effects fading to a still significant 0.15 standard deviations four years after the study concluded (Taylor et al., 2018; Stern et al., 2024). The Funda Wande program, which provides early learning resources and teaching assistants, reported improvements of up to 0.56 standard deviations in mathematics and 0.26 in literacy (Ardington, 2024). In the Western Cape, the Back on Track Programme, a COVID-19 recovery initiative, improved language outcomes among Grade 7 isiXhosa learners by up to 0.41 standard deviations (van der Berg et al., 2025).

These studies represent some of the largest effect sizes recorded in well-identified early learning interventions in South Africa (Wills, 2025), and are large in comparison to the median effect size from educational interventions of 0.1 standard deviations (Evans & Yuan, 2022). While the aforementioned interventions have collectively impacted over 200 000 learners[2], they are presently unavailable to most learners and grade repetition remains the primary remediation strategy employed in South African public schools.

2.1.2 Repetition in South Africa

Grade repetition is widely employed as a remediation strategy in South Africa, designed to give learners additional time to master the required curriculum content (Department of Basic Education, 2011b), and, potentially, to incentivise greater effort among learners seeking to avoid repetition. The Department of Basic Education (2011b) specifies minimum achievement levels for Grades 1 to 11 that learners must meet to be promoted to the next grade. Failure to meet these guidelines should result in the learner being retained.

The South African schooling system is divided into phases, with the Foundation Phase (Grade R to Grade 3) and the Intermediate Phase (Grade 4 to Grade 6) being most relevant to this study. Policy stipulates that learners may repeat a grade at most once per phase, and teachers are instructed to ensure that learners "receive the necessary support in order to progress to the next grade" (Department of Basic Education, 2011b). However, the structures and mechanisms to provide such support are not specified, and there are no formal support systems for repeating learners.

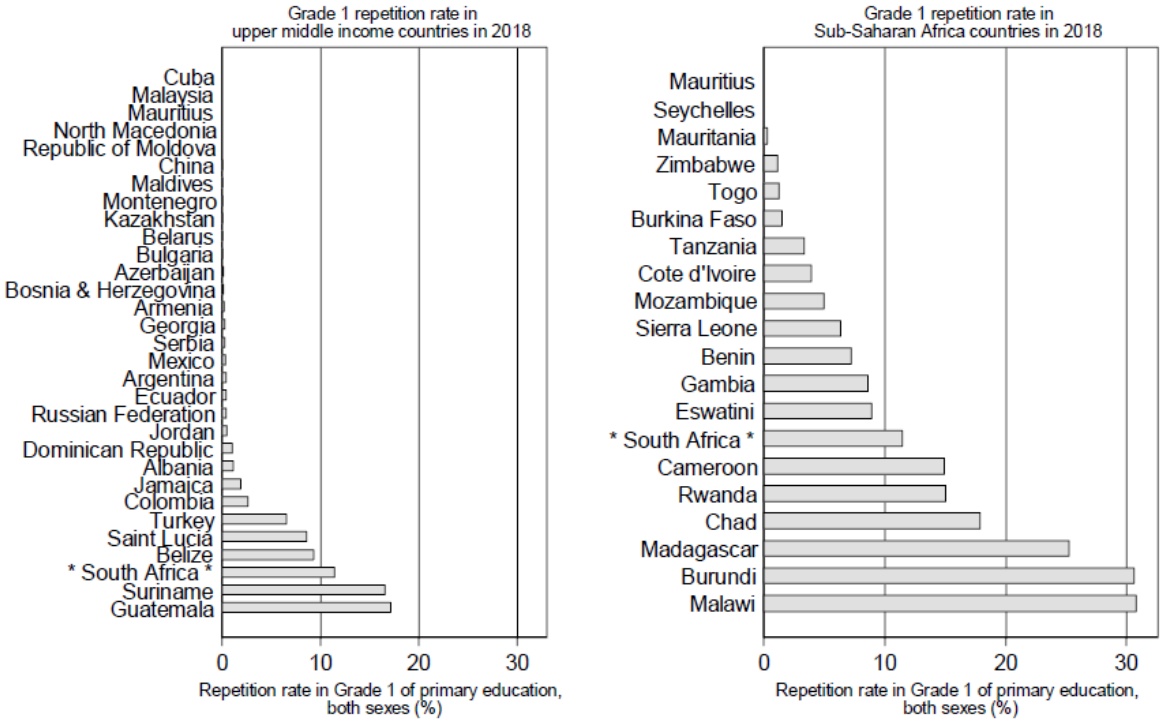
Figure 1 shows Grade 1 repetition rates for select countries including South Africa. Repetition rates in South Africa are high in relation to both other upper-middle income countries and neighbouring countries (Wills, 2023). Repetition rates decreased during the COVID-19 pandemic

[2] Funda Wande has more than 180 000 learners enrolled in program schools in the Western Cape alone (Funda Wande, 2023).

due to increased leniency, but are returning to pre-pandemic levels (van der Berg et al., 2023) which were estimated at 12% for Grade 1 and 11% for Grade 4 in 2019 (Gustafsson, 2023). Due to the policy of at most one repeat per phase, in conjunction with Grade 1 and Grade 4 being the start of the Foundation and Intermediate Phase respectively[3], these grades have the highest repetition rates in their respective phases, with the majority of learners repeating at the first opportunity and then possibly being progressed through later grades[4] in that phase before the next repetition opportunity.

There is a significant pro-female gender gap throughout the education system in South Africa (Spaull & Makaluza, 2019) and this is borne out in repetition outcomes, with males more likely to repeat than females (Branson & Lam, 2010; van der Berg et al., 2019). Repetition is almost twice as high in Quintile 1 schools compared to Quintile 5 schools (van der Berg et al., 2019). Repetition rates also differ by race, parental education, and household income (Branson & Lam, 2010).

Figure 1. Grade 1 repetition rates in selected countries



Source: Wills (2023).

2.1.3 Measuring the impact of repetition in South Africa

To my knowledge, there have been no causal studies estimating the impact of repetition in South Africa. However, Wills (2023) used panel data on Grade 1 to 4 Setswana reading proficiency to estimate the relationship between early grade repetition and reading scores. The observational

[3] Although Grade R marks the beginning of the Foundation Phase, repetition at this level is almost exclusively limited to learners who are underage (Böhmer, 2025).
[4] Although some learners repeat more than once per phase, against recommendations.

analysis indicates that both Grade 1 and later repetition have negative impacts on reading scores. However, when negative selection into Grade 1 repetition is controlled for (using a comparison group of Grade 2 and 3 repeaters, who were also matched on baseline characteristics), the same-grade analysis suggests that Grade 1 repetition has a positive impact on reading outcomes. Across all measures, repeating Grade 2 or 3 was found to be associated with more negative outcomes compared to repeating Grade 1.

In another observational study, this time in the Western Cape province, van der Berg et al. (2019) found that learners who repeated Grade 3 made significant gains on standardised language and mathematics tests in their repeating year (+14 and +21 percentage points respectively). Furthermore, older learners benefitted less from repetition of Grade 3 than their correct-age-for-grade counterparts (but this is possibly a selection effect). Repetition in Grade 9 was associated with almost no improvement in test outcomes, consistent with international evidence that earlier repetition is better than later repetition.

Also studying the associations between grade repetition and subsequent test score outcomes for learners in the Western Cape province, Selkirk (2025) found repetition in Grades 3, 6 and 9 to be associated with lower test score results in subsequent grades, with indications that early repetition is less harmful than later repetition. However, as stated by the author, these results are likely to be biased towards finding negative impacts of repetition due to negative selection into repetition.

An indirect measure of factors that are associated with repetition is the impact of being overage-for-grade. Most South African learners start school at the appropriate age (Böhmer, 2025); consequently, grade repetition is the primary reason learners become increasingly overage-for-grade as they progress through the education system. By age 14, only 58% of children remain in school and in the correct grade (van der Berg et al., 2019). Being overage is a strong predictor of further repetition (Branson & Lam, 2010), dropout in later grades (Branson et al., 2014), and poorer National Senior Certificate (NSC) results among those who do complete Grade 12 (Wills et al., 2024). Moreover, there is evidence of a random component to repetition, which is particularly pronounced in poorer schools (Lam et al., 2011), suggesting that any harmful effects of repetition may be amplified among disadvantaged learners.

2.2 Identifying the causal impact of repetition

The impact of repetition is notoriously difficult to identify due to latent factors, such as motivation, ability and home environment, that impact both repetition and later outcomes. Studies that rely on matching-on-observables are likely to be biased against repetition as the observed controls, however comprehensive, frequently omit the most important factors, thereby biasing the

estimator (Heckman, 1979; Angrist & Pischke, 2009). This is especially true when repetition decisions are at the discretion of the teacher, such as was the case in the United States up to the late 1990s (Jacob & Lefgren, 2004; Manacorda, 2012).

Quasi-experimental approaches, such as instrumental variables (IV) and RD designs, can produce unbiased causal estimates of the treatment effect under credible identifying assumptions. Both approaches estimate a local average treatment effect (LATE) that pertains to the compliers – those individuals whose treatment status is affected by the relevant instrument or discontinuity (Angrist & Imbens, 1995; Angrist et al., 1996). In the case of RD designs, the estimand represents the average treatment effect for those learners whose treatment status is impacted by their position just above or below the cutoff (Hahn et al., 2001; Lee & Lemieux, 2010). The LATE of repetition, as estimated by RD designs, is highly policy-relevant, since it represents the treatment effect at the margin, precisely where policy interventions are typically targeted (Lee, 2008; Angrist & Pischke, 2009).

RD designs can only be implemented when treatment assignment is determined, at least in part, by an explicit rule that creates a discontinuous change in treatment probability. In the education context this rule is typically a test score threshold below which learners are required to repeat a grade. In the context of an RD design, this test score result is called the running variable. The rule does not need to be applied perfectly; it is sufficient that the probability of treatment changes discontinuously at the cutoff. When compliance with the rule is imperfect, the design is referred to as a fuzzy RD, and treatment status becomes a discontinuous but not deterministic function of the running variable (Hahn et al., 2001; Imbens & Lemieux, 2008). Identification of the LATE in a fuzzy RD requires two main assumptions: (i) continuity of potential outcomes at the cutoff, and (ii) monotonicity, meaning that crossing the threshold does not reduce the probability of treatment for any individual (Lee & Lemieux, 2010; Cattaneo et al., 2019).

The second assumption is typically satisfied in the context of grade repetition, as it merely requires that no student becomes less likely to repeat when their test score falls below the promotion threshold. The first assumption – continuity of potential outcomes at the cutoff – cannot be tested directly, but it is commonly supported by evidence that there is no precise sorting or manipulation around the threshold. When learners or teachers cannot finely influence the running variable, those just below and just above the cutoff can be considered comparable in both observed and unobserved characteristics, except for their treatment status (Lee, 2008; Lee & Lemieux, 2010).

Manipulation in the running variable can be formally assessed using a density test, initially developed by McCrary (2008), which detects discontinuities in the distribution of the running variable at the cutoff. However, the absence of manipulation is neither a necessary nor a sufficient

condition for the continuity of potential outcomes (Cattaneo & Titiunik, 2022). As an additional diagnostic, researchers often examine discontinuities in predetermined covariates at the cutoff: if baseline characteristics are continuous, this provides supportive (though indirect) evidence for the validity of the continuity assumption (Imbens & Lemieux, 2008; Lee & Lemieux, 2010).

Repetition outcomes can be measured using a same-age approach (comparing learners of the same age in different grades) or a same-grade approach (comparing learners in the same grade but of different ages). Same-age analysis is biased against repeaters, since they would have been exposed to less advanced material than their same-age peers who are a grade ahead after the repetition, and also possibly due to differences in average rates of learning in different grades (Schwerdt et al., 2017).

The same-grade approach is biased in favour of repeaters since it includes maturation effects (as repeaters are one year older in each grade after the repetition), and thus any observed improvements may be the result of maturation rather than the additional year of schooling. However, both approaches have practical relevance, with same-age analyses more closely measuring the learning impact of repetition independently of the maturation effect, and same-grade analyses being more relevant if stakeholders are interested in learning levels after completion of a specific grade (Schwerdt et al., 2017). Of the 42 studies in Valbuena et al.'s (2021) meta-analysis of studies that measure the impact of grade repetition and "control for endogeneity", 34 (81%) use a same-grade approach in at least one analysis.

2.3 International evidence on grade repetition

2.3.1 Early non-causal evidence

Historical reviews of the literature – spanning the 1970s (Jackson, 1975), 1980s (Holmes, 1989), and 1990s (Jimerson, 2001) – consistently concluded that grade repetition adversely affects both academic and socio-emotional outcomes for repeating learners. Jimerson (2001) found the evidence against repetition so compelling that he urged "researchers, educational professionals, and legislators to abandon the debate regarding social promotion and grade retention in favour of a more productive course of action in the new millennium." However, the studies included in these reviews typically controlled for observed characteristics only, leaving estimates vulnerable to bias arising from unobserved differences between repeaters and promoted learners. Moreover, the evidence base reviewed was drawn almost entirely from the United States, where promotion and retention decisions were largely at the discretion of individual teachers, making repetition particularly susceptible to selection on unobservables.

Nevertheless, these reviews provide important insights. Jackson (1975) discusses three experimental studies (from the 1940s) which investigated the impact of grade repetition by

randomly assigning repetition or promotion to sets of matched learners who were experiencing academic difficulty. The studies used a same-age comparison which favours promoted learners; despite this, only one result was significantly in favour of promotion, while slightly more of the non-significant results (22) favoured repetition over promotion (17).

Holmes (1989) grouped the 63 studies in his meta-analysis according to whether they used a same-age or same-grade approach, finding that same-age studies frequenty estimated negative impacts of repetition on academic outcomes, while same-grade studies frequently estimated positive impacts. Repetition in later grades was consistently found to have a more negative (or less positive) impact than earlier repetition. Studies which reported positive results often involved additional remedial support, and the retained students in these studies typically struggled only in a single subject (reading) and were generally more academically able – measured by IQ – than the average retained student. This suggests that repetition may be more effective for more able learners.

Jimerson (2001) provides a review of research on repetition between 1990 and 1999, summarising 20 studies in terms of their conclusion (whether repetition has a positive or negative impact), sample size (frequently below 50), comparison type (same-age or same-grade), outcome type (achievement or socioemotional), retention and outcome grade or age, and a list of controls used for matching (since almost all of the studies in the review used matching-on-observables). However, despite the identification of comparison groups these studies suffered from endogeneity: repetition and these later outcomes of interest are simultaneously determined by unobserved characteristics, and these are not wholly controlled for through post-hoc creation of comparison groups (Jacob & Lefgren, 2004). The author acknowledges the weaknesses in the methodologies of the studies included in the review but nonetheless argues that the "confluence" of results warrants action. However, if the causal effect is poorly identified in all studies, and they all suffer the same direction of bias, then consistency of the results is relatively meaningless in terms of identifying the causal impact of repetition.

A recent review by Valbuena et al. (2021) provides a comprehensive synthesis of the quasi-causal evidence on grade repetition up to 2020. The review includes studies employing RD designs as well as less well identified studies that rely on matching-on-observables. Their synthesis shows that repetition is associated with positive short-term effects on test scores: 68% of studies report positive or strongly positive impacts, 11% find no significant effect, and 21% identify negative effects (own calculations using Table A2 in Valbuena et al. (2021)). However, many of the studies reporting positive results only track learners for a limited number of years, as most do not have data suitable for longer-term analyses. This highlights the need for more long-term evaluations

of grade repetition, a gap in the international literature that remains difficult to address given the data requirements for tracking learners over extended periods.

Of the 42 studies in the review, only three draw on data from low- and middle-income countries (LMICs). In Mexico, Cabrera-Hernandez (2022) finds that abolishing repetition policies for Grades 1 to 3 reduced dropout and had no effect on standardised test scores, challenging the notion of a "threat effect" of repetition. This stands in contrast to evidence from Colombia, where Ferreira Sequeda et al. (2018), using a similar methodology but exploiting the re-introduction of early-grade repetition, show that retention improves reading but not mathematics outcomes. Using panel data from Senegal and exploiting variation in promotion thresholds across schools, Glick and Sahn (2010) find that early-grade repetition increases dropout.

In summary, strong evidence on the effects of grade repetition from LMICs is sparse and mixed, and none of these studies use an RD design.

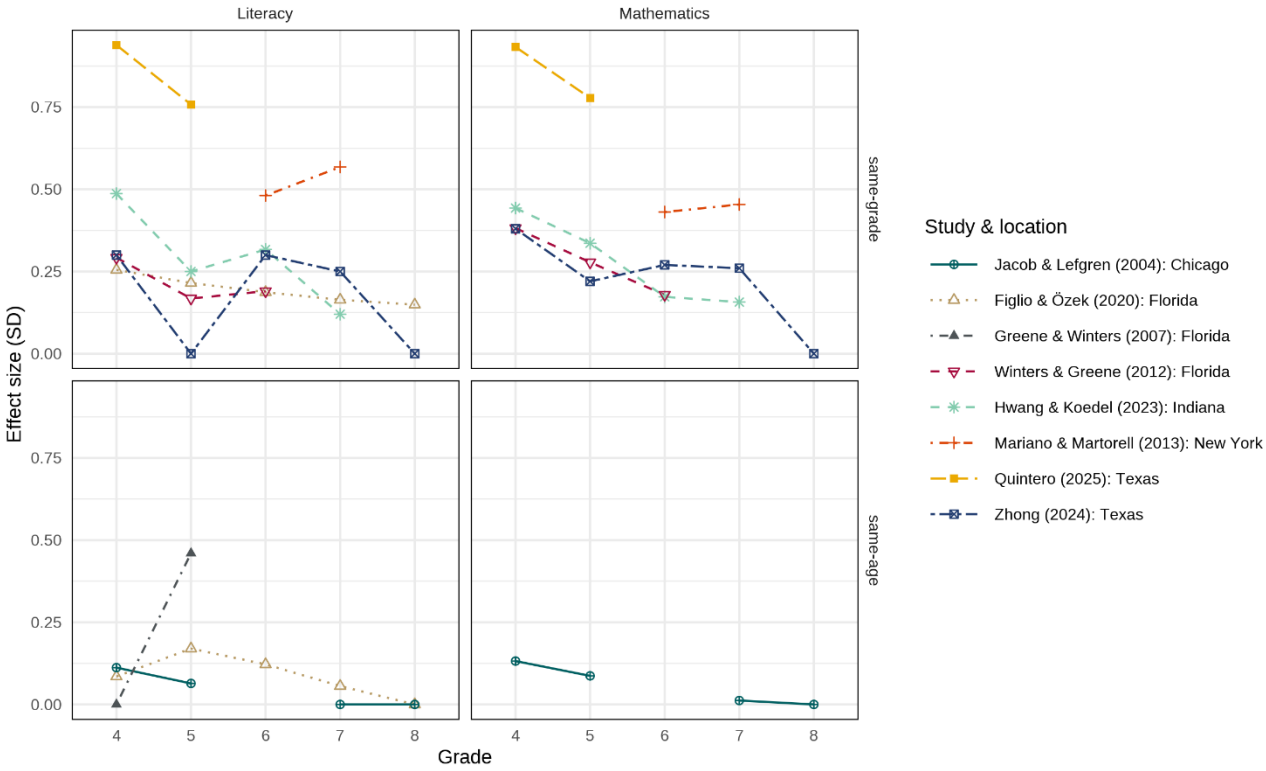
2.3.2 Evidence from studies using regression discontinuity designs

Recent research, conducted almost entirely in the United States, addresses the identification challenge inherent in estimating the impact of repetition by employing regression discontinuity (RD) designs that exploit strict, test-score-based retention policies implemented on a state-by-state basis from the late 1990s onward. Rather than estimating the average effect of repetition across all learners, RD designs identify the causal impact of repetition on those who score just below the retention threshold. The widespread use of state-wide standardised testing further facilitates consistent measurement of learner outcomes. In the US context, repetition is often accompanied by summer school and other remedial supports, distinguishing it from repetition practices in countries such as South Africa, where such supports are typically absent.

In contrast to the early observational literature – which generally reported negative effects of repetition – RD studies frequently find positive impacts on test scores, particularly when repetition occurs in the early grades (before Grade 5). The size and direction of these effects depend on both the analysis type (same-age versus same-grade) and the grade in which repetition takes place. Figure 2 . I collated the results from all comparable[5] RD studies I could locate, and plotted them in, alongside the results from this study. In these studies, the repeating grade is always the grade before the first reported effect.

[5] These are defined as all RD studies that estimate the impact of primary-school grade repetition on test scores and report same-grade effects in standard deviations.

Figure 2. Results of RD studies which report effects in standard deviations, by analysis type



Sources: Jacob and Lefgren (2004), Figlio and Özek (2020), Winters and Greene (2012), Greene and Winters (2007) Hwang and Koedel (2023), Mariano and Martorell (2013), Quintero (2025), and Zhong (2024). All non-zero effect sizes are significant at least at the 10% level; effect sizes not statistically different than zero at the 10% level are set equal to zero.

For the purposes of this discussion, I focus on same-grade analyses of repetition in primary school, noting that studies examining repetition in middle school or later (Grade 5 or higher) often find no impacts or negative impacts on test scores and increases in dropout rates (Jacob & Lefgren, 2004; Jacob & Lefgren, 2009; Larsen & Valant, 2024; Mariano et al., 2024). Same-age analyses typically report small or null effects that fade out to zero over time (Jacob & Lefgren, 2004; Roderick & Nagaoka, 2005; Schwerdt et al., 2017; Figlio & Özek, 2020). A notable exception is Greene and Winters (2007), who document positive same-age effects of Grade 3 repetition (in Florida) that grow over the two years following the retention decision.

In contrast, same-grade RD analyses consistently identify positive impacts of primary school repetition in the grade immediately following the retention year (Winters & Greene, 2012; Mariano & Martorell, 2013; Schwerdt et al., 2017; Figlio & Özek, 2020; Hwang & Koedel, 2023; Quintero, 2025). Fadeout, in which large short-term gains diminish over time, is well documented in a range of intervention settings (Bailey et al., 2020) and evident in the effects of repetition in subsequent grades. Nonetheless, in each of the aforementioned studies the estimated (same-grade) impacts remain both practically and statistically significant for the full duration of follow-up. Schwerdt et al. (2017) provide one of the longest follow ups on test score impacts, tracking learners in Florida for up to seven grades after Grade 3 repetition. They report large and

statistically significant effects persisting through Grades 9 and 10, the final grades observed in their study[6].

Two RD studies have examined the impact of repetition on English learners specifically, making them highly relevant to the South African context where English (the typical destination language of instruction) is a second language for most learners. Figlio and Özek (2020) show that Grade 3 repetition in Florida substantially improves the English skills of English learners in the short term and increases the likelihood of taking college courses in the long term. In Texas, Quintero (2025) finds that Grade 3 repetition has no effect on dropout among English language learners and in fact increases the probability of graduating on time. However, despite these positive impacts on high school outcomes, the author finds no evidence that these gains translate into improvements in tertiary education outcomes or earnings.

Even if early-grade repetition has a positive impact on test scores in the medium-term (which is consistently the case when same-grade analysis is applied – see Figure 2), if this effect fades out completely or increases dropout then it may have overall negative impacts. Several RD studies investigate the impact of early-grade repetition on dropout. Jacob and Lefgren (2009) find that Grade 6 repetition (in Chicago) does not increase dropout, but Grade 8 repetition does. Schwerdt et al. (2017) find that Grade 3 repetition has no impact on dropout in Florida. On the other hand, Eren et al. (2017) disentangle the effects of summer school from those of repetition, finding that, independent of summer school, "potential repetition" (scoring just below the test score threshold, whether actually repeating or not) in both Grade 4 and Grade 8 may increase later dropout.

Zhong (2024) also estimates the effect of Grade 3 repetition in Texas using a same-grade analysis on a comprehensive administrative dataset covering all learners. The author finds that Grade 3 repetition fails to consistently raise short-term test scores and reduces earnings in adulthood. However, in this setting, promotion decisions can be overturned through a parental appeals process. If the likelihood of a successful appeal is correlated with unobserved characteristics such as socioeconomic status (SES), ability, or motivation, this could generate discontinuities in potential outcomes at the promotion cutoff. In particular, higher-SES parents may be more successful in securing exemptions from retention, leading to negative selection among the compliers – those whose treatment status is determined by the cutoff. Such selection would violate the continuity assumption underlying the regression discontinuity design and could bias the estimated LATE. This mechanism may partly explain why Zhong's test-score results differ from those of other same-grade RD studies.

[6] These results are not plotted in Figure 2 because the effect sizes are not reported in standard deviation units.

Most of the RD studies discussed above do not disentangle the effects of summer school, and additional remedial supports offered to repeaters in the United States, from those of repetition itself, limiting their relevance to the South African context. Only Jacob and Lefgren (2004) and Mariano and Martorell (2013) separate these mechanisms. Mariano and Martorell (2013) show that the positive effect of an additional year of schooling is substantially larger than the effect of summer school, which mitigates this concern to some extent.

However, Berne et al. (2025) argue that much of the apparent benefit attributed to repetition may instead reflect these non-repetition supports, which are typically provided to all students in the United States who fail to meet promotion criteria. Using a similar RD design to the studies above, they estimate the effect of being flagged for Grade 3 repetition in Michigan while also examining the outcomes of students who fall below promotion thresholds in districts where repetition is not implemented. They find comparable effects in both settings, indicating that being flagged for repetition may influence outcomes through channels other than repetition itself. This pattern violates the exclusion restriction and suggests that prior estimates may be upwardly biased, with the observed gains driven largely by factors associated with being flagged for repetition rather than by repetition per se.

Manacorda (2012) examines grade repetition in Grades 7 to 9 in Uruguay and finds that it increases dropout rates and lowers subsequent test scores. This study is particularly relevant because there is clear evidence of manipulation in the running variable at the promotion cutoff – a feature also observed in the present analysis. To address this concern, the author implements a "worst-case" correction, assuming that the most successful students (in terms of post-repetition outcomes) among those near the cutoff are precisely those whose scores were manipulated. This approach yields a lower-bound estimate of the causal effect. Even under this conservative assumption, the results indicate negative impacts of repetition. However, as in several other RD studies, the comparison group of "just-promoted" students excludes individuals who later repeat a grade. This restriction likely biases the estimates against repetition, as the retained group is effectively compared to a subset of promoted students who are systematically more persistent and higher-performing.

2.3.3 Grade 1 repetition

Few causal studies estimate the impact of Grade 1 repetition on test scores, and, to my knowledge, no regression discontinuity studies exist, likely reflecting the absence of test-score-based promotion policies in this grade in many countries. I therefore draw on studies that rely on matching-on-observables approaches in this discussion.

A substantial share of what we know about Grade 1 repetition comes from a Texas dataset that includes administrative test scores and 72 comprehensive background variables for 734 children

at risk of repetition, who were demographically representative of the population from which they were drawn (Wu et al., 2008). Using propensity-score matching to estimate the effect of Grade 1 repetition on test scores over the subsequent four years, Wu et al. (2008) found, using a same-grade comparison, that Grade 1 repetition significantly increased both mathematics and reading scores in Grade 2. Although the effects faded in later grades, the scores of repeaters remained significantly higher than those of comparable non-repeaters after four years. Using the same dataset but focusing on high-stakes Grade 3 test scores, Hughes et al. (2010) found positive associations between Grade 1 repetition and both mathematics and reading outcomes, although the reading result was only marginally significant. Several analyses report negative effects of Grade 1 repetition on test scores when using a same-age comparison (Hong & Yu, 2007; Hong et al., 2008; Wu et al., 2008).

Using the same dataset to examine psychosocial outcomes, Wu et al. (2010) found that Grade 1 repetition produced large and sustained improvements in academic self-efficacy and a sense of school belonging. They also found that peer acceptance increased during the repetition year but declined sharply thereafter, suggesting a potential negative social impact. This pattern may constitute one mechanism through which fadeout in academic outcomes arises.

Hughes et al. (2018) followed the same 734 learners after fourteen years to examine the impact of repeating any of Grades 1–5 on dropout. Early grade repetition was associated with an increased dropout, which is unsurprising given that repeaters were matched with learners who did not repeat any of these grades. Nonetheless, the results indicate that early-grade repetition may carry longer-term risks despite short-term improvements in test scores.

Hwang and Cappella (2018) used propensity score matching on a nationally representative dataset from the United States to estimate the impact of repetition in either Grade 1 or Grade 2. Using a same-age analysis, the authors find repetition to be associated with poorer reading outcomes by eighth grade.

Alet et al. (2013) estimate the impact of repetition in Grade 1 or Grade 2 in France using a large administrative panel and a multi-stage modelling approach to address selection. They find that repetition in either grade is associated with a 0.52 standard deviation increase in Grade 3 achievement, but that the effect turns negative by Grade 6. Although the modelling strategy may not fully account for negative selection, the results demonstrate that short-run gains may not persist and may even reverse in the medium term, highlighting the importance of studying longer-term outcomes.

When evaluated using a same-grade comparison, one channel through which grade repetition may increase learning outcomes is maturation. Repeaters are approximately one year older than their peers, which supports both cognitive and non-cognitive development, a particularly

important factor for younger learners (Blair, 2002). Consequently, the effects of repeating Grade 1 may be larger than those of repeating in later grades, as the maturation component is relatively more substantial at early ages. Consistent with this interpretation, McEwan and Shapiro (2008) exploit a regression discontinuity design based on exact birth dates and find that delaying entry into Grade 1 by one year substantially reduces Grade 1 repetition and increases Grade 4 and Grade 8 test scores by 0.3 standard deviations, with particularly pronounced effects for boys. These findings underscore the significant role of maturation at the start of formal schooling.

In summary, the existing evidence suggests that Grade 1 repetition tends to increase test scores in the short term; however, medium- and long-term risks, including higher dropout rates and declining academic outcomes, may persist despite these initial gains.

3 DATA

I use a comprehensive administrative dataset covering six South African provinces (Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West) from 2017 to 2023. The data include nearly all learners enrolled in these provinces, with information on demographics, school marks, and attendance. Unique learner identifiers enable longitudinal tracking and the construction of a balanced panel, though 28% of learners are dropped due to either attriting from the panel, or due to missing marks in the intervening years. These dropped learners are not missing completely at random (Rubin, 1976), since their average academic performance is lower (where observed) than retained learners. The dropped learners score, on average, two percentage points lower in mathematics achievement, suggesting that the sample may be biased against the weakest learners and the results may not adequately reflect outcomes for these learners.

The school marks in this dataset are comprised of each term's assessment scores; these assessments are both set and marked at the school level[7]. Comprehensive assessment guidance – which includes the format of the assessments, topics to be assessed, difficulty levels, and mark allocations – is provided to educators in the curriculum documents (see, for example, the curriculum document for Intermediate Phase Mathematics (Department of Basic Education, 2011a)). However, the assessments are not formally standardised, and school marks are therefore not comparable across schools. This will be dealt with by using school fixed effects in the analysis.

[7] There are some exceptions, where standardised or "common" assessments are used, but these typically occur in High School and do not comprise the majority of assessments.

To enhance the comparability of the treatment effects, the outcome variables – Grades 2 to 5 and Grades 5 to 7 Home Language (HL), Mathematics (MTH), and First Additional Language (FAL) marks – are converted to z-scores within the grade and year, with mean zero and standard deviation one[8]. The test scores in the repeating year are not converted as they must be raw scores for the cutoff to be meaningful.

Two panels are created for this analysis: the Foundation Phase Panel, which is used to measure the impact of Grade 1 repetition, and the Intermediate Phase Panel, which is used to measure the impact of Grade 4 repetition. These two panels consist of correct age[9] learners in Public Ordinary Schools who entered Grade 1 (or Grade 4) in 2017, 2018, or 2019 and reached Grade 4 (or Grade 7) by 2023. The restriction that learners must be the correct age at the start of Grade 4 does bias the Intermediate Phase Panel towards stronger learners who are already less likely to repeat (since they managed to get to Grade 4 without becoming overage through repetition).

Table 1 illustrates progression for the 2017 cohort: 481 606 learners began Grade 1 in 2017, of whom 420 261 advanced to Grade 2 in 2018, while 61 345 repeated Grade 1 – the primary treatment group of interest. By 2020, 358 034 learners had reached Grade 4 on schedule (without repetition), with most others doing so by 2021. Patterns for the 2018 and 2019 Grade 1 cohorts are similar, though the 2019 cohort is slightly right censored, with learners in the Foundation Phase Panel having to reach Grade 4 with at most one repeat. Analogous trends are observed in the Intermediate Phase Panel, with the progression of the 2017 Intermediate Phase cohort in Table 1. The Intermediate Phase cohorts are smaller than the Foundation Phase cohorts, mainly due to the restriction that learners must be the correct age at the start of the phase.

[8] This conversion is done within the raw DDD dataset, not within the balanced panels.
[9] Correct age is defined here as 5.5 (or 8.5) years or older, and younger than 7 (or 10) years old, at the start of Grade 1 (or Grade 4).

Table 1. Transition matrix for the cohorts who started Grade 1 (or 4) in 2017

	Foundation Phase Panel (Grade 1 repetition)					Intermediate Phase Panel (Grade 4 repetition)				
	Grade 1	Grade 2	Grade 3	Grade 4	Total	Grade 4	Grade 5	Grade 6	Grade 7	Total
2017	481 606 (100%)				481 606 (100%)	369 190 (100%)				369 190 (100%)
2018	61 345 (12.7%)	420 261 (87.3%)			481 606 (100%)	27 727 (7.5%)	341 463 (92.5%)			369 190 (100%)
2019	1 146 (0.2%)	97 572 (20.3%)	382 888 (79.5%)		481 606 (100%)	531 (0.1%)	41 920 (11.4%)	326 739 (88.5%)		369 190 (100%)
2020		5 779 (1.2%)	117 793 (24.5%)	358 034 (74.3%)	481 606 (100%)		2 093 (0.6%)	48 525 (13.1%)	318 572 (86.3%)	369 190 (100%)
2021			10 258 (2.1%)	113 314 (23.5%)	123 572 (25.7%)			2 990 (0.8%)	47 628 (12.9%)	50 618 (13.7%)
2022				10 258 (2.1%)	10 258 (2.1%)				2 990 (0.8%)	2 990 (0.8%)

Source: Balanced panel derived from administrative learner-level data.

Table 2 presents the sample size for each cohort (the year the learner started the phase). The Grade 4 repetition rates in this sample are lower than the national average for two main reasons. First, the sample includes only learners who are the correct age at the start of Grade 4, who are on average academically stronger than the full cross-section of Grade 4 learners. Second, the sample does not cover all nine provinces, which further contributes to the lower observed repetition rate. The 2019 cohort also has a lower repetition rate than the other two cohorts, as learners in this cohort have only one opportunity to repeat before they reach Grade 4 (or 7) and are thus slightly stronger than the average cross sectional group of Grade 4 (or 7) learners. The increase in the sizes of the cohorts from 2017 to 2019 is due primarily to improvements in data collection over the years, and to a growing school-age population.

Table 2. Cohort counts and repetition rates

	Foundation Phase Panel (Grade 1 repetition)			Intermediate Phase Panel (Grade 4 repetition)		
Cohort	N learners	N schools	Grade 1 repetition rate (%)	N learners	N schools	Grade 4 repetition rate (%)
2017	481 606	11 593	12.7	369 190	11 469	7.5
2018	516 488	12 602	12.4	421 246	12 434	7.9
2019	519 175	13 124	10.1	434 155	12 903	7.4

Source: Balanced panel derived from administrative learner-level data.

4 EMPIRICAL APPROACH

4.1 Fuzzy regression discontinuity design

To estimate the causal effect of repetition on test scores, I employ a fuzzy RD design, exploiting South Africa's grade promotion policy, which recommends repetition for learners scoring below 50% in Home Language or below 40% in Mathematics or First Additional Language in Grades 1 and 4 (DBE, 2011b). In practice, retention decisions are imperfect: many learners below the cutoff are promoted, while a few who meet the requirements may repeat for unrelated reasons (e.g., poor attendance), thus resulting in a fuzzy, rather than sharp, RD design. I construct for each learner and subject an index measuring the distance to the relevant pass threshold; the minimum of these values defines each learner's *minimum result index*, which determines whether they meet the promotion requirements. This approach of reducing several distinct outcomes, each with their own cutoff, into a single running variable is sometimes referred to as the "normalized-and-pooled" RD treatment effect (Cattaneo et al., 2024).

I have followed Cattaneo et al. (2024) in converting the running variable so that it conforms to the standard FRD convention that treatment occurs when the value of the running variable is greater than, or equal to, zero. I have done so by first adding 1 to each original *minimum result index* and then multiplying by -1. This transformed score is used in all estimations and in the tests for manipulation in the running variable. However, for interpretability of the discontinuity graphs, and to conform with the literature of FRD studies on grade repetition, I use the un-transformed *minimum result index* in all graphs.

Let X_i denote the transformed running variable (transformed *minimum result index*) for learner i, D_i an indicator for whether the learner repeated the first grade in the phase (Grade 1 or Grade 4 for the respective panel), and $c = 0$ the cutoff. Y_{ijg} is the end-of-year mark of learner i in grade g in subject j, with $g = 2, 3$ or 4 for the Foundation Phase Panel, and $g = 5, 6$ or 7 for the Intermediate Phase Panel; and $j = HL,\ MTH$ or FAL (Home Language, Mathematics, and First Additional Language subjects, respectively). The parameter of interest can be interpreted as a Wald estimator following Hahn et al. (2001):

$$\tau_{FRD} = \frac{\lim_{x \downarrow c} \mathbb{E}[Y_{ijg} \mid X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_{ijg} \mid X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[D_i \mid X_i = x] - \lim_{x \uparrow c} \mathbb{E}[D_i \mid X_i = x]} \qquad [1]$$

The numerator is the size of the discontinuity in the outcome at the cutoff, and the denominator is the size of the discontinuity in the probability of repetition (the first stage); the ratio therefore identifies the LATE for compliers at the cutoff under standard RD assumptions.

The running variable in this context is discrete (since the test scores are recorded as integers), but there are almost 100 distinct values (or "mass points") and the sample size is very large, with thousands of observations at each distinct value of the running variable close to the cutoff. The first-choice nonparametric local polynomial approach to RD estimation first introduced by Hahn et al. (2001) can be applicable to discrete running variables with sufficient density of the discrete values close to the cutoff, and for sufficiently large sample size, but it is important to consider that it can also fail due to low density of the running variable (Cattaneo et al., 2024). I therefore present parametric estimates alongside the nonparametric estimates.

Noting concerns about the validity of higher order polynomials in the local polynomial approach (Gelman & Imbens, 2019), and the recommendation from Cattaneo and Titiunik (2022) to default to linear polynomials for nonparametric estimation, I estimate local-linear models (using the *rdrobust* package in R, as described in Calonico et al. (2017) with inference following the methods from Calonico et al. (2014) and Calonico et al. (2020)). I use the mean squared error-optimal[10] automatic bandwidth selection procedure for fuzzy RD designs in that package, using the masspoints option set to "adjust" to account for discreteness in the running variable.

For the parametric estimations I use quadratic polynomials using a two-stage least squares approach following Cameron and Trivedi (2005) (implemented using the *feols* function from the *fixest* package (Berge et al., 2021)), with different functional form specifications on either side of the cutoff (Lee & Lemieux, 2010). The analysis is restricted to a sample around the cutoff within the bandwidth used in the nonparametric estimations for comparability. Let $Z_i = \mathbf{1}\{X_i \geq 0\}$ be the indicator for the transformed index X_i being above the cutoff $c = 0$. Let D_i denote an indicator equal to one if learner i repeated Grade 1 (or Grade 4), and zero otherwise. To allow for flexible functional forms on either side of the cutoff, I specify the control function as piecewise polynomial terms: linear in X_i for $X_i < 0$ and quadratic in X_i for $X_i \geq 0$. The outcome variable is denoted Y_{ijg}, as previously defined, and α_s are the school fixed effects.

First stage:

$$D_i = \pi_0 + \pi_1 Z_i + \pi_2 X_i 1\{X_i < 0\} + \pi_3 X_i 1\{X_i \geq 0\} + \pi_4 X_i^2 1\{X_i \geq 0\} + \alpha_s + u_i \qquad [2.1]$$

Second stage:

$$Y_{ijg} = \beta_0 + \beta_1 \widehat{D}_i + \beta_2 X_i 1\{X_i < 0\} + \beta_3 X_i 1\{X_i \geq 0\} + \beta_4 X_i^2 1\{X_i \geq 0\} + \alpha_s + \epsilon_{ijg} \qquad [2.2]$$

Under standard RD assumptions, the inclusion of group (e.g., school) fixed effects is not required for identification: as with other covariates, they serve primarily to improve efficiency without affecting the consistency or size of the treatment effect (Lee & Lemieux, 2010). Nevertheless, it

[10] This is implemented by specifying bwselect = msecomb2.

is valid to include group fixed effects in both parametric and nonparametric RD models, and the resulting estimators remain consistent (Calonico et al., 2019). In the fixed effects models, the estimated LATE represents the average effect of repetition within schools, controlling for school-level heterogeneity, which is important given that marks in this dataset are known to be inconsistent across schools.

In parametric specifications, fixed effects can be included directly via the *fixest* package in R. Nonparametric RD methods, however, do not have built-in support for fixed effects. However, I approximate school fixed effects in the nonparametric models by residualising the outcome and treatment variables. Because the inclusion of school fixed effects is equivalent to including school indicators as covariates, and covariate adjustment can be implemented in a nonparametric RD estimation by residualising with respect to these covariates (Cattaneo et al., 2023), I residualise both the treatment and outcome within each school. These residuals are then used in the nonparametric RD estimation, thereby accounting for school-level heterogeneity in outcomes while maintaining the standard identification assumptions of the RD design. These residualised outcome ($\tilde{Y}_{ijg}$) and treatment ($\tilde{D}_{ijg}$) variables are estimated as

$$\tilde{Y}_{ijg} = Y_{ijg} - \hat{\alpha}_s, \qquad \tilde{D}_i = D_i - \hat{\gamma}_s$$

where $\hat{\alpha}_s$ and $\hat{\gamma}_s$ are the estimated school fixed effects. Equations 2.1 and 2.2 are then re-estimated with $\tilde{Y}_{ijg}$ and $\tilde{D}_i$ replacing Y_{ijg} and D_i respectively.

4.2 Manipulation in the running variable

The key assumptions of the FRD design are continuity of the potential outcomes at the cutoff and the monotonicity of the treatment assignment. Monotonicity is comfortably satisfied: it seems unlikely that learners would be induced to repeat only if they are ineligible to do so according to the assignment rule. However, the first assumption is tenuous in this context, due to the practice of adjusting learners marks in South Africa. Although mark adjustments have never been national policy in the Foundation Phase[11], it is widely recognised that South African teachers often directly increase learner marks (at the point of submission of the final marks to the provincial education departments) to meet promotion thresholds.

These mark adjustments are evidenced in the "heaping" or discontinuity observed in the density functions of the results on one side of the promotion threshold and may be viewed as either manipulation in the running variable, or as group-specific measurement error (Bartalotti et al., 2021). If viewed through the lens of manipulation and if these adjustments are done at random,

[11] Mark adjustments were nationally mandated in the Senior Phase from 2015 (DBE, 2015), but not in the Foundation Phase.

estimates may remain unbiased (Cattaneo & Titiunik, 2022). However, it is more plausible that teachers selectively adjust marks for learners whom they perceive to have higher ability, or to the next-ranked learner below the pass threshold. McCrary (2008) provides a test for manipulation of the running variable, which assumes a continuous running variable and can break down in the presence of a discrete running variable. To check for evidence of manipulation in the running variable I use the *rddensity* test in R (Cattaneo et al., 2020), as this test yields more reliable inference than the original McCrary (2008) test.

The expected effect of the manipulation (or group-specific measurement error) on the estimator is uncertain: in the absence of manipulation, the discontinuity in the probability of treatment would be expected to be smaller (since many more non-repeaters would have had failing marks and would thus not be following the treatment rule). The discontinuity in the outcome would also be smaller, since more non-repeating learners would now be on the same side of the cutoff as the repeating (treated) learners, and (if repetition has a positive impact) they do worse without the repetition, thereby pulling down the average outcome for learners with a negative *minimum result index*. The net effect of the manipulation on the estimator depends on the relative sizes of these decreases in the relevant discontinuities and cannot be determined *a priori*.

To address potential mark adjustment, I identify schools where the distribution of the running variable shows no evidence of manipulation and define these as the Low Mark Adjustment (LMA) subsample. Schools are classified as LMA based on the smoothness of the running variable's density within a window around the promotion cutoff. A data-driven procedure is used to select the optimal window width, minimum school size, and smoothness parameters by testing multiple combinations and retaining the configuration that maximises sample size while passing the *rddensity* test for manipulation in the running variable. Results are presented for both the Full Sample and the LMA Subsample. The validity of the LMA analysis depends on the extent to which mark adjustment is truly absent – since passing the *rddensity* test does not guarantee the complete elimination of adjustment practices – and on whether the causal impact of repetition differs systematically between low-adjustment schools and the broader sample.

Within the LMA school sample, there is evidence suggesting adjustments in the outcome variables. If repeaters are systematically more likely to have their marks adjusted upwards – which is plausible, given that learners are officially allowed to repeat no more than once per phase – this could lead to upward-biased estimates of the effect of repetition. To address this potential bias, I construct a set of subsamples[12] of LMA schools following the same methodology

[12] I construct a separate subsample for each outcome individually. Restricting the sample of LMA schools to those that pass the *rddensity* test for all outcomes results in a prohibitively small analysis sample.

originally used to identify the LMA Subsample but further restricting each subsample to schools that do not exhibit evidence of outcome manipulation. I refer to this set of subsamples as the low mark adjustment–outcomes (LMA-O) Subsamples and present the results for this set of subsamples in the Appendix.

5 DESCRIPTIVE RESULTS

I first present descriptive results for the Full Sample, before discussing the characteristics of the Low Mark Adjustment (LMA) Subsample.

5.1 Full Sample

Table 3 presents the mean characteristics of repeaters and non-repeaters in both the Foundation and Intermediate Phase Panels. Grade repetition is strongly associated with ethnicity: African learners are substantially more likely, and White learners considerably less likely, to repeat a grade. Gender differences are also pronounced. In the Foundation Phase Panel, females are 18 percentage points less likely than males to repeat Grade 1, and in the Intermediate Phase Panel, they are 26 percentage points less likely to repeat Grade 4.

Repeaters in the Foundation Phase Panel are on average 1.6 months younger than non-repeaters. This pattern reflects that, among age-appropriate learners (those who entered school within the expected 1.5-year age window), relatively younger learners face a higher probability of grade retention (Böhmer, 2025). In contrast, in the Intermediate Phase Panel, the age differential is smaller and reverses in sign. This reversal likely reflects both selection effects – since only learners who have progressed to Grade 4 without prior repetition are included – and the diminishing relative disadvantage of being younger as learners advance through the school system (Böhmer, 2025).

Nearly all Grade 1 repeaters (95%) failed their Home Language (HL) subject, indicating that difficulties in HL are the primary barrier to promotion in the Foundation Phase. This pattern suggests that grade repetition is used as a remedial mechanism to address foundational literacy challenges. In contrast, among Grade 4 learners in the Intermediate Phase Panel, failures are distributed more evenly across subjects, implying a broader range of academic difficulties at this stage.

Test score differences between repeaters and non-repeaters narrow substantially across all subjects following Grade 1 in the Foundation Phase Panel, consistent with repetition improving subsequent academic performance. A similar convergence in scores is observed in the Intermediate Phase Panel, though the magnitude of the improvement is proportionally smaller.

Subsequent repetition – occurring in Grades 2 or 3 (Foundation Phase Panel) and Grades 5 or 6 (Intermediate Phase Panel) – is surprisingly common among learners who have already repeated once: 8.3% in the Foundation Phase Panel and 4.8% in the Intermediate Phase Panel. These rates exceed expectations given official policy, which stipulates that learners may not repeat more than once per phase (Department of Basic Education, 2011b). Notably, in the Intermediate Phase Panel, the rate of Grade 5 or 6 repetition is the only characteristic that does not differ between learners who repeated Grade 4 and those who were promoted, suggesting that once repetition occurs in the Intermediate Phase, it does not confer a clear advantage in avoiding future grade failure.

As anticipated, given the known mark adjustment practices of teachers, the densities of the running variables exhibit significant heaping just above the cutoff in both the Foundation and Intermediate Phase Panels (Figure 3). This heaping is particularly pronounced in the Intermediate Phase Panel, suggesting that a larger proportion of learners in this phase fail to meet the pass requirements but have their results adjusted upward. This pattern poses a more severe threat to the validity of the research design in the Intermediate Phase Panel.

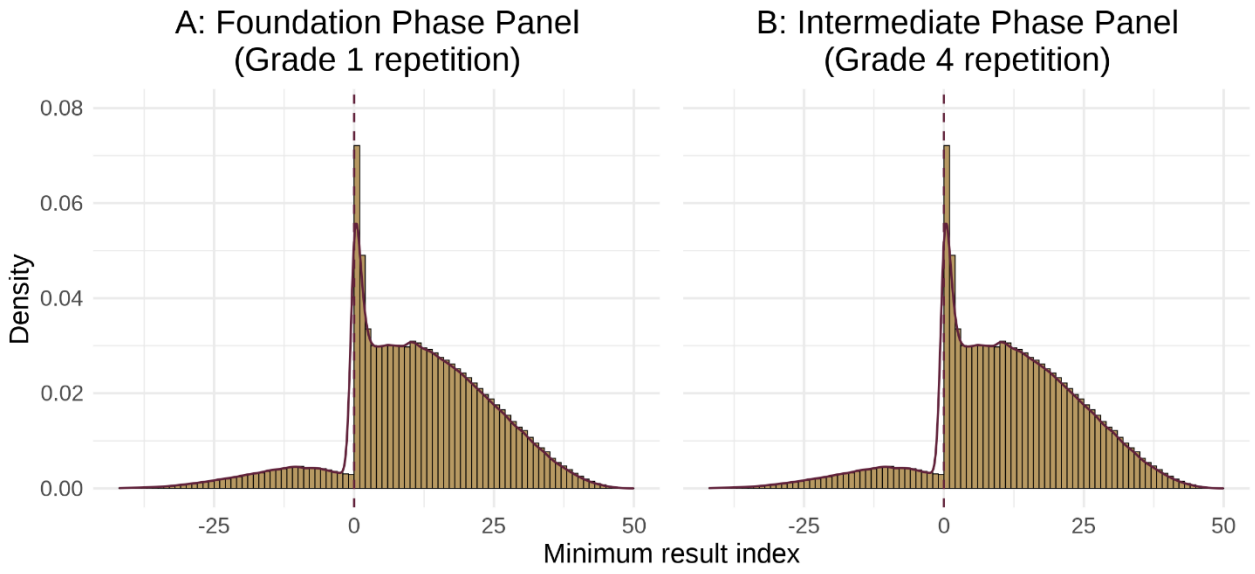
However, when discontinuities in the baseline covariates are examined (Table 4), there is limited evidence that baseline covariates are not evenly distributed on either side of the cutoff. While many of the estimated discontinuities are statistically significant (due to the large sample size, even within the bandwidths), they are so small in the Foundation Phase Panel as to have very little practical significance. In the Intermediate Phase Panel, the discontinuities are in the range of 1 to 3 percentage points, but this remains practically small. Nevertheless, the absence of discontinuities in baseline covariates does not guarantee the continuity of potential outcomes (Cattaneo & Titiunik, 2022) and this continuity assumption remains under scrutiny given the evidence of mark adjustments.

Table 3. Characteristics of learners who repeated or were promoted in Grade 1 or 4

	Foundation Phase Panel (Grade 1 repetition)					Intermediate Phase Panel (Grade 4 repetition)				
	1. Repeated		2. Did not repeat		Difference	1. Repeated		2. Did not repeat		Difference
	Mean / SE	N	Mean / SE	N	(1) - (2)	Mean / SE	N	Mean / SE	N	(1) - (2)
African/Black	0.959 [0.0027]	177 514	0.937 [0.0029]	1 339 755	0.022	0.974 [0.0031]	93 121	0.925 [0.0019]	1 131 470	0.049
Asian/Indian	0.001 [0.0007]	177 514	0.008 [0.0001]	1 339 755	-0.006	0.001 [0.0007]	93 121	0.009 [0.0001]	1 131 470	-0.008
Coloured	0.031 [0.0015]	177 514	0.023 [0.0028]	1 339 755	0.008	0.021 [0.0016]	93 121	0.027 [0.0019]	1 131 470	-0.006
White	0.008 [0.0022]	177 514	0.031 [0.0006]	1 339 755	-0.023	0.004 [0.0025]	93 121	0.038 [0.0004]	1 131 470	-0.034
Female	0.339 [0.0005]	177 514	0.519 [0.0013]	1 339 755	-0.181	0.298 [0.0007]	93 121	0.559 [0.0019]	1 131 470	-0.262
Age at start of Grade 1 (or 4)	6.001 [0.0020]	177 514	6.131 [0.0023]	1 339 755	-0.131	9.323 [0.0017]	93 121	9.233 [0.0022]	1 131 470	0.090
Failed HL (<50%)	0.950 [0.0003]	177 514	0.013 [0.0011]	1 339 755	0.937	0.746 [0.0003]	93 121	0.013 [0.0032]	1 131 470	0.733
Failed MTH (<40%)	0.512 [0.0001]	177 514	0.004 [0.0029]	1 339 755	0.508	0.621 [0.0003]	93 121	0.011 [0.0034]	1 131 470	0.610
Failed FAL (<40%)	0.445 [0.0001]	177 514	0.004 [0.0037]	1 339 755	0.442	0.639 [0.0003]	93 121	0.009 [0.0036]	1 131 470	0.630
Failed all three subjects	0.320 [0.0001]	177 514	0.002 [0.0031]	1 339 755	0.318	0.345 [0.0001]	93 121	0.003 [0.0031]	1 131 470	0.342
Grade 1 (or 4) minimum result index	-14.683 [0.0646]	177 514	19.455 [0.0622]	1 339 755	-34.138	-14.279 [0.0584]	93 121	13.288 [0.0674]	1 131 470	-27.567
HL1 (or HL4) result (first attempt)	35.839 [0.0645]	177 514	70.163 [0.0651]	1 339 755	-34.324	40.825 [0.0622]	93 121	67.109 [0.1020]	1 131 470	-26.284
MTH1 (or MTH4) result (first attempt)	39.918 [0.0760]	177 514	74.197 [0.0881]	1 339 755	-34.279	35.988 [0.0846]	93 121	59.940 [0.0930]	1 131 470	-23.951
FAL1 (or FAL4) result (first attempt)	41.888 [0.0899]	177 514	70.307 [0.1161]	1 339 755	-28.419	35.751 [0.0760]	93 121	62.450 [0.0922]	1 131 470	-26.700
HL2 (or HL5) result	59.744 [0.0656]	177 514	68.114 [0.0873]	1 339 755	-8.370	56.679 [0.0593]	93 121	67.327 [0.0960]	1 131 470	-10.648
HL3 (or HL6) result	57.855 [0.0623]	177 514	69.312 [0.0975]	1 339 755	-11.457	57.326 [0.0584]	93 121	68.331 [0.1006]	1 131 470	-11.005
HL4 (or HL7) result	55.883 [0.0623]	177 514	67.221 [0.0964]	1 339 755	-11.339	57.446 [0.0589]	93 121	67.984 [0.0989]	1 131 470	-10.538
MTH2 (or MTH5) result	61.610 [0.0757]	177 514	69.756 [0.0970]	1 339 755	-8.145	51.596 [0.0798]	93 121	59.274 [0.0935]	1 131 470	-7.678
MTH3 (or MTH6) result	57.569 [0.0729]	177 514	68.473 [0.0999]	1 339 755	-10.904	51.968 [0.0759]	93 121	60.343 [0.0982]	1 131 470	-8.374
MTH4 (or MTH7) result	52.550 [0.0758]	177 514	63.713 [0.1010]	1 339 755	-11.163	48.392 [0.0798]	93 121	59.090 [0.1013]	1 131 470	-10.699
FAL2 (or FAL5) result	57.547 [0.0787]	177 514	65.623 [0.0940]	1 339 755	-8.076	50.500 [0.0743]	93 121	62.400 [0.0942]	1 131 470	-11.899
FAL3 (or FAL6) result	55.535 [0.0719]	177 514	66.498 [0.0969]	1 339 755	-10.962	51.169 [0.0687]	93 121	63.890 [0.0976]	1 131 470	-12.722
FAL4 (or FAL7) result	50.739 [0.0731]	177 514	62.504 [0.0951]	1 339 755	-11.765	50.429 [0.0696]	93 121	63.218 [0.0975]	1 131 470	-12.789
Repeated Grade 2 / 3 (or 5 / 6)	0.083 [0.0010]	177 514	0.120 [0.0013]	1 339 755	-0.038	0.048 [0.0006]	93 121	0.050 [0.0011]	1 131 470	-0.002†

Source: Balanced panel derived from administrative learner-level data. Robust standard errors ("SE") are clustered at the school level. All differences are significant at the 1% level unless marked by a † (only "Repeated Grade 5 / 6").

Figure 3. Manipulation in the running variable



Source: Balanced panel derived from administrative learner-level data. Notes: Histograms are constructed with a bin width of 1. Kernel density estimates are plotted using Gaussian kernels with data-driven bandwidth selection. The cutoff at 0 is marked with a dashed line. The *minimum result index* in each case relates to the repetition grade.

Table 4. Discontinuities in baseline covariates

Variable	Number of observations	CER-Optimal Bandwidth	Parametric (linear) estimator			Non-parametric (local-linear) estimator		
			RD estimator	p-value	95% CI	RD estimator	p-value	95% CI
Foundation Phase Panel (Grade 1 repetition)								
Female	L = 35 414, R = 125 545	L = 5.45, R = 4.61	0.004	0.522	[-0.01, 0.02]	0.003	0.842	[-0.02, 0.02]
Age at start of Grade 1 (or 4)	L = 35 414, R = 155 032	L = 5.14, R = 5.37	0.019	0.000	[0.01, 0.03]	0.023	0.001	[0.01, 0.04]
African/Black	L = 66 474, R = 210 375	L = 9.01, R = 7.49	0.003	0.386	[-0.00, 0.01]	0.003	0.139	[-0.00, 0.02]
Asian/Indian	L = 58 725, R = 155 032	L = 8.32, R = 5.17	−0.002	0.000	[-0.00, -0.00]	−0.002	0.000	[-0.01, -0.00]
Coloured	L = 43 473, R = 384 421	L = 6.14, R = 12.41	−0.011	0.000	[-0.02, -0.01]	−0.009	0.053	[-0.02, 0.00]
White	L = 35 414, R = 155 032	L = 5.65, R = 5.08	0.006	0.000	[0.00, 0.01]	0.006	0.074	[-0.00, 0.01]
Intermediate Phase Panel (Grade 4 repetition)								
Female	L = 36 094, R = 226 188	L = 7.49, R = 4.19	0.011	0.063	[-0.00, 0.02]	0.017	0.002	[0.01, 0.05]
Age at start of Grade 1 (or 4)	L = 36 094, R = 262 663	L = 7.05, R = 5.46	0.010	0.044	[0.00, 0.02]	0.010	0.448	[-0.01, 0.02]
African/Black	L = 30 771, R = 226 188	L = 6.49, R = 4.53	0.029	0.000	[0.02, 0.04]	0.032	0.000	[0.03, 0.05]
Asian/Indian	L = 11 349, R = 299 417	L = 2.78, R = 6.45	−0.003	0.000	[-0.00, -0.00]	−0.004	0.000	[-0.01, -0.00]
Coloured	L = 30 771, R = 299 417	L = 6.16, R = 6.22	−0.022	0.000	[-0.03, -0.02]	−0.025	0.000	[-0.03, -0.02]
White	L = 30 771, R = 262 663	L = 6.89, R = 5.66	0.001	0.696	[-0.00, 0.00]	−0.001	0.013	[-0.01, -0.00]

Source: Balanced panel derived from administrative learner-level data. Notes: Robust p-values and confidence intervals are presented, derived from standard errors which are clustered at the school level. The data-driven bandwidth selector is adjusted for mass points and is run on a random sample of 100 000 observations for efficiency. CER-optimal = Coverage Error Rate optimal (Calonico et al., 2020), which minimises coverage error for inference.

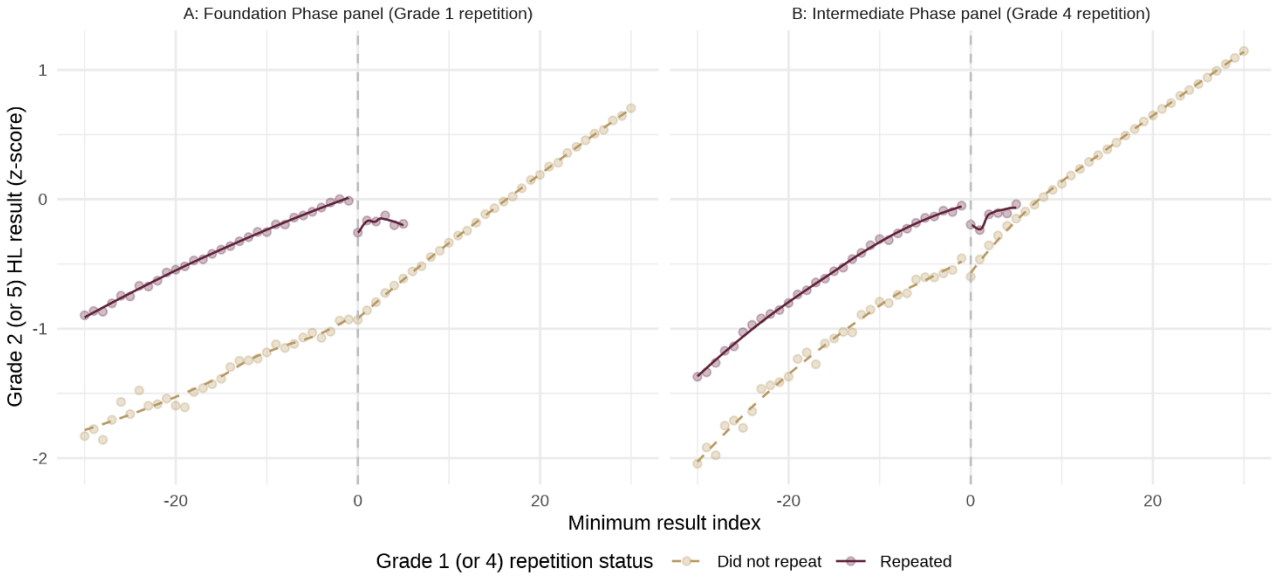
As another falsification test of the continuity of potential outcomes, I examine the discontinuity in an outcome variable (Home Language), disaggregated by treatment status (Figure 4). Amongst the non-repeating learners, there is significant support to the left of the cutoff, since many learners score below the cutoff yet do not repeat; however, there are very few learners who fall above the cutoff yet repeat (fewer than 0.1% of learners). In the Foundation Phase Panel, the curve for the non-repeaters is smooth at the cutoff, which is reassuring for continuity of potential outcomes.

In the Intermediate Phase Panel there is some dropoff just to the right of the cutoff, creating a small discontinuity. This may be caused by the very significant mark adjustments that occur in the Intermediate Phase Panel, whereby many learners who actually achieved a *minimum result index* below or significantly below zero, and therefore whose average performance in the outcome is much lower than a learner who actually scored a *minimum result index* of zero, have their *index* artificially adjusted to zero, which pulls down the average outcome at that point. This suggests that manipulation in the running variable may have a stronger impact on the estimations in the Intermediate Phase Panel.

On the other hand, the curve for the repeaters exhibits a significant discontinuity, which suggests that those repeaters whose *minimum result index* is above the cutoff do differ from the other repeaters on factors that are unobserved during the repeating grade. That is, when learners repeat despite achieving passing grades, they differ on factors that are not observed in the data at the point of the repetition decision. Thus, it is likely that potential outcomes are not continuous

for this subset at the cutoff. Fortunately, these learners represent less than 0.5% of each sample and are unlikely to influence the results.

Figure 4. Discontinuity in Grade 2 (or 5) Home Language result, by treatment status

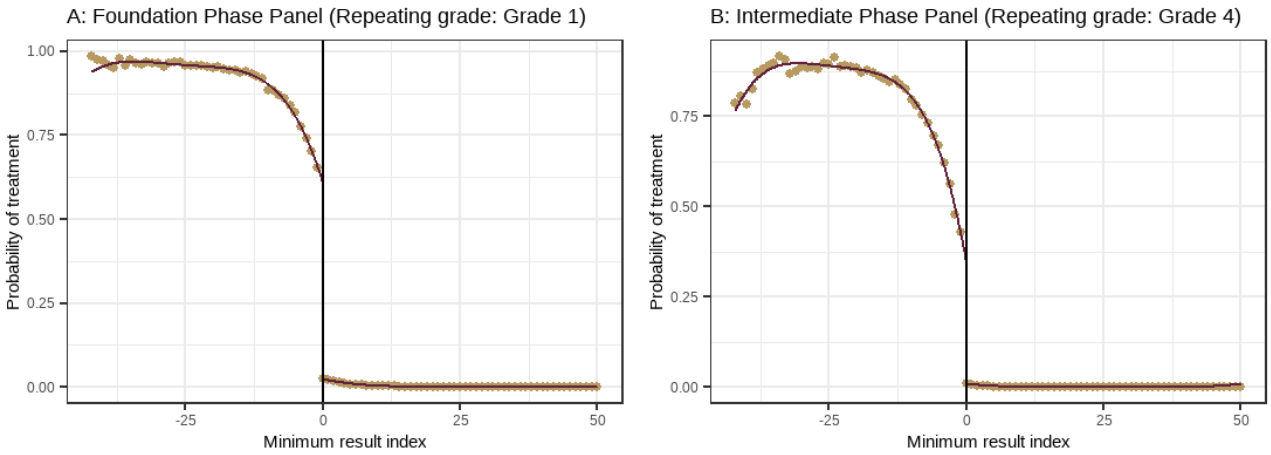


Source: Balanced panel derived from administrative learner-level data. Notes: Each point represents the local mean of the outcome variable within bins of the running variable. The fitted lines are obtained using a local polynomial regression of degree 2 with an epanechnikov kernel on either side of the cutoff. The *minimum result index* in each case relates to the repetition grade.

Notwithstanding the reservations regarding the continuity assumption and turning towards the requirement of a strong first stage, Figure 5 shows that there is a significant discontinuity in the probability of treatment in both the Foundation and Intermediate Phase Panels. Both panels display a fuzzy discontinuity, with a sharp decline in treatment probability as the running variable approaches the cutoff from the left (this decline is more pronounced in the Intermediate Phase Panel). This pattern reflects one mechanism through which learners who fail to meet the promotion thresholds are nonetheless advanced: some are promoted without any mark adjustment, which generates the observed fuzziness but does not threaten the validity of this design. A second mechanism, which is not visible in Figure 5 and is problematic for this study, arises when schools adjust marks upwards so that learners appear to satisfy the promotion requirements. Evidence of this second mechanism is shown in Figure 3, which displays clear heaping of the running variable to the right of the cutoff.

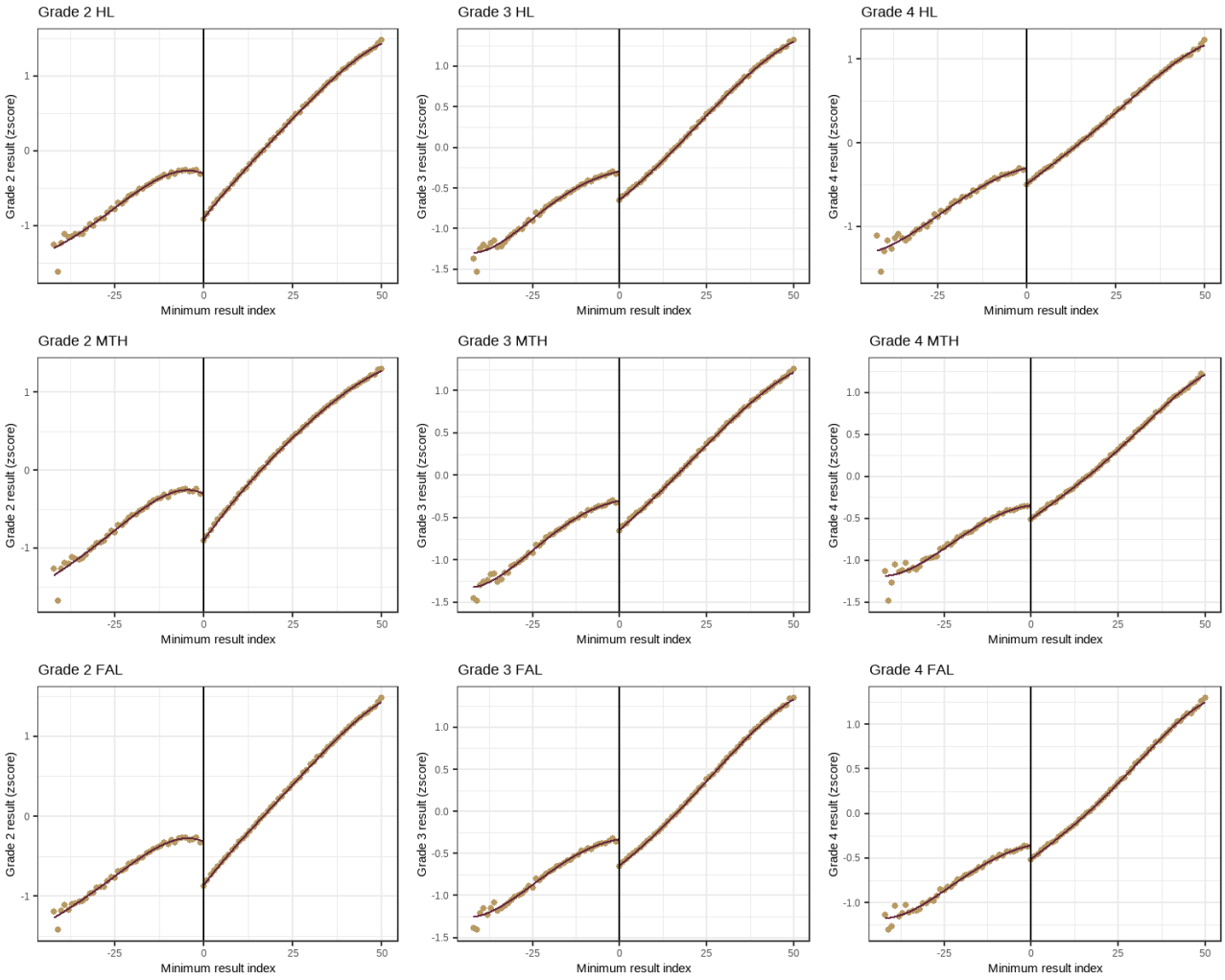
There are visible discontinuities in the outcome variables in the three grades following the repetition, and in all three subjects, in both the Foundation Phase (Figure 6) and Intermediate Phase (Figure 7) Panels. In the Foundation Phase Panel the discontinuities are largest in Grade 2 and substantially smaller in Grade 4, which is suggestive of fadeout. In the Intermediate Phase Panel, the initial impact is smaller, and the fadeout is less pronounced.

Figure 5. Discontinuity in the probability of treatment



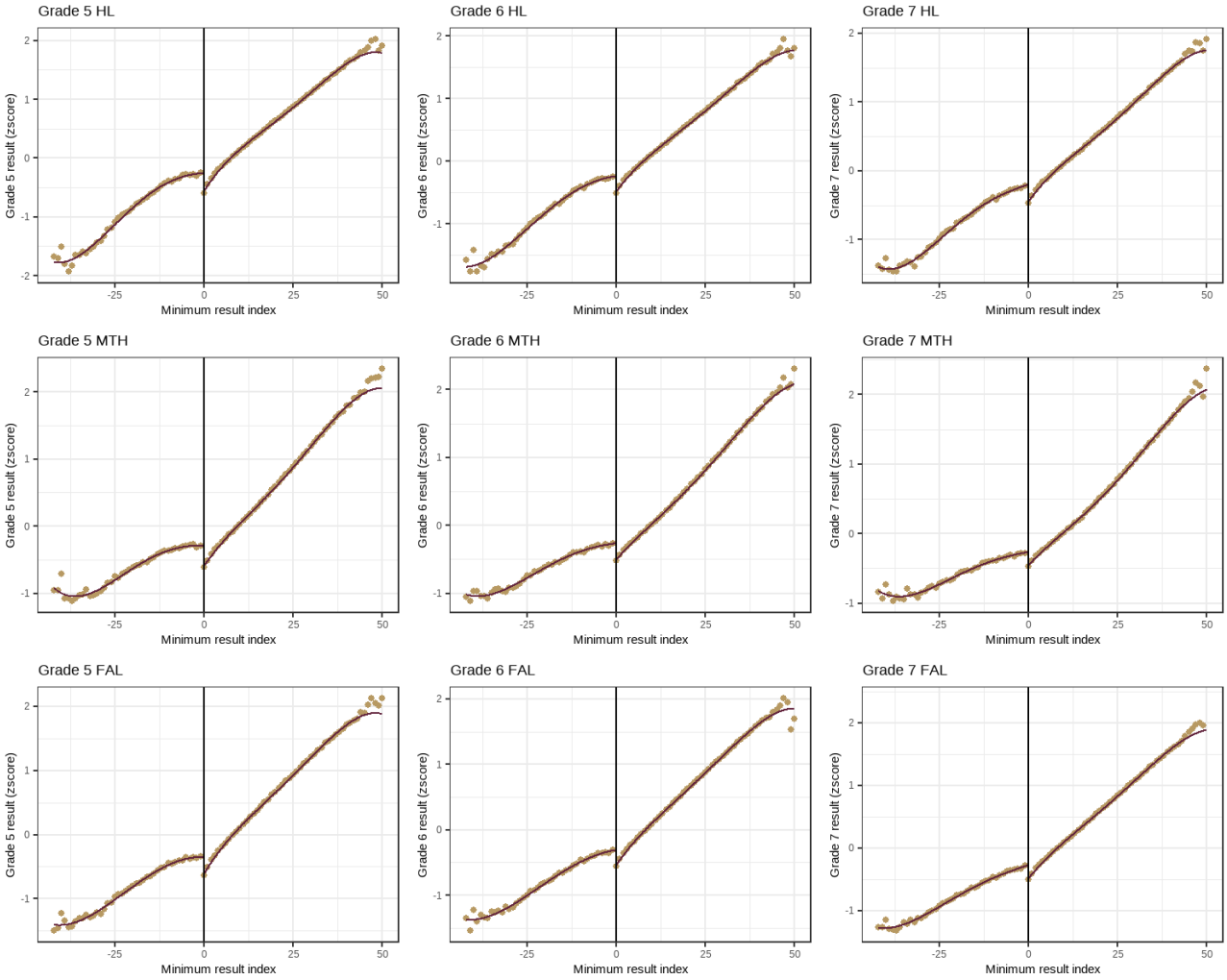
Source: Balanced panel derived from administrative learner-level data. Notes: Plots are generated using the *rdplot* package in R with default settings, except that masspoints are adjusted for. The data are partitioned into evenly spaced bins on each side of the cutoff, and local polynomial fits are estimated separately within each bin. The vertical line indicates the cutoff of the running variable. The *minimum result index* in each case relates to the repetition grade.

Figure 6. Discontinuities in Grade 2, 3 and 4 outcomes in the Foundation Phase Panel (Grade 1 repetition)



Source: Balanced panel derived from administrative learner-level data. Notes: Plots are generated using the *rdplot* package in R with default settings, except that masspoints are adjusted for. The data are partitioned into evenly spaced bins on each side of the cutoff, and local polynomial fits are estimated separately within each bin. The vertical line indicates the cutoff of the running variable. HL = Home Language, MTH = Mathematics, FAL = First Additional Language. The *minimum result index* in each case relates to Grade 1.

Figure 7. Discontinuities in Grade 5, 6 and 7 results in the Intermediate Phase Panel (Grade 4 repetition)

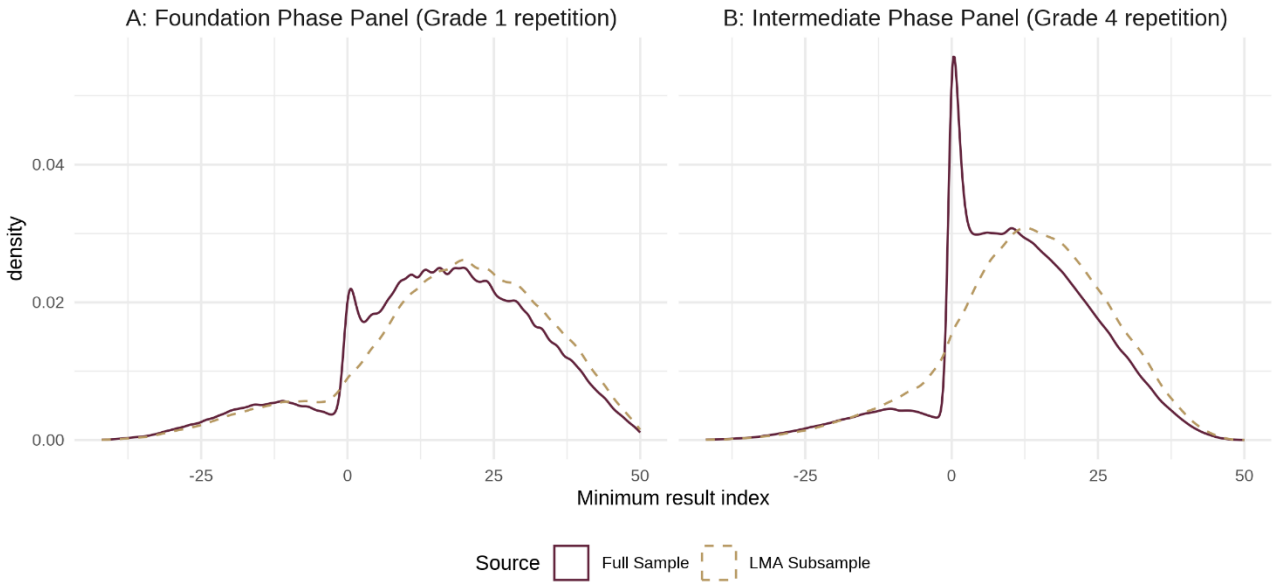


Source: Balanced panel derived from administrative learner-level data. Notes: Plots are generated using the *rdplot* package with default settings, except that masspoints are adjusted. The data are partitioned into evenly spaced bins on each side of the cutoff, and local polynomial fits are estimated separately within each bin. The vertical line indicates the cutoff of the running variable. HL = Home Language, MTH = Mathematics, FAL = First Additional Language. The *minimum result index* in each case relates to Grade 4.

5.2 Low Mark Adjustment Subsample

In both the Foundation and Intermediate Phase Panels, it was possible to identify a subset of schools that have smooth density curves (Figure 8) and in which the *rddensity* test (Cattaneo et al., 2020) fails to reject the hypothesis of no manipulation in the running variable at the 10% level (the p value of the test on the LMA samples is 0.34 in the Foundation Phase Panel and 0.99 in the Intermediate Phase Panel). The density plots indicate that learners in the LMA subsample perform better on the *minimum result index*, as indicated by the rightward shift in the density curves.

Figure 8. Density of the running variable: learners in the LMA Subsample vs Full Sample



Source: Balanced panel derived from administrative learner-level data. Notes: Kernel density estimates are plotted using Gaussian kernels with data-driven bandwidth selection. The *minimum result index* in each case relates to the repetition grade.

The numbers of learners and schools in the LMA subsample are reported in Table 5. LMA schools were more readily identified in the Foundation Phase Panel, where 28.6% of learners in appropriately sized schools from the full panel were located in LMA schools. The corresponding figure for the Intermediate Phase Panel is 9.2%. While this difference could reflect an artefact of the identification method, it is more plausibly attributable to the greater extent of mark manipulation in the Intermediate Phase Panel, which makes it inherently more difficult to identify LMA schools. The relevance of any results derived from the LMA subsamples depends on their representativeness – outside the manipulation region, where they are, by definition, unrepresentative – and on whether similar returns to repetition can reasonably be expected in these schools as in the Full Sample.

Table 5. Learner counts in low mark adjustment (LMA) schools

	Foundation Phase Panel (Grade 1 repetition)					Intermediate Phase Panel (Grade 4 repetition)				
Cohort	N learners in LMA schools	LMA as % of eligible learners	LMA as % of all learners	LMA as % of eligible schools	LMA as % of all schools	N learners in LMA schools	LMA as % of eligible learners	LMA as % of all learners	LMA as % of eligible schools	LMA as % of all schools
2017	125 727	28.3	26.1	33.1	21.7	30 337	9.2	8.2	12.3	7.0
2018	136 442	28.5	26.4	33.2	21.4	34 462	9.1	8.2	12.0	6.8
2019	138 093	29.0	26.6	33.5	21.2	36 129	9.4	8.3	12.2	6.8
Total	**400 262**	**28.6**	**26.4**	**33.5**	**20.2**	**100 928**	**9.2**	**8.2**	**12.2**	**6.4**

Source: Subsample of balanced panel derived from administrative learner-level data. Notes: "Eligible" here refers to schools with at least 50 learners in the panel, or to learners in these schools.

Learners in the LMA subsample do not differ substantially from those in the Full Sample in terms of observable characteristics, as shown by the mean values in Table 6. The repetition rate is identical in LMA and remaining schools within the Foundation Phase Panel, and only 1.8 percentage points higher in the Intermediate Phase Panel. In the Foundation Phase Panel, a similar proportion of learners are progressed despite failing grades, whereas in the Intermediate Phase panel, LMA schools progress 3.5 percentage points more failing learners than the remaining schools. The incidence of suspected mark adjustment is approximately three times higher in non-LMA schools in the Foundation Phase and five times higher in non-LMA schools in the Intermediate Phase. Academically, learners in LMA schools perform better, scoring on average 2.6 points higher on the minimum result index in the Foundation Phase Panel and 1.8 points higher in the Intermediate Phase Panel.

The two groups are equally balanced in terms of gender and age. African learners are slightly less likely, and White learners slightly more likely, to be in LMA schools. Quintile 1 schools are more likely to be classified as LMA than not, and Quintile 3 and 4 schools are slightly less likely, with no difference in Quintile 2 or 5 schools. Province is an important determinant of LMA status of a school, with learners in Gauteng and Limpopo schools less likely, and Mpumalanga and North West schools more likely, to be classified as LMA. In summary, there are some differences in observed characteristics of learners in LMA schools compared to those in non-LMA schools. The differences in important characteristics like demographics and repetition rates are small, but whether there are unobserved differences that affect repetition remains open to debate.

Table 6. Characteristics of learners in low mark adjustment schools v. rest of panel (schools with at least 50 observations)

	Foundation Phase Panel (Grade 1 repetition)					Intermediate Phase Panel (Grade 4 repetition)				
	1. LMA schools		2. Rest of panel[1]		Difference	1. LMA schools		2. Rest of panel[1]		Difference
	Mean / SE	N	Mean / SE	N	(1) - (2)	Mean / SE	N	Mean / SE	N	(1) - (2)
Mean school size[2]	143.721 [1.5707]	2 785	181.086 [1.8509]	5 521	-37.365***	114.561 [1.2185]	881	156.816 [2.3799]	6 326	-42.256***
Repeated Grade 1 (or 4)	0.117 [0.0013]	400 262	0.114 [0.0022]	999 774	0.003	0.089 [0.0009]	100 928	0.071 [0.0034]	992 020	0.018***
% failing learners progressed/promoted	0.017 [0.0003]	400 262	0.010 [0.0007]	999 774	0.006***	0.051 [0.0005]	100 928	0.016 [0.0028]	992 020	0.035***
% of marks suspected of adjustment[3]	0.011 [0.0005]	400 262	0.037 [0.0002]	999 774	-0.026***	0.017 [0.0008]	100 928	0.068 [0.0007]	992 020	-0.051***
Minimum result index	17.685 [0.1000]	400 262	15.022 [0.1549]	999 774	2.663***	13.239 [0.0764]	100 928	11.430 [0.2584]	992 020	1.809***
Female	0.499 [0.0006]	400 262	0.499 [0.0009]	999 774	-0.000	0.541 [0.0007]	100 928	0.538 [0.0027]	992 020	0.003
Age when starting Grade 1 (or 4)	6.122 [0.0024]	400 262	6.116 [0.0040]	999 774	0.006	9.250 [0.0019]	100 928	9.237 [0.0056]	992 020	0.013**
African/Black	0.923 [0.0031]	400 262	0.940 [0.0062]	999 774	-0.017**	0.910 [0.0034]	100 928	0.923 [0.0117]	992 020	-0.014
Asian/Indian	0.006 [0.0008]	400 262	0.008 [0.0011]	999 774	-0.002	0.009 [0.0008]	100 928	0.009 [0.0021]	992 020	-0.000
Coloured	0.016 [0.0022]	400 262	0.030 [0.0022]	999 774	-0.014***	0.010 [0.0020]	100 928	0.031 [0.0018]	992 020	-0.021***
White	0.053 [0.0020]	400 262	0.021 [0.0056]	999 774	0.032***	0.070 [0.0026]	100 928	0.036 [0.0109]	992 020	0.034***
School Quintile 1	0.290 [0.0064]	400 262	0.241 [0.0097]	999 774	0.048***	0.305 [0.0057]	100 928	0.223 [0.0176]	992 020	0.083***
School Quintile 2	0.252 [0.0068]	400 262	0.248 [0.0096]	999 774	0.004	0.250 [0.0061]	100 928	0.240 [0.0163]	992 020	0.010
School Quintile 3	0.255 [0.0073]	400 262	0.288 [0.0102]	999 774	-0.033***	0.239 [0.0067]	100 928	0.283 [0.0169]	992 020	-0.044**
School Quintile 4	0.077 [0.0056]	400 262	0.112 [0.0069]	999 774	-0.035***	0.073 [0.0052]	100 928	0.116 [0.0116]	992 020	-0.043***
School Quintile 5	0.127 [0.0054]	400 262	0.111 [0.0086]	999 774	0.016	0.133 [0.0056]	100 928	0.139 [0.0153]	992 020	-0.006
Eastern Cape	0.140 [0.0050]	400 262	0.154 [0.0066]	999 774	-0.014*	0.139 [0.0044]	100 928	0.129 [0.0119]	992 020	0.009
Gauteng	0.171 [0.0075]	400 262	0.242 [0.0098]	999 774	-0.071***	0.123 [0.0071]	100 928	0.261 [0.0148]	992 020	-0.138***
KwaZulu-Natal	0.198 [0.0059]	400 262	0.174 [0.0087]	999 774	0.024**	0.207 [0.0055]	100 928	0.184 [0.0156]	992 020	0.023
Limpopo	0.196 [0.0065]	400 262	0.230 [0.0086]	999 774	-0.034***	0.155 [0.0058]	100 928	0.225 [0.0137]	992 020	-0.070***
Mpumalanga	0.181 [0.0053]	400 262	0.110 [0.0096]	999 774	0.071***	0.217 [0.0047]	100 928	0.112 [0.0171]	992 020	0.105***
North West	0.115 [0.0045]	400 262	0.090 [0.0072]	999 774	0.024***	0.158 [0.0041]	100 928	0.088 [0.0145]	992 020	0.070***

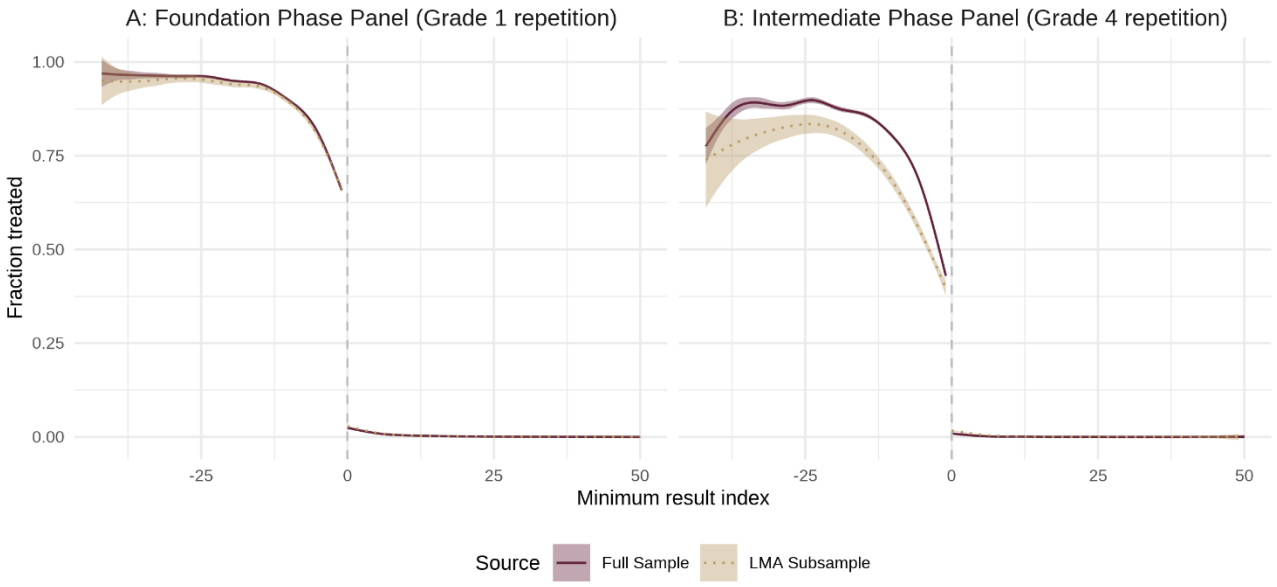
Source: Balanced panel derived from administrative learner-level data. Notes: [1]Only learners in sufficiently large schools (at least 50 learners in the school in the panel) are included, for comparability. [2]School size refers to the number of learners in the school in the panel, in any of the cohorts (2017, 2018 or 2019). "N" in this row refers to the number of schools. [3]Marks are suspected of adjustment if they are within 2 units of the relevant subject cutoff and a linear model using term marks predicts final marks below cutoff. See Section 9.2 of the Appendix for further details regarding the linear estimation. Robust standard errors ("SE") are clustered at the school level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

5.3 Comparison of discontinuities

Figure 9 presents the discontinuities in treatment status for the Full Sample and the LMA Subsample together. In the Foundation Phase (Panel A), the graphs of the discontinuities in the probability of treatment are virtually identical in the two samples. However, in the Intermediate Phase (Panel B) the discontinuity is slightly smaller in the LMA Subsample.

The discontinuities in one outcome variable (Grade 2 or 5 Home Language results – HL2 or HL5) are shown in Figure 10, for both samples (graphs of the remaining outcome variables may be found in Figure A 1 and Figure A 2 in the Appendix). In the Foundation Phase (Panel A), the size of the discontinuity in the LMA Subsample is visually indistinguishable from that in the Full Sample. Along with the identical discontinuity in the probability of treatment, this suggests that, at least for the HL2 outcome, the estimated treatment effects are likely to be very similar for both samples in the Foundation Phase.

Figure 9. Discontinuity in the probability of treatment, by data source

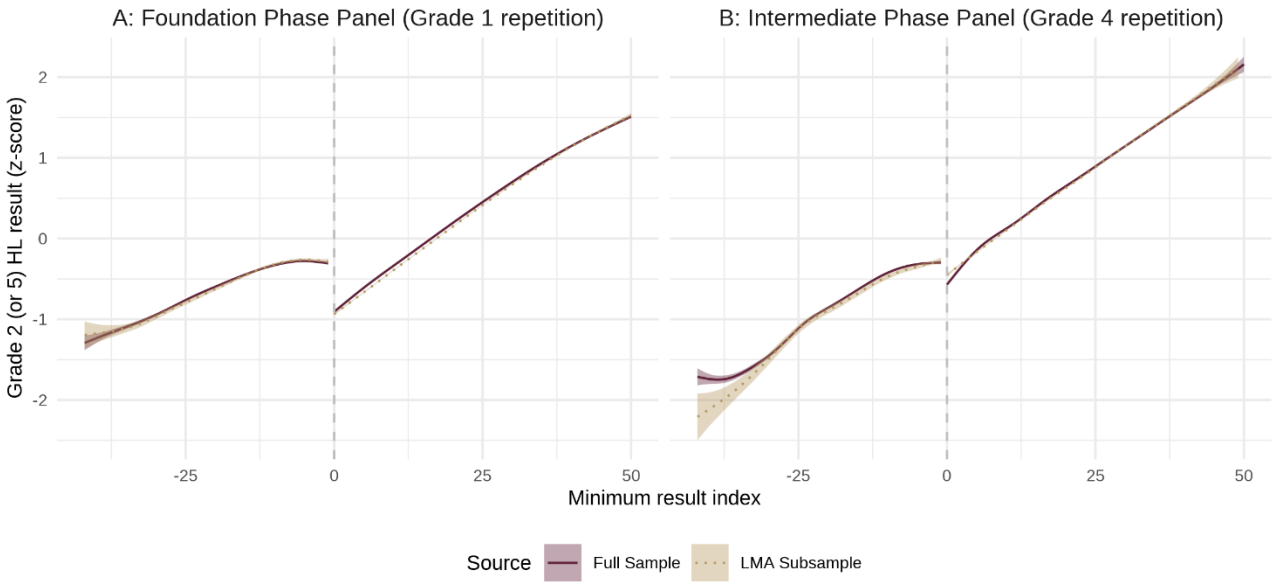


Source: Balanced panel derived from administrative learner-level data. Notes: Lines show group-specific smooth fits estimated via GAM splines, providing a flexible approximation of the local relationship around the cutoff (this was done to plot both discontinuities on one graph; an option that is unavailable in *rdplot*).

In the Intermediate Phase (Panel B) the magnitude of the discontinuity in the outcome variable is slightly smaller in the LMA Subsample than in the Full Sample, primarily due to the absence of the taper observed to the right of the cutoff in the Full Sample. This difference likely reflects the mark adjustment phenomenon: in the Full Sample, many learners just above the cutoff had in fact achieved a lower minimum result index, resulting in lower-than-expected outcomes for learners who appear to meet the threshold. In LMA schools, where mark adjustment is reduced or potentially absent, this taper is not observed. However, because the discontinuity in the probability of treatment is also smaller, the net effect on the estimator remains ambiguous.

In the LMA-O Subsample, a set of further subsamples of the LMA Subsample in which the the relevant outcomes also do not exhibit evidence of manipulation, the discontinuities are very similar to those in the LMA Subsample, although in the Foundation Phase the outcome discontinuity in the LMA-O Subsample tracks the Complete Sample more closely than the LMA Subsample does (see Figure A 3 and Figure A 4).

Figure 10. Discontinuity in an outcome variable (Home Language), by data source



Source: Balanced panel derived from administrative learner-level data. Lines show group-specific smooth fits estimated via GAM splines, providing a flexible approximation of the local relationship around the cutoff (this was done to plot both discontinuities on one graph; an option that is unavailable in *rdplot*).

6 ESTIMATION RESULTS

6.1 Grade 1 repetition

The analysis shows that Grade 1 learners whose marks were just below the threshold for repetition performed better after repeating than similar learners who just passed. The main fixed effects estimates suggest gains of 1.1 standard deviations in Grade 2, 0.6 in Grade 3 and 0.3 in Grade 4. This section details the estimation approaches that support these findings.

Table 7 reports the estimated impacts of Grade 1 repetition from the school fixed effects models, expressed as percentages of a standard deviation. Estimates from models without school fixed effects are similar, particularly for the parametric specifications, and are presented in Table A 1 in the Appendix. Although nonparametric estimators are common in the RD literature, they are more exposed to bias when the running variable is manipulated, since the triangular kernel places the greatest weight on observations closest to the cutoff, where manipulation is most likely. For this reason, I focus on the parametric estimates, which are also typically more conservative in this analysis. Across estimation approaches the results are both qualitatively consistent and quantitatively similar, indicating that the findings are robust to estimator choice and to bandwidth decisions, which vary by sample and subject. The corresponding MSE-optimal bandwidths (Calonico et al., 2020) and sample sizes are reported in Table A 3 of the Appendix.

Estimates from the Full Sample, which raises internal validity concerns due to potential mark adjustments in the running variable, suggest that Grade 1 repetition increases Grade 2 Home Language (HL) scores by 1.1 standard deviations, Grade 3 HL scores by 0.6 standard deviations and Grade 4 HL scores by 0.3 standard deviations. This medium-term effect, observed four years after the repetition year, remains both statistically and practically significant. Results from the LMA Subsample, which may face limitations in external validity, are strikingly similar: 1.1 standard deviations in Grade 2, 0.5 in Grade 3 and 0.3 in Grade 4.

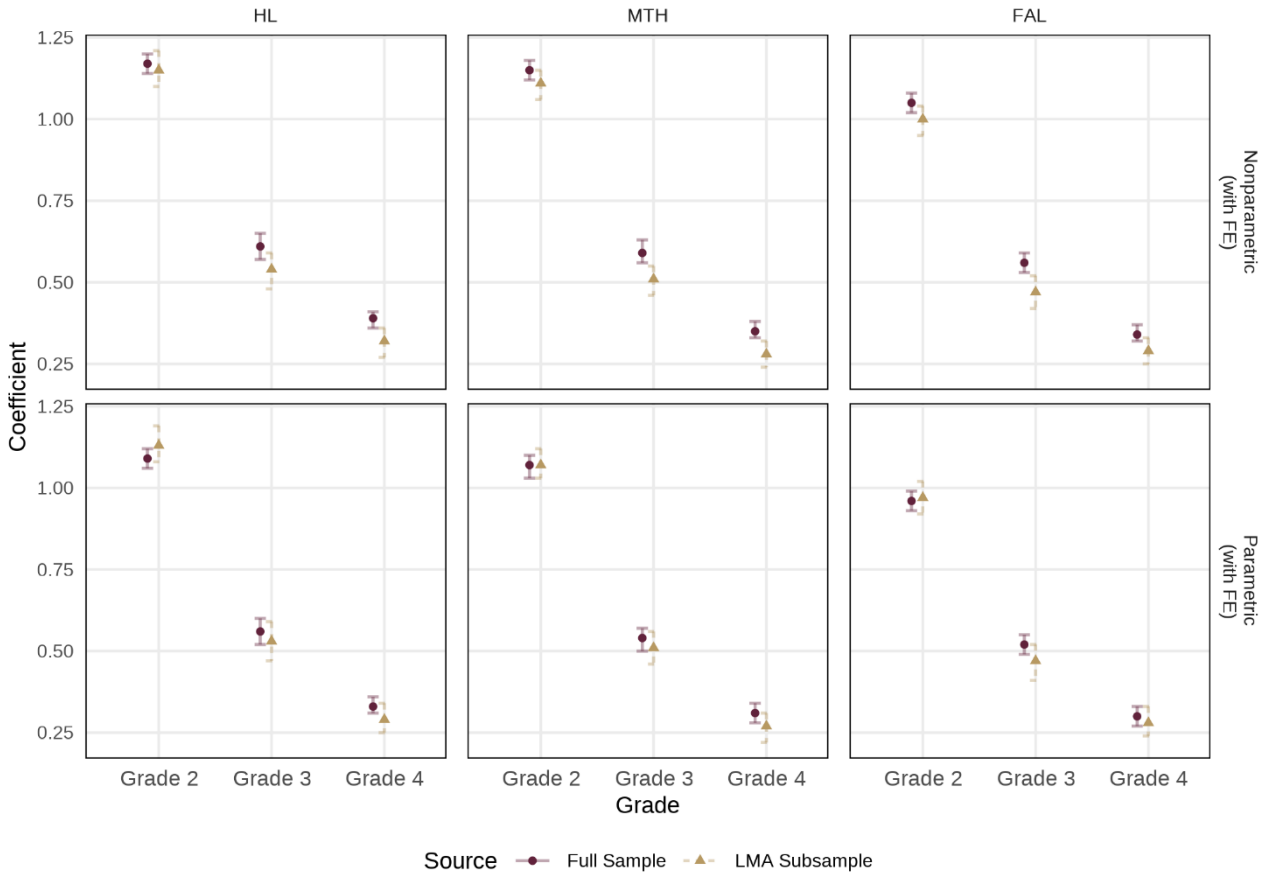
The estimated effects on Mathematics are closely aligned with the HL results, while the impacts on First Additional Language are slightly smaller. Figure 11 presents a coefficient plot from the fixed effects models and shows that the estimates from the Complete Sample and the LMA Subsample are statistically indistinguishable at the 5 per cent level.

Table 7. Estimated effect of Grade 1 repetition

Source	Grade 2		Grade 3		Grade 4	
	Parametric (with FE)	Nonparametric (with FE)	Parametric (with FE)	Nonparametric (with FE)	Parametric (with FE)	Nonparametric (with FE)
HL						
Full Sample	1.09 (0.016) F = 7 554	1.17 (0.015) F = 5 421	0.56 (0.019) F = 6 590	0.61 (0.018) F = 4 595	0.33 (0.014) F = 9 609	0.39 (0.013) F = 6 490
LMA Subsample	1.13 (0.028) F = 2 997	1.15 (0.026) F = 2 888	0.53 (0.028) F = 2 939	0.54 (0.027) F = 2 847	0.29 (0.025) F = 3 390	0.32 (0.023) F = 3 201
MTH						
Full Sample	1.07 (0.016) F = 7 554	1.15 (0.015) F = 5 438	0.54 (0.018) F = 6 893	0.59 (0.017) F = 4 824	0.31 (0.015) F = 8 561	0.35 (0.013) F = 6 089
LMA Subsample	1.07 (0.025) F = 3 309	1.11 (0.023) F = 3 099	0.51 (0.025) F = 3 367	0.51 (0.023) F = 3 137	0.27 (0.022) F = 3 548	0.28 (0.020) F = 3 237
FAL						
Full Sample	0.96 (0.016) F = 7 602	1.05 (0.015) F = 5 002	0.52 (0.015) F = 8 578	0.56 (0.013) F = 5 831	0.30 (0.014) F = 8 578	0.34 (0.013) F = 5 968
LMA Subsample	0.97 (0.025) F = 3 133	1.00 (0.023) F = 3 014	0.47 (0.027) F = 2 997	0.47 (0.025) F = 2 866	0.28 (0.021) F = 3 548	0.29 (0.020) F = 3 269

Source: Balanced panel derived from administrative learner-level data. Notes: Standard errors (clustered at the school level) in parentheses; robust and bias-corrected for nonparametric estimates. The F-statistic refers to the first-stage discontinuity in treatment assignment for fuzzy RD (nonparametric) estimates, and the conventional first-stage 2SLS F-statistic for parametric estimates. Bandwidths selected using the MSE-optimal procedure (Calonico et al., 2020), based on a random subsample of 100,000 observations; treatment estimates use the full sample. FE = school fixed effects.

Figure 11. Estimated effect of Grade 1 repetition with 95% confidence intervals



Source: Balanced panel derived from administrative learner-level data. Notes: Error bars represent 95% confidence intervals. FE = school fixed effects.

6.2 Grade 4 repetition

Grade 4 repetition also produces sizeable improvements in later achievement for repeaters who score just below the promotion threshold, compared to those who score just above. The immediate estimated effects are smaller than those observed for Grade 1 repetition, but there is less fadeout and the effect after four years is very similar in both phases. The results are broadly consistent across specifications and samples, although concerns about manipulation of the running variable are more pronounced in this panel. The remainder of this section sets out these results and explains how estimator choice affects their interpretation.

Table 8 reports the estimated effects of Grade 4 repetition for the models that include school fixed effects. Estimates from the corresponding models without fixed effects, shown in Table A 2 in the Appendix, are broadly consistent. The Grade 4 results display greater variability between the specifications with and without fixed effects (compared to the Grade 1 results), which reflects higher between-school heterogeneity in assessment practices, or in mark adjustment practices. The estimates are also more sensitive to model specification, which is expected given the stronger evidence of mark adjustment in this panel and the heightened vulnerability of the

nonparametric estimates to manipulation in the running variable due to the use of a triangular kernel. The MSE-optimal bandwidths and the associated sample sizes are provided in Table A 4 in the Appendix.

In the Full Sample – and focusing conservatively on the parametric estimates that are smaller and less sensitive to manipulation in the running variable – Grade 4 repetition increases Grade 5 Home Language (HL) performance by 0.8 standard deviations, Grade 6 HL by 0.6 standard deviations and Grade 7 HL by 0.5 standard deviations. Internal validity concerns are more pronounced in this sample due to stronger evidence of manipulation in the running variable. Consistent with this, the LMA Subsample yields slightly smaller treatment effects of 0.6 standard deviations in Grade 5, 0.5 in Grade 6 and 0.4 in Grade 7. The LMA estimates are far more stable across specifications, indicating that mark manipulation affects this sample less.

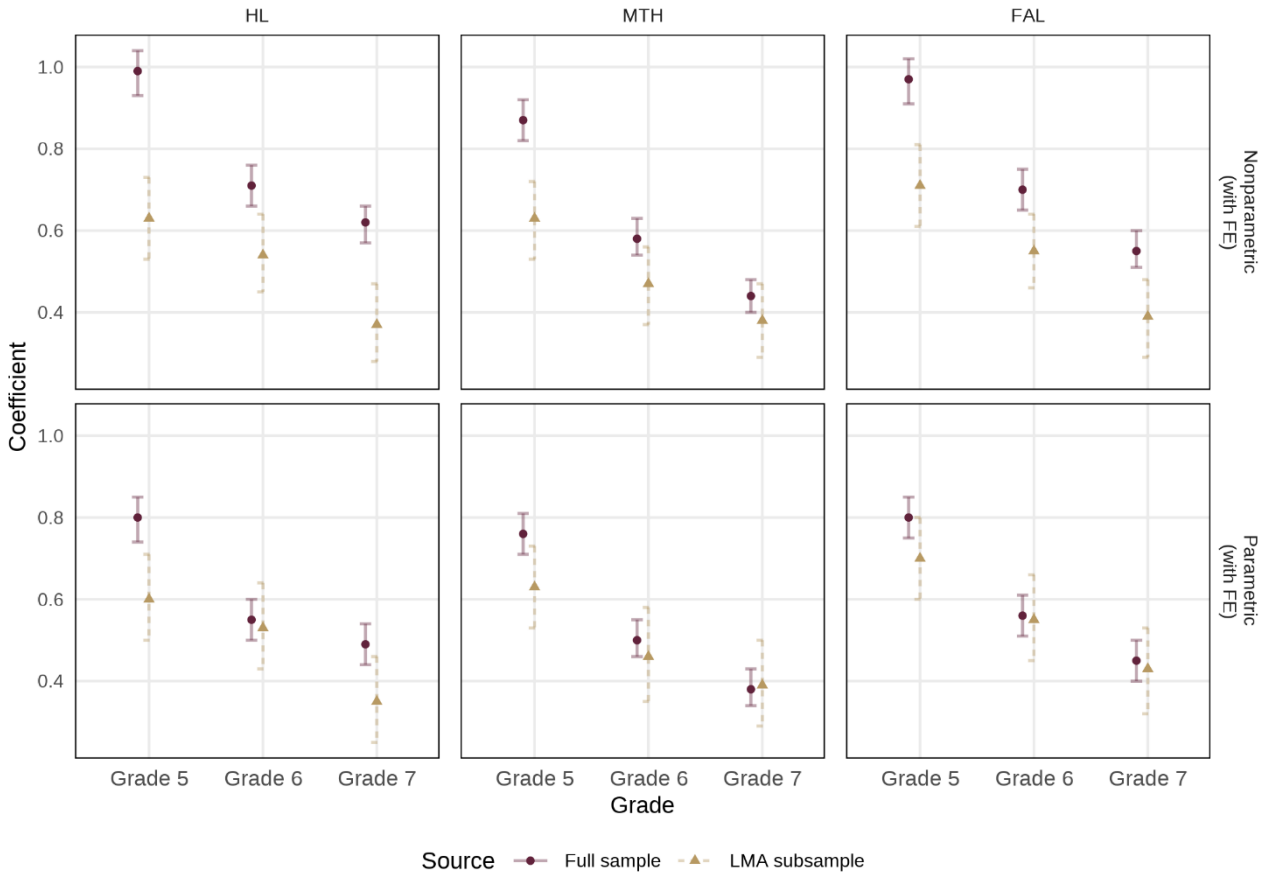
Table 8. Estimated effect of Grade 4 repetition

Source	Grade 5		Grade 6		Grade 7	
	Parametric (with FE)	Nonparametric (with FE)	Parametric (with FE)	Nonparametric (with FE)	Parametric (with FE)	Nonparametric (with FE)
HL						
Full Sample	0.80 (0.027) F = 2 367	0.99 (0.030) F = 1 308	0.55 (0.026) F = 2 499	0.71 (0.027) F = 1 375	0.49 (0.024) F = 2 762	0.62 (0.024) F = 1 556
LMA Subsample	0.60 (0.052) F = 449	0.63 (0.050) F = 491	0.53 (0.054) F = 449	0.54 (0.047) F = 491	0.35 (0.054) F = 447	0.37 (0.047) F = 482
MTH						
Full Sample	0.76 (0.026) F = 2 322	0.87 (0.026) F = 1 390	0.50 (0.024) F = 2 553	0.58 (0.023) F = 1 555	0.38 (0.022) F = 2 673	0.44 (0.020) F = 1 675
LMA Subsample	0.63 (0.052) F = 450	0.63 (0.048) F = 492	0.46 (0.058) F = 447	0.47 (0.049) F = 481	0.39 (0.055) F = 449	0.38 (0.046) F = 490
FAL						
Full Sample	0.80 (0.026) F = 2 367	0.97 (0.029) F = 1 301	0.56 (0.025) F = 2 499	0.70 (0.025) F = 1 419	0.45 (0.024) F = 2 408	0.55 (0.024) F = 1 507
LMA Subsample	0.70 (0.051) F = 459	0.71 (0.050) F = 493	0.55 (0.052) F = 450	0.55 (0.047) F = 493	0.43 (0.052) F = 450	0.39 (0.046) F = 494

Source: Balanced panel derived from administrative learner-level data. Notes: Standard errors (clustered at the school level) in parentheses; robust and bias-corrected for nonparametric estimates. The F-statistic refers to the first-stage discontinuity in treatment assignment for fuzzy RD (nonparametric) estimates, and the conventional first-stage 2SLS F-statistic for parametric estimates. Bandwidths selected using the MSE-optimal procedure (Calonico et al., 2020), based on a random subsample of 100,000 observations; treatment estimates use the full sample.

Figure 12 highlights that the nonparametric estimates differ meaningfully across samples, while the parametric estimates are more consistent and converge across samples by Grade 7. Treatment effects are similar across subjects, with slightly smaller impacts in mathematics than in language.

Figure 12. Estimated effect of Grade 4 repetition with 95% confidence intervals



Source: Balanced panel derived from administrative learner-level data. Notes: Error bars represent 95% confidence intervals. FE = school fixed effects.

The smaller differences between the Full and LMA Samples in the parametric estimates, combined with the stability of the LMA estimates across specifications, suggests that the nonparametric results in the Full Sample are likely upwardly biased by manipulation near the cutoff. Although nonparametric estimators are typically preferred for their flexibility, here the potential bias introduced by manipulation is more consequential. The closer alignment between parametric and nonparametric estimates within the LMA Subsample supports the view that the parametric model is appropriately specified. For these reasons, I prefer the parametric estimates, which are consistent across the Full and LMA Samples.

6.3 Robustness checks

The estimations already presented are robust to the different optimal bandwidths estimated for each outcome and sample (see Table A 3 and Table A 4 in the Appendix). However, even in the

LMA Sample, while there is no evidence of manipulation in the running variable, manipulation in the outcome variables remains evident. If this adjustment occurs equally amongst repeaters and non-repeaters (in terms of both frequency and magnitude), then it would not bias the estimation results. However, this assumption may not be plausible, especially since learners are only permitted to repeat once per phase: learners that repeat Grade 1 may not repeat again, and this may make them more likely to have their marks adjusted.

Therefore, I also estimate treatment effects on a set of subsamples with low mark adjustment in both the running variable and the relevant outcome (the LMA-O Subsamples). The results, reported in the Appendix in Table A 1 (Grade 1 repetition) and Table A 2 (Grade 4 repetition), are highly consistent with those reported for the Full and LMA Samples, thereby adding credibility to the validity of these estimates. This consistency supports the interpretation that these results represent a credible causal estimate based on a near-complete panel of learners spanning six provinces in South Africa.

7 DISCUSSION

7.1 Validity of results

The regression discontinuity design identifies a local average treatment effect, which in this context reflects the impact of grade repetition for learners whose attainment lies immediately below the promotion threshold. This is an important subgroup, yet it is not representative of the full distribution of learner marks. The estimates cannot be interpreted as an average treatment effect for all learners, and they do not inform the effect of repetition for very weak learners who fall far below the cutoff.

A further challenge is that many learners in the comparison group (those just above the threshold) subsequently repeat a grade. This dynamic treatment is not explicitly incorporated into the analysis. If subsequent repetition improves outcomes for these learners, the estimates presented here may understate the effect of Grade 1 and Grade 4 repetition relative to a comparison of repetition versus no repetition throughout primary school.

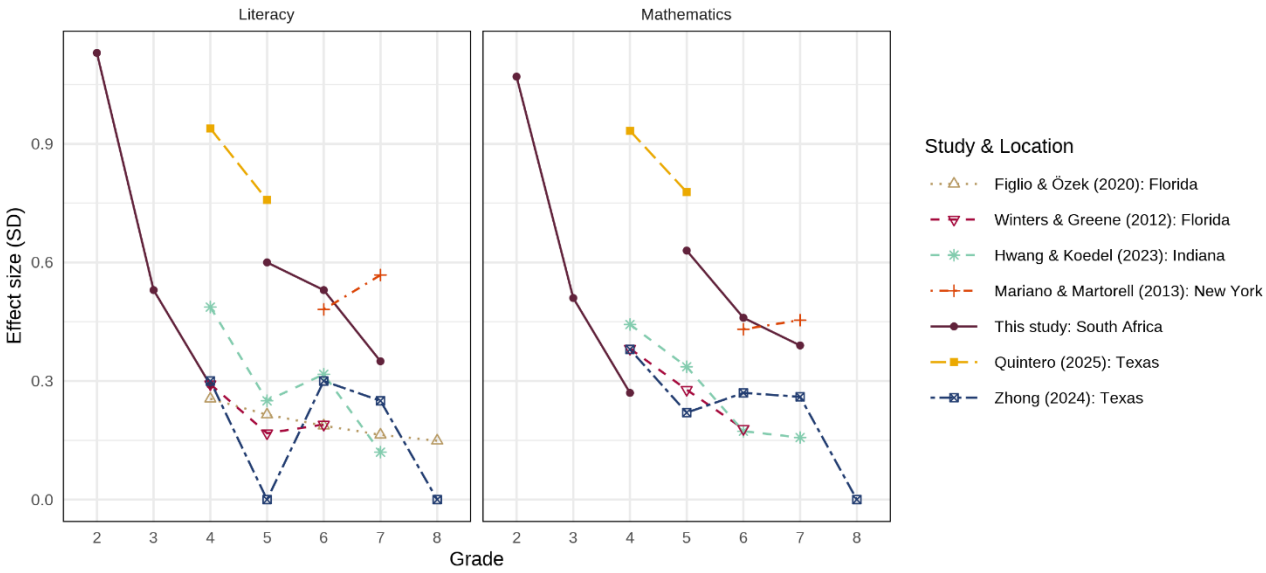
The robustness of the estimates across samples, specifications, bandwidths and subjects increases confidence in the causal interpretation. Significant concerns nonetheless remain. The study relies on school-assigned marks that are not standardised and are unlikely to be comparable across schools. School fixed effects address heterogeneity in outcome measures but cannot address heterogeneity in the measurement of the running variable.

In addition, manipulation of the running variable around the cutoff is a substantive threat to causal identification, despite the good balance in baseline covariates indicated in Table 4. Although the LMA Subsample was constructed to limit this concern, the density of the running variable is not perfectly smooth despite passing standard density tests, and the external validity of this subsample remains open to further debate.

7.2 Contextualisation within the literature

This analysis shows that Grade 1 repetition increases Grade 2 Home Language (HL) achievement by 1.1 standard deviations. The effect almost halves by Grade 3 to 0.6 SD and halves again to 0.3 standard deviations by Grade 4. Effects in Mathematics and First Additional Language are of similar magnitude. These findings align with international RD evidence documenting large immediate gains from early grade repetition followed by partial fadeout over time (Winters & Greene, 2012; Mariano & Martorell, 2013; Schwerdt et al., 2017; Figlio & Özek, 2020; Hwang & Koedel, 2023; Quintero, 2025).

Figure 13. Comparison of results from RD studies, including this study



Sources: The effect sizes are extracted from the studies as stated. Notes: Only studies which used a same-grade approach and report results in standard deviations are included. Where multiple specifications are presented, I selected results that use fixed effects. The results for this study are the parametric estimates from the LMA Subsample, using Home Language results in place of Literacy. Within each study. the repeating grade is always the grade before the first reported effect.

The estimated effects of Grade 1 repetition in this study are considerably larger than those reported in other same-grade RD studies that express results in standard deviations. I collated the results from all similar[13] RD studies I could locate and plotted them in Figure 13, alongside

[13] These are defined as all RD studies that estimate the impact of primary-school grade repetition on test scores and report same-grade effects in standard deviations.

the results from this study. In these studies, the repeating grade is always the grade before the first reported effect.

The fadeout between Grades 2 and 3 observed in this study is steeper than the fadeout at the equivalent time after repetition in most other studies. The younger age of repeaters may explain part of this pattern. The additional year of maturation inherent in a same-grade analysis is likely to have a larger relative effect on the cognitive and non-cognitive development of Grade 1 learners than on older learners (Blair, 2002). The rate of fadeout between Grades 3 and 4 is similar to that observed in international literature.

The effects of Grade 4 repetition are less of an outlier, with treatment effects on Home Language (Literacy) outcomes estimated at 0.6 SD in Grade 5, 0.5 SD in Grade 6 and 0.4 SD in Grade 7. These values remain large but fit more closely with the range observed internationally, and rates of fadeout are in line with other RD studies. Many studies document the slowing of fadeout over time, yet almost all maintain a downward trajectory with no indication of stabilising. This pattern raises the concern that, with longer follow-up, effects could approach zero despite the large impacts observed in the short- and medium-term.

Although treatment effects are not directly comparable across phases (or studies) due to differing score distributions, the evidence suggests that Grade 4 repetition produces medium-run gains similar in magnitude to the four-year effects of Grade 1 repetition. This finding contrasts with international evidence that earlier repetition tends to produce more positive impacts (Valbuena et al., 2021).

It is well established that later repetition, particularly after Grade 5, can increase dropout rates (Jacob & Lefgren, 2009; Manacorda, 2012; Mariano et al., 2024). There is also evidence that early-grade repetition may raise the likelihood of later dropout even while improving short-term test scores (Hughes et al., 2018) – although other research suggests that early-grade repetition does not increase dropout and in fact improves outcomes through high school (Schwerdt et al., 2017; Figlio & Özek, 2020). An extension of the longitudinal dataset used in this study is needed to examine whether Grade 1 and Grade 4 repetition affect dropout in South Africa and to clarify its long-run consequences.

To my knowledge, this is the first RD study to estimate the causal impact of Grade 1 repetition on test scores, and the first RD analysis of early grade repetition conducted in a middle-income country setting. This is also the first RD study in a setting that does not provide additional support to repeaters, and the positive results may mitigate the concerns of Berne et al. (2025) that the impact of repetition in the United States is largely due to the additional supports provided to repeaters in that context, rather than the repetition itself. The findings align with the international RD literature, which reports positive short-run and medium-run effects of early-grade repetition

on test scores when evaluated using a same-grade approach. The evidence presented here suggests that these positive impacts remain sizeable after four years, although further work is needed to understand their longer-term trajectory.

7.3 Policy recommendations

This study provides clear evidence that Grade 1 and Grade 4 repetition cause improved academic outcomes for learners just below promotion thresholds in at least the three grades following repetition. These findings lend support to existing repetition guidelines (Department of Basic Education, 2011b) and to the practice of allowing repetition in the early grades. However, it is not practical to interpret these results as a call to increase repetition rates (by, for instance, ending the practice of promoting some learners whose results do not meet promotion thresholds). Grade repetition increases class sizes, which are already very large in Grade 1 (Gustafsson & Mabogoane, 2012). Without additional teachers to maintain class sizes, further increasing repetition rates in the early grades could be harmful.

The data used in this study do not track learners long enough to determine whether the observed gains persist, fade out to zero or raise the risk of dropout. Some international studies that find positive short or medium-term effects on test scores also document increased dropout years later (Eren et al., 2017; Hughes et al., 2018). Repetition also carries broader social and economic costs, including delayed entry into the labour market, which must be considered when assessing whether it is a cost-effective remedial strategy.

Most learners who repeat Grade 1 do so because they fail their home language subject, which suggests that repetition is frequently used to address early language difficulties. While the effects of repetition found in this study are much larger than those found in other effective remedial interventions in South Africa (Wills, 2025), the costs, both direct and indirect, may be higher. Policymakers should weigh the costs and benefits of repetition against alternative language remediation strategies, including structured pedagogy programs (Stern et al., 2024). Complementary language support for repeaters may also be worthwhile given evidence that such support enhances outcomes (Valbuena et al., 2021). These additional supports in literacy may be especially effective for English learners, as demonstrated by the large positive effects of repetition found in studies that focus on this subpopulation (Figlio & Özek, 2020; Quintero, 2025).

A key mechanism behind the large Grade 1 repetition effects is likely maturation. Repeaters are a year older in Grade 2, and at this young age an additional year supports especially rapid cognitive and non-cognitive development (Blair, 2002). The relative developmental gain from an extra year of age is much larger for six- and seven-year-olds than for learners closer to ten, which helps explain the stronger immediate effects observed for Grade 1 repetition. The gains

associated with repetition at this stage may therefore reflect developmental readiness as much as additional instructional time. An alternative to higher repetition rates is to ensure that learners who are below the minimum school-starting age are genuinely school-ready before entering Grade 1 (McEwan & Shapiro, 2008; Böhmer, 2025). This approach offers the developmental advantages of greater maturity without requiring a repeated year. Such a policy must, however, guarantee access to alternative early-learning opportunities that include nutrition programs.

Sustaining and expanding the administrative data used in this study is critical. This would enable the study of longer-term outcomes, specifically fadeout and dropout, and provide a more complete understanding of the consequences of grade repetition in South Africa. Furthermore, while the work of estimating the cost of repetition has begun (van der Berg et al., 2019), further refinement is needed to accurately assess its cost–benefit ratio.

8 REFERENCES

Adams, G. 2020. The impact of the quintile funding system in reducing apartheid-inherited inequalities in education.

Alet, É., Bonnal, L. & Favard, P. 2013. Repetition: Medicine for a short-run remission. *Annals of economics and statistics,* 111/112):227–250.

Angrist, J. & Imbens, G. 1995. Identification and estimation of local average treatment effects. National Bureau of Economic Research Cambridge, Mass., USA.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association.,* 91(434):444–455.

Angrist, J. D. & Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Ardington, C. 2024. Funda wande limpopo workbooks evaluation. Endline report. April 2024. Cape Town: SALDRU, University of Cape Town.

Ardington, C., Wills, G. & Kotze, J. 2021. Covid-19 learning losses: Early grade reading in south africa. *International journal of educational development.,* 86(102480.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R. & Yeager, D. S. 2020. Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest,* 21(2):55–97. https://doi.org/10.1177/1529100620915848

Bartalotti, O., Brummet, Q. & Dieterle, S. 2021. A correction for regression discontinuity designs with group-specific mismeasurement of the running variable. *Journal of Business & Economic Statistics,* 39(3):833–848. https://doi.org/10.1080/07350015.2020.1737081

Berge, L., Krantz, S., McDermott, G. & Berge, M. L. 2021. Package 'fixest'.

Berne, J. S., Jacob, B. A., Weiland, C. & Strunk, K. O. 2025. The impacts of grade retention policy with minimal retention. Edworkingpaper no. 25-1188. *Annenberg Institute for School Reform at Brown University*.

Blair, C. 2002. School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American psychologist,* 57(2):111.

Böhmer, B. Understanding the roles of age, gender, and grade r attendance in academic progression and performance from grades 1 to 4. ESSA Centenary Conference, 2025 Somerset West.

Branson, N., Hofmeyr, C. & Lam, D. 2014. Progress through school and the determinants of school dropout in south africa. *Development Southern Africa.,* 31(1):106–126.

Branson, N. & Lam, D. 2010. Education inequality in south africa: Evidence from the national income dynamics study. *Studies in Economics and Econometrics,* 34(3):85–109. https://doi.org/10.1080/10800379.2010.12097211

Cabrera-Hernandez, F. 2022. Leave them kids alone! The effects of abolishing grade repetition: Evidence from a nationwide reform. *Education economics.,* 30(4):339–355.

Calonico, S., Cattaneo, M. D. & Farrell, M. H. 2020. Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal,* 23(2):192–210.

Calonico, S., Cattaneo, M. D., Farrell, M. H. & Titiunik, R. 2017. Rdrobust: Software for regression-discontinuity designs. *The Stata Journal,* 17(2):372–404.

Calonico, S., Cattaneo, M. D., Farrell, M. H. & Titiunik, R. 2019. Regression discontinuity designs using covariates. *The review of economics and statistics.,* 101(3):442–451.

Calonico, S., Cattaneo, M. D. & Titiunik, R. 2014. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica : journal of the Econometric Society.,* 82(6):2295–2326.

Cameron, A. & Trivedi, P. 2005. *Microeconometrics: Methods and applications.* Cambridge, UK: Cambridge University Press.

Cattaneo, M. D., Idrobo, N. & Titiunik, R. 2019. *A practical introduction to regression discontinuity designs: Foundations.* Cambridge (UK): Cambridge University Press.

Cattaneo, M. D., Idrobo, N. & Titiunik, R. 2024. *A practical introduction to regression discontinuity designs: Extensions.* Cambridge (UK): Cambridge University Press.
Cattaneo, M. D., Jansson, M. & Ma, X. 2020. Simple local polynomial density estimators. *Journal of the American Statistical Association,* 115(531):1449–1455. https://doi.org/10.1080/01621459.2019.1635480
Cattaneo, M. D., Keele, L. & Titiunik, R. 2023. Covariate adjustment in regression discontinuity designs. *Handbook of matching and weighting adjustments for causal inference,* 153–168.
Cattaneo, M. D. & Titiunik, R. 2022. Regression discontinuity designs. *Annual Review of Economics,* 14(Volume 14, 2022):821–851. https://www.annualreviews.org/content/journals/10.1146/annurev-economics-051520-021409
Department of Basic Education 2011a. Curriculum and assessment policy statement (caps): Mathematics intermediate phase grades 4–6. Pretoria: Department of Basic Education.
Department of Basic Education 2011b. 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. Pirls 2021: South african preliminary highlights report. Pretoria: Department of Basic Education.
Department of Basic Education 2024a. Department of basic education: Annual report 2023-2024. Pretoria: Department of Basic Education.
Department of Basic Education 2024b. South africa 2023 trends in international mathematics and science study (timss) highlight report. Pretoria: Department of Basic Education.
Department of Basic Education 2025. Strategic plan 2025–2030. Pretoria: Department of Basic Education. https://www.education.gov.za/Portals/0/Documents/Reports/2025/2025-2030%20DBE%20Strategic%20Plan_Compress.pdf
Eide, E. R. & Showalter, M. H. 2001. The effect of grade retention on educational and labor market outcomes. *Economics of Education Review,* 20(6):563–576. https://www.sciencedirect.com/science/article/pii/S0272775700000418
Eren, O., Depew, B. & Barnes, S. 2017. Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics.,* 153(9–31.
Evans, D. K. & Yuan, F. 2022. How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis,* 44(3):532–540. https://doi.org/10.3102/01623737221079646
Ferreira Sequeda, M., Golsteyn, B. H. H. & Parra-Cely, S. 2018. The effect of grade retention on secondary school performance: Evidence from a natural experiment. IZA Discussion Papers.
Figlio, D. & Özek, U. 2020. An extra year to learn english? Early grade retention and the human capital development of english learners. *Journal of Public Economics,* 186(104184. https://www.sciencedirect.com/science/article/pii/S0047272720300487
Funda Wande 2023. Annual report. Online. https://fundawande.org/wp-content/uploads/2025/03/WT3228-Funda-Wande-Annual-Report-2023-1-1.pdf
Gelman, A. & Imbens, G. 2019. Why high-order polynomials should not be used in regression discontinuity designs. *Journal of business & economic statistics.,* 37(3):447–456.
Glick, P. & Sahn, D. E. 2010. Early academic performance, grade repetition, and school attainment in senegal: A panel data analysis. *The World Bank economic review,* 24(1):93–120.
Greene, J. P. & Winters, M. A. 2007. Revisiting grade retention: An evaluation of florida's test-based promotion policy. *Education Finance and Policy,* 2(4):319–340.
Gustafsson, M. 2023. Grade promotion, repetition and dropping out 2018 to 2021. Pretoria: Department of Basic Education. https://www.education.gov.za/Portals/0/Documents/Reports/2023/Flows%20Through%20Grade%20Report%202023.pdf?ver=2023-10-26-141503-740
Gustafsson, M. & Mabogoane, T. 2012. South africa's economics of education: A stocktaking and an agenda for the way forward. *Development Southern Africa,* 29(3):351–364. https://doi.org/10.1080/0376835X.2012.706033

Hahn, J., Todd, P. & Van der Klaauw, W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica,* 69(1):201–209.
Hanushek, E. A. & Woessmann, L. 2012. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of economic growth.,* 17(4):267–321.
Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society,* 153–161.
Heckman, J. J. 2006. Skill formation and the economics of investing in disadvantaged children. *Science.,* 312(5782):1900–1902.
Holmes, C. T. 1989. Grade level retention effects: A meta-analysis of research studies, in L. A. Shepard, M. L. S. (ed.) *Flunking grades: Research and policies on retention.* London: The Falmer Press.
Hong, G. & Yu, B. 2007. Early-grade retention and children's reading and math learning in elementary years. *Educational evaluation and policy analysis,* 29(4):239–261.
Hong, G., Yu, B., Kalil, A., García Coll, C. & Foster, E. M. 2008. Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental psychology,* 44(2):407–421.
Hughes, J. N., Chen, Q., Thoemmes, F. & Kwok, O.-m. 2010. An investigation of the relationship between retention in first grade and performance on high stakes tests in third grade. *Educational evaluation and policy analysis,* 32(2):166–182.
Hughes, J. N., West, S. G., Kim, H. & Bauer, S. S. 2018. Effect of early grade retention on school completion: A prospective study. *Journal of Educational Psychology,* 110(7):974.
Hwang, N. & Koedel, C. 2023. Helping or hurting: The effects of retention in the third grade on student outcomes. *Educational Evaluation and Policy Analysis,* 47(1):65–88. https://doi.org/10.3102/01623737231197639
Hwang, S. H. J. & Cappella, E. 2018. Rethinking early elementary grade retention: Examining long-term academic and psychosocial outcomes. *Journal of research on educational effectiveness,* 11(4):559–587.
Imbens, G. W. & Lemieux, T. 2008. Regression discontinuity designs: A guide to practice. *Journal of econometrics.,* 142(2):615–635.
Jackson, G. B. 1975. The research evidence on the effects of grade retention. *Review of educational research.,* 45(4):613–635.
Jacob, B. A. & Lefgren, L. 2004. Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics,* 86(1):226–244.
Jacob, B. A. & Lefgren, L. 2009. The effect of grade retention on high school completion. *The American Economic Journal,* 1(3):33–58.
Jimerson, S. R. 2001. Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review,* 30(3):420–437.
Kika, J. C., Crouch, L. A., Dulvy, E. N. & Thulare, T. D. 2022. Early grade reading in south africa. The World Bank.
Köhler, T. 2024. A paradox of progress: Rising education and unequal labour market returns in post-apartheid south africa. Stellenbosch: Resep.
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. https://www.sciencedirect.com/science/article/pii/S0304387810000490
Larsen, M. F. & Valant, J. 2024. The long-term effects of grade retention: Evidence on persistence through high school and college. *Journal of Research on Educational Effectiveness,* 17(4):615–646. https://doi.org/10.1080/19345747.2023.2240323
Lee, D. S. 2008. Randomized experiments from non-random selection in u.S. House elections. *Journal of econometrics.,* 142(2):675–697.
Lee, D. S. & Lemieux, T. 2010. Regression discontinuity designs in economics. *Journal of economic literature,* 48(2):281–355.
Manacorda, M. 2012. The cost of grade retention. *The Review of Economics and Statistics,* 94(2):596–606.
Mariano, L. T. & Martorell, P. 2013. The academic effects of summer instruction and retention in new york city. *Educational Evaluation and Policy Analysis,* 35(1):96–117.

Mariano, L. T., Martorell, P. & Berglund, T. 2024. The effects of grade retention on high school outcomes: Evidence from new york city schools. *Journal of Research on Educational Effectiveness,* 1–31. https://doi.org/10.1080/19345747.2023.2287607

McCrary, J. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics,* 142(2):698–714.

McEwan, P. J. & Shapiro, J. S. 2008. The benefits of delayed primary school enrollment: Discontinuity estimates using exact birth dates. *The Journal of Human Resources,* 43(1):1–29. http://www.jstor.org.ez.sun.ac.za/stable/40057337

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T. & Harrell, F. E. 2006. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology.,* 59(10):1092–1101.

Nyamunda, J. 2024. Assessing educational outcomes in south africa relative to economically comparable countries: A comparative analysis. *South African Journal of Education,* 44(si1):S1–S12.

OECD 2025. Oecd economic surveys: South africa 2025. Paris: OECD Publishing.

Quintero, D. 2025. The effects of third-grade retention on multilingual students: A gateway or a gatekeeper? : Annenberg Institute, Brown University.

Roderick, M. & Nagaoka, J. 2005. Retention under chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis,* 27(4):309–340.

Rubin, D. B. 1976. Inference and missing data. *Biometrika,* 63(3):581–592. http://www.jstor.org/stable/2335739

Rubin, D. B. 1987. *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Schwerdt, G., West, M. R. & Winters, M. A. 2017. The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from florida. *Journal of Public Economics,* 152(154–169.

Selkirk, R. 2025. Grade repetition in south africa: Impacts on achievement, flows, and school completion. PhD. Stellenbosch: University of Stellenbosch.

South Africa 2006. South african schools act, 1996 (act no 84 of 1996) amended national norms and standards for school funding. *In:* Education, D. O. (ed.). Pretoria.

Spaull, N. & Makaluza, N. 2019. Girls do better. *Agenda,* 33(4):11–28. https://doi.org/10.1080/10130950.2019.1672568

Stern, J. M. B., Jukes, M. C. H., Cilliers, J., Fleisch, B., Taylor, S. & Mohohlwane, N. 2024. Persistence and emergence of literacy skills: Long-term impacts of an effective early grade reading intervention in south africa. *Journal of Research on Educational Effectiveness,* 1–22. https://doi.org/10.1080/19345747.2024.2417288

Taylor, S., Cilliers, J., Prinsloo, C., Fleisch, B. & Reddy, V. 2018. Improving early grade reading in south africa. *3rd Grantee Final Report. New Delhi: International Initiative for Impact Evaluation*.

Valbuena, J., Mediavilla, M., Choi, Á. & Gil, M. 2021. Effects of grade retention policies: A literature review of empirical studies applying causal inference. *Journal of Economic Surveys,* 35(2):408–451.

van Buuren, S. 2018. *Flexible imputation of missing data.* Boca Raton: CRC Press.

van Buuren, S. & Groothuis-Oudshoorn, K. 2011. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software,* 45(3):1 – 67. https://www.jstatsoft.org/index.php/jss/article/view/v045i03

van der Berg, S. & Gustafsson, M. 2019. Educational outcomes in post-apartheid south africa: Signs of progress despite great inequality, *South african schooling: The enigma of inequality: A study of the present situation and future possibilities.* Springer.

van der Berg, S., Hoadley, U., Galant, J., van Wyk, C. & Böhmer, B. 2022. Learning losses from covid-19 in the western cape: Evidence from systemic tests. *Research on Socio Economic Policy (Resep), Stellenbosch University February*.

van der Berg, S., van Wyk, C. & van Biljon, C. 2025. The impact of the 2023 back-on-track programme on learning. Stellenbosch: Research on Socio-Economic Policy, Stellenbosch.

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. www.ekon.sun.ac.za/wpapers/2019/wp132019

van der Berg, S., Wyk, C. v., Gustafsson, M., Meyer, H., Chari, A., Biljon, C. v., Lilenstein, A., Selkirk, R. & McCallum, J. 2023. What rich new education data can tell us. Stellenbosch: Research on Socio-Economic Policy, Stellenbosch University. https://resep.sun.ac.za/wp-content/uploads/2023/12/ReSEP-MSDF-2023-Report_WEB-UPDATED-1.pdf

van Staden, S. & Gustafsson, M. 2022. What a decade of pirls results reveal about early grade reading in south africa: 2006, 2011, 2016, in Spaull, N. & Pretorius, E. (eds.) *Early grade reading in south africa.* Cape Town: Oxford University Press.

von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., Bookbinder, A., Bezirhan, U. & Yin, L. 2024. Timss 2023 international results in mathematics and science. Boston College: TIMSS & PIRLS International Study Center.

Wills, G. 2023. Early grade repetition in south africa: Implications for reading. Stellenbosch: Research on Socio-Economic Policy, Stellenbosch University.

Wills, G. 2025. Gender and equity in south african education. Stellenbosch: Research on Socio-Economic Policy, Stellenbosch University.

Wills, G., Selkirk, R. & Kruger, J. 2024. School completion, the matric and post-school transitions in south africa. Stellenbosch: Resep.

Winters, M. A. & Greene, J. P. 2012. The medium-run effects of florida's test-based promotion policy. *Education finance and policy.,* 7(3):305–330.

Wu, W., West, S. G., Hughes, J. N. & Graesser, A. C. 2010. Effect of grade retention in first grade on psychosocial outcomes. *Journal of educational psychology,* 102(1):135–152.

Wu, W., West, S. G., Hughes, J. N. & Harris, K. R. 2008. Effect of retention in first grade on children's achievement trajectories over 4 years: A piecewise growth analysis using propensity score matching. *Journal of educational psychology,* 100(4):727–740.

Zhong, J. 2024. Early grade retention harms adult earnings.

9 APPENDIX

9.1 Additional tables and figures

Figure A 1. Discontinuities in Grade 2, 3 and 4 outcomes in the Foundation Phase Panel (LMA Subsample)

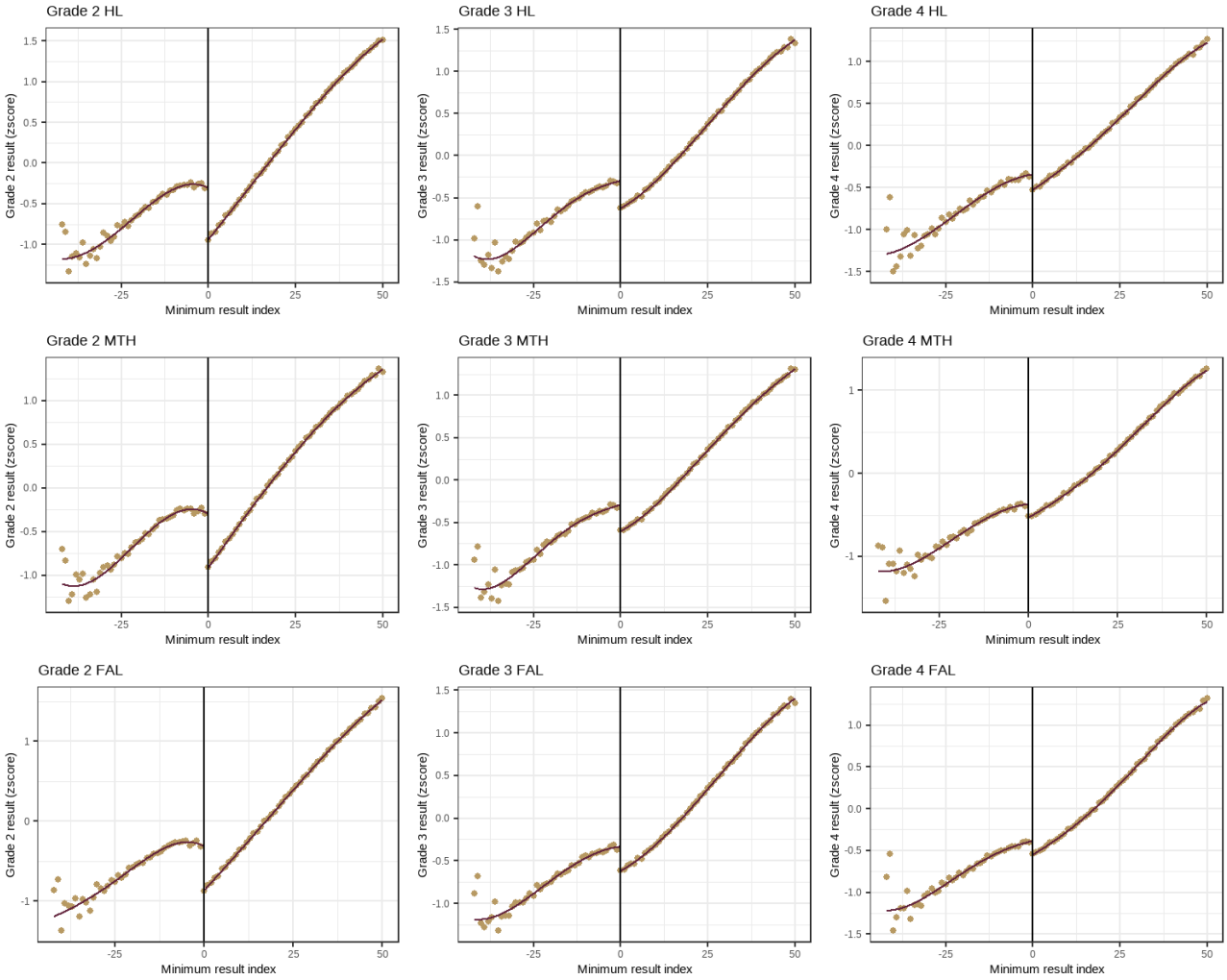


Figure A 2. Discontinuities in Grade 5, 6 and 7 outcomes in the Intermediate Phase Panel, LMA Subsample

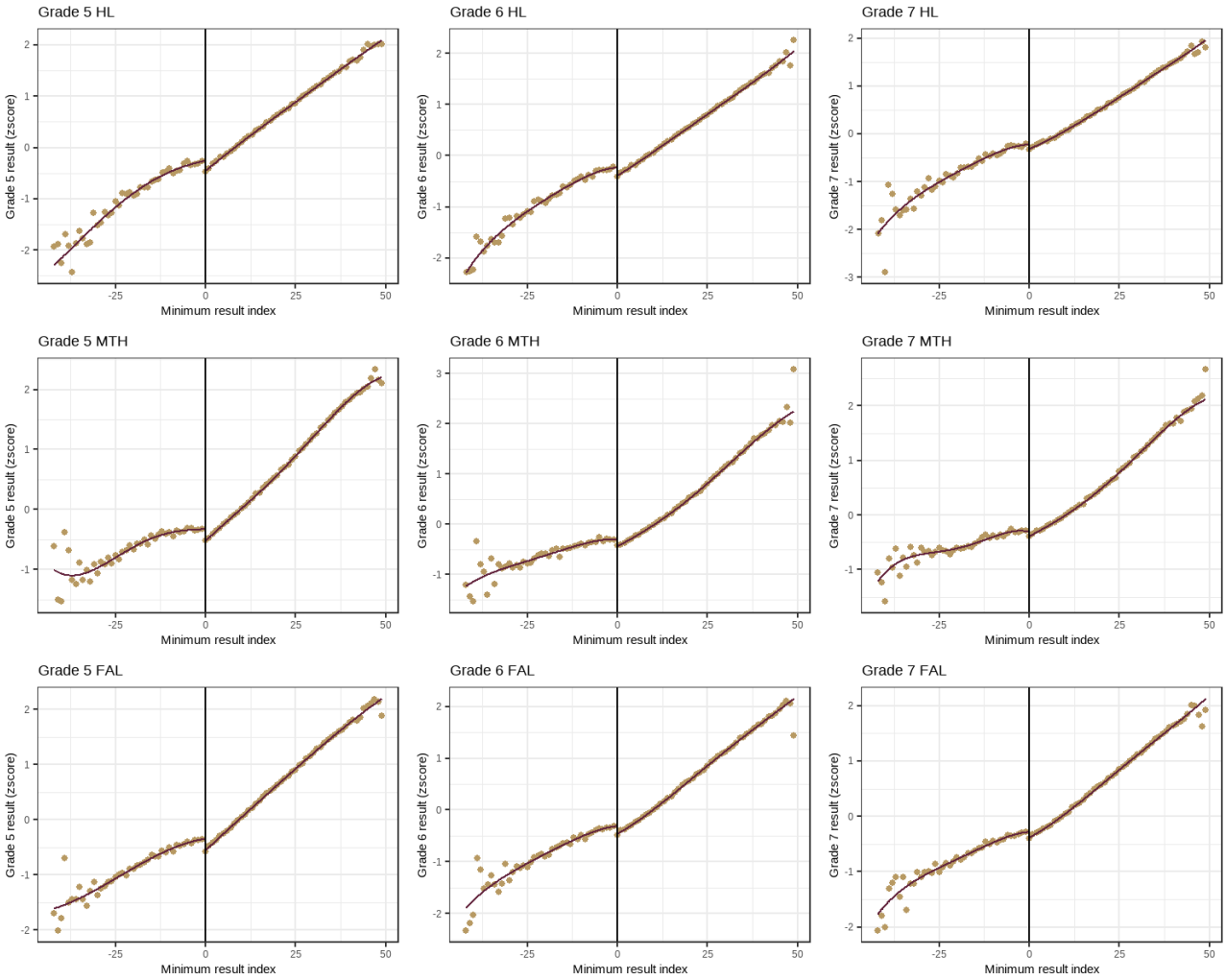


Figure A 3. Treatment discontinuities in all three samples

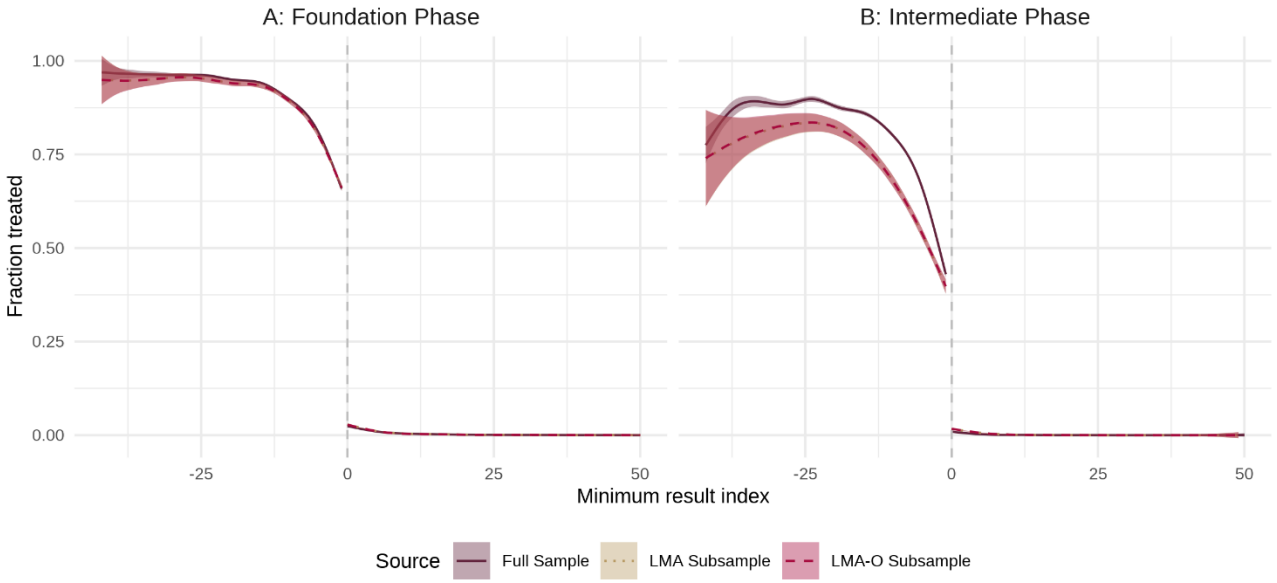


Figure A 4. Outcome discontinuities in all three data sources

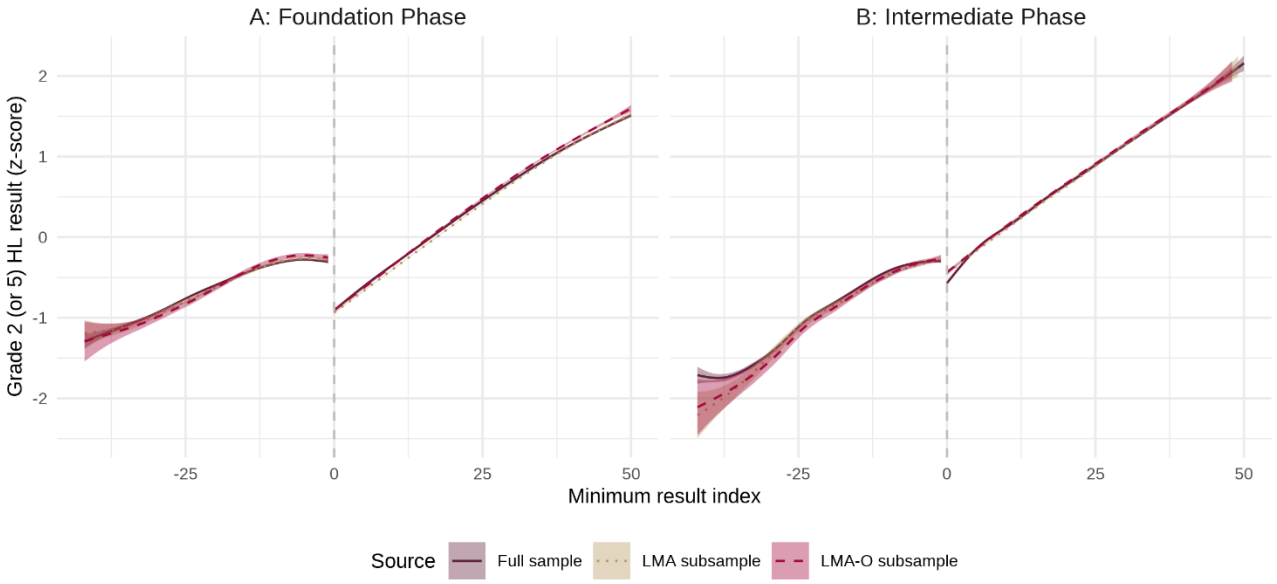


Table A 1. Estimated effect of Grade 1 repetition (all specifications and samples)

Source	Grade 2				Grade 3				Grade 4			
	Parametric (no FE)	Parametric (with FE)	Nonparam (no FE)	Nonparam (with FE)	Parametric (no FE)	Parametric (with FE)	Nonparam (no FE)	Nonparam (with FE)	Parametric (no FE)	Parametric (with FE)	Nonparam (no FE)	Nonparam (with FE)
HL												
Full Sample	1.02 (0.018) F = 6 767	1.09 (0.016) F = 7 554	1.03 (0.016) F = 6 191	1.17 (0.015) F = 5 421	0.58 (0.021) F = 5 966	0.56 (0.019) F = 6 590	0.59 (0.019) F = 5 610	0.61 (0.018) F = 4 595	0.32 (0.018) F = 8 624	0.33 (0.014) F = 9 609	0.33 (0.017) F = 7 355	0.39 (0.013) F = 6 490
LMA Subsample	1.12 (0.029) F = 2 896	1.13 (0.028) F = 2 997	1.12 (0.031) F = 2 416	1.15 (0.026) F = 2 888	0.55 (0.029) F = 2 860	0.53 (0.028) F = 2 939	0.54 (0.033) F = 2 371	0.54 (0.027) F = 2 847	0.32 (0.027) F = 3 264	0.29 (0.025) F = 3 390	0.33 (0.032) F = 2 723	0.32 (0.023) F = 3 201
LMA-O Subsample	1.12 (0.041) F = 3 160	1.18 (0.040) F = 3 315	1.14 (0.046) F = 1 061	1.21 (0.038) F = 1 281	0.53 (0.054) F = 3 236	0.58 (0.053) F = 3 380	0.51 (0.059) F = 775	0.57 (0.046) F = 957	0.28 (0.053) F = 3 246	0.26 (0.050) F = 3 386	0.27 (0.062) F = 919	0.26 (0.045) F = 1 062
MTH												
Full Sample	1.04 (0.019) F = 6 767	1.07 (0.016) F = 7 554	1.06 (0.018) F = 6 204	1.15 (0.015) F = 5 438	0.60 (0.021) F = 6 222	0.54 (0.018) F = 6 893	0.60 (0.020) F = 5 777	0.59 (0.017) F = 4 824	0.28 (0.019) F = 7 689	0.31 (0.015) F = 8 561	0.29 (0.018) F = 6 793	0.35 (0.013) F = 6 089
LMA Subsample	1.08 (0.026) F = 3 171	1.07 (0.025) F = 3 309	1.09 (0.031) F = 2 602	1.11 (0.023) F = 3 099	0.53 (0.027) F = 3 241	0.51 (0.025) F = 3 367	0.52 (0.032) F = 2 644	0.51 (0.023) F = 3 137	0.30 (0.025) F = 3 413	0.27 (0.022) F = 3 548	0.29 (0.031) F = 2 850	0.28 (0.020) F = 3 237
LMA-O Subsample	1.12 (0.032) F = 2 906	1.11 (0.030) F = 3 057	1.11 (0.036) F = 1 906	1.12 (0.028) F = 2 258	0.49 (0.034) F = 3 252	0.46 (0.033) F = 3 356	0.50 (0.041) F = 1 675	0.48 (0.030) F = 1 994	0.32 (0.038) F = 3 317	0.29 (0.034) F = 3 454	0.32 (0.048) F = 1 364	0.30 (0.031) F = 1 463
FAL												
Full Sample	0.94 (0.019) F = 6 730	0.96 (0.016) F = 7 602	0.96 (0.018) F = 6 176	1.05 (0.015) F = 5 002	0.53 (0.017) F = 7 666	0.52 (0.015) F = 8 578	0.52 (0.016) F = 6 645	0.56 (0.013) F = 5 831	0.27 (0.018) F = 7 666	0.30 (0.014) F = 8 578	0.27 (0.018) F = 6 795	0.34 (0.013) F = 5 968
LMA Subsample	0.98 (0.027) F = 2 998	0.97 (0.025) F = 3 133	0.98 (0.032) F = 2 498	1.00 (0.023) F = 3 014	0.49 (0.028) F = 2 896	0.47 (0.027) F = 2 997	0.48 (0.032) F = 2 408	0.47 (0.025) F = 2 866	0.30 (0.024) F = 3 413	0.28 (0.021) F = 3 548	0.29 (0.030) F = 2 874	0.29 (0.020) F = 3 269
LMA-O Subsample	1.01 (0.032) F = 2 872	1.01 (0.030) F = 2 990	1.01 (0.038) F = 1 811	1.03 (0.029) F = 2 172	0.50 (0.032) F = 3 002	0.50 (0.031) F = 3 155	0.48 (0.037) F = 1 724	0.49 (0.027) F = 2 072	0.33 (0.039) F = 3 246	0.32 (0.035) F = 3 386	0.31 (0.048) F = 1 228	0.31 (0.032) F = 1 428

Table A 2. Estimated impact of Grade 4 repetition (all specifications and samples)

Source	Grade 5				Grade 6				Grade 7			
	Parametric (no FE)	Parametric (with FE)	Nonparam (no FE)	Nonparam (with FE)	Parametric (no FE)	Parametric (with FE)	Nonparam (no FE)	Nonparam (with FE)	Parametric (no FE)	Parametric (with FE)	Nonparam (no FE)	Nonparam (with FE)
HL												
Full Sample	0.91 (0.037) F = 1 755	0.80 (0.027) F = 2 367	0.94 (0.033) F = 1 674	0.99 (0.030) F = 1 308	0.68 (0.035) F = 1 852	0.55 (0.026) F = 2 499	0.72 (0.032) F = 1 708	0.71 (0.027) F = 1 375	0.62 (0.034) F = 2 044	0.49 (0.024) F = 2 762	0.64 (0.030) F = 1 870	0.62 (0.024) F = 1 556
LMA Subsample	0.61 (0.058) F = 418	0.60 (0.052) F = 449	0.61 (0.083) F = 367	0.63 (0.050) F = 491	0.54 (0.059) F = 418	0.53 (0.054) F = 449	0.55 (0.080) F = 367	0.54 (0.047) F = 491	0.32 (0.062) F = 418	0.35 (0.054) F = 447	0.35 (0.083) F = 357	0.37 (0.047) F = 482
LMA-O Subsample	0.60 (0.071) F = 416	0.61 (0.063) F = 450	0.61 (0.106) F = 227	0.63 (0.065) F = 311	0.61 (0.081) F = 417	0.59 (0.071) F = 452	0.62 (0.121) F = 175	0.62 (0.072) F = 234	0.42 (0.082) F = 415	0.41 (0.074) F = 446	0.39 (0.109) F = 189	0.39 (0.068) F = 249
MTH												
Full Sample	0.84 (0.034) F = 1 758	0.76 (0.026) F = 2 322	0.86 (0.030) F = 1 697	0.87 (0.026) F = 1 390	0.62 (0.032) F = 1 939	0.50 (0.024) F = 2 553	0.63 (0.029) F = 1 798	0.58 (0.023) F = 1 555	0.47 (0.032) F = 2 051	0.38 (0.022) F = 2 673	0.47 (0.029) F = 1 899	0.44 (0.020) F = 1 675
LMA Subsample	0.64 (0.056) F = 418	0.63 (0.052) F = 450	0.65 (0.077) F = 369	0.63 (0.048) F = 492	0.48 (0.063) F = 418	0.46 (0.058) F = 447	0.48 (0.082) F = 357	0.47 (0.049) F = 481	0.39 (0.059) F = 416	0.39 (0.055) F = 449	0.39 (0.082) F = 365	0.38 (0.046) F = 490
LMA-O Subsample	0.63 (0.071) F = 416	0.62 (0.066) F = 448	0.65 (0.101) F = 212	0.63 (0.063) F = 285	0.43 (0.080) F = 420	0.44 (0.074) F = 455	0.45 (0.105) F = 204	0.45 (0.066) F = 273	0.33 (0.077) F = 414	0.38 (0.069) F = 443	0.31 (0.111) F = 230	0.33 (0.060) F = 302
FAL												
Full Sample	0.81 (0.034) F = 1 755	0.80 (0.026) F = 2 367	0.83 (0.031) F = 1 667	0.97 (0.029) F = 1 301	0.64 (0.033) F = 1 852	0.56 (0.025) F = 2 499	0.67 (0.029) F = 1 754	0.70 (0.025) F = 1 419	0.52 (0.034) F = 1 859	0.45 (0.024) F = 2 408	0.56 (0.031) F = 1 740	0.55 (0.024) F = 1 507
LMA Subsample	0.68 (0.054) F = 423	0.70 (0.051) F = 459	0.69 (0.074) F = 377	0.71 (0.050) F = 493	0.52 (0.056) F = 418	0.55 (0.052) F = 450	0.54 (0.075) F = 371	0.55 (0.047) F = 493	0.41 (0.055) F = 418	0.43 (0.052) F = 450	0.39 (0.080) F = 372	0.39 (0.046) F = 494
LMA-O Subsample	0.64 (0.072) F = 421	0.68 (0.068) F = 457	0.66 (0.096) F = 255	0.67 (0.066) F = 327	0.46 (0.078) F = 416	0.53 (0.073) F = 444	0.48 (0.092) F = 208	0.51 (0.062) F = 276	0.43 (0.079) F = 413	0.46 (0.074) F = 445	0.42 (0.120) F = 202	0.43 (0.067) F = 288

Table A 3. Bandwidths and sample size counts: Grade 1 repetition

Source	Grade 2	Grade 3	Grade 4
HL			
Full sample	$BW_L = 9.5$, $BW_R = 10.4$ $N_L = 66\,474$, $N_R = 311\,152$	$BW_L = 6.7$, $BW_R = 7.9$ $N_L = 43\,473$, $N_R = 210\,375$	$BW_L = 14.0$, $BW_R = 13.2$ $N_L = 107\,307$, $N_R = 415\,967$
LMA subsample	$BW_L = 8.2$, $BW_R = 8.7$ $N_L = 21\,045$, $N_R = 40\,771$	$BW_L = 7.1$, $BW_R = 8.5$ $N_L = 18\,642$, $N_R = 40\,771$	$BW_L = 11.7$, $BW_R = 12.1$ $N_L = 27\,576$, $N_R = 73\,097$
LMA-O subsample	$BW_L = 11.0$, $BW_R = 11.3$ $N_L = 25\,274$, $N_R = 62\,614$	$BW_L = 11.1$, $BW_R = 10.9$ $N_L = 27\,292$, $N_R = 55\,224$	$BW_L = 11.2$, $BW_R = 11.2$ $N_L = 27\,292$, $N_R = 62\,614$
MTH			
Full sample	$BW_L = 9.5$, $BW_R = 10.6$ $N_L = 66\,474$, $N_R = 311\,152$	$BW_L = 7.6$, $BW_R = 8.4$ $N_L = 50\,242$, $N_R = 242\,303$	$BW_L = 11.9$, $BW_R = 12.3$ $N_L = 83\,789$, $N_R = 384\,421$
LMA subsample	$BW_L = 10.1$, $BW_R = 11.1$ $N_L = 25\,529$, $N_R = 63\,594$	$BW_L = 11.6$, $BW_R = 10.1$ $N_L = 27\,576$, $N_R = 56\,100$	$BW_L = 14.2$, $BW_R = 12.1$ $N_L = 33\,439$, $N_R = 73\,097$
LMA-O subsample	$BW_L = 8.8$, $BW_R = 10.8$ $N_L = 20\,838$, $N_R = 55\,224$	$BW_L = 12.3$, $BW_R = 8.4$ $N_L = 29\,354$, $N_R = 40\,145$	$BW_L = 12.2$, $BW_R = 12.6$ $N_L = 29\,354$, $N_R = 71\,948$
FAL			
Full sample	$BW_L = 9.7$, $BW_R = 8.3$ $N_L = 66\,474$, $N_R = 242\,303$	$BW_L = 11.2$, $BW_R = 12.0$ $N_L = 83\,789$, $N_R = 344\,360$	$BW_L = 12.0$, $BW_R = 11.4$ $N_L = 83\,789$, $N_R = 344\,360$
LMA subsample	$BW_L = 9.3$, $BW_R = 9.4$ $N_L = 23\,136$, $N_R = 47\,603$	$BW_L = 8.7$, $BW_R = 8.2$ $N_L = 21\,045$, $N_R = 40\,771$	$BW_L = 14.6$, $BW_R = 12.4$ $N_L = 33\,439$, $N_R = 73\,097$
LMA-O subsample	$BW_L = 8.4$, $BW_R = 8.5$ $N_L = 20\,838$, $N_R = 40\,145$	$BW_L = 9.6$, $BW_R = 11.3$ $N_L = 22\,903$, $N_R = 62\,614$	$BW_L = 11.9$, $BW_R = 11.9$ $N_L = 27\,292$, $N_R = 62\,614$

Table A 4. Bandwidths and sample size counts: Grade 4 repetition

Source	Grade 5	Grade 6	Grade 7
HL			
Full sample	$BW_L = 7.3$, $BW_R = 5.5$ $N_L = 36\,094$, $N_R = 262\,663$	$BW_L = 8.1$, $BW_R = 6.7$ $N_L = 41\,203$, $N_R = 299\,417$	$BW_L = 10.4$, $BW_R = 7.5$ $N_L = 52\,386$, $N_R = 336\,127$
LMA subsample	$BW_L = 11.7$, $BW_R = 11.9$ $N_L = 9\,372$, $N_R = 25\,931$	$BW_L = 11.8$, $BW_R = 11.8$ $N_L = 9\,372$, $N_R = 25\,931$	$BW_L = 10.7$, $BW_R = 10.9$ $N_L = 8\,871$, $N_R = 22\,901$
LMA-O subsample	$BW_L = 11.6$, $BW_R = 12.3$ $N_L = 9\,305$, $N_R = 28\,713$	$BW_L = 12.0$, $BW_R = 12.0$ $N_L = 9\,744$, $N_R = 28\,713$	$BW_L = 11.7$, $BW_R = 11.5$ $N_L = 9\,305$, $N_R = 25\,651$
MTH			
Full sample	$BW_L = 7.9$, $BW_R = 6.7$ $N_L = 36\,094$, $N_R = 299\,417$	$BW_L = 9.3$, $BW_R = 8.6$ $N_L = 46\,869$, $N_R = 372\,971$	$BW_L = 10.8$, $BW_R = 9.2$ $N_L = 52\,386$, $N_R = 409\,836$
LMA subsample	$BW_L = 12.1$, $BW_R = 11.8$ $N_L = 9\,814$, $N_R = 25\,931$	$BW_L = 10.7$, $BW_R = 10.8$ $N_L = 8\,871$, $N_R = 22\,901$	$BW_L = 11.6$, $BW_R = 10.9$ $N_L = 9\,372$, $N_R = 22\,901$
LMA-O subsample	$BW_L = 12.5$, $BW_R = 11.1$ $N_L = 9\,744$, $N_R = 25\,651$	$BW_L = 12.2$, $BW_R = 13.0$ $N_L = 9\,744$, $N_R = 31\,784$	$BW_L = 10.1$, $BW_R = 10.1$ $N_L = 8\,807$, $N_R = 22\,641$
FAL			
Full sample	$BW_L = 7.2$, $BW_R = 5.4$ $N_L = 36\,094$, $N_R = 262\,663$	$BW_L = 8.7$, $BW_R = 6.9$ $N_L = 41\,203$, $N_R = 299\,417$	$BW_L = 8.4$, $BW_R = 8.0$ $N_L = 41\,203$, $N_R = 372\,971$
LMA subsample	$BW_L = 13.2$, $BW_R = 13.4$ $N_L = 10\,263$, $N_R = 32\,132$	$BW_L = 12.4$, $BW_R = 11.4$ $N_L = 9\,814$, $N_R = 25\,931$	$BW_L = 13.0$, $BW_R = 11.0$ $N_L = 9\,814$, $N_R = 25\,931$
LMA-O subsample	$BW_L = 12.8$, $BW_R = 14.1$ $N_L = 9\,744$, $N_R = 34\,908$	$BW_L = 10.6$, $BW_R = 11.6$ $N_L = 8\,807$, $N_R = 25\,651$	$BW_L = 11.3$, $BW_R = 11.0$ $N_L = 9\,305$, $N_R = 22\,641$

9.2 Imputation approach to addressing mark adjustment

Initially I had attempted to correct for the measurement error induced by the mark adjustments by using auxiliary data (learner term results) and applying multiple imputation techniques (Rubin, 1987; van Buuren, 2018). Ultimately, however, this approach introduced inconsistencies in the Grade 4 repetition estimates which led me to abandon the analysis. I nonetheless provide details of the approach, as well as the results.

I assume that the adjustment occurs only for marks that are initially below the promotion threshold, is upward in direction, and is limited to values within two percentage points[14] above the cutoff. To identify likely adjusted scores, I first estimate an OLS regression with each outcome variable (Grade 1 (or 4) subject result index) for each province, year started Grade 1 (or 4), and quintile combination[15], with age, gender, ethnicity and all Grade 1 (or 4) term 1 to term 3 marks as covariates. Observed values between zero and two but with predicted values below zero are classified as potentially adjusted. These suspected values are set to missing and imputed using the same covariates, along with all variables from the main estimation model (Equation 1), to avoid attenuation bias (Moons et al., 2006). Additionally, approximately 10% of term marks are missing, and these are imputed in a first stage (using the other term marks). Imputation is conducted via the mice package in R (van Buuren & Groothuis-Oudshoorn, 2011), employing Bayesian linear regression and producing five imputed datasets. I then re-estimate the equations using the imputed datasets.

To address possible adjustment of the running variable, multiple imputation is applied to predict results that were identified as potentially adjusted. For the purpose of these descriptive statistics, I use the mean outcomes across the five imputed datasets. The imputation substantially reduced but did not eliminate evidence of manipulation, as seen by the sharpness of the density functions around zero (Figure A 5); presence of manipulation is confirmed in both the Foundation and Intermediate Phase Panels with the rddensity test. This is due, at least in part, to the nature of Bayesian regression; the random component, in conjunction with the presence of "heaping" in the training data, means that smoothness is not achieved (in comparison to deterministic linear regression, which does produce smooth distributions, but which cannot be used for inference).

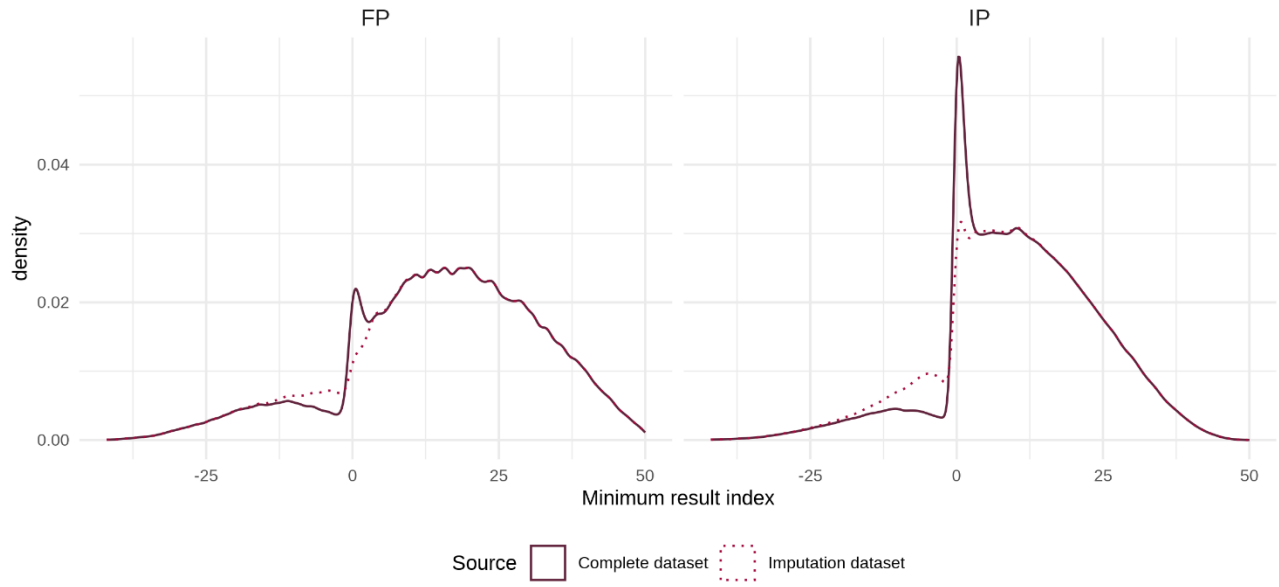
Table A 5. Counts of imputed values

	Foundation Phase Panel			Intermediate Phase Panel		
Cohort	N imputed values	Imputed values as % of eligible values	Imputed values as % of all values	N imputed values	Imputed values as % of eligible values	Imputed values as % of all values
2017	15 359	33.8	3.2	25 445	36.6	6.9
2018	15 835	34.0	3.1	27 613	35.7	6.6
2019	15 173	34.9	2.9	27 595	36.1	6.4
Total	**46 367**	**34.2**	**3.1**	**80 653**	**36.1**	**6.6**

Notes: "Eligible" values are the individual subject result indices that lie between 0 and 2.

[14] This is based on a visual inspection of the spike in the distribution of the running variable.
[15] I estimated 6 provinces × 3 cohorts × 5 quintiles = 90 linear models using OLS, to estimate the most accurate models.

Figure A 5. Density of the running variable, imputed v. complete panel



Source: DDD panel and mean values across the five imputed datasets. Notes: Kernel density estimates are plotted using Gaussian kernels with data-driven bandwidth selection. FP = Foundation Phase, IP = Intermediate Phase.

Results were only eligible for imputation if they were between zero and two percentage points above the cutoff, and thus the characteristics of learners whose marks were imputed are compared to the characteristics of learners whose marks were not imputed and whose *minimum result index* was of a magnitude that made them eligible for imputation, reported in Table A 6. In the FP panel, 23% of results in this interval were imputed, compared to 26% in the IP panel. Males are more likely to have results imputed (+4 p.p. in FP and +8 p.p. in IP), and there is very little practical difference in imputation rates by age and ethnicity. In the FP panel Quintile 3 learners are 2 p.p. more likely to have their results imputed, and Quintile 5 learners 2 p.p. less likely; in the IP, panel Quintile 5 learners are 2 p.p. less likely to have their results imputed, and imputation rates are similar in other school quintiles. Learners in Gauteng are 3 p.p. less likely to have their results imputed in the Foundation Phase Panel, with only small differences in other provinces; in the IP panel Eastern Cape and Gauteng learners are about 3 p.p. less likely to have their results imputed, and learners in Limpopo are close to 5 p.p. more likely.

Table A 6. Characteristics of learners whose marks were imputed due to suspected adjustment, v. those whose marks were not imputed (only learners with a subject index between 0 and 2)

	FP (Repeating grade: Grade 1)					IP (Repeating grade: Grade 4)				
	1. Results imputed		2. Rest of panel (0-2)		Difference	1. Results imputed		2. Rest of panel (0-2)		Difference
	Mean / SE	N	Mean / SE	N	(1) - (2)	Mean / SE	N	Mean / SE	N	(1) - (2)
Female	0.373 [0.0018]	46 367	0.380 [0.0025]	89 139	-0.007**	0.368 [0.0016]	80 653	0.405 [0.0021]	142 744	-0.037***
Age when starting Grade 1 (or 4)	6.047 [0.0022]	46 367	6.044 [0.0029]	89 139	0.003	9.279 [0.0017]	80 653	9.265 [0.0022]	142 744	0.014***
African/Black	0.959 [0.0024]	46 367	0.961 [0.0028]	89 139	-0.002	0.959 [0.0018]	80 653	0.965 [0.0026]	142 744	-0.005***
Asian/Indian	0.003 [0.0003]	46 367	0.002 [0.0003]	89 139	0.000	0.004 [0.0003]	80 653	0.003 [0.0005]	142 744	0.000
Coloured	0.030 [0.0022]	46 367	0.028 [0.0027]	89 139	0.002	0.030 [0.0016]	80 653	0.023 [0.0025]	142 744	0.007***
White	0.007 [0.0007]	46 367	0.007 [0.0007]	89 139	-0.000	0.006 [0.0007]	80 653	0.009 [0.0005]	142 744	-0.003***
School Quintile 1	0.274 [0.0057]	46 367	0.286 [0.0066]	89 139	-0.013***	0.277 [0.0058]	80 653	0.290 [0.0068]	142 744	-0.013***
School Quintile 2	0.261 [0.0058]	46 367	0.259 [0.0068]	89 139	0.002	0.269 [0.0058]	80 653	0.267 [0.0071]	142 744	0.002
School Quintile 3	0.301 [0.0065]	46 367	0.291 [0.0079]	89 139	0.010**	0.289 [0.0063]	80 653	0.285 [0.0076]	142 744	0.004
School Quintile 4	0.091 [0.0044]	46 367	0.088 [0.0053]	89 139	0.003	0.089 [0.0042]	80 653	0.086 [0.0048]	142 744	0.003
School Quintile 5	0.072 [0.0040]	46 367	0.075 [0.0045]	89 139	-0.003	0.075 [0.0035]	80 653	0.072 [0.0042]	142 744	0.003
Eastern Cape	0.269 [0.0060]	46 367	0.280 [0.0072]	89 139	-0.011**	0.183 [0.0048]	80 653	0.208 [0.0055]	142 744	-0.025***
Gauteng	0.151 [0.0059]	46 367	0.172 [0.0061]	89 139	-0.021***	0.177 [0.0057]	80 653	0.171 [0.0071]	142 744	0.006
KwaZulu-Natal	0.195 [0.0050]	46 367	0.179 [0.0062]	89 139	0.017***	0.208 [0.0051]	80 653	0.187 [0.0064]	142 744	0.021***
Limpopo	0.202 [0.0052]	46 367	0.190 [0.0063]	89 139	0.012***	0.252 [0.0058]	80 653	0.244 [0.0070]	142 744	0.008*
Mpumalanga	0.105 [0.0042]	46 367	0.103 [0.0054]	89 139	0.002	0.094 [0.0040]	80 653	0.099 [0.0046]	142 744	-0.005
North West	0.077 [0.0034]	46 367	0.076 [0.0040]	89 139	0.001	0.086 [0.0039]	80 653	0.091 [0.0044]	142 744	-0.005**

Source: DDD panel and imputed datasets. Notes: FP = Foundation Phase, IP = Intermediate Phase. Stars indicate significance at the *** 1% ** 5% and * 10% critical level.

Figure A 6. Grade 1 repetition effects (Imputation datasets)

	Grade 2		Grade 3		Grade 4	
	Parametric (with FE)	Nonparametric (with FE)	Parametric (with FE)	Nonparametric (with FE)	Parametric (with FE)	Nonparametric (with FE)
HL						
Treatment effect (τ)	1.09	1.07	0.47	0.39	0.21	0.13
Standard error	(0.029)	(0.040)	(0.032)	(0.050)	(0.031)	(0.042)
Sample size	L = 96 834, R = 204 922	L = 96 834, R = 204 922	L = 96 834, R = 143 942	L = 96 834, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942
BW	L = 9.0, R = 8.8	L = 9.0, R = 8.8	L = 9.4, R = 7.0	L = 9.4, R = 7.0	L = 7.9, R = 7.6	L = 7.9, R = 7.6
F-statistic	3 071	2 215	2 841	1 793	2 616	1 944
MTH						
Treatment effect (τ)	1.06	1.05	0.43	0.33	0.18	0.11
Standard error	(0.031)	(0.039)	(0.031)	(0.047)	(0.033)	(0.041)
Sample size	L = 76 718, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942
BW	L = 8.7, R = 8.7	L = 8.7, R = 8.7	L = 7.9, R = 7.2	L = 7.9, R = 7.2	L = 8.2, R = 8.2	L = 8.2, R = 8.2
F-statistic	2 616	2 214	2 616	1 906	2 616	2 135
FAL						
Treatment effect (τ)	0.95	0.90	0.40	0.34	0.18	0.10
Standard error	(0.036)	(0.042)	(0.029)	(0.049)	(0.034)	(0.041)
Sample size	L = 76 718, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942	L = 76 718, R = 143 942
BW	L = 9.2, R = 9.2	L = 9.2, R = 9.2	L = 7.2, R = 6.8	L = 7.2, R = 6.8	L = 10.5, R = 8.5	L = 10.5, R = 8.5
F-statistic	2 616	2 397	2 616	1 667	2 616	2 347

Source: Datasets derived from DDD panel.

Figure A 7. Grade 4 repetition effects (Imputation datasets)

	Grade 5		Grade 6		Grade 7	
	Parametric (with FE)	Nonparametric (with FE)	Parametric (with FE)	Nonparametric (with FE)	Parametric (with FE)	Nonparametric (with FE)
HL						
Treatment effect (τ)	0.12	0.07	0.07	0.06	0.07	0.04
Standard error	(0.058)	(0.059)	(0.059)	(0.071)	(0.052)	(0.071)
Sample size	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084
BW	L = 8.5, R = 9.0	L = 8.5, R = 9.0	L = 8.5, R = 8.2	L = 8.5, R = 8.2	L = 8.5, R = 8.2	L = 8.5, R = 8.2
F-statistic	1 077	998	1 031	914	1 031	1 034
MTH						
Treatment effect (τ)	0.36	0.38	0.20	0.07	0.10	0.06
Standard error	(0.068)	(0.085)	(0.064)	(0.078)	(0.053)	(0.064)
Sample size	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084
BW	L = 8.6, R = 9.4	L = 8.6, R = 9.4	L = 8.6, R = 9.4	L = 8.6, R = 9.4	L = 8.6, R = 9.4	L = 8.6, R = 9.4
F-statistic	1 031	971	1 031	1 008	1 031	1 010
FAL						
Treatment effect (τ)	0.13	0.08	0.10	0.07	0.10	0.02
Standard error	(0.062)	(0.060)	(0.066)	(0.069)	(0.050)	(0.068)
Sample size	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084	L = 85 542, R = 269 084
BW	L = 8.5, R = 9.0	L = 8.5, R = 9.0	L = 8.5, R = 9.4	L = 8.5, R = 9.4	L = 8.5, R = 9.4	L = 8.5, R = 9.4
F-statistic	1 031	1 001	1 031	964	1 031	1 024

Source: Datasets derived from DDD panel.