ESSAYS ON THE PROMOTION OF FOUNDATIONAL LITERACY AND NUMERACY IN DEVELOPING COUNTRIES

A Dissertation

Presented to

The Faculty of the Curry School of Education and Human Development

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Daniel Rodriguez-Segura

April 2022

Daniel Rodriguez-Segura
Education Policy
School of Education and Human Development
University of Virginia
Charlottesville, Virginia

APPROVAL OF THE DISSERTATION

This dissertation, ("Essays on the Promotion of Foundational Literacy and Numeracy in Developing Countries"), has been approved by the Graduate Faculty of the School of Education and Human Development in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Beth E. Schueler (Co-Chair)

_____
Daphna Bassok (Co-Chair)

_____
Vivian C. Wong

_____
Isaac M. Mbiti (Frank Batten School of Leadership and Public Policy)

_____Date

DEDICATION

Le quiero dedicar esta tesis a mis papás, Isabel y Arturo, y mi abuelo, Javier. Ellos siempre han puesto todo lo que han tenido, y hasta lo que no, para darme todo lo que he necesitado. Sin ellos, la formación que me dieron y su apoyo incondicional, nunca hubiera podido llegar a donde estoy hoy. No sé cómo pagarles por todo lo que han hecho por mí, y no creo que sea posible. Por eso, sólo espero que estén orgullosos de lo que he podido hacer con lo mucho que me han dado. Gracias, los amo mucho.

Le quiero dedicar esta tesis también a la mejor hermana del mundo, Cristina que, entre risas, peleas y cocinazos, ilumina mi vida.

To my wonderful wife Caroline, whose love and support have made me incredibly happy for the past five years. I am so, so glad that you started that conversation at Mudhouse, and that through our mutual support, we were both able to finish our PhDs. I can't wait for the rest of our lives together. I love you.

Finally, to the two most loving dogs in the world, Lucy and Rafita, who, through cuddles, barks and lickies, know exactly how to cheer me up every day.

ACKNOWLEDGEMENTS

I was able to complete this dissertation, and my doctoral studies more broadly, because of the help that I received from many individuals and institutions. First, I am deeply indebted to Beth Schueler, my wonderful PhD advisor. Throughout my time at UVA, Beth gave me fantastic guidance, feedback, and support. Beth encouraged me to consider the big picture and how it all fits together, while also making sure that I was thinking hard about all the important details. A very large part of why I was able to successfully complete a dissertation that I am proud of is because of Beth's tireless devotion to my personal and professional growth. Every round of insightful and compassionate feedback helped me get better, and made me a more thoughtful and well-rounded researcher. I am also especially grateful to Beth for taking a chance on me despite the small overlap in our main research interests, and for being so supportive in actively finding projects that we would both find valuable, interesting, and rewarding.

I would also like to thank my dissertation committee – Daphna Bassok, Vivian Wong, and Isaac Mbiti – for all their time and willingness to support my work. Their feedback, comments, and advice significantly strengthened this dissertation, and my work beyond it. Ben Castleman was also a thoughtful and generous advisor during my first years in the program. Ben helped me better understand how to frame research questions, how to filter the promising ideas from the not-so-promising ones, and how to interrogate the data to extract valuable insights.

My classmates within the Education Policy program were also a constant source of support and encouragement. I am particularly grateful to Brian Kim, Walter Herring, and Arielle Boguslav, whose friendship and partnership helped me navigate some of the highs and lows of this process. Katharine Meyer was an outstanding mentor throughout my first years in the program.

I am also grateful for the support of many institutions throughout this journey. The School of Education and Human Development at the University of Virginia was a wonderful place to complete a PhD in a kind and interdisciplinary environment. The RISE Programme allowed me meet brilliant researchers from all over the world, and provided a generous platform for me to disseminate my work. I would like to also thank the research partners that generously shared their data and became valuable thought-partners through my doctoral studies. Cooperative for Education in Guatemala, NewGlobe in Kenya and Nigeria, the Costa Rican Ministry of Education, Building for Tomorrow in Uganda, and Twaweza in Tanzania all helped me access datasets to conduct several research projects throughout my studies. I hope that my work was at least a little bit as valuable to them, as their support was to me.

Finally, I have been the beneficiary of incredibly generous financial support from different organizations and institutions throughout the past 15 years, which has allowed me to further my education to a degree that 14-year old me would have never imagined. I am eternally grateful for the Founders' Scholarship Fund at Lincoln School in Costa Rica, the Johnson Scholarship at Washington and Lee University, the financial support

from the Graduate School of Arts and Sciences at the University of Virginia, as well as the Dean's Fellowship at the School of Education and Human Development at the University of Virginia. I cannot imagine my "counterfactual" had these institutions and the people behind them not placed a bet on me when they did. I hope I can give back as much as I have been given.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DISSERTATION OVERVIEW

Primary Education in the Developing World

Primary schools in developing countries serve a large share of the global population. Nine of every ten children under 15 live in a low- and middle-income country (LMIC), and 91% of them attend primary school (World Bank, 2021a, 2021b). These figures are expected to further increase in coming decades, as the population growth rate in these contexts is four times faster than that in high-income countries (World Bank, 2021c). In all, over half a billion children – 8% of the world population – are *currently* enrolled in a primary school in a developing country (UNESCO, 2019), and a much larger share of the global population will have been through primary school in a LMIC at some point in their lives. Given that education has long been considered an engine towards economic growth and social mobility (Chetty et al., 2020; Cunha and Heckman, 2007; Heckman, 2006; Montenegro and Patrinos, 2014), especially in contexts where extreme poverty and inequality are rampant, improving the quality of basic education can be both an effective policy goal in itself, and also a wide-reaching tool to improve human welfare more broadly (Hanushek and Woessmann, 2007; Ozturk, 2001). Yet, education policymakers working in these contexts face pernicious constraints in terms of the institutional capacity, initial human capital of the populations they serve, and the evidence base on which they can draw to design thoughtful policies. As such, creating research that is grounded in LMIC contexts is a valuable step towards improving

educational quality, and through this dissertation, I offer three chapters on the issue of learning in developing countries.

More specifically, this dissertation focuses on foundational literacy and numeracy (FLN) outcomes in primary schools in LMIC, and aims to build on the body of evidence on improving and measuring these skills. Strengthening these skills has received increasing research and policy attention in recent decades as one of the most pressing policy goals to improve educational quality in developing contexts more broadly (Evans and Hares, 2021). There are at least three key reasons why FLN should be front and center in the agendas of education policymakers and researchers in developing contexts – namely, the weak current levels of learning, the potential gains in equity achieved through stronger FLN, and the broader systematic diagnostics policymakers can obtain by measuring these skills.

The "Learning Crisis" and Weak FLN Outcomes Around the World

There is ample and growing documentation that the modal experience for a child in the developing world is "schooling without learning" (Pritchett, 2013), or the fact that in spite of the large increases in school enrollment around the world, most children never end up acquiring FLN. For instance, while primary school enrollment in India was at 97% in 2019, up from 79% in 1971 (World Bank, 2021d), only little over a quarter of all grade 3 children in rural India can read at the grade 2 level, and by grade 8, 27% of children in rural India still cannot read at the grade 2 level (ASER, 2019). This experience is likely to be worsened by the pandemic-induced school closures, as preliminary evidence from the Indian state of Kartanaka suggests that as the share of grade 3 rural students being able to read a grade 2 passage fell from 19% in 2018 to 10% in 2020 (ASER, 2021). This "learning crisis", as it is called in policy circles, is

particularly worrying due to the extent to which FLN shapes later stages in people's academic and personal lives. For those in LMIC that leave the education system early, there is evidence suggesting that fewer than half of young adults who only reached primary school are functionally literate (Kaffenberger and Pritchett, 2020). This implies that governments and families invested valuable and limited resources on sending these children to school, with little return in terms of foundational skills for the majority of them. On the other hand, for those that do persist in the educational system, acquiring these skills can put them on a higher learning trajectory which ensures that they have a strong academic performance in later grades (Bau et al., 2021; Carter et al., 2020). Therefore, regardless of students' paths after the early grades, addressing the current gaps in foundational literacy and numeracy across much of the developing world is a promising path to improve life outcomes for most students in these contexts.

Pursuing Equitable *and* Efficient Policies by Targeting FLN

In the process of designing policies, decisionmakers typically face a dire trade-off between elements that promote "equity" and elements that promote "efficiency" (Okun, 2015). However, interventions that focus on improving FLN may be among the rare set of policies that can simultaneously address both. On the equity side, early-grade interventions in developing countries – such as those that focus on FLN – are likely to reach a wider range of socioeconomic backgrounds than other educational programs implemented in later grades. While this is the case in most educational settings, the significantly higher rates of dropout in LMIC make this phenomenon particularly pronounced. For instance, in Kenya and Tanzania, countries where two chapters of this dissertation are based, for every child from the poorest income quintile in the first two grades of primary school, there are 0.5 and 0.8 children from the richest income quintile

respectively. By the first two years of secondary school, this ratio becomes 2 and 4.4 children from the richest income quintile, respectively. In Tanzania, by the last grade of secondary school, fewer than 1% of all students come from the poorest income quintile (DHS, 2014, 2015-16). These differences in socioeconomic composition also overlap with large achievement gaps between socioeconomic groups. For instance, Spaull and Kotze (2015) suggest that the gap between the wealthiest income quintile and the three poorest quintiles in South Africa grows by over one grade-level between third and ninth grade. Yet, there is a growing body of evidence which suggests that interventions that simply aim to improve FLN *averages* can indeed also reduce learning *inequality* in the early grades, both in terms of achievement and socioeconomic gaps, and especially so when baseline performance is weak as it is in much of the developing world (Crouch et al., 2021; Rodriguez-Segura et al., 2021; Asim, 2020). Therefore, policies that act in the early grades to improve FLN are likely to be significantly more progressive than interventions targeted at later grades once the student body has a wealthier composition.

Given that FLN underpin much of children's educational and cognitive development, strengthening FLN can also lead to more productive citizens and stronger economic growth as these students join the workforce, advancing the "efficiency" side of policy planning. For instance, literacy is correlated with higher incomes (Hanushek, 2015;; Valerio et al., 2016), greater educational benefits for the next generation (Banerji et al., 2017, Andrabi et al. 2012 or Abuya et al. 2015), better health outcomes (Mathew, 2012; Mensch et al., 2019; Taylor et al., 2016), and higher agricultural productivity (Paltasingh and Goyari, 2018). Similarly, the development of these skills can lead to more cohesive societies with higher civic engagement, as skills such as reading with

comprehension and performing mathematical operations for real-world scenarios are key steps towards becoming a self-sufficient citizen and member of society. In aggregate, these gains are likely to increase the economic well-being of individuals and countries as a whole. As policymakers consider areas to invest, they should be aware that the promotion of FLN can yield gains on both sides of the dreaded policy trade-off between gains in equity and efficiency.

FLN as a Policy Target and a Diagnostic Tool

The fact that educational systems worldwide have a universal need to foster FLN makes these skills promising candidates to act as standardized benchmarks for the overall performance of educational systems, and as metrics to set clear learning goals for specific interventions. In practical terms, using FLN to track educational progress has at least three tangible benefits. First, as countries move away from education as a signaling or credentialing system and towards education being a path to acquire tangible skills, directly tracking the acquisition of these skills —as opposed to indirect outcomes like school enrollment— is a valuable way to ensure that educational systems indeed deliver on their promise. Second, using educational inputs such as school construction or expenditures as benchmarks for performance may be, at best, a noisy indicator of educational outputs. This is in part due to the fact there is now strong evidence that the mere availability of these inputs, without accompanying systemic or pedagogical changes, does not lead to improved academic performance (Masino and Niño-Zarazúa, 2016; Murnane and Ganimian, 2017; Evans and Mendez Acosta, 2021). Instead, monitoring outputs like learning can also incentivize a more informed and deliberate use of inputs, as ensuring mastery of FLN requires a minimum level of cohesion and alignment of these inputs with stakeholders and institutional policies (Azevedo et al.,

2021b). Finally, using learning as an outcome measure can also pressure policymakers to focus on equity. There is now extensive evidence that educational systems in LMIC, through implicit and explicit incentives, tend to cater to high-performing students (Glewwe et al., 2009; Glewwe and Muralidharan, 2016; Pritchett and Beatty, 2015; among others). Therefore, tracking inputs like "textbooks" may not reflect how these inputs are used, particularly along the lines of equity. Instead, monitoring learning requires policymakers to focus on the full range of students, as the large lower tails in the current distributions of learning in LMIC make it so that moving the overall sample average may be easier by focusing on the large number of low-achieving students than on the few high achieving students (Crouch et al., 2021; Rodriguez-Segura et al., 2021). In this sense, using system-level measures of performance on FLN can create strong incentives to get these systems to use resources more effectively, equitably, and ultimately to deliver adequate learning for all students.

These potential benefits of using FLN as trackers of learning have already moved some policymakers and donor organizations to adopt harmonized metrics of FLN achievement for national benchmarks of educational progress. For instance, the World Bank developed the notion of "learning poverty" (Azevedo et al., 2020a) – a metric which captures deficits in reading achievement and school enrollment. This specific metric was created to both bring attention to the extent of the "learning crisis", and to be able to set actionable goals for education projects with this indicator as a target. This indicator was developed to mirror important global commitments such as the Sustainable Development Goal 4 (United Nations, 2021), which have placed strong and equitable FLN outcomes at the forefront of global educational goals. Importantly, Azevedo et al.

(2021a) also show that the "learning poverty" metric is a good proxy for aggregate future performance and contemporaneous learning in other subjects. Furthermore, although the concept of "learning poverty" has grown in prominence and popularity, it only serves as an example of the use of FLN to track overall educational performance. In fact, learning poverty is not the only option that policymakers have pursued, and for instance, policymakers in Liberia and Kenya have also used a well-known metric of reading fluency, "correct words per minute", as another system-wide metric of educational achievement. These differing metrics illustrate that FLN can be a meaningful way to track progress and align incentives towards the needs of all students.

This Dissertation

As I argue here, foundational literacy and numeracy are pivotal for both child development and the broader strengthening of educational systems starting in the early grades. Investments in FLN can promote student well-being in the short- and long-run, and can do so while also advancing goals of equitable educational development. As such, in this dissertation I study topics around the promotion and measurement of foundational literacy and numeracy in LMIC through three studies in Costa Rica, Tanzania, and Kenya.

My first two chapters investigate how curriculum reforms affect learning, particularly the acquisition of FLN, in the early primary grades. These two chapters respond to the broad need to find large-scale interventions that can effectively improve FLN even by governments with relatively weak implementing capacity. Much research – such as the early research agenda around the "Teach at the Right Level" movement – has centered on documenting the principles of better content alignment with students' proficiencies through targeted and relatively small-scale interventions. However, there is

not much evidence to date on how curriculum reforms play out when deployed at a large-scale. This is partly due to the relative rarity of this scale for educational intervention, but also due to the difficulty of rigorously evaluating interventions that almost by definition are rolled out all at once. To fill this gap, I study nationwide curriculum reforms aimed at improving FLN and implemented by the governments of Costa Rica and Tanzania in 2014 and 2015 respectively. In both cases, I leverage fairly unique nationally representative panel data, and the exogenous variation created by the grades that were targeted and the timing of the policies to derive causal estimates of the effects of the reforms on student outcomes. Interestingly, these policies differed significantly in terms of the level of expectations from teachers and complexity levels for a successful implementation as designed. In turn, these differences led to very different achievement outcomes in both countries, yielding negative unintended consequences in Costa Rica, and positive effects on learning in Tanzania. Below, I describe these two chapters in greater depth, and provide some ideas on the potential policy features that led to their main results.

My first chapter is titled "Strengthening early literacy skills through social promotion policies? Intended and unintended consequences in Costa Rica." This chapter was published in the International Journal of Educational Development (Rodriguez-Segura, 2020). Historically, grade repetition has been between 4-7% for grade 1 students in Costa Rica, a higher rate than for any other primary school grade. This pattern is typically attributed to children not attaining sufficient literacy skills to move on to grade 2. In 2014, the government implemented a curricular reform with the goal of improving early literacy skills for children. This reform was coupled with a social promotion policy,

abolishing grade repetition in grade 1, with the goal of allowing students to develop their reading skills during grades 1 and 2 before they had any high-stakes outcomes linked to their reading proficiency. The curriculum reform outlined new instructional approaches teachers were expected to follow as they taught early literacy, but the policy did not have any accountability mechanisms to ensure changes were implemented. Ultimately, the social promotion portion of the policy was the only feature that ended up being closely followed by teachers. Using difference-in-differences methods, I find that instead of allowing children more time to catch up, this policy simply delayed grade repetition for many students, leading to increases in grade repetition once they were in grades 2 and 3. This pattern was most pronounced for the most socioeconomically disadvantaged communities. This paper highlights the role of high-quality support for teachers and schools as they implement a complex instructional change.

My second chapter is called "Back to the Basics: Curriculum Reform and Student Learning in Tanzania", co-authored with Dr. Isaac Mbiti. In response to very low learning levels in Tanzania, and a curriculum that was "overloaded" with subjects for grades 1 and 2, the Tanzanian government passed a curriculum reform which narrowed the instructional scope teachers were expected to cover so they could focus on early literacy and numeracy. This reform mandated that the other subjects were left to either be taught during Kiswahili class as readings, or postponed until grade 3. Observational data shows the policy increased instructional time for early literacy and numeracy by about 2 hours in a 15 hour-week. Difference-in-difference estimates illustrate that the reform improved performance in math and Kiswahili by approximately 0.2 SD, and also decreased the overall dropout rate. Larger gains were (noisily) correlated with teacher

training on the new curriculum. Four years after, the policy still caused more students to stay in school, although this was accompanied by decreased passing rates in the standardized national test in grade 4 – highlighting that even though curriculum reforms can yield higher learning outcomes at-scale, they may not be sufficient towards helping all children meet certain desired benchmarks. To the best of our knowledge, this is the first evaluation of a successful curriculum reform at a national level in a LMIC, serving as empirical support for the large but mostly theoretical literature on the benefits of better aligning the curriculum with students' needs, and as evidence that such a reform can indeed be carried out by governments at a national scale.

My third chapter is called "Assessors influence results: Evidence on enumerator effects and educational impact evaluations", and it is co-authored with Dr. Beth Schueler. This chapter is a methodological paper which investigates the issue of "enumerator effects" — inconsistent practices between the interviewers who administer questionnaires that can lead to increased measurement error— in a phone-based assessment on foundational numeracy. The issue of enumerator effects is particularly relevant for the measurement of FLN, as internationally-validated exams like the Early Grade Reading Assessment (EGRA) — adapted for over 65 countries (Dubeck and Grove, 2015) —, or the International Common Assessment of Numeracy (ICAN), used in the nationally representative Annual Status of Education Report (ASER) in India and Uwezo in East Africa, and follow this approach. In turn, these statistics have been widely used in important pieces that have laid out the landscape of education in developing countries in recent years (World Bank, 2018; Pritchett, 2013). Therefore, understanding the extent of "enumerator effects" in educational assessments of FLN, and how they might bias the

results emerging from these data collection efforts is important as it could shape, to some degree, our understanding of the state of FLN in the developing world.

In fact, it is well-documented that enumerator effects can be a key source of error in survey data collection. However, it is currently less understood whether this is a problem for academic assessments or performance tasks. Furthermore, even if researchers have a reason to suspect that enumerator effects do exist in educational assessments, less is known about how they might bias potential estimates of interest, like treatment effects in impact evaluations. This is especially true when the assignment of assessors is carried out at a level of disaggregation higher than the student, so there are more opportunities for measurement error to cluster around certain aggregate units like classes or schools. Our contribution in this paper is that we find evidence that the foundational numeracy assessment was indeed prone to enumerator effects, and we use simulation to show that these effects were large enough to lead to spurious results at a troubling rate in the context of impact evaluation when assessors are not assigned at the level of the student. In all, this paper highlights the need to thoughtfully weigh the logistical and empirical trade-offs when assigning enumerators at different levels of aggregation, and proposes some alternative solutions like norming or allowing for enumerator fixed-effects to minimize the extent of the bias that enumerator effects might introduce.

The Limits of Policies which Focus on FLN

As LMIC aim to strengthen their current systems by promoting foundational literacy and numeracy, policymakers need to also be mindful of some of the potential pushback towards these policies. First, focusing on tangible educational inputs, like textbook delivery or school construction, has clear political advantages (e.g., Williams, 2017 or Ejdemyr et al., 2018) but it is less clear if politicians reap similar political

rewards from focusing on education quality. For instance, Harding and Stasavage (2014) find that politicians in Kenya benefit less from electoral promises around the improvement of educational quality relative to those focusing on educational inputs, and Sandholtz (2021) finds that after a large national education reform in Liberia, politicians were rewarded and punished in their local elections accordingly with the improvement, or lack thereof, in the quality of education as a result of the reform. This trade-off may be starker when the intended policies targeted at educational quality compete for significant resources with input-augmenting policies. This, in turn, could be an advantage of policies like curriculum reforms which mostly use already existing resources in the system. Still, when deciding which educational policies are implemented, policymakers may have to balance promoting their future political careers, and developing policies that advance educational quality. The current evidence suggests that these policies, contrary to the broadly popular input-enhancing programs, can be at best a "high risk, high reward" endeavor for politicians, and in essence could weaken their motivation to pursue this type of policy.

Another drawback of policies that focus on FLN is that, as any other policy, there are opportunity costs. Especially when states have weak implementing capacity, and they can only run a few programs at a time, the focus on FLN should not distract from other important policies that are equally needed to build strong educational systems. For instance, schools in the developing world still experience high dropout rates in upper primary – clearly an issue that is just as worthy of the attention of policymakers. Similarly, as the name indicates, FLN are simply the *foundation* of learning in other subjects, and as such, educational systems must ensure that once students have mastered

FLN, they can keep learning in the higher grades. Even beyond learning and enrollment, schools must also foster socioemotional learning (SEL) to ensure that children can successfully navigate their environment and personal relationships once they leave the educational system. In this sense, the two policies that I study in chapters 1 and 2 of this dissertation are fairly narrow, as neither focused explicitly on outcomes beyond foundational learning. Yet, there are other areas of student development that are as valuable as FLN, and therefore, policymakers in developing contexts need to carefully balance the promotion of FLN while also minding these other crucial tasks, all the while staying within their tight resource constraints – no easy feat.

Conclusion

In all, this dissertation aims to build on the body of evidence that explores ways to improve and measure foundational literacy and numeracy in the developing world. The contrast between a mostly successful and an unsuccessful curriculum reform in two contexts begins to populate a largely missing literature on rigorous evaluations of nationwide early grade reforms. Similarly, these two chapters in tandem can also shed light on the curricular features that could maximize the likelihood of success for this type of program. On the other hand, measuring FLN outcomes is the first required step towards improving them. We find that the issue of "enumerator effects" could introduce significant measurement error into the quantification of FLN in LMIC. Therefore, researchers need to carefully consider the logistical and empirical trade-offs of different approaches to assign enumerators to subjects, and the implications that this choice might have on the accuracy and precision of their results and policy recommendations. Ultimately, the creation and dissemination of research is a valuable step towards the

design of better-informed policies, and as a result, I hope that this dissertation is a small move in the right direction towards effectively promoting FLN in developing contexts.

CHAPTER 1

Strengthening Early Literacy Skills Through Social Promotion Policies? Intended and Unintended Consequences in Costa Rica

(Daniel Rodriguez-Segura)

Abstract – Social promotion policies (SPP) are often justified as allowing students to learn at their own pace while avoiding the consequences of grade repetition, particularly in settings where grade repetition is a pervasive feature of the educational system. The 2014 passing of a SPP for first graders in Costa Rica was designed to give students more time in a low-stakes environment to develop the basic literacy skills required for subsequent grades. Using a difference-in-differences approach and the universe of schools in Costa Rica from 2010-2017, I explore the downstream effects of the SPP on enrollment and grade repetition outcomes in later grades. I find that while the policy unsurprisingly lowered grade repetition for first graders, it also increased grade repetition rates for second and third graders by 77% and 24% respectively, likely due to the presence of students who did not reach the basic literacy standards, and yet passed as a result of the new policy. These negative, unintended consequences were mostly borne by school communities of low socioeconomic status. The paper calls for caution and coherence across grades in the design of SPP, along with better tailored policies for students at risk of repeating grades due to low achievement levels.

INTRODUCTION

Grade repetition is a commonly-used policy worldwide to deal with students that do not meet the minimum standards for the grade they are enrolled in. In fact, 32.2 million pupils across the world repeated a grade in 2010 (UNESCO, 2012). In Latin America, the average pupil expects to finish 9.5 years of education, and almost one full year of these would be spent repeating a grade (UNESCO, 2012). As evidence-based education policy in developing countries is moving towards personalization and customization to target the full range of achievement levels through efforts like tracking (Duflo et al., 2011), technology-aided designs (Muralidharan et al., 2019), and Teach at the Right Level interventions (TaRL, 2019), grade repetition lies on the opposite side of the customization spectrum. In other words, even if grade repetition as a policy is simple to implement, it still expects that students will learn once they have seen exactly the same material for a second time, without tailoring the curriculum to the students' specific deficits, or modifying potential system-wide deficiencies that may have led to the failure in the first place. Whether grade repetition indeed improves short-term and long-term learning outcomes in practice is an empirical question, and one which likely varies by context and subgroup. What is clear is that it is a very costly policy for the government and families alike. For example, in the context of this paper, the Costa Rican government spends about USD 2900 per year for every primary school student, and families spend an additional USD 545 per year in out-of-pocket education expenses (UNESCO, 2012), more than a minimum-wage worth of monthly income. In other contexts, Burundi spends more than 15% of its annual basic education budget on repetition, and South Africa spends over USD 750 million per year on repetition for primary school (Minardi et al.,

2020) Therefore, concerns about high rates of grade repetition in schools are well-justified from a cost-benefit perspective, as the benefits would need to be very large to achieve a reasonable ratio with respect to its costs.

Grade repetition and social promotion policies (SPP) vary widely across geographies, and yet the causal evidence of its effects on student outcomes is sparse, particularly in developing contexts. Without establishing causality, longitudinal analyses in Mexico (Gibbs & Heaton, 2014) and Bangladesh (Sabates, Hossain, et al., 2010) have shown that repeaters are also less likely to complete educational cycles than their counterparts who do not repeat grades. In turn, Glick and Sahn (2010), and André (2009) both use quasi-experimental approaches to get at the causal relationship between grade repetition and school dropout in Senegal, both finding a negative effect of the latter on the former. Similarly, Manacorda (2012) leverages a rule which establishes the threshold number of subjects which high school students can fail before having to automatically repeat grades in Uruguay. In this case, the author also finds that grade repeating is conducive to higher dropout rates and up to five years after the student repeated the grade.

The literature is similarly scant regarding the effect of grade repetition policies on learning outcomes. Among the few papers that explore this relationship, Gomes-Neto and Hanushek (1994) find that repetition does increase overall learning, and that schools in the poor, north-eastern region of Brazil used grade repetition to raise overall school scores. They find that students who repeat do learn more, and tend to go from below-average in the grade they repeat to above average the next year. However, the authors are concerned about the very high price tag that involves this positive average effect of grade

repetition on learning. Using their panel data, they proceed to simulated what the learning levels would be if a policy of social promotion was implemented, finding that learning outcomes would not differ as much, and the financial burden to the system of grade repetition is alleviated. They conclude by acknowledging that a SPP may be a good alternative to the status quo of high grade repetition rates, but only a second-best policy response to actually improving the quality of primary schools.

Other related work studying the relationship of grade repetition policies on learning is by Ahsan, Banerjee, and Hari (2018). This paper analyzes the effect of a national SPP, the opposite of a grade repetition policy, between first and eighth grade on learning outcomes in India. One of the main strengths of this paper is that the authors can measure learning outcomes directly at the individual-level for in-school and out-of-school children across the whole country. By employing a difference-in-differences approach, they find that the policy improved reading schools by 2.5% and math scores by 5%, and that the gains are concentrated at the bottom of the distribution. While after testing several hypotheses, the authors cannot pin down the precise mechanism, they speculate that a more relaxed, low-stakes, and learning-friendly environment might contribute to this. A crucial difference between the paper studying the Indian policy and this current paper is that Ahsan, Banerjee, and Hari focus on the learning outcomes for the grades for which social promotion is still in place, as opposed to this paper which also studies the downstream effects after students have left the grades with social promotion policies. Therefore, even if learning increased in grades 1-8 in India under the SPP, it is still an open question whether these gains were large enough to equip them to meet the standards of higher grades.

The context of the present study is the Costa Rican public education system. As a matter of background, education has been a major priority for the Costa Rican government since the social reforms of the 1940s. The country's Constitution guarantees that a minimum of 8% of its GDP must go towards education, and its government spends more as a share of GDP than countries like Sweden, Norway, or Denmark (OECD Data). Public schools in Costa Rica have no fees for preschool through the secondary level, and schools are mostly funded and managed centrally by the Ministry of Education (MEP), although minor decisions on allocation of some funds and human resources can be made by local school districts. While the country has relatively high literacy and enrollment rates (World Development Indicators, 2019), it still shows inefficiencies in learning outcomes. For instance, its students achieved similar scores in mathematics on the Programme for International Student Assessment (PISA) as Mexico and Chile in 2018, even though these countries spend almost half as much on education as Costa Rica (OECD, 2018). Furthermore, the problem can be seen from the lens of equity: while the top income decile in Costa Rica performs at the median OECD level, the bottom decile performs 32% percent lower than the OECD median (OECD, 2015). Even though enrollment in private schools in Costa Rica in 2010 was only 8.5% of all students, these institutions are generally thought of being of higher quality and catering to groups of higher socioeconomic status compared to public schools[1], hence allowing for part of the gap in achievement to emerge early on (World Development Indicators, 2016).

[1] Unfortunately, there is no standardized test or measure of achievement available through which this can be quantified, particularly for primary schools.

A chronic issue within the Costa Rican educational system is the high repetition rates in primary school (see Figure 1[2]). As a response to the prevalence of learning gaps leading to high rates of grade repetition, particularly in the first grade, the Costa Rican Ministry of Education passed a policy in 2014 which instituted social promotion for first graders throughout all public schools in the country. The goal of this policy was to give students two years, first and second grade, to achieve the minimum literacy and numeracy standards to succeed in subsequent grades. The policy came with a curricular change for how Spanish would be taught in first grade, even though the extent to which this was well implemented is an open question. For instance, by the second month of the academic year when this was first implemented, only 55% of all teachers had received any materials or training on the curricular change (Mena, 2014). Furthermore, this curricular change was not accompanied by any enforcement or accountability mechanisms to increase take-up of the new approach to teaching literacy skills.

The weak enforcement of the new curriculum, and the inclusion of SPP only for one grade beg the questions of whether this policy change indeed managed to increase learning and reduce overall grade repetition. In practice, I will proxy adequate learning progress by looking at the ripple effects in enrollment and grade repetition in subsequent years to try to understand the consequences of the SPP. To the best of my knowledge, this paper is the first to study the downstream enrollment and repetition effects of a SPP from

[2] Note that grade repetition rates fluctuate between 2000 and 2017, and that there is a slight decreasing trend throughout this period. These fluctuations respond to different policy stances of the ruling party over the five different administrations that this figure covers. Interestingly, in spite of the broad fluctuation, the year-to-year patterns in grade repetition rates are very similar across grades. This is in contrast to targeted policy changes for a specific grade, like the policy studied in the current paper. However, no grade repetition policy of any kind was implemented for primary school grades between 2010-2017, the period spanned by the data used for the main analyses, besides the main policy discussed in this paper.

the point of view of a country's full educational system. Finally, this paper is the first to address the heterogenous downstream impact of SPP by socioeconomic background of school communities.

Research Questions

The specific research questions that I address in this paper are the following:

1. Was the SPP actually executed for first grade students, i.e., did grade repetition rates decrease among first graders?
2. Did the policy affect grade repetition rates for subsequent grade levels ("downstream effects")?
3. Did the policy affect intra-annual desertion, both for first graders and all other grades?
4. Did the downstream effects vary by the socioeconomic status of school communities?

Policy Context and Change

Curriculum and promotion policy before 2014. First grade is in practice the time when most Costa Rican children learn how to read and write. In spite of scattered policy calls to move this process to an earlier point in the educational process, access to pre-primary education stands at 78%, which is still far from universal coverage (World Development Indicators, 2016). Even among those that do attend pre-school, the quality of education drastically varies by the geography and institution.

Up until 2014, the national curriculum stated the mastery of basic literacy skills as one of the crucial outcomes of first grade, as taught in Spanish class. Furthermore,

students are also required to take subjects like Math, Science, Social Studies, English, among others. While literacy skills are expected to develop within Spanish class, by end of the year, the assessments and homework for the other subjects are mostly also in written-form, requiring basic literacy skills to complete them. Therefore, if a student is struggling to pick up the literacy skills in Spanish, this will certainly have spillovers to all other subjects. The development of the basic literacy skills is critical not only because of its value in and of itself but also because without these, children run the risk of falling behind all other subjects.

Depending on a student's learning outcomes at the end of a school year, the Costa Rican educational system establishes three potential outcomes for this student. The first is "approved" (*aprobado*), which means that the student met the minimum of all learning standards for that grade. This usually happens automatically when all grades across all subjects are above a 65%, although teachers typically have some discretion to modify grades to pass students that for some reason fell below this threshold. The second status is "held back" (*reprobado*), which means that a student is far from meeting learning outcomes in three or more subjects, usually measured as three or more grades being below the 65% threshold. However, teachers also have some discretion to hold back students if they consider them to be socioemotionally immature for their age, or for some other valid reason, usually also cleared with other school staff. The consequence for a held back student is that they have to repeat the grade during the following academic year, if they want to continue their education. Finally, the third status is "deferred" (*aplazado*), which means that students fell short of the learning outcomes in less than three subjects. In this case, students have the opportunity to take a test at the end of the

year in each of the subjects that was below as 65%, and if the student clears all tests, then they are allowed to progress to the next grade. If the student were to fail one or both of the exams, they must retake them again at the end of the break between academic years. If the student were to fail this second exam again, they must repeat the year. For 2010, the baseline year of this study, approximately 86% of all first graders were approved at the end of the academic year, 6% were held back, and 8% were deferred, although the publicly available data do not allow us to discriminate the test outcomes for deferred students.

Grade repetition policies as a policy tool. Grade repetition has been a commonly-used policy in Costa Rican primary schools to deal with students showing learning deficits. As Figure 1 shows, there has been some volatility in the number of students that are held back and deferred throughout time, but with all grades generally moving in similar patterns across years and at non-zero rates. Interestingly, the policy is heterogeneously used across all six primary grades, with the highest rates of held back students typically being in the first grade, at least up until 2014. Contrarily, the grade with the lowest rate of held back students is sixth grade, the last grade of primary, and the grade during which students leave the institution to go to high school.

There are at least two glaring inefficiencies in the structure of the Costa Rican education policies regarding grade repetition. Given the high levels of per-pupil-spending in the country, the first inefficiency is the high financial cost of keeping another student in primary school due to learning deficits, especially if there are cheaper policies that could be implemented to help these students catch up. For instance, interventions like the "once-off catch-up" programs that target children at risk of repetition through summer

classes to get them ready for the next year have displayed promising, cost-effective ways to deal with lagging students in Ethiopia and Mozambique (Akyeampong et al., 2018). The second inefficient aspect of the current policy is that once a student is deemed to have large learning deficits in a subject at the end of a school year, the current grade repetition policy forces this student to repeat *all* subjects the following year, even if the learning deficit was in a single subject. One can imagine cases in which a first-grade student who initially struggled with reading experienced a real catch up in basic literacy skills, and yet has to repeat the grade because they could not quite catch up in other subjects that are not as predictive of future success, or that they can catch up on while also progressing through grades as expected. This is particularly worrying since the number of subjects that students must take from the first grade in Costa Rica is high, increasing the chances of falling behind at least one subject.

Move towards social promotion policies. With the transition to a new, more left-leaning government in 2014, the Ministry of Education (MEP) completed a policy shift towards SPP for the first grade in an attempt to address concerns regarding the high repetition rates in this grade. The motivation behind this policy change is best expressed in the words of the then Minister of Education, Leonardo Garnier: "some people say that this is giving the students a free pass, but what we see all over the world is that the process of learning how to read and write is very different for every child. Grade repetition stops the learning process" (Ross, 2014). Specifically, the MEP designed a new Spanish curriculum in 2013 for first and second graders called the "Programa de estudio de Español de I Ciclo" (MEP, 2013), and implemented it in 2014 for the first time. The main policy change consisted of an effective ban on grade repetition, deferred and held

back students, for first graders in public schools, except for some extreme cases of immaturity, absenteeism, or poor socioemotional skills. The curriculum reform unofficially combined first and second grade, aiming to give students two years to fully acquire basic literacy skills without having to repeat the grade, as opposed to only one year as it was the case before the reform. The curricular reform did not effectively change the learning goals and standards for each grade, but rather the consequences of not meeting these goals in the first grade. In particular, before and after the policy was passed, the expectation for early literacy levels at the end of "*primer ciclo*" (i.e. grades first through third) was still a high degree of reading fluency, and comprehension, up to the understanding of nuanced sub-text information within children-appropriate readings (MEP, 2013). The MEP branded this as meant to address the high repetition rates in the first grade by "incorporating the latest neuroscience and pedagogical findings applied to learning how to read" (MEP, 2013). Crucially for the identification strategy used in this paper, the policy did not change grade repetition policies for all other grades besides the first grade, nor did it imply a change in the content expected to be covered in any subsequent grade.

The implementation of the policy was far from perfect. The dissemination of the new curriculum was mostly done by giving teachers a booklet with the new approaches they should have been using to teach literacy skills. Furthermore, there were reports that one month into the school year, 45% of all teachers had not received any guidance. Even as of 2018, only 69% of all teachers had received any training on the new curriculum, and 83% of those that received training described it as "insufficient" (Meléndez, 2019; Estado de la Educación, 2019). Similarly, 82% of all teachers noticed an uptake in the number

of formal and informal "requests for additional support" for students since the passing of the policy, where 80% of all requests are from second graders, and 96% are related to basic literacy skills (Meléndez et al., 2019, Estado de la Educación, 2019). In this sense, the design of the policy relied on the premise that what students struggling with literacy skills needed was more time with the same treatment, in spite of the clear diminishing returns to classroom instruction of literacy skills in a context where the average class size is 19.4 students per teacher. The perceived increase in requests for additional support to improve literacy skills of second graders may be evidence to contradict the premise that more class time for students before facing higher stakes would help them all master the basic literacy skills. In fact, the policy did not consider individualizing instruction or more tailored practices for these students. Finally, the policy was designed with weak to no accountability measures to ensure that the curriculum was actually being implemented, as opposed to teachers just retaining their previous practices, and only following the new directions regarding grade promotion in first grade.

Data and Methods

Data available. The data available to study the system-wide effects of this policy consist of data disaggregated at the school-level for the universe of schools in Costa Rica from 2010 to 2017. These data are broken down by grade, and include total enrollment at the beginning of the year, enrollment at the end of the year, the number of students that dropped out during the year ("intra-annual desertion"), and the number of students which passed the grade, were deferred, or were held back. Given how the structure of the policy classifies end-of-year outcomes for students, students that are approved can be interpreted as those students that cleared all the requirements, while held back students are the lowest

performers that substantially struggled to meet the learning standards for their grade. The deferred students in turn can be interpreted as the "marginal" students: those that struggled to meet the standards, but that have a last chance to meet them prior to the promotion determination.

The data can also be merged in with school-level covariates, such as geographic location, school district affiliation, numbers of teachers by gender, rural/urban classification, public/private classification, some infrastructure data between 2011-2017, and student health outcomes such as the percentage of students that are underweight. These health outcomes are particularly relevant as I will use them as school-level proxies for socioeconomic status given the lack of recent and frequent data on local poverty rates. Instead, I make sure that that the share of underweight students within a school is a valid proxy, using census data on poverty rates at the district level from 2011.

Unfortunately, there is not yet data to identify long-run outcomes, given the recency of the policy. Furthermore, the MEP does not provide individual-level data publicly, and since there are no high or low stakes tests built into the system until the 11th grade, there are no comparable data available on standardized-test scores.

Methods. The identification strategy for all of the research questions consists of a difference-in-differences approach[3]. This strategy seeks to estimate the causal effect of the policy on the outcomes of interest. In particular, the first difference will be the pre and post policy difference: the pre-period consisting of 2010-2013, and the post period from 2014-2017. The second difference is between cohorts affected by the policy, and

[3] For a visual inspection that the parallel trends assumption holds, see Figure 2. This figure shows the yearly rates of the main outcome variables for grades 1, 2, and an average of the comparison group (grades 5 and 6), weighted by their respective enrollments.

cohorts that barely missed being affected by the policy. The first cohort affected by the policy was the cohort that entered first grade in 2014, so any cohort that was in first grade in 2014 and after 2014, was a treated cohort throughout all of their years in school in our data. Fifth and sixth grades were the only two primary school grades, according to the Costa Rican educational system, whose cohorts were never covered by the policy in the data. In other words, the earliest cohort affected by the policy was the cohort which entered first grade in 2014, and whose children under regular grade progression would be in fourth grade in 2017, the last year of the data. Therefore, fifth and sixth grade will jointly serve as the comparison group throughout this analysis. The key assumption behind this comparison group is that students in these cohorts are not systematically different, other than in age, from the cohorts that were affected by the policy. By using cohorts across grades from the same schools, I ensure that the demographic characteristics in treated and comparison groups are similar, and given that there were no major changes in enrollment rates in Costa Rica since the mid 2000s, this assumption is likely to hold. Furthermore, finishing primary school, in one way or the other, is almost universal in Costa Rica (96%), so any concern that students that make it to fifth and sixth grade are systematically different from those in the earlier grades is not a major issue in this context[4]. Specifically, I estimate the following model for school i, grade j, and year t:

$$y_{ijt} = \beta_0 + \beta_1(\text{Treated cohort})_j + \beta_2(\text{Treated cohort*Post})_{jt} + \mathbf{B}(\text{School covariates})_i + \alpha + \gamma + \theta \qquad [1]$$

[4] I also run a robustness test of all results for first through third grades, where I group fourth graders into the comparison group along with fifth and sixth graders, as fourth graders would be a less self-selected group than fifth and sixth graders, but would also be the least affected cohort by the policy (only being a treated cohort in 2017). This does not significantly change the results or their interpretation, so I decide to simply use as comparison group the two grades (fifth and sixth) which had no cohorts affected by the policy in the data whatsoever.

Where y is the outcome variable, which will be one of four different options: count of students held back, count of students deferred, count of students deferred and held back, and count of students that dropped out throughout the year. I can also express the outcomes as percentages of total enrollment instead of counts, but I will prefer using counts to ease the interpretation of the coefficients since the results are substantively unaffected by this decision. I estimate this model separately for each target grade i.e., the specific grade I try to understand changes in, and the comparison group. In other words, I first run the model using only first grade and the comparison group (fifth and sixth grades). Then I substitute first grade with second grade, and repeat it for third, and fourth grade. The model is estimated for grades 1-3, what is known in Costa Rican education as "first cycle" (*primer ciclo*), and for grade 4 as well, the last grade with a treated cohort for which I can observe outcomes in the data.

The model includes a dummy variable for whether the grade was part of the treated group or not, i.e., whether the grade is fourth grade or below. Most importantly, the model has an interaction term between the treated variable and the indicator for the post variable, which is attached to my estimator of interest, $\beta_{3.}$ The post variable indicates whether a specific cohort is or was at some point exposed to the SPP. In other words, it is 1 for first graders in 2014 onwards, for second graders in 2015 onwards, and so on. The coefficient β_3 should be interpreted as the causal effect of the change in policy on the outcome of interest. Finally, I also include fixed effects at different levels: year fixed-effects α to capture additional changes over time that are unrelated to the policy shift, province fixed-effects γ to soak up some geographic variation, and school district fixed-effects θ to absorb some of the administrative-level variation. I show on the first

two columns of Table 1 that the results do not vary due the inclusion of these fixed-effects. The table showcases this only for one outcome for simplicity's sake, but this holds for all outcomes. Finally, standard errors are clustered at the grade level.

I also estimated the model separately by socioeconomic status of each school. Given the lack of direct socioeconomic data at the school-level, I proxy this using the share of underweight children in each school. I use district-level poverty rates to check that the share of underweight students is a valid proxy for socioeconomic status. In particular, I merge district-level poverty rates for 2011 onto school-level information on the share of underweight students, and regress district poverty rates on the mean share of underweight students in schools within the district in 2011, weighted by school enrollment. The coefficient that emerges from this is 1.1, significant at the <0.01 level, meaning that at least in aggregate there is broadly a one-to-one relationship between district-level poverty and the share of underweight children in school. This provides confidence in my use of the share of underweight children as a school-level proxy for socioeconomic status. In particular, I classify schools below the nation median share of underweight children as "low SES", and all other schools as "above low SES".

Results

Adherence to the policy. I display the primary results to the first three research questions in Table 1[5]. The first key result is that the policy was indeed executed by schools, as evidenced by the negative signs and high significance on the coefficients for all the grade repetition variables in the Grade 1 (top) panel, columns 1-4. This is not

[5] Note that Figure 3 standardizes all coefficients to standard deviations to make effects across grades more comparable.

surprising and can be thought of as adherence to the policy: it is expected that first graders would be less likely to be deferred or held back as the policy explicitly ordered schools to socially promote at that grade level. Specifically, the number of first graders that were held back diminished by 85%, and the number of deferred students diminished by 90%[6]. In other words, while schools mostly stopped using grade repetition as a way to deal with first graders that did not achieve their learning goals, this was even more prevalent for the students that would have been deferred or the "marginal students", than for students that would have been directly held back, or the lowest achievers.

Downstream effects. The second key insight is displayed in columns 1-4 for the panels with results for second, third, and fourth grade. The statistical significance of the coefficients across the board suggests that the introduction of the SPP for first graders changed downstream results for students as these cohorts progressed to later grades. In particular, it increased the number of deferred and held back in second grade by 61% and 116% respectively, and 27% and 11% in the third grade. Finally, by the fourth grade, all the signs reverse again and they are now negative: the policy led to a decrease of 26% in the number of held back students and 5% in the number of deferred students. Throughout all grades, the coefficients on deferred students, and on held back students are statistically different from each other, at the 0.001 significance level.

This is strong evidence to suggest that this policy had very large effects in the composition of student outcomes post-first grade. The high increase in the number of students that failed second grade as a result of this policy suggests that at least for some

[6] This is also displayed in Figure 2, as the trend for all outcomes for first grade significantly drops after 2014.

students, the policy simply delayed grade repetition. In other words, students that would have been held back or deferred in the first grade under the previous policy, are now shielded from this by the new policy, but are held back or deferred in the second grade or in the third grade, when the policy still allows educators to do so. As a side result of the large increases in grade retention rates in second and third grade, the policy resulted in a much more selected cohort by the fourth grade. This explains the change in sign of the results for this grade. While the fourth grade is the last year covered by the data, this is suggestive evidence that the policy simply changed the timing of when some students would repeat a grade.

However, the actual magnitudes of the point estimates suggest that the policy did not just result in a "net zero" effect in grade repetition within the system as a whole. In particular, the policy overall lowered grade repetition, deferred or held back students, on average by 2.31 first graders per school and 0.28 fourth graders per school, while it also increased grade repetition on average by 1.32 second graders per school and 0.40 third graders per school. Scaling these numbers up by the respective enrollments in each grade and the number of schools for the cohort that was in first grade in 2014, the policy overall decreased grade repetition in first and fourth grade by 9,680 students and increased grade repetition in second and third grade by 6,421, for a net reduction in grade repetition of 3,259 students for this cohort from 2014-2017[7].

[7] The point estimate for first grade is -2.31 students per school and for fourth grade is 0.28 students per school. Multiplying these numbers by the total number of schools at their respective baselines, which were 3,740 in 2014 and 3719 in 2017, this leads to a reduction of 9,680. On the other hand, the point estimate for second grade was 1.32, and 0.40 for the third grade. There were 3734 schools at second grade baseline (2015) and 3731 at third grade baseline (2016), so an overall increase of 6,421. Therefore, the net calculation is the decrease in grade repetition in the first and fourth grade of 9,680 students minus the increase for second and third graders of 6,421, for a net decrease of 3,259 students for the system as a whole.

Intra-annual desertion. The other important outcome variable that this policy may have affected is intra-annual desertion. Even though education is mandatory in Costa Rica up until the ninth grade, children do drop out of school before then, and the law may or may not be enforced to get them back to school. This happens for the mostly in lower secondary, and through "inter-annual dropout", i.e., between school years (as opposed to within school years, or "intra-annual dropout"), yet it does exist for primary school to a smaller extent[8]. For perspective, approximately 466 first graders (0.1%) dropped out of school during the school year in 2013, so while this is not a major systemic problem, these are still 466 future citizens that will not have an education in the future. According to Table 1, this policy only really affected grades 1, 2, and 3 lowering intra-annual desertion in first grade on average by 0.05 students per school, and increasing it in the second and third grade on average by 0.01 students per school. On net, the policy decreased intra-annual desertion by 112 students[9], even after accounting for the students in the second and third grades for which it may have led to desertion.

Heterogeneity by socioeconomic status. Since academic achievement levels are strongly correlated with socioeconomic status, it is worth exploring whether this policy, which was explicitly targeted towards lower performing first graders, affected different socioeconomic groups differently. I previously described in Section IV how I classify schools into the "low SES" and "above low SES" groups. Using this classification, I run

[8] For instance, in 2013 the average inter-annual dropout for grades 1-6 was 1.8%, but 10.9% for grades 7-9 (Ministry of Education, 2020). Unfortunately, I cannot explore inter-annual dropout as an outcome (only intra-annual dropout) due to the lack of student-level linkage data across years.

[9] The point estimate for first grade is -0.05 students per school, and there were 3740 schools at first grade baseline (2014), so a decrease of 187 students for the whole country in the first grade. Similarly, the point estimate for second grade is 0.01, and there were 3734 schools at second grade baseline (2015), so this means an increase of 37.3 students in the second grade.

the same models as before for grades 1 and 2, but separately for each socioeconomic group. The results, benchmarked against the baseline mean, are displayed in Table 2. Firstly, the policy was similarly executed for schools of both SES groups, as displayed by the relatively low differences in the effect on grade repetition for first graders. However, the unintended negative consequences are very much born by the poorest schools, as displayed by the difference in effects for second graders. Taking dropout rates for the second grade as an example, the increase for the poorest schools was 29%, but it was precisely 0% for all other schools. In other words, the policy may have increased intra-annual desertion in the second grade by 0.02 students per low SES school, or 37[10] students in the whole country, but did not change intra-annual desertion for schools classified as above low SES. A similar pattern is seen for all indicators, where second graders in poor schools are much more heavily impacted by this policy than their counterparts in other schools. Interestingly, this gap is wider for the absolute low performers, as proxied by the effect on held back students, than on the marginal students. This suggests that it was particularly in poor schools where children who really struggled in the first grade, and would have otherwise been immediately held back, were shielded from this by the policy, but were then let into the second grade without the proper support to master the learning standards for this grade where grade repetition was still commonly used as a policy to deal with low performers.

[10] The point estimate was a 29% increase over the baseline of 0.07 students per school, meaning an increase of 0.02 students per school, over 1848 poor schools.

Discussion

The main point that this paper raises is well synthesized by the World Bank report on learning in developing countries: “Ensuring that the parts of an education system work together is as important as ensuring alignment toward learning” (World Bank, 2018). The adoption of a new curriculum or a new policy is likely to be ineffective if all of parts of said system are not aligned to work well with this change. In the case of the SPP in Costa Rica, while the policy did achieve the goal of lowering grade repetition rates in the first grade, it nevertheless had unintended consequences for other grades whose learning goals and standards were not aligned with the change in policy for the first grade. Especially concerning is the fact that school communities of low socioeconomic status bore the burden of the unintended consequences of policy by simply delaying grade repetition for many of their students without actually achieving the expected learning outcomes. Even if the policy lowered grade repetition by 3,259 students per cohort in the system overall, it still failed the 4,929 second and third graders for whom the policy simply delayed grade repetition and ended up still repeating a subsequent grade. The fact that these unintended consequences where experienced by communities of low socioeconomic status is problematic, given the high levels of educational inequality in Costa Rica. Not just this, but the high price tag of grade repetition more broadly leaves plenty of room to find more cost-effective interventions which actually improve learning outcomes of students at-risk of having to repeat a subsequent grade by targeting instruction to their specific deficits, while still letting them progress through grades at a standard pace. Finally, this analysis should serve as a cautionary tale for other countries considering a SPP in earlier grades, such as South Africa at the time this article was written (Parent 24, 2019). Even if social promotion is implemented, this paper urges policymakers to design

a platform that supports the needs of struggling students even throughout the grades with social promotion, so that their transition into later grades without the SPP is smoother and more conducive to learning in those grades as well.

TABLE 1.1

Regression Results of Difference-in-differences Estimation for the Effect of Social Promotion Policy on End-of-year Enrollment Status Outcomes, by Grade.

	Deferred or held back (1)	Deferred or held back (2)	Held back (3)	Deferred (4)	Intra-annual dropout (5)
			Grade 1		
DiD estimate	-2.31***	-2.31***	-0.92***	-1.39***	-0.05***
	(0.10)	(0.11)	(0.04)	(0.07)	(0.00)
Mean (2013)	2.57	2.57	0.94	1.64	0.12
SD (2013)	5.63	5.63	2.56	3.77	0.60
School-level controls	No	Yes	Yes	Yes	Yes
Grade, province, year, and school district fixed effects	No	Yes	Yes	Yes	Yes
Observations	89,448	89,328	89,328	89,328	88,659
			Grade 2		
DiD estimate	1.32***	1.32***	0.57***	0.75***	0.01***
	(0.06)	(0.06)	(0.02)	(0.03)	(0.00)
Mean (2013)	1.72	1.72	0.49	1.23	0.05
SD (2013)	4.24	4.24	1.47	3.27	0.36
School-level controls	No	Yes	Yes	Yes	Yes
Grade, province, year, and school district fixed effects	No	Yes	Yes	Yes	Yes
Observations	89,448	89,328	89,328	89,328	88,711
			Grade 3		
DiD estimate	0.40***	0.40***	0.04***	0.37***	0.01**
	(0.04)	(0.04)	(0.02)	(0.02)	(0.00)
Mean (2013)	1.67	1.67	0.35	1.33	0.05
SD (2013)	4.18	4.18	1.10	3.46	0.29
School-level controls	No	Yes	Yes	Yes	Yes
Grade, province, year, and school district fixed effects	No	Yes	Yes	Yes	Yes
Observations	89,448	89,328	89,328	89,328	88,767
			Grade 4		
DiD estimate	-0.28**	-0.28**	-0.15**	-0.13*	-0.00***
	(0.05)	(0.05)	(0.02)	(0.03)	(0.00)
Mean (2013)	3.09	3.09	0.57	2.52	0.07
SD (2013)	7.48	7.48	1.78	6.17	0.40
School-level controls	No	Yes	Yes	Yes	Yes
Grade, province, year, and school district fixed effects	No	Yes	Yes	Yes	Yes
Observations	89,448	89,328	89,328	89,328	88,720

Notes. Outcome variables listed on top row. Units are student counts, per school per year. Rows labeled "DiD estimate" show the coefficient of the interaction between the dummies for treated graded, and a dummy for the years post intervention. Significance levels: * p<0.10, ** p<0.05, *** p<0.01

TABLE 1.2
Point Estimates of Subgroup Analysis by Average School Socioeconomic Status.

Outcome: Holding back and deference rates

		Socioeconomic status		
		Low SES	Above low SES	Low-above low
Grade	1	-92%	-93%	2%
	2	85%	65%	20%

Outcome: deference rates

		Socioeconomic status		
		Low SES	Above low SES	Low-above low
Grade	1	-42%	-35%	-7%
	2	29%	17%	12%

Outcome: Holding back rates

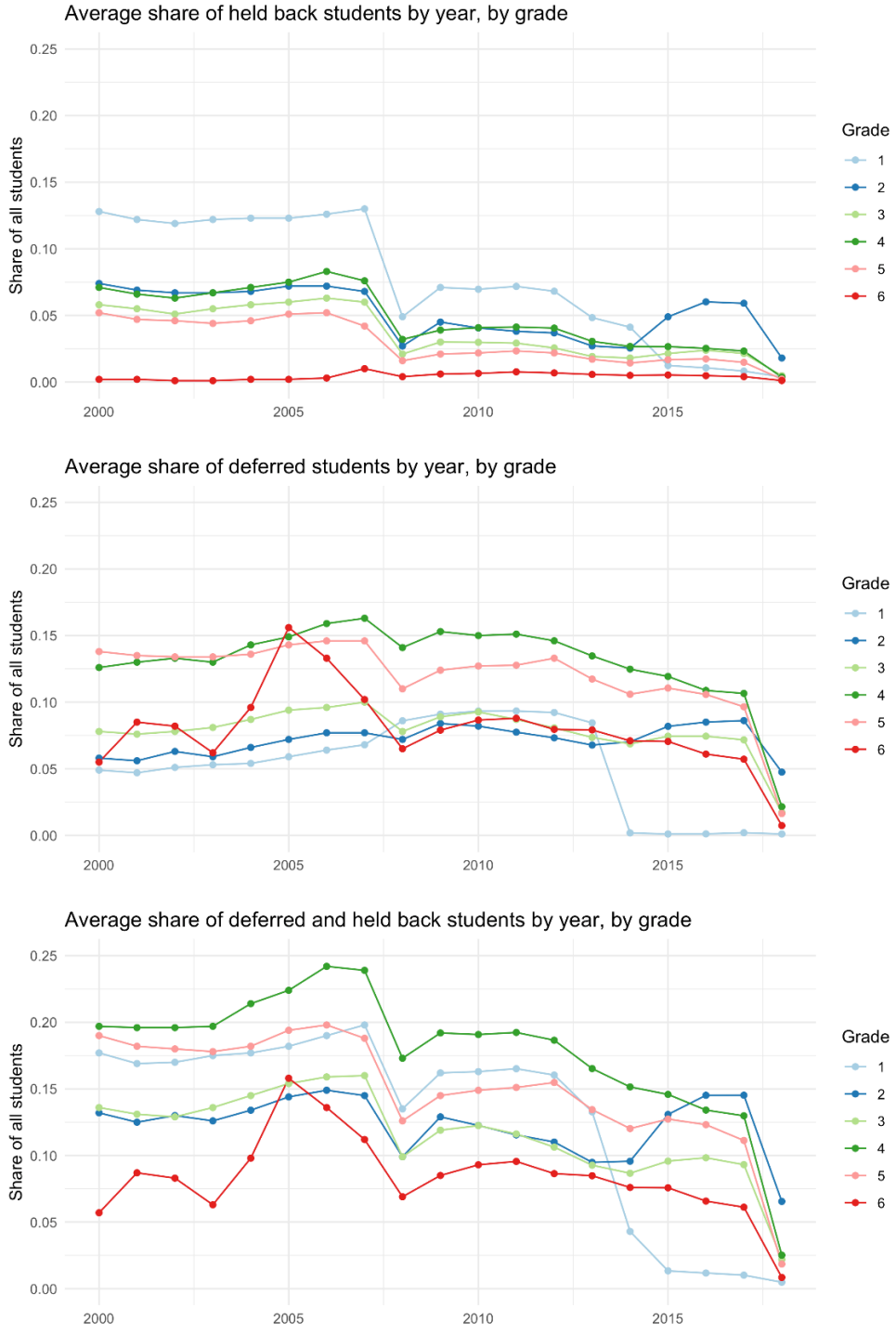
		Socioeconomic status		
		Low SES	Above low SES	Low-above low
Grade	1	-101%	-102%	1%
	2	128%	103%	26%

Outcome: Dropout rates

		Socioeconomic status		
		Low SES	Above low SES	Low-above low
Grade	1	-43%	-44%	2%
	2	29%	0%	29%

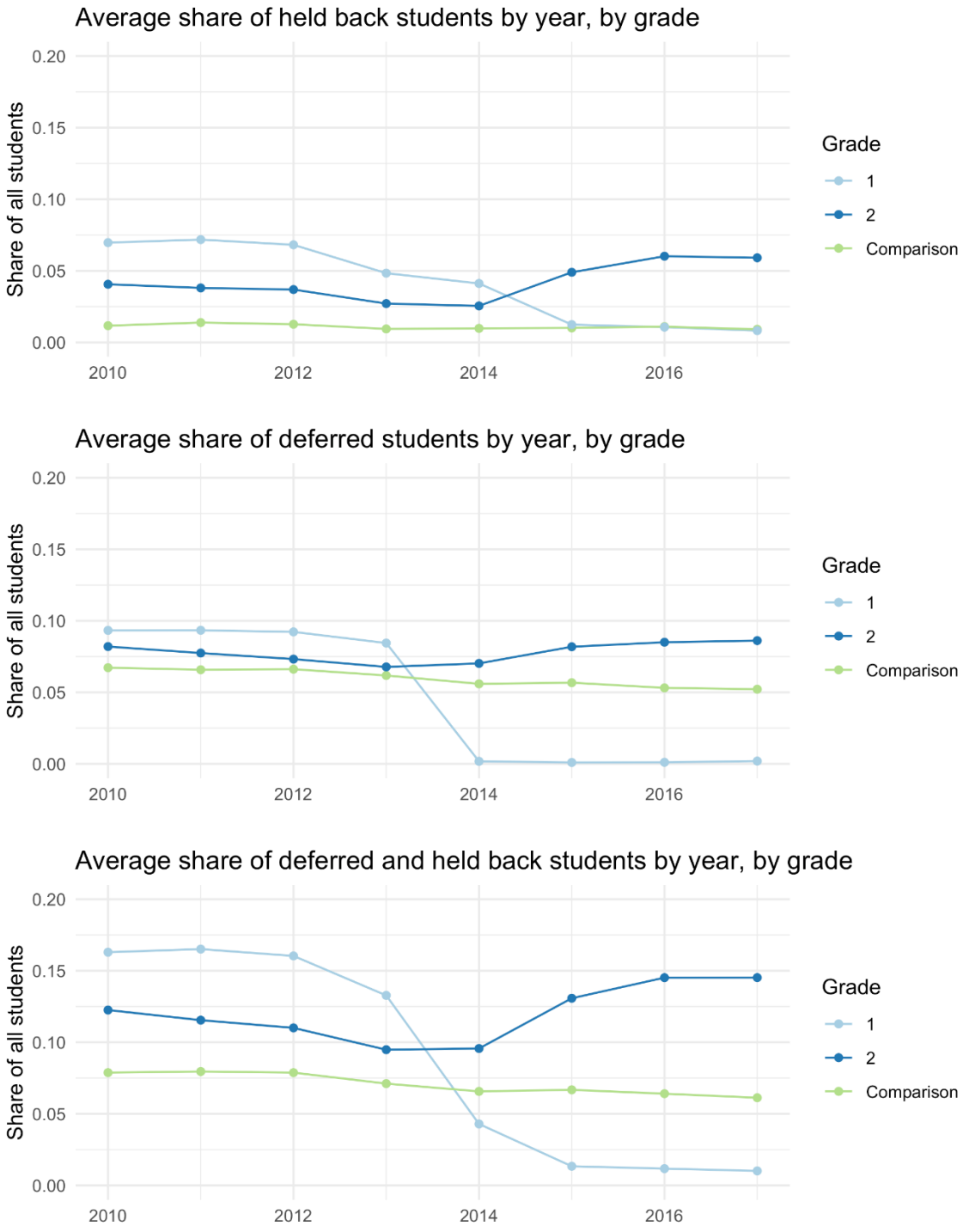
Notes. Numbers shown are the point estimate of the difference-in-differences estimation (ran separately by socioeconomic status) using counts as the unit of the outcome variable, divided by the 2013 mean of each outcome, also in counts as the unit of the outcome variable. Schools with an average share of underweight children below the national median are classified as schools of low socioeconomic status, and all other schools are grouped into “above low SES”.

FIGURE 1.1. *End-of-year Enrollment Status Outcomes by Grade. (2000-2017)*



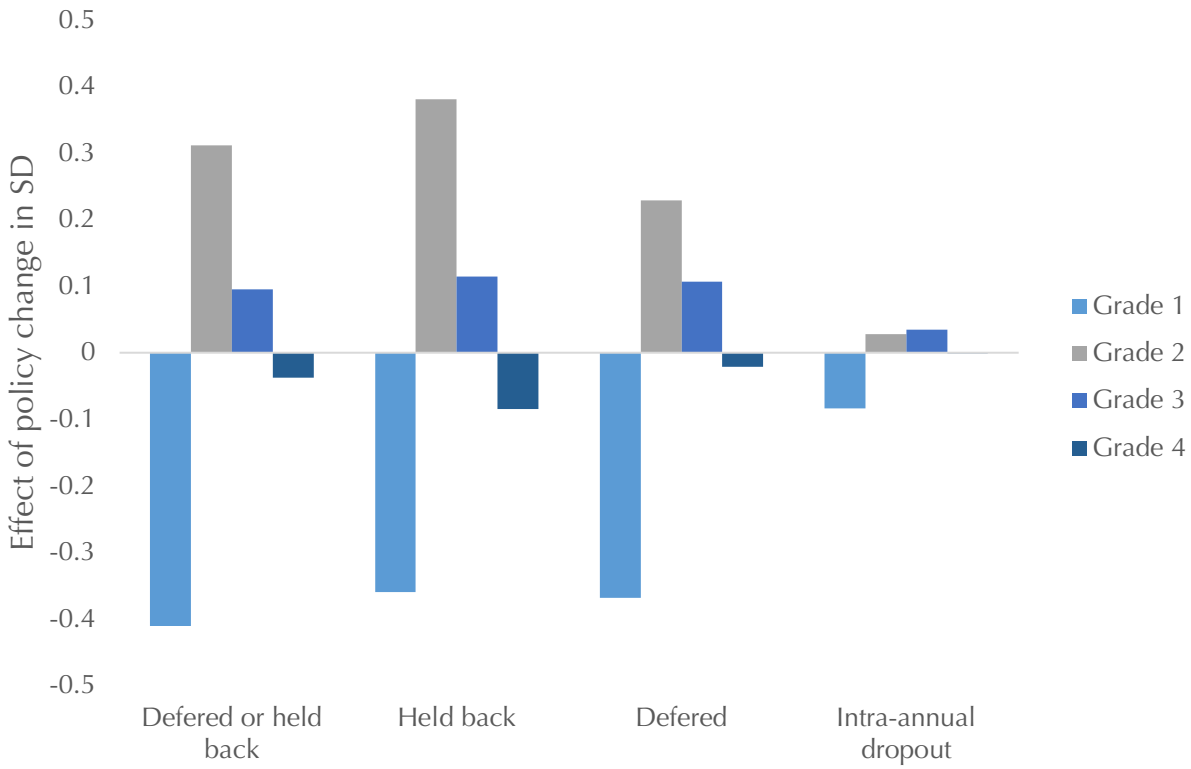
Notes. Yearly rates calculated as the average of each outcome across all public schools, weighted by enrollment.

FIGURE 1.2. *End-of-year Enrollment Status Outcomes for Grades 1, 2, and the Comparison Group (2010-2017)*



Notes. Yearly rates calculated as the average of each outcome across all public schools, weighted by enrollment. Note that the yearly, weighted average of grades 5 and 6 is the comparison group for all the regressions.

FIGURE 1.3. *Point Estimate Results from Table 1 Expressed in Standard Deviation Units, by Outcome and Grade.*



Notes. Each bar represents a difference-in-differences estimate, expressed in standard deviation units of the outcome variable in 2013 (pre-reform for all grades). All bars represent effects with levels of significance at least at the 10% level, mirroring those shown in Table 1.

CHAPTER 2

Back to the Basics:
Curriculum Reform and Student Learning in Tanzania

(Daniel Rodriguez-Segura and Isaac Mbiti)

Abstract –In 2015, the Tanzanian government implemented a curriculum reform that focused instruction in grades 1 and 2 on the "3Rs" —reading, writing, and arithmetic. Consequently, almost 80% of the instructional time in these grades was mandated towards foundational literacy in Kiswahili and numeracy skills. Other subjects such as English were no longer taught. Using student-level panel data, we evaluate the effect of this policy on learning outcomes using a difference-in-differences approach which leverages the variation in the timing of implementation across grade levels and cohorts impacted by the policy. We find that the policy increased learning by around 0.20 standard deviations in Kiswahili and math test scores one year after the start of the reform. Timely teacher training on the new curriculum was associated with even larger effects. Evaluating longer term outcomes, we find suggestive evidence that the reform decreased the dropout rate of children up to four years later. However, this was also accompanied with lower average passing rates in the national grade 4 examination due to compositional changes as low-performing students became less likely to dropout.

INTRODUCTION

Curricula are a key input of any educational system. Ideally, an educational system's intended curriculum determines the material mandated to be taught in school and the desired instructional approaches. Yet in practice, curricula in many developing countries are often too expansive or "overambitious" relative to their education system's capacity (Pritchett, 2013). There are also concerns that these curricula favor children from advantaged backgrounds (Glewwe, Kremer, and Moulin, 2009). Because teachers have incentives to cover the entire syllabus, scholars have hypothesized that a wide curriculum could encourage teachers to either increase the pace of instruction beyond the rate of student learning, to focus their attention on the students who can keep up, or both (World Bank, 2017; Pritchett and Beatty, 2015; Muralidharan and Zielenkiak, 2014). This, in turn, could be a potential explanation of why the progression of student learning in many developing countries is slow, with very few students demonstrating appropriate grade-level competencies, and the majority of them having a mastery level several grades below where they should be (World Bank, 2018; Pritchett and Beatty, 2015).

The Tanzanian education system, which we study, exhibited many of the characteristics associated with overburdened education sectors for the past two decades. These included a curriculum featuring numerous subjects, low levels of learning relative to international benchmarks, slow learning progression, and trademarks of systems burdened by expansive curricula (Ministry of Education, 2015; USAID, 2015). For instance, prior to 2015, students in early primary school (Grades 1 through 3) were taught eight different subjects, including Information and Communications Technology (ICT) and agriculture. The learning profiles were flat with foundational numeracy and literacy

skills gained slowly over time (Jones, Ruto, Schipper, and Rajani, 2014). Consequently, most students fell behind the prescribed curriculum– only 31 % of grade 3 students were proficient at the grade 2 level (Jones, Ruto, Schipper, and Rajani, 2014), and the majority of grade 4 students had not mastered grade 3 material (World Bank, 2017).

Faced with the growing evidence on the low learning levels in early grades, the Tanzanian government enacted the 3Rs reform (Reading, wRiting, and aRithmetic), also known as the "3Ks" in Kiswahili, for grades 1 and 2. This reform was enacted in the 2015 school year and narrowed the scope of the grade 1 and 2 curriculum such that 80% of the instructional time would focus on the 3Rs, with all literacy focused on Kiswahili rather than English as it had been previously. English, which was taught as a subject in the first two grades was removed from the curriculum and reintroduced starting at Grade 3. While proponents of narrow curricula argue such reforms are likely to benefit most learners, the potential benefits might not materialize due to state capacity constraints, such as teacher training. Further, the potential benefits on numeracy and literacy may come at the expense of non-focal subjects, and the reforms may constrain the potential of high performing students who likely benefit from a faster pace. In fact, in wealthier contexts like the United States, "curriculum narrowing" often comes with negative connotations linked to the unintended consequences of test-based accountability and the excessive focus on a handful of tested subjects. The slower pace may also generate a compositional effect if students who would have fallen behind under the expansive regime are less likely to drop out because of the reform.

Despite the ubiquity of overambitious curricula in developing contexts, there is limited causal evidence on the potential for content reducing curriculum reforms to

improve student learning outcomes. This is partly due to the challenges of credibly estimating the casual impact of a nationwide reform which affects all students simultaneously, and the lack of adequate data. The reforms that have been studied have focused evaluating different targeted instruction models (e.g., Banerjee et al., 2017, Muralidharan et al., 2019), and changes in the language of instruction (e.g., Ramachandran, 2017; Seid, 2019, and Laitin et. al., 2019). In this paper, we examine the consequences of this 2015 Tanzania curriculum reform using a unique student-level panel dataset of students from grades 1-3, drawn from a large nationally representative randomized control trial (Mbiti et al., 2019 and Mbiti et al. 2021). To estimate longer run effects, we use administrative data on national test scores in grades 4 and 7 (the last grade in primary school). This allow us to explore the effects of the reform on both passing rates in national assessments four years after it was implemented, as well as the relative change in the number of test-takers in grade 4 (which acts like a proxy for downstream enrollment.)

We identify the impact of the 3Rs reform using a difference-in-difference strategy that takes advantage of the variation in student exposure to the reform by grade level. Specifically, we compare test score outcomes among students in the first two grades (treated grades) to the test scores among third graders (comparison grade), pre- and post-reform (2014 compared to 2015). To explore longer term implications of the reform, we use the administrative data on grade 4 and 7 national test scores to estimate the effect of the program on learning and school enrollment four years after the reform was first implemented. In particular, we compare outcomes for pupils in grade 4 in 2018 —the cohort of students who was in grade 1 in 2015— to the outcomes of pupils in grade 7 that

the same year, as grade 7 pupils did not experience the curriculum reform during the period covered by our data. Given that we have access to the universe of national test scores for these assessments, we also explore whether the number of test-takers increased as a result of the policy to proxy for school enrollment outcomes four years later.

We find that the nationwide curricular reform produced moderate average gains in numeracy and literacy by approximately 0.20 SD. These gains were equivalent to a reduction in pre-policy learning gaps between the top and bottom wealth quintiles of 28% in both math and in Kiswahili. Similarly, we can rule out negative effects on English, a subject de-emphasized by the reform, smaller than -0.02 SD. We also find that learning grains were larger in schools that received timely teacher training on the new curriculum, providing suggestive evidence for the importance of proper implementation, especially for a governmental curricular reform of this scale. In the longer-term, the policy increased the number of students taking the fourth-grade national test by 16%, suggesting that the policy improved student retention and grade progression. The improvement in student grade progression was also accompanied by decreases in the passing rate of these assessments. However, these decreases were comparable in magnitude or smaller than what would be expected given the overall increase in the number of test takers, suggesting that there was still an increase in the aggregate level of learning in Tanzania as a result of the reform.

Our study makes three distinct contributions. First, it is one of the few studies that examines the causal impact of a narrower curriculum on learning outcomes in a developing country. Despite the recognition of overcrowded curriculums in developing countries (Atuhurra and Alinda, 2018; Atuhurra and Kaffenberger, 2020; Pritchett and

Beatty, 2015), there is limited causal evidence on the potential impact of reducing the required instructional content. The literature that estimates the causal impact of curriculum reforms on learning has generally focused on the effects of (large-scale) changes to the language of instruction in schools from colonial languages of instruction to local languages (or mother tongue). The debate on language policy is less relevant for primary education in Tanzania because, unlike other countries in regions, the language of instruction in primary schools has been Kiswahili (rather than English) since the 1960s. Overall, the evidence on the effectiveness of reforming the language of instruction on student learning in mixed. For instance, Ramachandran (2017), Seid (2019), Laitin et. al (2019), Brunette et. al (2019) and Kerwin and Thornton (2020) find that that these reforms improve learning outcomes, whereas as Piper et. al (2018) and Chicoine (2019) show that such reforms fail to improve learning. Our study fills this gap by using a credible identification strategy, coupled with student panel data to show that such reforms can improve student learning in early grades. Further, we show that the learning improvements across almost all measured sub-domains of numeracy and literacy (for example, two-digit addition and word recognition).

Second, we use administrative data from 2015-2018 to examine the longer run impact of the reform, which has been broadly hard to quantify in the international education literature due to the lack of appropriate data. Our results show that the student's fully exposed to the reform were more likely to take the fourth-grade exam. This could reflect the improved retention and grade progression effects of the reform. However, the differences-in-differences estimates on learning are negative, potentially

reflecting the compositional change in the sample towards a lower-performing student body on average.

Third, we use our data to examine potential mechanisms. We focus on implementation – specifically the extent to rollout of teacher training on the curriculum. Developing nations often face challenges implementing policies, programs, and reforms at scale, muting their potential beneficial effects (Banerjee et. al, 2017; Bold et. al, 2018). Students in schools with at least one teacher trained in the 3R reforms had better learning gains compared to the counterparts in schools with no trained teachers, although this difference was not statistically significant.

Our work contributes to the literature on the potential for curriculum reforms that narrow the instructional content to improve learning outcomes when implemented in contexts in which the curriculum has previously been overcrowded or overly ambitious. These reforms are arguably extremely relevant for developing country contexts where instructional time is limited due to teacher absenteeism (World Bank, 2018) and the inclusion of multiple (potentially tangential) subjects can crowd out the teaching of core competencies such as the 3Rs. Even outside of primary and secondary education systems in developing countries, similar debates are ongoing regarding the potential deleterious effects of an overcrowded curriculum in medical schools in developed countries (Slavin and D'Eon, 2021). In this way, this study offers some initial evidence on the potential benefits, and unintended side effects, of such a reform.

Context

National curricula often reflect the political priorities, historical roots, and sociocultural environment in which schools operate. For example, the curricula in many developing often contain features from past colonial institutions, such as retaining

English or French as the language of instruction (Mwiria, 1991; Malisa and Missedja, 2019; Erling and Hultgren, 2017). On this dimension, Tanzania has been an exception relative to its neighbors – the language of instruction in primary schools has been, for the most part, Kiswahili rather than English since at least 2007 (Sa, 2007).

Primary school in Tanzania comprises seven grades. While the net enrollment rate in primary school increased from 53% in 2000 to 80% in 2014, in 2015 only 35% of third graders and 72% of grade 7 students could pass exams measuring second grade standards (Twaweza, 2017). These low levels of learning are coupled with large geographic and socioeconomic disparities. For instance, the urban-rural gap in 2015 was about 0.5 standard deviations in test-based math and Kiswahili performance, while the gap between the top and bottom wealth quintiles was about 0.7 standard deviations (Twaweza, 2015).

As of 2013, the Tanzanian curriculum for grades 1-2 was an archetypical overambitious curriculum, consisting of eight subjects, including "Vocational Skills", "Information and Communication Technology", and "Personality"[11]. In the words of the Ministry of Education and Vocational Training,

> "The Curriculum for Standard I and II was overloaded with subjects, causing teachers to overemphasize the teaching of subject content and placing less emphasis on the development of the basic skills and competences in Reading, Writing and Arithmetic that are necessary in order for learners to effectively learn content." (Tanzanian Government Policy Report, 2016)

Twaweza, a well-known East African civil society organization, speaks directly to this issue and reports that, "the learning expectations implied by the curriculum are that

[11] The inclusion of "technical" subjects from an early age appears to be a colonial legacy (Malisa and Missedja, 2019), as Mwiria (1991) also describes: "in keeping with a colonial ideology which stressed the role of the African as that of service to the white man, technical and agricultural (as opposed to academic) education were recommended for Africans by the missionaries, colonial authorities and external educational commissions".

children rapidly master basic reading skills in both English and Kiswahili, as well as basic numeracy skills up to multiplication[...] Contrary to curriculum expectations, the data show that many children in Tanzania do not master these basic skills quickly" (Twaweza, 2017). Furthermore, they highlight that although by the end of third grade students are expected to have mastered basic numeracy and literacy, students continue to develop these skills in later years (Twaweza, 2017). In sum, government agencies and external observers were in agreement that before the 3Rs reform, the curricular expectations and students' learning levels were clearly misaligned due to the presence of a typical overambitious curriculum.

Policy Reform

In response to the weak learning levels, the Government of Tanzania implemented the Big Results Now in Education (BRN) Initiative in 2013. The BRN policy included nine reforms, ranging from the mandated public release of within-district school rankings to infrastructure improvement to teacher training, all of which were rolled out at different times. The different policy changes are described in greater depth in Appendix D. In general, these policy changes did not overlap in terms of content or grades targeted by the reforms that we study here and therefore do not confound our main estimates. However, we still conduct empirical checks for potential heterogeneity based on these other reforms in Section V.

One of these policy changes, and the focus of this study, centered on curricular reform for grades 1-2, implemented in 2015. The aim of this reform was to strengthen the "3Rs" reading, writing, and arithmetic by allocating a larger share of the existing instructional time to numeracy and literacy. The new de jure allocation of time was such that 80% of the instructional time was supposed to be spent working on the three core

skills through the school subjects of math and Kiswahili. The remaining 20% of time was allocated for the other subjects. The school day was not extended, and therefore the policy entailed a re-allocation rather than an increase in class time. English was officially removed from the grade 1-2 curricula, in an effort to focus on literacy in Kiswahili, Tanzania's national language. Under the revised 2015 curriculum, English is only taught starting in grade 3[12]. Finally, content from some of the other subjects that had been removed from the new curriculum was incorporated into the curriculum via Kiswahili reading passages on science or social studies.

In practice, the change in the time allocation for numeracy and literacy did not increase all the way to the mandated 80% of instructional time by 2015 – yet, the change was sizable, as we show in Table 1[13]. We use data from class observations and government documents such as the policy report in Tanzanian Ministry of Education (2015) to estimate that before the reform, roughly 45% to 60% of the time was devoted to the "3R", including English lessons[14]. Using similar class observation data from 2015, we place the lower bound of the increase in instructional time for 3Rs at 1.3 hours, or ~14% from a base of 9.2 hours in the observational data. In other words, after the reforms, 70% of the total instructional time was devoted to the 3Rs, on average. This estimate includes English lessons, which should have technically not been taught post reform but that we

[12] While teachers indeed reported following this change post-reform in observational and survey data, in practice, we observe that this change did not happen as suddenly, and while the observed time of English instruction decreased ~40% post-reform, it was still being taught for about 2 hours per week (see Figure 1).

[13] The numbers displayed on this figure come from observational data collected for the original studies, only available for 2014 and 2015. While this data displays a large degree of missingness, especially in 2014, we were not able to find any reliable source quantifying precisely the allocation of time for the core subjects pre-reform.

[14] This is in rough agreement with the 9.2 hours (61% of the time) that we quantify on average for grades 1-2 for 2014 in the observational data.

still detect in our observational data. When we consider only math and Kiswahili, the focus of the policy, the increase in instructional time in total for both subjects in grades 1-2 was 2.4 hours, or 39% from a base of 6.2 hours per week in the observational data (that is, 57% of the total instructional time would have been devoted to the 3Rs, as opposed to the mandated 80%). In turn, this increase of 39% in instructional time towards the 3Rs serves as our upper estimate of the effect (i.e., the "first stage") of the policy in practice. The midpoint between these two bounds is an increase of 1.9 hours per week during a week that expects 15 hours of instruction. In other words, our mid-range estimate is that 12 additional percentage points, or almost two additional hours per week, were devoted to the 3Rs as a result of the reform.

An important factor to understand how this particular reform "slowed down" the pace of the curriculum is to understand which curricular inputs changed. In other words, if the pace of curricula is defined as "materials covered" over "time allocated to these materials", the slowing down of the pace of an overambitious curricula could happen through a decrease in the amount of material which is expected to be covered in class, an increase in the time allocated to this material, or both. Although the policy documents and curriculum descriptions do not explicitly mention which avenue was pursued by the Tanzanian government, we do not find any evidence that the expected amount of material to be covered within the "3Rs" changed in any way (World Bank, 2015, 2016, 2017; Ministry of Education, Science and Technology, 2016). Instead, from our teacher observation data in 2014 and 2015 and the main official policy description (Ministry of Education, Science and Technology, 2016), it seems like this reform slowed down the pace of the curriculum almost entirely through the channel of time re-allocation.

The spirit of the policy reform was aligned with best practices to improve learning levels that researchers and donor institutions have advocated for, and that have been effective in other interventions which have sought to better align instruction with student learning levels like "Teach at the Right Level" (for instance, in Banerjee et al., 2017). However, there are several reasons why it is not a certainty that learning would increase after implementing curricular reform at a national level. First of all, as expected in contexts with weaker state capacity, the implementation of the policy was not standardized, and not all teachers and schools received the same materials and degree of government support (Komba and Shukia, 2021). For instance, while 93% of schools claimed to have changed curricula to 3R in 2015, 4% of these also claimed to still teach English and Kiswahili in the grade 2, something that was explicitly contrary to the policy[15]. Similarly, not all teachers received the training on time: only 37% of all the teachers in our sample received the training, and 96% of these got it in 2015, after the school year had started. The distribution of materials was similarly scattered: from our survey data, we estimate that 4 of every 10 teachers in our sample do not have any textbooks that reflected the 3R curricular changes, and even among those that do, the kind of materials varied. Half of the teachers with textbooks that reflected the 3R curricular reform had them for writing and math, but only one third had books for reading[16]. Secondly, even when the underlying mastery of skills of socioeconomically

[15] Still, the rollout of the teacher training was gradual, but it ended up covering almost all schools: in 2014 only 14% of all schools had at least one teacher that had received training on the 3R curriculum, but by 2015, 99% of schools claimed to have at least one teacher who had received the training.

[16] In spite of all these logistical challenges, according to our survey data, 72% of all head teachers think that the implementation of the 3R reform went "well" or "very well". Note that most of these figures were calculated using the full experimental sample, not just the control schools, as it is the case with most of the other figures in this paper (see the Data section for more details on the sample). Tests of statistical

disadvantaged children in LMIC is improved through educational interventions, work such as Dillon et al. (2017) shows that these gains may not translate into gains in formal test scores, displaying the potential gap between not only curriculum and children's knowledge, but also children' knowledge and performance on assessments. Because of these reasons, the study of this particular reform is valuable to begin to understand whether the best practices of curricular alignment can indeed be implemented at scale by mostly government entities, and eventually be reflected in traditional learning measurements.

Research Design

Main learning panel. The main data source for this project was collected through the KiuFunza I and KiuFunza II projects, conducted in Tanzania between 2013 and 2016. These projects were randomized controlled trials studying school incentives and teacher bonuses respectively (Mbiti et al. 2019; Mbiti et al., 2021). Both studies included a core set of 180 schools from 10 districts. For this study, we focus on students in the 60 randomly selected schools which served as the control schools for the original RCT studies. Since the original experimental sample of Mbiti et al. (2019) consisted of a set of nationally representative public schools, and the current paper uses a subset of randomly-selected schools from this sample (that is, the control group from the RCTs), the current sample also consists of a sub-sample of nationally representative schools. It is worth noting that this control group did not receive any of the incentives that treatment schools did, and served solely to benchmark the effects of the other interventions. In other words,

significance showed no systematic differences between these two groups, so we chose to use the full sample to increase the precision of these numbers.

these schools would have been exposed only to the same policy and input changes as all other schools in Tanzania over this period[17].

Within these 60 schools, we have a longitudinal panel of grade 1-3 students for three years, 2014-2016[18] (although our main specification uses 2014-2015 for reasons that will be detailed in the next section), where students were assessed at the end of each school year with grade-specific assessments. The initial sampling of these students was such that 10 students from each grade were randomly selected and tested within each of the 60 schools. Once selected into the sample, these students were then followed for the duration of the panel until they reached the last grade surveyed, or they left the school for any reason. From the 3,000 unique students who were recruited as part of our study within the 2014-16 period, we end up with learning outcomes for 2833 of them. Since these schools did not receive any exclusive or targeted intervention that other school in different parts of the country did not also receive, the attrition in the sample does not threaten the representativeness of the dataset, as schools outside of the panel would have also been expected to display similar patterns of attrition. For most of the students in this panel, we also have information on household characteristics, and non-financial educational inputs at the household level – although these covariates were only collected in 2015 and 2016.

The tests measuring learning outcomes were designed and administered by Twaweza, who was simultaneously also in charge of the broader Uwezo initiative across

[17] For more information on where schools were located, and how these districts compare to the rest of the country, please see the Appendix ("c. More contextual details").

[18] We also have data for 2013, although due to changes in the assessments, and the lack of item-level data, we decide to only use it for robustness checks. In particular, in 2013 the difficulty of the grade 2 and 3 tests was deemed to not provide enough discriminating power. Therefore, the difficulty was adjusted for the following year, and then kept constant for the remaining three years.

East Africa which aims to document at-scale learning levels in foundational numeracy and literacy for children under 17. The tests were low-stakes exams, used purely for research purposes. Every year of the study, the students took a grade-specific test in math, English and Kiswahili. The test provided item-level data by subject (e.g., Kiswahili) and by sub-topic (e.g., reading words in Kiswahili) The assessments were similar across years, which was partly done by developing test booklets which kept the same items "in spirit" across years, but whose digits or words were modified each subsequent year. Appendix E shows examples of math and Kiswahili questions for all four years.

For our primary outcomes, we use the continuous test scores obtained from these assessments. We show two different scoring approaches, one scoring the tests as a raw percentage of the total number of items per subject, grade and year, and another using item-response (IRT) for each test at the level of the subject, grade, and year (e.g., English for grade 1 in 2014). We use the IRT scores as our main test scores because IRT scoring can place weights differentially by item to maximize discriminating power, but for the most part, none of our results are sensitive to this choice. We standardize these scores within subject and year for all three years.

As an additional robustness check, we also attempt to create a second set of outcome scores by incorporating the 2013 scores so that we are able to examine a longer pre-treatment trend. However, these scores come from assessments that were different between 2013 and the rest of the years, harming the comparability of these scores with those from other three years. Furthermore, we do not have access to item-level data for this year. So, to incorporate this additional baseline year, we use the actual test booklets

to manually flag questions within the 2014-2016 test booklets that most resembled those asked in 2013 for all subjects and grades, with the goal of creating a "pseudo-2013" test out of the 2014-16 assessments. We then created a percentage score as an outcome for each grade, subject, and year, considering only those items that made the 2014-2016 most resemble the 2013 assessments. Finally, this outcome is also standardized by subject and grade against the pooled sample from all years. As an additional robustness check, we repeat this exercise using the 2014 booklets to find equivalent questions in 2015 and 2016 so that we also have access to a "pseudo-2014" measure. In a sense, this approach not only serves as a robustness check by adding a baseline year (in the case of the "pseudo-2013"), but by ensuring the comparability of assessments across years by manually picking items that most resemble each other across time.

Leveraging the item-level data, we also create two other sets of outcomes of interest. First, we identify which specific sub-skills each student is mastering each year. In particular, we follow the approach of international assessments like Uwezo and determine that if a student can answer over half of all questions for a given sub-skill correctly, they are flagged as having mastered that sub-skill. Second, we leverage the outcomes for these sub-skills to label each student as having achieved specific "grade 1-" or "grade 2-proficiency" in each of the three subjects. We define "minimum grade-level proficiency" based on the curricular expectation pre-reform, and as such, these are mastering addition by grade 1, and multiplication by grade 2. For Kiswahili and English, these consist of reading sentences by grade 1, and reading paragraphs by grade 2. Both of these measures allow us to speak to policy effects on more concrete units of policy-relevance like grade-level proficiency and mastery of key numeracy and literacy skills.

In terms of school and teacher data, we have some information on school facilities, management practices, and school income and expenditures. Appendix Figure 1 describes these schools in our sample as of 2013. Enumerators also surveyed all teachers (about 1,500) who taught the students in our focal grades (grades 1, 2, 3) and focal subjects (math, English and Kiswahili), and collected data on individual teacher characteristics such as education and experience, as well as effort, teacher satisfaction, and teaching practices (e.g., whether teachers tried "tracking" within their classrooms).

Other achievement data. The main learning data from Mbiti et al. (2019) and Mbiti et al. (2021) has two key strengths in that (1) it has item-level information, which allows us to decompose treatment effects by sub-skills driving the changes, and (2) it samples the same schools and children across time, reducing the extent to which differences across time are simply due to random sampling variation. We complement this main learning data with two additional measures of student achievement that do not have these advantages but that do allow us to examine achievement trends over a longer period of time. First, we use Uwezo learning data from 2010-2017 (excluding 2016, as Uwezo data was not collected this year). Uwezo is a large-scale national, citizen-led data collection effort led by civil society organization Twaweza as a tool to benchmark learning outcomes in East Africa, including Tanzania through a low-stakes assessment that aims to be as representative of the whole country as possible. These data sets are publicly available, and cover children of roughly ages 5-17. Like our main outcomes, Uwezo tests cover English, Kiswahili, and math, and in fact, the test booklets administered for the Mbiti et al. studies (2019, 2021) are modelled after the Uwezo tests. Uwezo data does not have item-level outcomes, but it rather places children at a given

"level" for each sub-skill within each subject (e.g., student *j* is at the addition level in math, at the letter level in English, and at the syllable level in Kiswahili). We transform these outcomes into numeric scores that are comparable for all years across the 2010-2017 Uwezo panel, and which allow us to compare these outcomes with the outcomes from the main panel of learning outcomes[19]. This secondary data set allows us to increase the number of observations in the estimations, and to explicitly test for pre-trends – which the single pre-period in the main learning data does not allow for. Having said this, we give preference to the data from Mbiti et al. (2019) and Mbiti et al. (2021) because, again, Uwezo does not provide item-level data which allows for a more consistent grading of the outcome, is not necessarily sampled in a consistent manner across years, is not statistically guaranteed to be nationally representative, and consists of repeated cross-sections of data collection, increasing the risk of random noise affecting cross-year comparisons[20].

The second additional source of achievement data consists of test scores for national examinations in grade 4 ("Standard Fourth National Assessment" or SFNA) and in grade 7("Primary School Leaving Examination" or PSLE). These are publicly available at the individual-level at https://www.necta.go.tz, and contain information about the universe of students in Tanzania, allowing us to understand what happened to school

[19] We realize that, qualitatively, the discrete changes from certain sub-skills to higher skills may not represent the same underlying change in learning outcomes. In other words, it may be cognitively more challenging to go from the level "nothing" to "counting" than to go from "addition" to "subtraction". However, for the purposes of this analysis and additional robustness checks, we needed to create a unified continuous score using the Uwezo data, and as such, this was the most transparent approach.

[20] Furthermore, the potential categories in which children could have been placed within the Uwezo dataset changed from 2014-2015, only to then return to the previous system in 2017. More specifically, in 2015, some skills were broken into finer levels of disaggregation, such as addition becoming "single-digit addition" and "double-digit addition". This forces us to make assumptions in terms of how to reconcile this data cross years, and our results may be affected by these empirical decisions.

enrollment by the time the students in our main panel reached grade 4. For the purposes of this analysis, we use scraped data on both tests from 2015-2018. The main goal of the grade 4 test scores is to understand the long-term effects that the policy reform had on grade 4 exam passing rates. The grade 7 students serve as a control group for the same period, given that even the oldest cohort to be affected by the policy would not have been in grade 7 until 2020, outside our period of study. The scores for both of these assessments are reported separately for each subject, and the outcomes are given in letter grades, where a student needs to score a C or above to pass the examination. Using these letter grades, we create binary variables flagging whether a child passed that subject or not. Given the anonymization of our main panel, we cannot link our initial learning panel with this administrative data base at the student level, but we still analyze these data at the level of student.

Empirical strategy. We exploit the variation in the timing of the policy introduction, and in the grades targeted by the curricular reform to estimate the impacts of the reform on learning outcomes through a difference-in-differences (DiD) framework. The intuition behind our identification strategy is that, absent the curriculum reform, the trend in performance for students in treated grades (1-2) would have remained similar to that of students in the untreated grade (3). Therefore, our preferred specification follows the structure of a classical two-period, two-group DiD strategy, like that found in Beatty and Shimshack (2011) and Carvalho and da Mota (2017). In this case, our first difference consists of the difference in learning levels, within each grade, before and after the reform. Our second difference is the difference between grades that were targeted by the reform (grades 1-2), and the grade that was not (grade 3), which is how we account for

the "secular trend" in the specification. In other words, we look at grade 1-2 outcomes before and after the reform, and account for trends in how learning levels changed over the same time period for other grades untreated by the reforms using the grade 3 data. In Appendix Figure 2 we display the different groups that are part of the identification strategy.

In particular, we estimate the following model:

$$Outcome_{ijgt} = \beta_0 + \beta_1(Treatment*Post_{gt}) + \lambda_g + (Post)_t + \varepsilon_{ijgt} \quad [2]$$

Where the "Outcome" refers to the learning or enrollment outcome for individual *i*, subject *j*, grade *g* in year *t*. We introduce grade-level fixed effects through λ_g, and $Post_t$ is an indicator variable which equals 1 for 2015. The coefficient of interest is β_1, attached to the "Treatment*$Post_{gt}$" term. Specifically, this variable equals 1 only when a child is in grade 1 or 2 in 2015 and the year is 2015 (after the reform was implemented).

For our main specification, we focus only on 2014-2015 data. Note that given the panel nature of the data, students who were in grade 3 in 2016 were affected by the policy when they were in grade 2 in 2015. Therefore, we give preference to the data from 2014-15, as opposed to also including 2016. Including the 2016 data with the current specification would group a cohort that was actually treated into the comparison group and would "contaminate" our comparison group. More specifically, if the reform had positive effects on learning, this approach might yield underestimates of any potential increases in learning as a direct result of the policy.

Similarly, notice that in our current specification, the students in 2014 who constitute the comparison group for grade 1 are also those who are the treated group for grade 2 in 2015. Contrary to the case described in the paragraph before, we do not believe

that this poses a threat to our identification strategy. This is because outcomes are observed at the end of each grade and therefore, the outcomes for this cohort when they were in grade 1 are observed in 2014 after completing grade 1 under the previous curriculum, and as such, our specification correctly groups them into the comparison group for grade 1 – those who did not receive the 3R curriculum in grade 1. In the same manner, we observe their end-of-year outcomes for grade 2 in 2015 – after having completed grade 2 under the 3R curriculum, and hence, are properly classified as part of our treatment group by the current specification. A similar argument can be made for those students in grade 2 in 2014, who are part of the comparison group for grade 2 in 2014, and then become the group which contributes "post" information for grade 3 in 2015. In all, we do not believe that this issue poses a challenge to our internal validity, as this specification properly groups students into their corresponding treatment and comparison classifications within each year. Given the relatively small number of groups, we present both robust standard errors, and also p-values emerging from wild-bootstrapped clustered standard errors at the grade-level. In general, our results are not sensitive to the empirical decisions described here.

The difference-in-differences identification strategy requires that we justify whether the parallel trends assumption holds in this case. In other words, our main the assumption is that, absent the curriculum reform, students in a treated would have experienced similar trends in performance to untreated grades. In the case of the long-term analyses, this assumption would imply that, absent the reform, the number of test-takers and passing rates in grade 4 would have experienced the same trends as the number of test-takers and passing rates in grade 7 within the same year. Unfortunately,

we cannot explicitly show parallel trends using our main data (that from Mbiti et al., 2019 and Mbiti et al., 2021), as we only have one year of consistent data from the period before the curriculum reform. We explore this issue by using 5 years of Uwezo data before the reform to visually check for differences in the income cohorts in this large-scale assessment. As mentioned before, Uwezo tests in Tanzania are very similar to the instruments used to collect the data used in the current paper, as they were developed by the same organization, around the same time, and with the same aim of measuring foundational knowledge in math, English, and Kiswahili. We visually show in Figure 1 that students in lower grades (1-4) do seem to move in the same trajectory in all three subjects, for the 5 years of data available in the pre-period.

Qualitatively, we argue that the "spirit" of parallel trends might not be met in at least two cases. First, there could be another policy that heterogeneously affects one of the grades in the sample and hence confounds our estimates. As previously discussed, we are not aware of any other policy of the kind for these grades between 2014 and 2015. We also believe that the parallel trends assumption may not hold if there is a change in the composition of the incoming cohorts, which may introduce selection bias in the estimates of our treatment effects. For this specific case, we are aware that in 2016, the Tanzanian government introduced the Fee-Free Basic Education (FFBE) policy, which made primary education more accessible to students of more disadvantaged socioeconomic status in grade 1. Specifically, we observe in the data that the pupil-to-teacher (PTR) increased from 87 to 122 (40%) for grade 1 between 2015 and 2016. However, the PTR for grades 2 and 3 remains constant at 79 and 32 respectively over the same period. However, since we focus on the 2014-15 period, this reform does not affect our main

estimates, nor do these cohorts reach grades 4 or 7 within the time window of our long-term analysis.

Results

The reform improved foundational literacy and numeracy in grades 1-2. We find that the curriculum reform had a positive and statistically significant effect on math and Kiswahili learning outcomes one year after the reform. As the first row of Table 2 shows, students experienced an increase of 0.19 SD in an index outcome that combines the two main subjects targeted by the reform, and an increase of 0.20 SD in each of these subjects when estimated separately. As shown in the other two rows of Table 3, these results are directionally the same, and even of larger magnitude, if one uses secondary measures that attempt to increase the comparability of the tests across years. Similarly, as Appendix Table 3 shows, these results are directionally the same when using outcomes from the Uwezo dataset, although the differences in how the outcomes are reported and the different time period do change the magnitude of the treatment effects (in this case, decreasing them closer to 0.1 SD). Although the reform de-emphasized English instruction, we find no evidence of large reductions in English test scores in our main outcomes. Our point estimates are positive and the standard errors are such that we can rule out negative effects smaller than -0.02 SD with 95% confidence. As we will explore further when we discuss the effects on sub-skills, we believe the improvement in English was due to spillover effects on basic skills transferrable from one language to the other.

Another way to understand these learning gains is to examine what happened to levels of minimum grade-level proficiency as a result of the policy. As shown in Figure 3, these results are not only meaningful in units of standard deviations but also in terms

of reaching minimum proficiency levels. For instance, the policy reform increased the likelihood of a student reaching grade 1 math proficiency by 40%, and it more than doubled the likelihood of a student reaching grade 2 math proficiency. Similarly, it increased the probability of reaching grade 1 proficiency in Kiswahili by 29% and grade 2 by 71%. Even in English, the probability of reaching grade 2 proficiency increased by 7 percentage points over a base of 2%. The large magnitudes of these relative increases across all three subjects are partly due to the significant positive effects on learning, but also due to the low baseline levels of learning achieved by pupils in the sample, and in Tanzania more broadly.

In terms of attrition, we estimate that the reform had a causal reduction in the attrition rate from the sample of 6 percentage points. Our data cannot track individual students across the universe of Tanzanian schools, so we cannot definitively claim that this attrition is equivalent to school dropout. In other words, there are two potential interpretations for the reduction in attrition among treated students. The first hypothesis is that the reform made students more likely to remain in the schools where they were enrolled at the start of our data panel. Although plausible, the reform did not target specific schools, so we do not have an ex-ante reason to believe that the relative quality and desirability of schools changed as a result of the reform. The second interpretation, and the one we favor, poses that this decrease in attrition among treated students was indeed linked to a decrease in dropout. Particularly when this hypothesis is coupled with the results we present below on longer-term outcomes, it appears the policy not only led to improved learning, but also higher enrollment retention of students.

Skills across the range of complexity improved as a result of the reform. We would also like to understand whether the policy had heterogenous effects on the different literacy and numeracy sub-skills (e.g., "reading words in English") that were assessed. This is a valuable exercise as it can provide evidence on the mechanisms through which the reform operated. For example, did the reform only improve basic skills but weaken the more complex sub-skills? Or did help students master higher order concepts, but not at improving the more foundational skills? Since we have access to item-level data which we can aggregate up to the level of these sub-skills for each student, and we leverage the fact each grade was tested on very similar topics and skills across years and estimate the effect of the policy on each sub-skill. We use our main difference-in-differences specification to estimate the effect of the policy on the likelihood of mastering each of the sub-skills that students were tested on.

Figure 2 shows the estimates of the policy on specific sub-skills by subject. For math, it is not clear that the level of complexity of the sub-skills moderated how much the reform affected these tasks. In fact, sub-skills across the whole spectrum of complexity benefited from the policy (e.g., inequalities, addition, and multiplication). If at all, this figure shows that the reform indeed strengthened the most foundational sub-skills at a similar rate as the more complex tasks. Similarly, there were gains across the spectrum of complexity in Kiswahili. Together, these two findings suggest that the improvements in learning spurred by the reform did not come at the expense of sub-skills at either end of the spectrum: they did not "over-simplify" the instruction such that only the most foundational skills were improved, nor did it only benefit pupils already mastering a certain level of proficiency.

Interesting, the two most basic English sub-skills that were assessed, meaning "recognizing letters" and "reading single words", also seem to have improved as a result of the policy. Even if the policy moved instruction away from English, if improvements in Kiswahili literacy were to have any spillover effects on other subjects, one would hypothesize that they would be in the most basic literacy skills of another language which uses the same script, but not necessarily as much in higher order English skills, as we find here. As such, de-emphasizing English during these two first grades did not lead to overall losses in English learning. This was partly due to the low baseline levels shown in Table 3, as students were close to floor of the assessment at this point. However, we also have suggestive evidence that another potential mechanism for the lack of decreases in English test scores was by creating a common foundation in Kiswahili to build upon and, as such, the policy was also able to speed up the acquisition of more advanced skills in a different subject.

The policy led to higher enrollment, but lower passing rates four years after first implemented, likely due to compositional effects. Next, we explore whether the policy had persistent effects on educational outcomes. We leverage the universe of standardized national test scores from 2015-2018 for grades 4 and 7 grade to explore whether the curriculum reform also led to changes in educational attainment in the longer term, which we show in Table 4. In particular, the first cohort to be fully under the new curriculum is those students in grade 1 in 2015, and who were our treated group in grade 4 by 2018. Therefore, when using these data and our main model, our treatment group consists of repeated cross-sections of grade 4 students, with grade 7 students serving as the comparison group. The pre-period consists of the 2015-2017 period, and the post

period 2018. Note that under this set up, those in grade 4 in 2017 were technically affected by the new curriculum when they were in grade 2. As such, the estimates that emerge from using the full sample here can be interpreted as underestimates of the true estimates. However, as a robustness check, we also display in the second row, the estimates resulting from the same specification but dropping 2017 for both grades 4 and 7 students, which would remove any potential (upwards) bias from the control group.

These results show two key patterns. First, the number of test-takers —a proxy for system-wide enrollment— increased by 16-17%. This is consistent with the decrease in attrition observed using the main panel data shown in Table 2. This result is also consistent with the hypothesis that learning and enrollment are linked to a certain extent, as either enrollment leads to higher learning (as Bau et al., 2021 might suggest), and/or higher learning leads to a higher likelihood to remain enrolled. Having said this, this increase in enrollment came with decreases in the passing rate of these national grade 4 examinations. In particular, using the baseline rates as a benchmark, the passing rate in math decreased 5-7%, the passing rate in English decreased 11-16%, and the passing rate in Kiswahili decreased 16-19%.

These long-term changes are suggestive of two facts. First, the increase in enrollment led to compositional changes in the universe of students reaching grade 4, particularly towards the inclusion of lower-performing students who would have otherwise dropped out of school by grade 4. Second, these decreases in passing rates do not negate the short-term learning gains in learning observed. At worst, these decreases in performance are comparable in magnitude to the increase in student enrollment, which suggests that aggregate learning and educational attainment still increased –if one

assumes that those who did not pass still learned something over these four years– relative to a counterfactual where the reform was not implemented and fewer children would have been enrolled in school.

For the most part, more disadvantaged groups of children benefited more from the policy. As described in Section III, rolling out nationwide curricular reforms in a large country like Tanzania is logistically challenging, and it is likely to yield heterogenous effects at the local- and individual-level due to variation in implementation across contexts. A key component of a curriculum reform of this scale is teacher training, as teachers must be aware and capable of implementing the expected instructional changes. In fact, weak teacher training was identified as one of the main reasons for the failure of a curricular reform aimed at improving early literacy outcomes in grade 1 in Costa Rica (Rodriguez-Segura, 2020). Therefore, we explore the extent to which teacher training may have moderated learning gains in this context. Teacher training was not randomly assigned at baseline, and its implementation varied across Tanzania depending largely on the regional entity in charge of imparting the training (Komba and Shukia, 2021). As such, we can only provide suggestive and correlational evidence for the issue of teacher training. Having said this, Table 6 shows the treatment effects of schools that had any teacher trained in the new 3R curriculum in 2014 - before the policy was actually implemented, and for those schools that did not. This table suggests that receiving teacher training was imprecisely correlated with larger treatment effects across the two subjects. Together, these results are suggestive that beyond informing teachers of a change in the allocation of time across subjects, training them on how to do it may be a key element to achieve larger learning gains through a reform of this type.

We also study whether certain demographic characteristics are correlated with heterogeneous gains in learning, as shown in Table 7. We observe that female and rural students drove most of the treatment effects for learning. In other words, groups that are typically considered more disadvantaged in this context benefited the most from the policy in terms of learning in literacy and numeracy. We do not observe any heterogeneity by grade, as the difference between the two grades is not substantively or statistically significant. In terms of attrition, the patterns are similar except for the difference between urban and rural students, as urban students drive most of the decrease in attrition. In all, while this sub-group analysis sheds light on some heterogeneous effects by demographic characteristics, it does not reveal that an exclusive sub-group benefited from the policy across the board. Instead, the gains seem to be, to some extent, distributed across different sub-groups, and if at all, they benefited disadvantaged groups more than their peers.

Other contemporaneous reforms do not appear to confound the effects. As described in Section III, this curriculum reform was only one part of a suite of reforms undertaken under the heading of Big Results Now - described in greater depth in the Appendix ("d. Description of other contemporary reforms"). None of these reforms targeted directly or differently our treated and comparison groups, and some of these reforms even happened after our period of analysis. However, we still empirically test whether we find some heterogeneity due to these reforms.

One of the other reforms that could be affecting our results is the Student Teacher Enrichment Programme (STEP - implemented in 2014). This policy trained teachers on how to identify struggling students and support them. The STEP training was rolled out

in selected districts, and 4 out of 10 of our districts were in this group. We run our main specification only on districts that were not STEP districts, and display this in Table 7. When broken down by whether a district was part of the STEP program, the results similar and the difference is not statistically significant. Therefore, it does not appear that the implementation of the STEP program is confounding our main treatment estimates.

There were two other school-wide reforms for which we check whether we have heterogeneous effects: the distribution of School Improvement Kits (including the "Mwaongozo" leadership training for head teachers), and the school grants disbursed by the Tanzanian government. The former deals with the quality of school management, and the latter deals with a fairer distribution system of school funding. Table 7 again shows the heterogeneity results for both of these school characteristics. Although the differences between the groups do not rise to be statistically significant, the magnitudes of the differences are medium-sized. Much like Mbiti et al. (2019) and Mbiti et al. (2021), we believe that, if these differences are indeed suggestive of treatment effect heterogeneity, the presence of adequate school resources may have augmented the effectiveness of the curricular reform, but not necessarily confounded the treatment effects, as none of these policies were targeted at specific grades. It is also worth noting that the implementation of most of BRN components were delayed due to the lack of funding. For instance, the capitation grant reform was only launched in 2016, the last period of our study.

Finally, these schools were explicitly chosen as control schools in the companion experimental evaluation, so by default, we ensure that they were not affected by the other interventions being rolled out by researchers. Other reforms, such as the school ranking program was the first component launched and one of the few that was consistently

implemented throughout our study period. However, this program focused on results in grade 7, which would be completely out of reach for even our oldest cohort.

Discussion

Our results suggest that the Tanzanian curricular reform of 2015 improved foundational literacy and numeracy for early grade students. The targeted restructuring of instruction within an overcrowded curriculum, coupled with low achievement levels at baseline, led to significant improvements in proficiency levels for early literacy and numeracy. These results are robust to the use various selections of items in the assessment, different definitions of the outcome variables, and do not seem to be fully driven by any of the other Big Results Now reforms. These findings provide empirical backing for the prior set forth by papers like Pritchett and Beatty (2015) or Muralidharan et al. (2019), which advocate for a realignment and simplification of curricula in LMIC to allow students to properly develop early literacy and numeracy. These results also describe a successful case study where such a reform was implemented and led at the national level by the government of a LMIC like Tanzania.

The strengthening of the foundational numeracy and literacy skills through this curricular reform in the earlier grades highlights the key role that curriculum design plays as a key input for educational systems. In particular, these results challenge policymakers to re-think the focus of curricula in developing countries, and their specific targets during the earlier years of education. In a sense, the poor learning outcomes in developing countries need not be fully explained by irreversible school and student characteristics, but also by the pedagogy of how the material is taught, and what is expected of students. Failing to meet educational standards could be both due to the student's low levels of

learning, but also due to the stringent, overambitious, and unrealistic standards that they are subject to. Interventions such as Teach at the Right Level (for instance, see Banerjee et al, 2017) or the current study show that thoughtful curricular design and pacing can lead to promising gains in learning.

Importantly, curriculum reforms like the one that we study could improve learning outcomes in the developing world without the need for any additional resources like more teachers or instructional time. The investments made by the Tanzanian government on this specific project mostly entailed the printing and distribution of materials like new textbooks, and the costs associated with the training of teachers on the new curriculum. While we were not able to get cost estimates for these components, our best prediction is that their costs are fairly small relative to the overall education budget for the country. The fact that this type of curriculum reform is likely close to being a fiscally-neutral intervention that also led to significant improvements in the learning levels across the country makes it particularly appealing for other LMICs that are similarly resource-constraint. Other types of interventions that have been typically identified as promising for educational systems in LMICs (World Bank, 2020), such as scripted lessons, merit-based scholarships, or customized instruction via edtech tools, tend to carry a larger price tag with them, which in turn makes them harder to implement even if they are equally as effective at raising learning. Beyond curriculum reforms, interventions that aim to improve pedagogical practices in LMICs tend to be relatively low-cost while also targeting potentially "low-hanging fruits" in the improvement of classroom instruction, and as such, they emerge as a rich field for future research and policy design.

We also find that the curriculum reform led to increased school enrollment, at least until grade 4. The decrease in student dropout as a result of a curriculum reform which focuses on strengthening FLN is both a welcome and unsurprising effect. A curriculum that is more tailored to most students' needs and does not focus (as much) on high performing students —likely from a high socioeconomic background— is likely to have larger effects on students that were more likely to leave school prematurely, as shown in the current study. Yet, this decrease in school dropout does not imply that the educational system has done its part with these new entrants. We also observe decreases in the passing rate of the grade 4national examination that matches very closely the increase observed in school enrollment. This fact suggests that even if the reform did lead to higher enrollment and learning gains in the short term for FLN, these gains were not enough for most of these new entrants to pass the grade 4 national assessments. These new students are likely to be from more disadvantaged backgrounds, and while the curriculum reform likely aligned classroom instruction closer to their achievement level, it may have not met all the educational needs of these children. Therefore, these results are indicative that while this type of curriculum reform may be beneficial and desirable, it may not be enough to ensure educational success in the long-run. Additional interventions, such as more individualized instruction or the revision of the curricula of the higher grades as well, may be needed to help these children keep succeeding later in their educational path.

Similarly, the effectiveness of a new curriculum, as well-designed as it may be, is likely to be dampened if all of parts of said educational system are not aligned to work well with this change, that is, if "implementation" is poor. For instance, in the case of the

3R curriculum reform, training even a single teacher per school ahead of implementing the reform was correlated with larger gains in learning. This is suggestive that, unsurprisingly, the quality of implementation of a new curriculum can moderate the effects of curricular reform policies. The World Bank makes this point on their Report on Learning: "if a country adopts a new curriculum that increases emphasis on active learning and creative thinking, that alone will not change much. Teachers need to be trained so that they can use more active learning methods, and they need to care enough to make the change because teaching the new curriculum may be much more demanding than the old rote learning methods" (World Bank, 2017). Even in LMIC with weak state capacity, well-designed, and well-implemented, programs can greatly improve literacy outcomes in developing countries (for instance Kerwin and Thornton, 2019; or Eble et al., 2020). We display a case in which a well-designed policy had the intended results in learning gains on average, but also where the results were likely magnified, at least correlationally, through better implementation at the local level.

Our study has several shortcomings and limits to what can be inferred from these results. First of all, our main results cover a very short time period. Therefore, we cannot ensure that the parallel trends assumption holds in the same data from which we draw our our main estimates. To address this, we use a different data set, Uwezo, and qualitative knowledge of the context to justify why parallel trends might hold. However, these options are only second best to a more comprehensive check for parallel trends in the same data set as the one we use for our treatment effects estimates. Second, the learning data collected at the beginning of each year was of poor quality, and as such, we cannot provide direct evidence on the heterogeneity of the effects by baseline performance. This

is a valuable area for future research to explore, as a potential worry with this type of curriculum reform is that it may affect high-performing students at the expense of low-performing students. Finally, we do not have strong metrics for the quality of implementation of the reform in each school, beyond information about teacher training on the new curriculum. Although we present some suggestive evidence that the quality of the implementation may moderate the effects of the policy, further research is needed on this issue, especially given other evidence (Komba and Shukia, 2021) highlighting that the policy was heterogeneously implemented across the country.

In all, our findings contribute to the literature on curricular reform in developing countries. More broadly, our results speak to the issue of adapting antiquated and "overambitious" curricula in developing countries to the current educational needs. Curricula affect all students within an educational system, and as such, well-designed and well-implemented curriculum reforms can be a valuable tool to boost educational outcomes at scale. The current study presents evidence of such a reform which was indeed successful, yet not perfect, at improving learning in a LMIC like Tanzania.

TABLE 2.1
Estimated Time Allocation in Hours Per Week Across Subjects and Grades Before and After the Reform

	Grade	Math	Kiswahili	English
2014 (pre-reform)	1	3.3	3.1	2.9
	2	3.0	3.0	3.0
	3	4.1	4.4	4.0
2015 (post-reform)	1	3.6	5.0	1.9
	2	3.6	5.0	1.9
	3	4.1	4.2	4.2
Percent change from 2014 to 2015	1	9.1%	61.2%	-35.5%
	2	20.0%	66.7%	-36.7%
	3	0.0%	-4.5%	5.0%

Notes. Figures derived from data on class observations by external enumerators.

TABLE 2.2.

Regression Estimates of the Causal Effect of the Curriculum Reform on Learning and Enrollment

	Aggregate, all three	Aggregate, Math+ Kiswahili	Math	English	Kiswahili	Attrition
Main estimates (IRT)	0.16** (0.06) [0.14] 3275	0.19** (0.06) [0.19] 3275	0.20** (0.07) [0.17] 3132	0.14 (0.08) [0.33] 3132	0.20* (0.09) [0.17] 3132	-0.06** (0.02) [0.12] 3023
Pseudo-2014 test	0.16** (0.06) [0.13] 3275	0.19** (0.06) [0.18] 3275	0.18** (0.05) [0.04] 3132	0.12 (0.09) [0.4] 3132	0.23* (0.1) [0.18] 3132	-
Pseudo-2013 test	0.20** (0.06) [0.04] 3275	0.25** (0.07) [0.09] 3275	0.29** (0.08) [0.15] 3132	0.13 (0.09) [0.4] 3132	0.25** (0.08) [0.41] 3132	-

Notes. Coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

TABLE 2.3
Regression Estimates of the Causal Effect of the Curriculum Reform on Achieving Minimum Proficiency Levels of Grades 1 and 2

	Math		English		Kiswahili	
	Control group	Estimates	Control group	Estimates	Control group	Estimates
Achieving G1 minimum proficiency	0.47 (0.50) N=1066	0.19*** (0.04) [0.18] N=3132	0.03 (0.18) N=1087	0.04 (0.02) [0.17] N=3132	0.41 (0.49) N=1087	0.12** (0.04) [0.14] N=3132
Achieving G2 minimum proficiency	0.10 (0.30) N=1066	0.17*** (0.03) [0.12] N=3132	0.02 (0.13) N=1087	0.07*** (0.01) [0.12] N=3132	0.21 (0.41) N=1087	0.15*** (0.03) [0.15] N=3132

Notes: coefficients from linear probability model. Robust standard errors of coefficients, and standard deviations of control group in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

TABLE 2.4.
Regression Estimates of the Causal Effect of the Curriculum Reform on Learning and Enrollment

	Passing rate in math pre-reform	Math	Passing rate in English pre-reform	English	Passing rate in Kiswahili pre-reform	Kiswahili	Number of takers pre-	Students in 4th grade
Main estimates	0.45	-0.03***	0.38	-0.06***	0.8	-0.13***	1,100,868	173,014***
		(0.00)		(0.00)		(0.00)		(55.7)
		[0.21]		[0.16]		[0.27]		[0.16]
		N=8,062,450		N=8,062,450		N=8,062,450		N=8,062,450
Main estimates, no 2017	0.41	-0.02***	0.36	-0.04***	0.78	-0.15***	1,048,686	176,784***
		(0.00)		(0.00)		(0.00)		(8.55)
		[0.21]		[0.22]		[0.16]		[0.27]
		N=5,953,025		N=5,953,025		N=5,953,025		N=5,953,025

TABLE 2.5.
Comparison of Point Estimates of Schools that Received Some Teacher Training in 2014, and those that did not

	Aggregate (all three)	Aggregate (Math+ Kiswahili)	Attrition
No teachers trained in 2014	0.13* (0.06) [0.00] 3076	0.14* (0.07) [0.00] 3076	-0.06** (0.02) [0.00] 3076
At least one teacher trained in 2014	0.31* (0.13) [0.25] 771	0.37** (0.14) [0.26] 771	-0.06 (0.04) [0.00] 771
Difference no teachers trained-at least one teacher trained	-0.18 (p=0.20)	-0.23 (p=0.14)	-0.01 (p=0.86)

Notes. Coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

TABLE 2.6

Heterogeneity of Results by Different Baseline and Demographic Characteristics

	Female	Male	Female-Male	Rural	Urban	Rural-Urban	G1	G2	G1-G2
Aggregate learning index (Math+Kiswahili)	0.23**	0.15*	-0.08	0.25***	-0.14	0.39***	0.17**	0.20***	-0.03
	(0.09)	(0.09)	p=0.55	(0.07)	(0.11)	p=0.00	(0.07)	(0.07)	p=0.64
	[0.15]	[0.15]		[0.14]	[0.64]		[0.16]	[0.16]	
	1641	1634		2729	546		2174	2197	
Attrition	-0.10***	-0.02	-0.08**	-0.04**	-0.16***	0.12**	-0.07***	-0.05**	-0.02
	(0.02)	(0.03)	p=0.03	(0.02)	(0.05)	p=0.04	(0.02)	(0.02)	p=0.16
	[0.05]	[0.37]		[0.13]	[0.16]		[0.16]	[0.16]	
	1641	1634		2729	546		2174	2197	

Notes. Coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

TABLE 2.7

Heterogeneity of Results by Whether Schools were Affected by Other Contemporaneous Reforms or Policies

	STEP	No STEP	STEP-No STEP	School Improvement Kit	No school Improvement Kit	Kit-No Kit	Below median financial support	Above median financial support	Below-Above
Aggregate learning index (Math+ Kiswahili)	0.20*	0.18*	-0.02	0.20**	0.09	0.11	0.12	0.26**	-0.14
	(0.09)	(0.09)	p=0.89	(0.07)	(0.17)	p=0.54	(0.09)	(0.09)	p=0.28
	[0.38]	[0.12]		[0.18]	[0.13]		[0.08]	[0.18]	
	1313	1962		2796	479		1714	1561	
Attrition	-0.10**	-0.04	-0.07*	-0.06**	-0.07	0.01	-0.05	-0.08**	0.04
	(0.03)	(0.02)	p=0.07	(0.02)	(0.05)	p=0.86	(0.02)	(0.03)	p=0.32
	[0.08]	[0.27]		[0.14]	[0.13]		[0.12]	[0.13]	
	1313	1962		2796	479		1714	1561	

FIGURE 2.1: *Visual Display of Parallel Trends Using Uwezo Data Set*

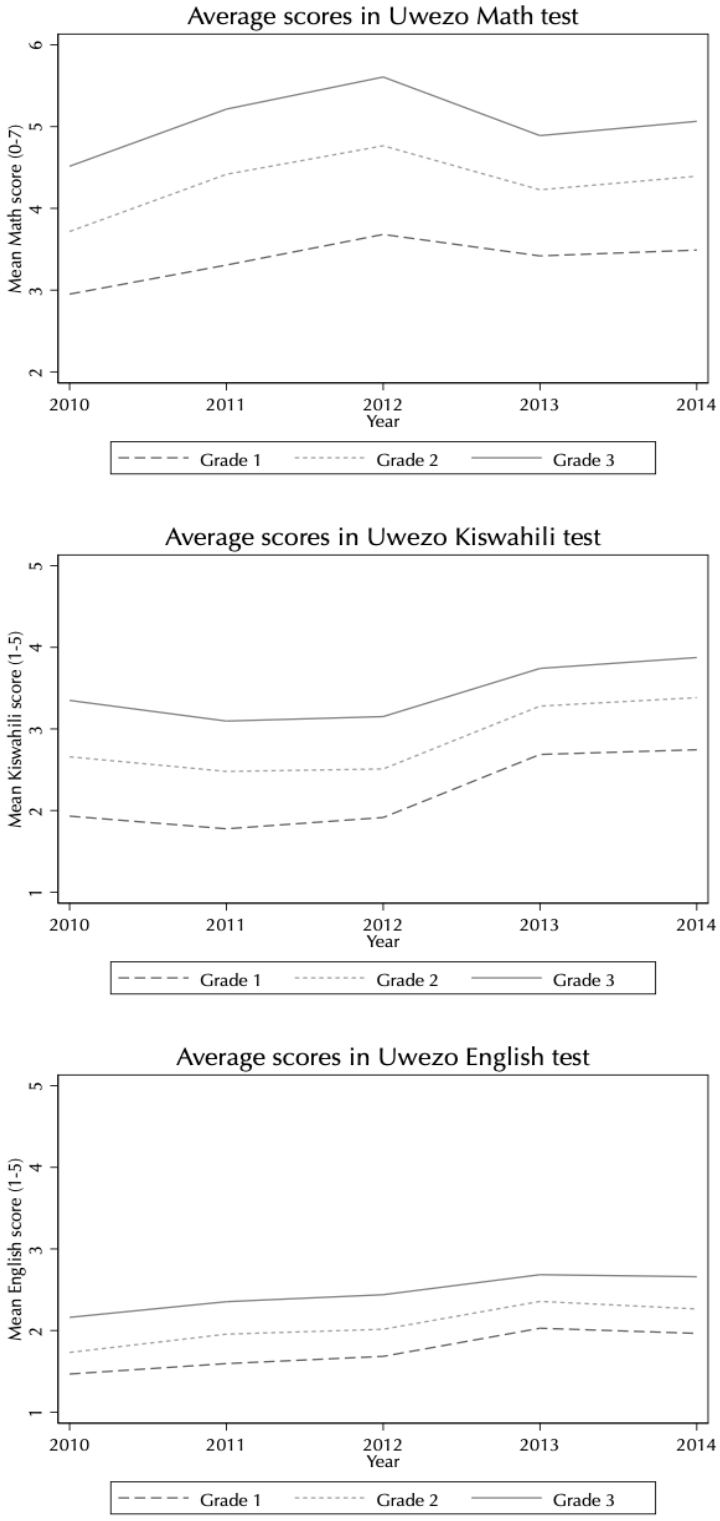
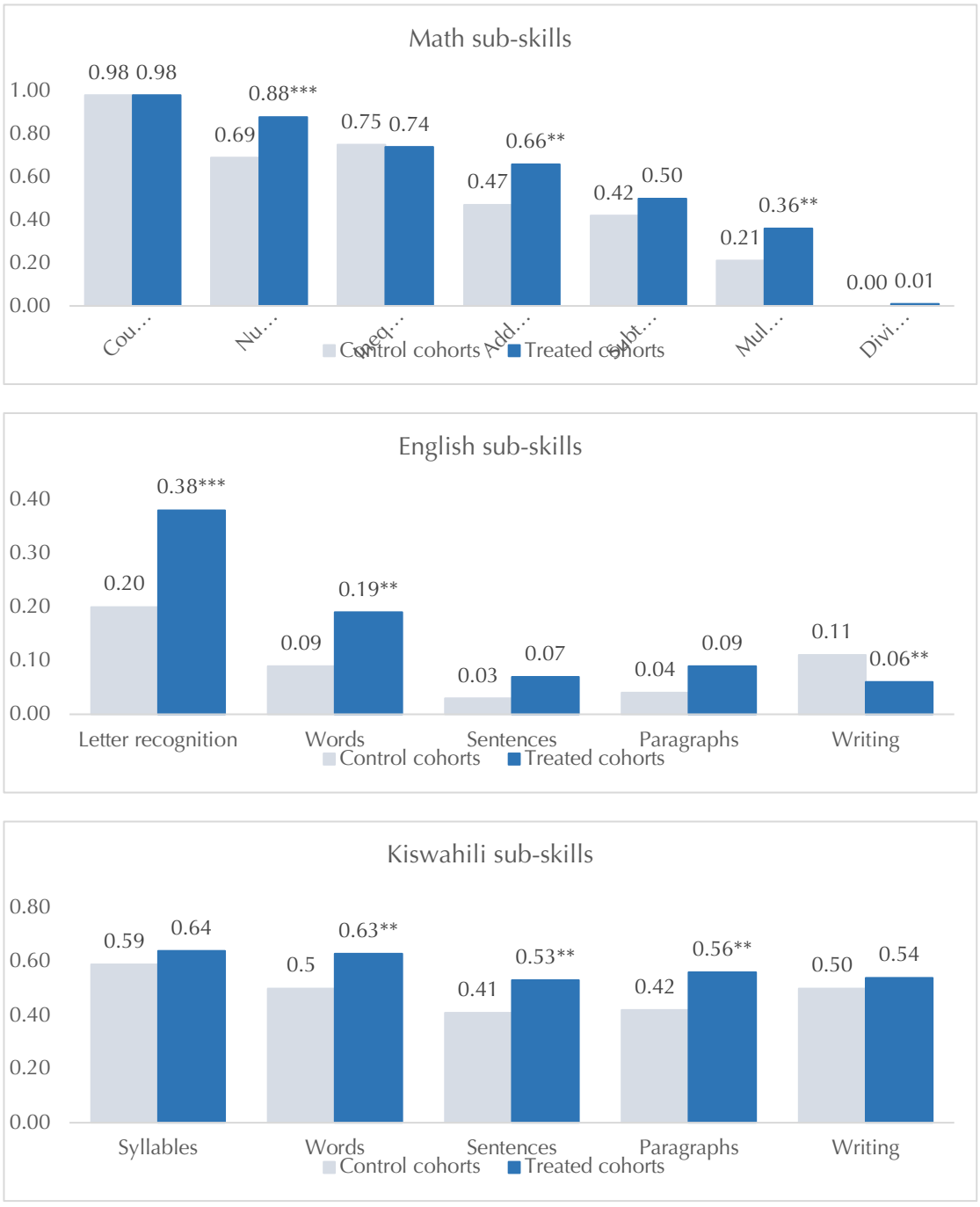


FIGURE 2.2: *Comparison Between Treated and Control Cohorts in the Probability of Mastering Different Sub-skills*



CHAPTER 3

Assessors influence results:
Evidence on enumerator effects and educational impact evaluations

(Daniel Rodriguez-Segura and Beth E. Schueler)

Abstract –A significant share of education and development research uses data collected by workers called "enumerators." It is well-documented that "enumerator effects"—or inconsistent practices between the individual people who administer measurement tools—can be a key source of error in survey data collection. However, it is less understood whether this is a problem for academic assessments or performance tasks. We leverage a remote phone-based mathematics assessment of primary school students and survey of their parents in Kenya. Enumerators were randomized to students to study the presence of enumerator effects. We find that both the academic assessment and survey was prone to enumerator effects and use simulation to show that these effects were large enough to lead to spurious results at a troubling rate in the context of impact evaluation. We therefore recommend assessment administrators randomize enumerators at the student level and focus on training enumerators to minimize bias.

INTRODUCTION

A significant share of research in low- and middle-income countries (LMICs) relies on data collected directly and on a one-on-one basis by workers called "enumerators"[21] (Lupu and Michelitch, 2018). This is particularly true in education research where surveys of teachers, parents, school leaders, and one-on-one assessments of student academic achievement are often conducted by teams of assessors. When collecting this type of data, individual enumerators can influence the recruitment of subjects, exercise significant discretion in how they interpret the information received, and shape how responses end up being coded in the data. In doing so, assessors can affect the quality of the data received by introducing measurement error through "enumerator effects", which happen when assessors record differential response rates or scores for similar populations of respondents (Olson et al., 2020). These inconsistent practices across enumerators or systematic variation in how respondents react to different enumerators can lead to erratic data, which could in turn yield spurious research results and unhelpful policy recommendations. Although the presence of enumerator effects has been well-documented in survey data (West and Blom, 2017; Di Maio and Fiala, 2018), researchers have not yet studied as much the extent to which they can affect educational assessments and the results of impact evaluations using these data.

In this paper, we study the presence of enumerator effects in a learning assessment of primary school children's early numeracy skills and an accompanying survey of their parents for over 2,500 students across 105 schools in Kenya, delivered by 20 individual assessors. We leverage the fact that this assessment was centrally

[21] Enumerators are sometimes referred to as "assessors" or "interviewers". We will use these terms interchangeably.

administered over the telephone, removing many logistical or geographic barriers to creating a fully interpenetrated design. In other words, administering this assessment and survey remotely allowed us to randomly assign enumerators to students at the individual-level. Importantly, this phone-based assessment (PBA) was the first set of outcomes collected after a randomized impact evaluation of a remote mathematics instructional intervention implemented in Kenya while in-person schooling was on hold due to the COVID-19 pandemic (Schueler and Rodriguez-Segura, 2021). This enables us to also estimate how much any enumerator effects could have biased the estimation of treatment effects in this impact evaluation had enumerators been assigned in a more typical manner. Finally, this PBA also included more traditional survey questions directed at parents, which lets us compare enumerator effects on educational assessments to those on survey questions, in a realm closer to the type of measures for which other researchers have previously studied enumerator effects and which are commonly used by social scientists within and beyond the field of education.

Motivation and Contribution

One-on-one surveys and assessments that are administered by teams of assessors are prevalent in education research. Questionnaires that are widely-used in research and policy planning like national censuses or Demographic and Health Surveys (DHS), and internationally-validated exams like the Early Grade Reading Assessment (EGRA) — adapted for over 65 countries (Dubeck and Grove, 2015) — or the nationally representative Annual Status of Education Report (ASER) in India follow this approach. Many important studies and reports have leaned on this type of data for either the framing of their research questions, or as direct outcome measures (e.g., Mbiti et al., 2019; Evans and Mendez Acosta, 2021; World Bank, 2018, Varly, 2020). However, this type of data

collected by teams of individual assessors on a one-on-one basis, is susceptible to enumerator effects, or the non-zero correlations among the responses collected by an individual assessor or interviewer (West and Blom, 2017.)

Enumerator effects could play a significant role in shaping assessment results and indeed the state of the literature as a whole through several potential channels. First, individual assessors could exercise more or less leniency in what is considered a "correct" answer, or they could vary how long children are actually given to answer questions. The level of leniency could be impacted, in part, by the assessor's familiarity with the content or the individual student being assessed. For instance, if an enumerator that is unfamiliar with fractions is asking a question on this topic and the answer sheet says that the correct answer for this question is "1/4", they might mark an answer of "0.25" wrong. Some enumerators may be more forthcoming than others when students ask for clarification about a given question in ways that systematically advantage or disadvantage the students they assess. Additionally, the same respondent may provide different answers depending on their perceptions of a given enumerator. For example, a student may be less likely to try hard on an academic assessment if the assessor is a stranger than a known teacher. Similarly, on survey-based measures, parents may be less likely to divulge personal information if they perceive the interviewer to be indiscrete. Even subtle differences in tone when reading a question could influence the response.

One potential negative implication of enumerator effects is that they could influence the accuracy and precision of estimates comparing between groups, such as a treatment effect estimate in the context of impact evaluation. Mechanically, higher intra-enumerator correlations increase the error introduced into these estimates through higher

variance in the responses (Olson et al. 2020). Therefore, the presence of enumerator effects can result in unwanted inflation of the estimate's variance, ultimately reducing the statistical power and precision of the results. Increased variance could influence the accuracy of the point estimates in a study by allowing for a wider range of probable averages for the treatment and control groups individually due to higher measurement error within each. This would in turn increase the probability of observing a difference (or lack of thereof) between the treatment and comparison groups that was simply due to the measurement error introduced by the enumerator effects, and not as a result of a "true" treatment effect. In fact, beyond the literature on enumerator effects, researchers have documented cases in which measurement error in survey data was introduced systematically differently for treatment and comparison groups (Baird and Özler, 2012; Blattman et al., 2016), potentially leading to biased estimated treatment effects if the researchers relied solely on survey data. Therefore, given the wider variance introduced by enumerator effects, and the previous documentation of how systematic differences in measurement errors across treatment and control groups might lead to spurious results in impact evaluations, it is reasonable to suspect that enumerator effects in survey and assessment data could lead to biased treatment estimates. Yet, the extent to which enumerator effects bias impact estimates is currently a gap in our empirical knowledge.

Another related concern introduced by enumerator effects is that the assignments of enumerators to clusters of respondents (e.g., classrooms, schools, districts) could lead to biased point estimates because of "unlucky bunching" in one group over the other (e.g., in the treatment versus control group in the context of an impact evaluation). For example, imagine an experiment where there are only two schools (one in each

experimental group), two assessors that will each be assigned to one school, and a "true" treatment effect that is null. If there are no systematic differences in how these enumerators record answers, then the estimate of the difference between the treatment and control schools (i.e., the estimated treatment effect) will be accurately null. However, if one of the assessors records systematically higher scores than the other, then the treatment effect will be either positive or negative, even if there is no true treatment effect. In this case, enumerator effects combined with the assignment of assessors at the school level will lead to biased estimates. Yet, even in the presence of enumerator effects, the less clustered the assignment of the assessors is, the less bias that one would expect enumerator effects to introduce into the treatment effect. For instance, in the previous example, if both assessors were randomly assigned to half of the students in both schools, the treatment effects would cancel out, and one would end up with the correct treatment estimate. In practice, due to logistical constraints, assessors are often assigned to clusters in such a way that could lead researchers to generate biased estimates in cases where enumerator effects are operating.

In the case of surveys, the presence of enumerator effects has been long known to social scientists, including in the context of developing countries (West and Blom, 2017), along three different lines of work. First, enumerator effects have been documented as a result of different observable characteristics of enumerators, subjects, or the interaction of both. For example, Adida et al. (2016) find evidence of enumerator effects in large-scale surveys across African 14 countries, especially when the ethnic group of the respondent and interviewer have a history of political competition. Similarly, Di Maio and Fiala (2018) find that in Uganda, although most observable characteristics of assessors yield

minimal enumerator effects, when enumerators are asking highly sensitive political preference questions, differences between enumerators account for over 30 percent of the variation in responses. Benstead (2014) and Blaydes and Gillum (2013) find that the perceived religiosity of the interviewer affects response patterns in surveys administered in Morocco and Egypt respectively, as respondents provided more "socially desirable" answers depending on the appearance of the interviewer. Secondly, enumerator effects have also been documented due to enumerators interpreting the content of questionnaires differently. For instance, Randall et al. (2013) show that the word "household" is difficult to translate into some languages, leading enumerators to venture into their own conceptual interpretation of the questions when presenting them to respondents, and as such, increases the potential for wider variance in the data. Finally, enumerator effects have also been documented appearing at different rates depending on the content of the survey. For example, Himelein (2016) finds interviewer effects in a survey in Timor Leste across subjective and objective questions, but with effects of larger size for subjective questions.

Although enumerator effects have been well studied in the context of survey research, there has been much less formal documentation of enumerator effects in educational assessments. This is true despite the fact that the characteristics of one-on-one educational assessments also make them susceptible to enumerator effects in that, they are often administered by teams of assessors assigned to assess clusters of students, and require discretion from assessors to mark questions, to properly allocate how much time allow for each question, to provide clarification to students on confusing questions, and to know when to stop an assessment. Part of the difficulty of studying enumerator

effects in education, and in development research more broadly, is logistical (Di Maio and Fiala, 2018; West and Blom, 2017). Specifically, to clearly isolate the extent of "enumerator effects" during a round of data collection, researchers would ideally create "fully interpenetrated designs" where the assignment of assessors to respondents is randomized at the individual-level[22] (West and Blom, 2017). In other words, fully interpenetrated designs randomly assign individual subjects to enumerators so that, in expectation, any major differences in response patterns obtained by individual enumerators would be due to the enumerators themselves and not due to differences in the respondent pool assigned to each enumerator.

In spite of the methodological desirability of fully-interpenetrated designs, the physical logistics of randomizing assessors to students at the individual-level can be challenging, especially for in-person assessments. In fact, there are two other recent papers that have studied enumerator effects in surveys through frameworks that approach a fully-interpenetrated designs in developing contexts. However, neither of these studies managed to reach the ideal level of individual-level randomization due to either the infeasible amounts of travel for enumerators this would have entailed (Di Maio and Fiala, 2018) or the logistical difficulty of enforcing individual-level assignments (Laajaj and Macours, 2017). In the case of Di Maio and Fiala's (2018) study, the authors were only able to randomly assign in-person enumerators to small geographical areas in Uganda within which it was still feasible for enumerators to travel. For the Laajaj and Macours (2017) study, the authors randomly assign enumerators to subjects at the individual level,

[22] In the absence of fully-interpenetrated designs, researchers in the past have had to instead use complicated hierarchical statistical models to isolate enumerator effects, ranging from basic random effects models, to extensions like cross-classified random effects. (Olson et al., 2020; Brunton-Smith et al., 2016).

but their compliance rate is only 75%. Because of similar constraints in educational assessment, assessors are rarely randomly assigned to respondents at the individual-level. As such the quantification of enumerator effects has been an elusive subject in the education literature.

Our paper offers two concrete contributions to the literature on enumerator effects, and their potential implications for educational assessments and impact evaluations. First, we document substantial "enumerator effects", or variation in how different assessors graded similar levels of performance, on a phone-based mathematics test, presenting some of the first evidence of enumerator effects on educational assessments. Through a fully-interpenetrated design in the administration of the assessment, we show that enumerator assignments explain 12 percent of the variation in the numeracy scores recorded. When we examine the survey questions, accounting for the enumerator assignment explains an even larger share of the variation in these responses than in the numeracy assessment. For example, enumerator assignment explains 32 percent of the variance in the likelihood of reporting a COVID-19-related income shock. We provide some evidence that younger teachers, teachers with fewer years at our partner's schools, teachers in charge of higher grades, and teachers at the same school as the child that they are assessing, record systematically higher math scores. This is suggestive evidence that enumerator effects, in this case, can be at least partially explained by observable characteristics of the assessors and the match along observables with their students.

Our second contribution is the quantification of the extent to which enumerator effects could yield spurious results in the context of a typical impact evaluation where

enumerators are assigned to classes or schools rather than randomly assigned to individual students due to logistical considerations. Through simulation work, we find that assigning enumerators to whole classes or whole schools would yield point estimates that are statistically different from those found in the original impact evaluation about 10 and 13 percent of the time, respectively. This is in contrast to what we find when we simulate enumerator assignment at the level of individual students, where we obtain statistically different results from those in the original impact evaluation less than 1 percent of the time. Following Evans and Yuan (2020), we estimate that had enumerators been assigned to whole schools for the companion RCT, there would have been more than a 1 in 10 chance of observing a treatment effect that was larger than the mean effect size in math, simply because of enumerator assignment at the school level. As a result, we ultimately recommend randomly assigning individual students to assessors, whenever feasible, to minimize bias. Our study also points to one advantage of phone-based assessments which is the relative ease with which assessors can be randomly assigned and bias therefore reduced compared to assessments administered in-person, particularly in the context of impact evaluation.

Study Design

Context. This project was conducted in partnership with the organization NewGlobe, which operates as a technical partner of government-led education programs and supports its own community schools in several LMIC. One of these networks of community schools is Bridge Kenya, which is the context for our study. Our sample covered students across 105 private schools in 29 of the 47 Kenyan counties, and in all eight of the areas previously considered "provinces." There is a wide range in the

socioeconomic characteristics of the locations covered. The local multidimensional poverty rate (at a 5 km radius from the school) ranges from 7.8% at the 10th percentile in our sample, to 59.2% at the 90th percentile. Although students in these schools and their families are relatively disadvantaged on a global scale, they are likely more socioeconomically advantaged and urban than typical families enrolled in Kenya's public schools. For example, nationally, 27 percent of families report the mother having no formal education while this is true for only one percent of our sample (Twaweza, 2014). Similarly, we estimate that the adult female literacy rate in the communities where pupils in our sample reside is 85%, compared to a national average of 79% (Bosco et al., 2017). The average student in our sample is 11.5 years old, as our sample consists of students in grades 3, 5, and 6. To be part of our sample, students needed to have access to a cellphone. However, this was not a very restrictive condition in this context, as the World Bank reports that there are 1.14 mobile subscriptions per person in Kenya. Finally, we show some additional descriptive statistics of our sample in Table 1.

The current study was part of a larger project conducted in Kenya, which comprised three different studies, all of which rely on overlapping sources of data. The first study consists of a randomized impact evaluation of individualized remote math instruction that teachers delivered by phone while schools were closed due to COVID-19 (Schueler and Rodriguez-Segura, 2021). The second project consists of the validation of phone-based assessments for early literacy and numeracy, and the exploration of the best uses for these assessments based on their psychometric properties. That second study concludes that the phone-based assessments administered as part of this project demonstrated evidence of validity when used to measure aggregate performance, such as

in the context of comparing a treatment and control group, but less evidence of validity for accurately tracking individual student-level performance (Rodriguez-Segura and Schueler, 2022). Finally, the third project consists of this study, which explores the presence and potential implications of enumerator effects in educational assessments. The impact evaluation study was pre-registered as AEARCTR-0006954 while the second study and current study were pre-registered in a separate pre-analysis plan related to educational measurement (AEARCTR-0006913).

Data collection and sampling. The collection of the phone-based assessment (PBA) data happened over 18 days in December of 2020 while face-to-face learning was on hold, after 9 months of school closures. These data were intended as one interim measure of learning outcomes for the impact evaluation of a remote instructional program via mobile phones and were needed because, at the time, it was unclear when students would return to school and take in-person assessments. The sample includes students in grades 3, 5, and 6 across all 105 schools. Due to budget constraints and response rate projections based on an earlier pilot, we selected a simple random sub-sample of students to be assessed from baseline performance blocks of students from all schools (6,295 students out of 8,319 in the impact evaluation sample were selected to be assessed). Of the 6,295 students on call lists, 2,644 were ultimately reached and assessed. Given the pace of enumerators and the rate at which successful assessments were completed, our initial target of 6,295 students proved too ambitions given the time constraints to collect the data. The sampling strategy and logistics are discussed in more detail by Rodriguez-Segura and Schueler (2022).

Assessors and assignment to students. Each of the 6,295 students on the call list was randomly assigned to one of 20 assessors. Randomization was accomplished at the student, rather than the class, school, or region level. Again, this allows us to attribute any differences in average scores across assessors to the assessors themselves rather than to differences in the performance or demographic characteristics of the different groups of students assessed by each enumerator. Among the 2,644 students that were ultimately assessed, the compliance rate to this random assignment was 98.8 percent. All results are robust to analyzing from the perspective of "assigned" assessors (in spirit, akin to an "intent-to-treat" analysis), or from the perspective of the "actual" assessor (similar in spirit to a treatment-on-the-treated analysis). The order in which assessors were asked to call students was also random. Strong protocols were in place to preserve this order, as assessors were centrally and simultaneously trained and had continuous guidance for the first few days of data collection. The only information assessors had at the time of calling was the student's name, grade, and school, but no information about their previous performance or socioeconomic characteristics that they could have used to selectively call students.

All 20 assessors who worked on the data collection of the PBA were teachers within the Bridge Kenya system, for whom our partner had data on the grade and school where they taught and their years of experience. Our partner recruited these assessors to work full-time between December 7th and 23rd for data collection. Although assessors were full time Bridge teachers, typically PBA assessors did not know the students who they called to assess prior to the PBA administration. Only for 10 students (fewer than 0.4 percent) did the randomly assigned assessor end up being the student's own in-person

teacher from the first two months of 2020. Assessors were paid by the day worked, and absenteeism was low. The minimum number of days worked for this wave were 9, the median 14, and the maximum 18. Assessors were aware that the PBA was a low-stakes assessment for the purpose of monitoring learning.

Numeracy assessment and survey questions. The phone-based assessment measured numeracy skills using 14 questions for students. These questions were divided into two sub-sections: the first 9 questions were part of the "core numeracy" sub-section, which asked the same questions to students from all three grades. This section included questions that ranged from counting, to basic operations, and finally a word problem. The second set of 5 questions were part of the "curriculum-aligned" sub-section, and these questions differed by grade. More specifically, these questions were designed so that they would assess concepts students would have been learning in class had in-person school been open.

The grading of each sub-section and of the assessment as a whole was done using a two-parameter item-response model, but we obtain almost identical results if we grade the exam through a simple percentage of the share of correct answers. All scores are standardized at the grade-level such that the mean for each grade is 0, and the standard deviation is 1. Rodriguez-Segura and Schueler (2022) provide a more rigorous psychometric exploration of the properties of this assessment.

After enumerators asked all 14 numeracy questions to students, they were instructed to also ask five survey questions to the parents or guardians of the children. These questions ranged in their degree of sensitivity: the less sensitive questions asked about the child's study habits and practices during school closures and these were asked

first. The more sensitive questions asking about parental education ("*What is the highest level of school that you or someone in your household has completed?*", with eight discreet answer choices ranging from "Some primary school" to "Post-graduate degree", and COVID-19-related shocks to the household's living conditions were left for the end ("*Finally, this has been a hard time for many families due to the coronavirus pandemic. In order to understand how this disruption has influenced children's learning outcomes, it would be helpful to know whether you have experienced any of the following since March when schools were closed*", with three potential answers including moving to a different home, health shocks, and changes to the household income). Conditional on a child starting the assessment, all children completed the assessment and all parents answered all five of the survey questions. The full assessment and survey are provided in Appendix A.

Administrative data and school information. We complement the phone-based assessment data with student-level administrative data provided by our partner. This data includes, for each pupil, individual-level covariates which include the gender, and age of the student, and school that they were enrolled at as of two months before the assessment. Similarly, we have access to three rounds of in-person baseline test scores in math, English, and Kiswahili collected before the phone-based assessment was administered and before school closures. These scores come from standardized tests administered across all 105 schools, in which, for any round of assessments, all students in a given grade took the same test. These data help us get a reliable and comparable measure of baseline achievement for all pupils.

School-level administrative data consists of the school's latitude and longitude, the total student enrollment, the pupil-teacher ratio in the overall student body and in the target grades (i.e., 3, 5, and 6), the female-male ratio in the overall student body and in the target grades, and the average principal, teacher, and student attendance rates. To learn more about the communities where these schools are located, we complement the administrative data with geospatial data containing information on community-level covariates. In particular, we use the GIS poverty rate raster layer from Tatem et al (2013), and the GIS adult (15-49) female literacy rate raster layer from Bosco et al (2017). We use each school's latitude and longitude of each school to create an average poverty rate and average adult female literacy rate for the 5-km circular area surrounding each school.

Methods

Broadly speaking, we are first interested in understanding whether there were systematic differences between the way individual assessors scored PBAs, including both the academic assessment and the survey responses. If so, we also want to understand whether observable characteristics of enumerators or the match along observable characteristics between enumerators and pupils—at least those for which we have data—predict higher or lower scores. Similarly, if we do observe enumerator effects, we want to quantify the extent to which this could become a problematic feature in the estimation of policy-relevant statistics like treatment effects. Below we outline our methodological plan for answering these three main research questions.

How large were the enumerator effects in this assessment? To explore whether enumerators recorded scores for similarly-performing students differently, in the past, the logistical challenges and research concerns of studying enumerator effects have typically

been addressed by assigning in-person assessors to small geographic areas where it is still feasible for assessors to move around (Lupu and Michelitch, 2018). Although this is typically the best choice under realistic logistical constraints for in-person measurement tools, it might still allow for a high rate of "unlucky draws" in the assignment of enumerators to subjects, which when coupled with heterogenous practices across enumerators in data recording, might lead to spurious estimates simply because of the assignment of enumerators to clusters of observations. To avoid this issue, researchers have proposed individual-level assignment of units to assessors, "fully-interpenetrated" designs (West and Blom, 2017), when feasible. In our study, the full randomization of enumerators to students and high compliance rates with these assignments allow us to explore the extent to which enumerator effects may be present in PBAs. In other words, we are able to isolate the effect of enumerators from differences in achievement or other characteristics among the students assessed by different enumerators. Randomization allows us to assume that the achievement levels are the same, on average, across all groups of students assessed by different enumerators.

Our first approach to exploring the presence of enumerator effects is to follow Di Maio and Fiala (2018), Himelein (2016), and Laajaj and Macour (2017) by running a multivariate linear regression for each of the outcomes of interest (i.e., assessment scores) on enumerator fixed effects as the independent variables. We include fixed effects for each of the enumerators (19 in total, as one is the reference group). From this regression, we obtain a set of R^2 statistics which display the extent to which simply accounting for each student's enumerator explains variation in the outcome. Since the assignment of enumerators to students was random, one would expect this R^2 statistics to be close to

zero if enumerators recorded scores in the same manner – in the same manner that enumerator assignment does not explain variation in our covariates (Table 1). Therefore, the higher the R^2 statistics, the stronger the evidence for the presence of enumerator effects.

The fully interpenetrated design also allows us to make predictions about the number of enumerators who are outliers in terms of the scores they recorded – that is, how many assessors recorded scores that were significantly different from everyone else's scores. In particular, given the individual-level assignment of assessors to students, we expect that the average score recorded by all assessors should be statistically equivalent for a large share for the assessors, in the absence of enumerator effects. For instance, we expect two of the 20 numeracy assessors to be statistically different from the rest at a confidence level of 90 percent. To address this, we test the extent to which each of our assessors is different from the rest, separately for numeracy and literacy. Specifically, for student i, and assessor m:

$$\text{Total score}_{ij} = \beta_0 + \beta_1 \text{Assessor}_m + e_{im} \quad [3]$$

Where Assessor_m is an indicator variable which takes the value of 1 if the assessor is assessor m, and 0 for all other assessors. This is repeated for each of the 20 assessors. The sets of β_1, with their respective confidence intervals, are recorded. Following, Von Hippel et al. (2016) and Von Hippel and Bellows (2018), we also use a Bonferroni correction for these confidence intervals to account for multiple hypothesis testing and to generate a null distribution against which we can compare the Bonferroni-corrected

confidence intervals[23] As a robustness check, we repeat this exercise also controlling for the baseline score of each pupil, their grade, and their school. In other words, even if after randomization of assessors, an assessor obtained an imbalanced draw along these characteristics, this robustness check would account for this.

Did observable enumerator characteristics predict higher scores? If we find evidence for the presence of enumerator effects, we also want to understand whether there are any observable characteristics of enumerators correlated with differential scores. This is particularly relevant as it could shed light on whether program managers could know a priori who among their enumerators might yield systematic differences, and hence select or train them accordingly. To tackle this question, we regress learning outcomes on different assessor characteristics, one covariate at a time, to test whether any of these characteristics predicts higher scores. We also leverage the fact that the allocation of students and enumerators was random, so that the match on their observable characteristics was also random. Hence, we can also explore the extent to which the match of student and teachers based on baseline student characteristics drives differential results. Specifically, for student i, assessor m, and observable X:

$$\text{Outcome}_{im} = \beta_0 + \beta_1(\text{Assessor and student match on X})_{im} + e_{im} \qquad [4]$$

Where the variable "Assessor and student match on X" takes the value of 1 if the assessor and student share the same characteristic, and 0 otherwise. These characteristics

[23] We follow this analytic approach while realizing that using Bonferroni-corrected confidence intervals and these null distributions are the most conservative approach to studying enumerator effects in this manner. In other words, if we detect enumerator effects through this approach, we would certainly detect them with looser approaches like tallying the share of instances in which β_1 displayed statistical significance, and checking if this share is lower than what one would expect by sheer chance, or using a null distribution that is 0 for all assessors.

include whether the assessor and the student are based in the same school, whether they are assigned to the same grade, or whether the teacher is assigned to a similar age group as the child's grade (i.e., lower or upper primary). Taken together, the results from this section can inform whether enumerator effects, if at all, can be reduced by targeting specific sub-groups of enumerators of instances of assessor-student matches.

Could enumerator effects bias point estimates from an impact evaluation?

We also seek to understand the extent to which scoring differences across enumerators could bias the estimation of metrics like treatment effects in the context of impact evaluation. In particular, we ask whether different realizations in the allocation of enumerators to individual or organizational units would yield significantly different results solely due to differences in how enumerators record scores. In particular, we explore whether these results would vary were the enumerators allocated at the class- or school-level, as this has historically been the more common approach to assigning assessors for in-person assessments. To do so, we leverage the main intent-to-treat (ITT) estimates of the effect of phone-based tutoring on PBA numeracy scores from the field experiment that prompted the collection of the PBA data analyzed here (Schueler and Rodriguez-Segura, 2021). For the field experiment, treatment was assigned at the school-level, and there were two different treatment arms (T1 and T2). Schueler and Rodriguez-Segura (2021) find average ITT effects on the numeracy PBA scores of non-statistically significant $\beta_1 = 0.04$ for T1 and $\beta_1 = -0.03$ for T2, both in standard deviation units.

To explore the sensitivity of these point estimates to enumerator effects, we first predict the counterfactual PBA scores that students would have received under different enumerator assignments. To do so, we start by running the following model for student i,

in grade j, at school k, assessed by enumerator m – who was the enumerator that actually assessed child i in the context of this study:

$$Y_{ijkm} = \beta_0 + \beta_1(\text{Baseline score})_{jjk} + \lambda_j + \mu_k + \eta_m + e_{ijkm} \quad [5]$$

Y_{ijkm} is the observed math score outcome, "Baseline score" represents the baseline in-person score for each student, and λ_j, μ_k, and η_m represent grade-, school-, and enumerator-fixed effects. This "calibrated model" yields a set of coefficients, all of which were estimates based on observed data, for each of these predictors. If one were to plug in a given student's covariates into this model and add the error term, one would obtain their actual PBA score.

After running this model with the, we proceed to randomly re-assign enumerators at different levels of aggregation to simulate alternate assignment scenarios as if we were starting the project from scratch. In particular, we randomly re-assign enumerators at the level of the students, then at the level of classes, and finally at the level of schools. These last two steps mirror alternative study designs where enumerators are not assigned in fully-interpenetrated designs, but rather are clustered within some natural organizational unit. The simulation at the student-level allows us to benchmark the outcomes obtained for these other two levels of enumerator assignment to students.

Using the calibrated model, meaning the coefficients obtained from the first model with observed data, we plug each student's actual baseline score, grade, and school, along with their newly simulated random assignment to an enumerator n, to create a predicted test score $\hat{Y}_{ijkn}$ for each student:

$$\hat{Y}_{ijkn} = \beta_0 + \beta_1(\text{Baseline score})_{ijk} + \lambda_j + \mu_k + \eta_n + e_{ijkn} \quad [6]$$

$\hat{Y}_{ijkn}$ represents the predicted score one would expect student i to have obtained, had they been assessed by enumerator n (using the fixed-effects obtained for each enumerator in the calibrated model) rather than their actual enumerator m. For this exercise, each student receives three different $\hat{Y}_{ijkn}$: one under a new assignment of assessors at the student-level, another at the class-level, and another at the school-level. We then generate the predicted ITT estimates of the treatment effect of remote tutoring under the different enumerator assignments using $\hat{Y}_{is}$ as the outcome. This model yields $\hat{B}_1$ and $\hat{B}_2$, which are the simulated treatment effects had this specific realization of enumerator assignment been the actual assignment. We can then test whether the $\hat{B}_1$ and $\hat{B}_2$ estimated with simulated enumerator assignments at the student-, class-, and school-level are statistically different from β_1 and β_2, that is – from the "true" ITT estimates observed in the field experimental study.

Finally, we repeat this simulation exercise 10,000 times for each of three levels of aggregation at which we assign assessors. This repeated exercise yields distributions of simulated treatment effects under the three levels of assignment of enumerators. Using each of these distributions, we tally the number of times that $\hat{B}_X$ was statistically different from β_x (in other words, the number of times the simulated treatment effect was different than the actual treatment effect). This number can be interpreted as the share of the time that we would have obtained spurious results when estimating a treatment effect simply as a result of the specific assignment of enumerators at different levels of aggregation.

Results

Enumerator effects were present and non-trivial in size for both academic and survey measures. We find that the scores and responses that individual enumerators

recorded did depend on enumerator assignment, and that size of these enumerator effects was non-trivial. Specifically, accounting solely for the assessor conducting the PBAs through a set of assessor fixed-effects explains 12 percent of the variance in the academic assessment score, (8 percent for core numeracy and 13 percent for curriculum-aligned scores). The enumerator effects were even larger when it came to the survey measures, where 32 percent of the variance for the likelihood of reporting of a COVID-19 related income shock, and 23 percent of the variance for self-reported parental education, were explained by differences in enumerator alone. To put these numbers into perspective, we can compare them against the findings of Di Maio and Fiala (2018) and Laajaj and Macours (2017), the two other papers in a developing context that also used some degree of randomization to quantify enumerator effects in surveys. The largest result presented by Laajaj and Macours is that enumerator assignment explains 9 percent of their variance in non-cognitive scores. Similarly, most of the results presented by Di Maio and Fiala (2018) report that enumerator assignment explains less than 5 percent of the variation for most questions, while their main finding is that for political questions, enumerator assignment explains around 30 percent of the variation in responses. Therefore, the enumerator effects that we find for the math scores on the PBA could be considered "medium-sized", while the magnitude of enumerator effects for the survey questions on the PBA are as large as Di Maio and Fiala (2018)'s largest finding – namely, politically sensitive questions in Uganda.

Another approach that we take to document the extent of enumerator effects in our data is to tally the number of assessors who report scores that are significantly different from a null distribution if there were no enumerator effects. We visually display

these results for numeracy, literacy, and fluency in Figure 1. Through this approach, we find that at the 95 percent level of significance, five (25 percent) of the numeracy assessors recorded average scores that were significantly different from what one would expect to happen by chance if enumerators had no effect on the responses. This finding remains identical when we control for baseline performance, and include grade- and school-level fixed effects in the model above to account for potential unintentional differences in the random allocation of enumerators to students. Furthermore, as a placebo test, we also test the extent this is caused by differences in baseline performance, by running the same model as above but with the baseline score in each subject as the outcome. In this case, no assessor is statistically different from the null distribution. We observe even larger effects for the survey questions. For instance, 11 (65 percent), and 16 (80 percent) of the assessors are different from the rest when reporting a COVID-19 income shock, and for self-reported parental education respectively. In sum, we find very strong evidence for the presence of "enumerator effects", or the systematically heterogenous recording of similar answers by different enumerators for phone-based assessments of foundational numeracy skills and survey-based measures of parent-reported student learning time, parental education, and economic disruptions.

Observable enumerator characteristics did predict some differences in academic and survey responses. Since we do find systematic differences in how enumerators record scores, we explore which assessor characteristics drive these effects, as this could inform enumerator selection or the extent to which enumerators could be provided with targeted reinforcement before the data collection to reduce bias. It could also provide suggestive evidence regarding the mechanisms through which enumerator

effects operate. We show results in Table 2. In general, assessors who taught higher grades during the school year recorded slightly higher scores, on average. Interestingly, the total number of students or the total number of students assessed on the same day do not predict differential assessment, which suggests that differential assessment is not driven by differential assessor pace.

A striking feature of Table 2 is that enumerator experience, measured in terms of days worked on the PBA data collection by the time a student was assessed, is correlated with higher recorded scores. Keep in mind, the order of calls was randomized. Although preserving the randomized order was not feasible 100 percent of the time, the actual order in which students were reached explains as little variance as the order assigned (which was randomized). To put the magnitude of this effect in perspective, the median number of days an assessor worked was 14. Between the first day, and the time enumerators were half-way done (day 7), the mean score had increased by 0.35 standard deviations. In fact, the mean over-estimation of in-class percentiles by PBA percentiles increased about 1.5 percentile every day. When the distribution by day worked is plotted, as in Figure 2, it is clear that there is a positive trend for the median, but the sharp increase in the mean is driven by a decrease in the lower tails as low scores become less frequent by day worked. In fact, the standard deviation of the numeracy score shrunk about 0.17 or 15 percent by day 7. Following the empirical strategy from the section on enumerator effects, we run a regression of total numeracy scores on day assessed-fixed-effects, which yields an R^2 of 0.05. In other words, the day of work for each assessor in which students were called explains 5 percent of the variance in scores, which is particularly interesting given that both of these assessment features were randomized at the individual level. In all, we find

evidence that assessors did grade students differently over time, becoming more lenient toward the lowest-performing students the longer they were on the job.

Since the match along observables between assessors and students was also randomly determined, we study whether these matches predict differential scoring practices and display results in Table 3. In general, the results are not always consistent across specifications which may not suggest non-random sorting but rather may simply reflect the relatively small samples for which there is a match on these dimensions. However, there are two results that appear robust. First, teachers assessing students in the same grade as the grade they teach tend to record higher scores. It is unclear whether this is because the teachers are more lenient, or the students perform better in these cases. Second, teachers who teach at the same school during the regular in-person school year as attended by the student being assessed, recorded lower levels of COVID economic disruptions, on average. Again, it is unclear whether this is because parents were less likely to disclose disruptions to teachers from their schools or enumerators were less likely to believe or record disruptions when reported by families attending the schools with which they were most familiar.

Enumerator effects can bias point estimates from an impact evaluation. Given the strong presence of enumerator effects, we also want to understand how effects of this magnitude could affect the estimation of treatment effects for a given impact evaluation, and the extent to which individual-level assignment of enumerators to students may reduce this bias on the estimands. To do so, we run the simulation described in the previous section, where we recreate the treatment effects that we see in a companion impact evaluation paper (Schueler and Rodriguez-Segura, 2021) under

different simulated levels of enumerator assignments. We present the distributions of the simulated treatment effects at all three levels of disaggregation in Figure 3, along with a summary of these results in Table 4. Although the distributions of treatment effects at all three levels of disaggregation are centered around similar values in all three cases, the dispersion grows with the level of enumerator assignment. In other words, the probability of observing a large—either positive or negative—treatment effect solely as a result of the level at which the enumerator assignment occurred, increases significantly the larger the unit of aggregation assigned to each assessor. For instance, the 90th percentile of treatment effects for phone-based tutoring was 0.02 SD at the student-level, 0.07 SD at the class-level, and 0.12 SD at the school level.

To benchmark these scores, we use Evans and Yuan (2020), who claim that the average RCT in international education has a mean effect size in math of 0.09 SD, and that the effect size for math at the 60th percentile is 0.10 SD. In other words, had we assigned enumerators at the school level for the companion impact evaluation study, there was over a 1 in 10 chance of observing a treatment effect that was larger than the mean effect size for RCTs in similar contexts, simply because of the realized assignment of enumerators. In fact, we quantify in Table 4 the number of times that we observe significantly different results from the "true" ITT estimates. While student-level assignment of assessors yields similar results over 99 percent of the time, class- and school-level assignment of assessors yields different results between 10 and 13 percent of the time – at a rate 13 times higher than that at the student-level. In sum, the combination of enumerator effects and enumerator assignment at a clustered level might lead to

potentially spurious evaluations of policies because of the random clustering of outcomes given the realized allocation of enumerators.

Discussion

In this paper, we present some of the first evidence from a fully interpenetrated study design documenting the presence of enumerator effects in educational assessments of academic achievement. We find significant differences, larger than what one would expect to arise by chance, in how numeracy scores were recorded across enumerators for similarly-achieving students. Consistent with previous research, we also document enumerator effects for parent survey measures that are even larger in magnitude than those observed for the academic tests. We find that the combination of enumerator effects of this magnitude, and a more aggregate level of assignment of assessors to students beyond the level of individuals, could yield spurious estimates of the estimands of interest, such as treatment effects in the context of an impact evaluation, at a worryingly high rate. In other words, we find evidence to support the claim that specific realizations of the assignment of enumerators to units can yield undesirable and heterogenous clustering of enumerators across treatment and control groups, which can in turn lead to differences in outcomes which are solely due to differences in how assessors recorded scores within their assigned clusters. While we explore the extent to which enumerator effects could bias treatment effect estimates, this phenomenon also has similar implications for a range of between-group comparisons beyond impact estimates.

In theory, the ideal solution to this issue would be the individual-level random assignment of enumerator to subjects, which in the case of phone-based assessments does not tend to pose significant logistical constraints. We show that this practice would, in

most cases, diminish the extent to which the results obtained from a survey or assessment are biased due to enumerator effects by ensuring that any bias introduced by particular assessors is randomly distributed across the population of respondents and not systematically clustered within any group of respondents. Therefore, our study reveals this major potential benefit of phone-based assessments, which have gained some popularity in recent years in large part due to pandemic-induced disruptions to in-person schooling and assessment. In another paper (Rodriguez-Segura and Schueler, 2022), we describe some of the potential and disadvantages of PBAs, like the significantly weaker correlations with baseline assessments compared to repeated in-person assessments. However, we also describe some of the advantages of PBAs, like being able to assess hard-to-reach populations at a fraction of the costs of in-person assessments, and the capability to use fully-interpenetrated designs. However, we also acknowledge that the exact model of enumerator randomization at the individual-level that we use in this paper is not always feasible when it comes to the collection of in-person field surveys and assessments. We propose several ideas to minimize bias from enumerator effects when randomization of enumerators to students is not possible.

The first approach we propose is to make logistical efforts to assign enumerators to the smallest clusters that are feasible for them to assess. This may be insufficient to eliminate all bias, as in the case of the remote tutoring field experiment, we see that the biggest difference in the rate of spurious effects due to enumerator effects is the shift from student- to class-level assignment (the next most disaggregate unit after students), and not from classes to schools. That said, this may reduce bias in some cases. Another potential compromise would be to still randomly assign enumerators to students at the

individual level, and group school visits within the same time window for an enumerator, if the travel distances between the natural clusters (e.g., schools) are not too prohibitive and the number of clusters is not too large. Then, the researcher could account for the day that each student was assessed in their models (via methodological approaches like day fixed-effects), while still being able to claim that there was some degree of full interpenetration in the study design. However, unlike phone-based assessments, this approach would not allow for the randomization in the overall order in which all students within a single enumerator were assessed.

The second approach we propose is to assign each assessor to a roughly equal number of clusters being compared to each other. For example, in the case of a field experimental study, assign each assessor to an equal number of treatment and control clusters. Alternatively, if the goal is to compare private to public schools, evenly distribute these two types of schools across assessors. This approach would be a first step towards offsetting the differential scoring of assessors in the estimation of treatment effects or differences between other groups of interest. Similarly, this approach would allow researchers to include assessor-level fixed effects in their specifications, as it would provide common support across both sides of the treatment effect estimation for each assessor. The inclusion of assessor fixed effects would then de-mean the outcome from each assessor's idiosyncratic scoring bias. However, this approach is only feasible when there are fewer assessors than clusters. There is also the opposite scenario, where the clusters are so large that they require more than one assessor. In this case, the random allocation of assessors to clusters seems particularly important to "dissolve" enumerator differences within clusters as much as possible.

Third, we find that, when it comes to academic assessments, whether or not the assessor teaches the grade of the student being assessed matters for the responses recorded. Therefore, assessment administrators would be wise to distribute these matches even across groups being compared. Relatedly, we observe that teachers from the same school as the family being surveyed record fewer economic shocks. This suggests that survey administrators should therefore carefully consider whether assessors are from the same community as the respondents when being tasked with asking sensitive questions, and again, attempt to spread these matches out across any groups being directly compared on the survey outcomes.

Fourth, there are potentially statistical adjustments that can be performed *after* data collection, and which could reduce some of the error introduced by the enumerator effects. For example, one could imagine a typical set up where enumerators are rigorously trained, and then randomly assigned to large clusters like schools to collect reading fluency data from grade 3 students. At the end of the data collection process, they would all be shown the same pre-normed footage of several children with different performance levels taking the same fluency assessment that the enumerators just administered. The enumerators are individually asked to score each of the students in those videos, without knowing what the actual normed score of each video is. Then, each enumerator's scores are compared to the pre-normed scores for the videos to understand the magnitude and direction of the bias for each assessor (e.g., "enumerator A's tendency was to record scores that were on average 0.5 SD above a pre-normed score, while enumerator B's tendency was on average only 0.1 SD below the pre-normed scores). Finally, this information could be used for post-data collection adjustments to the

observed scores. Questions regarding ex-post statistical adjustments based on additional data collection, to the best of our knowledge, have remained largely unexplored in the literature, and as such, this is an area that is ripe for future research.

In terms of logistical considerations, we also recommend robust training of assessors. Training in general might have several goals, like reducing the extent to which enumerators make data entry errors, or maximizing the probability of subjects agreeing to participate in the interview. However, if the skills that are being trained are heterogeneously distributed across enumerators at the beginning of the training, pushing all these goals forward can also contribute to the reduction of enumerator effects. In this sense, there are valuable publicly available resources to incorporate best-practices into enumerator training (World Bank DIME, 2022). Having said that, researchers and practitioners may also find that there is only a certain amount of training that can sustainably take place given the financial and time constraints of a project, and as such, it is critical to incorporate into the training the most effective practices to successfully train enumerators.

The training of enumerators does not need to end at the beginning of the data collection process, and in fact, we believe that constant "norming" of assessors throughout the life cycle of the process can be beneficial to reduce enumerator effects. In particular, in this study we also find that assessors' behavior changes over time – which in this case manifested as fewer low numeracy scores recorded in the later stages of the data collection process suggesting either increased leniency toward lower achieving students by enumerators over time or a greater ability to detect achievement among students typically perceived as low performing with greater experience. Regardless of the

mechanism, to avoid this phenomenon, other literatures have considered the practice of "norming" (Cohen and Goldhaber, 2016), that is, the conduction of frequent exercises that align enumerators amongst themselves, and with an ideal scoring behavior over time. Additionally, future research should explore the extent to which norming can minimize enumerator effects.

We also see implications for researchers drawing on previously published studies or datasets that rely on one-on-one assessments for which assessors were assigned at high levels of aggregation, such as at the school, district, county, or country level. Firstly, such estimates should be interpreted with caution and may be over- or under-estimated due to the presence of enumerator effects. For areas of study in which this is the only method that has previously been used to gather outcome data, researchers should prioritize new studies that are not susceptible to this form of bias. Additionally, on topics for which there are puzzling discrepant results across studies, enumerator effects could be one explanation for these differences if there is variation in the level of enumerator assignment across studies. Secondly, for publicly available data sets, we recommend that the maintainers of these data also include certain de-identified information about the enumerators as part of the data set. For each observation, researchers should know, to the extent possible, who the enumerator who collected this data was. Then, this information should then be able to be linked to additional information about the enumerator that may potentially explain some of the enumerator effects that could manifest in this data set, if at all. For instance, for a data sets on early-grade reading fluency, understanding each enumerator's educational background and experience collecting early-grade literacy data would be a valuable addition to the data set.

Our hope is that this study is a starting point for future research on the implications of enumerator effects in other types of assessments, subjects, modes of assessments, and contexts. For instance, one question that remains from our study is the extent to which we observed enumerator effects in this case because the assessment happened over the phone. Paradoxically, it was the fact that the assessment was conducted over the phone what allowed us to isolate the enumerator effects in a clean empirical way. Future research should test whether these findings replicate with in person assessments and assessments of additional content areas such as literacy. Another important line of research is the extent to which different types of training and norming of assessors could reduce the extent to which enumerator effects can manifest in educational assessments. In other words, we do not know whether the enumerator effects that we observe here happened because of the training that they received, or in spite of it.

In all, the study of quantification of enumerator effects in educational assessments, particularly in LMIC, remains a nascent area of study. However, we present evidence making the case that researchers should pay closer attention to how these systematic differences between enumerators might be affecting their results, and ultimately, the conclusions that can be drawn from their studies.

TABLE 3.1
Sample Description by Baseline Covariates

		Mean / (SD)	Share of variance explained by assessor assignment
Student characteristics	Student is female	0.50	0.3%
		(0.50)	
	Student's age	11.5	1.2%
		(1.7)	
	Latest standardized score math pre-school closures, by grade	0.02	1.0%
		(0.98)	
	Latest standardized score Kiswahili pre-school closures, by grade	0.02	0.9%
		(1.00)	
	Latest standardized score English pre-school closures, by grade	-0.00	0.7%
		(0.98)	
School characteristics	Total school enrollment	273.1	0.7%
		(73.0)	
	Attendance rate of school principal	0.89	0.4%
		(0.26)	
	Attendance rate of teachers	0.84	0.9%
		(0.15)	
	Attendance rate of pupils	0.51	0.7%
		(0.15)	
Community characteristics	Population within a 5km radius	261753.8	0.9%
		(376719.2)	
	Average female literacy within a 5km radius	0.85	0.9%
		(0.14)	
	Average poverty rate within a 5km radius	0.33	0.9%
		(0.17)	
	Distance to nearest cell tower (km)	0.42	0.6%
		(0.80)	
	Observations	2552	

Notes: the column displaying the share of the variance of each covariate that is explained by assessor assignment corresponds to the R^2 resulting from regressing each covariate on a set of fixed effects for each assessor. This follows the first methodological approach outlined in the "Methods" section.

TABLE 3.2
Assessor Predictors of Outcomes Collected Through Phone-based Assessment

		Math score		Time reported studying		Reports using books		Parental education		Reports COVID-related income shock	
	Mean	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Years at Bridge as a teacher	6.25	-0.02***	-0.02**	0.07***	0.07***	-0.01**	-0.01**	-0.03***	-0.04***	0.03***	0.02***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Assessor grade taught	1.41	0.02***	0.03***	0.12***	0.12***	-0.05***	-0.05***	0.07***	0.07***	0.00	0.00
		(0.01)	(0.01)	(0.02)	(0.02)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Total students assessed	139.36	0.00***	0.00***	0.01***	0.01***	0.00***	0.00***	0.00	0.00	0***	0***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Number of students assessed by assessor on the same day	12.85	0.00	0.00	0.04***	0.04***	0.00	0.00	0.00	0.00	0.00	0.00
		(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Total days worked	14.04	0.00	0.00	-0.1***	-0.09***	0.01*	0.01*	-0.01	-0.01	0.04***	0.04***
		(0.01)	(0.01)	(0.02)	(0.02)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Enumerator experience (days)	7.26	0.05***	0.05***	-0.04***	-0.04***	0.01***	0.01***	0.01***	0.01***	0.02***	0.02***
		(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
School FEs, Grade FEs, baseline performance		N	Y	N	Y	N	Y	N	Y	N	Y
Outcome mean		-0.010		2.650		0.610		2.090		0.750	
Outcome SD		0.990		2.020		0.490		0.710		0.430	
Observations		2552	2552	2552	2552	2552	2552	2509	2509	2552	2552

Notes. Each coefficient comes from running a regression of the outcome on each assessor characteristic (predictors). * p<0.10, ** p<0.05. *** p<0.01.

TABLE 3.3

Effect of Matching Characteristics Between Assessor and Student Characteristics

		Math score		Time reported studying		Reports using books		Parental education		Reports COVID-related income shock	
	Mean	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Same school	0.11	0.14**	0.06	-0.48***	0.71**	0.18***	-0.02	-0.04	-0.11	-0.28***	-0.22***
		(0.06)	(0.11)	(0.1)	(0.28)	(0.03)	(0.05)	(0.04)	(0.07)	(0.03)	(0.06)
Same grade	0.04	0.24**	0.19*	0.14	-0.15	-0.25***	0.00	0.43***	-0.03	-0.16***	-0.01
		(0.09)	(0.11)	(0.21)	(0.21)	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)	(0.05)
Same age group of students, and class taught by assessor (lower primary vs. upper primary)	0.26	-0.02	-0.01	-0.31***	-0.29***	0.09***	-0.02	-0.07**	0.02	-0.04*	-0.04**
		(0.04)	(0.05)	(0.07)	(0.07)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
School FEs, Grade FEs, baseline performance		N	Y	N	Y	N	Y	N	Y	N	Y
Outcome mean		-0.010		2.650		0.610		2.090		0.750	
Outcome SD		0.990		2.020		0.490		0.710		0.430	
Observations		2552	2540	2552	2540	2552	2540	2509	2497	2552	2540

Notes. * p<0.10, ** p<0.05. *** p<0.01.

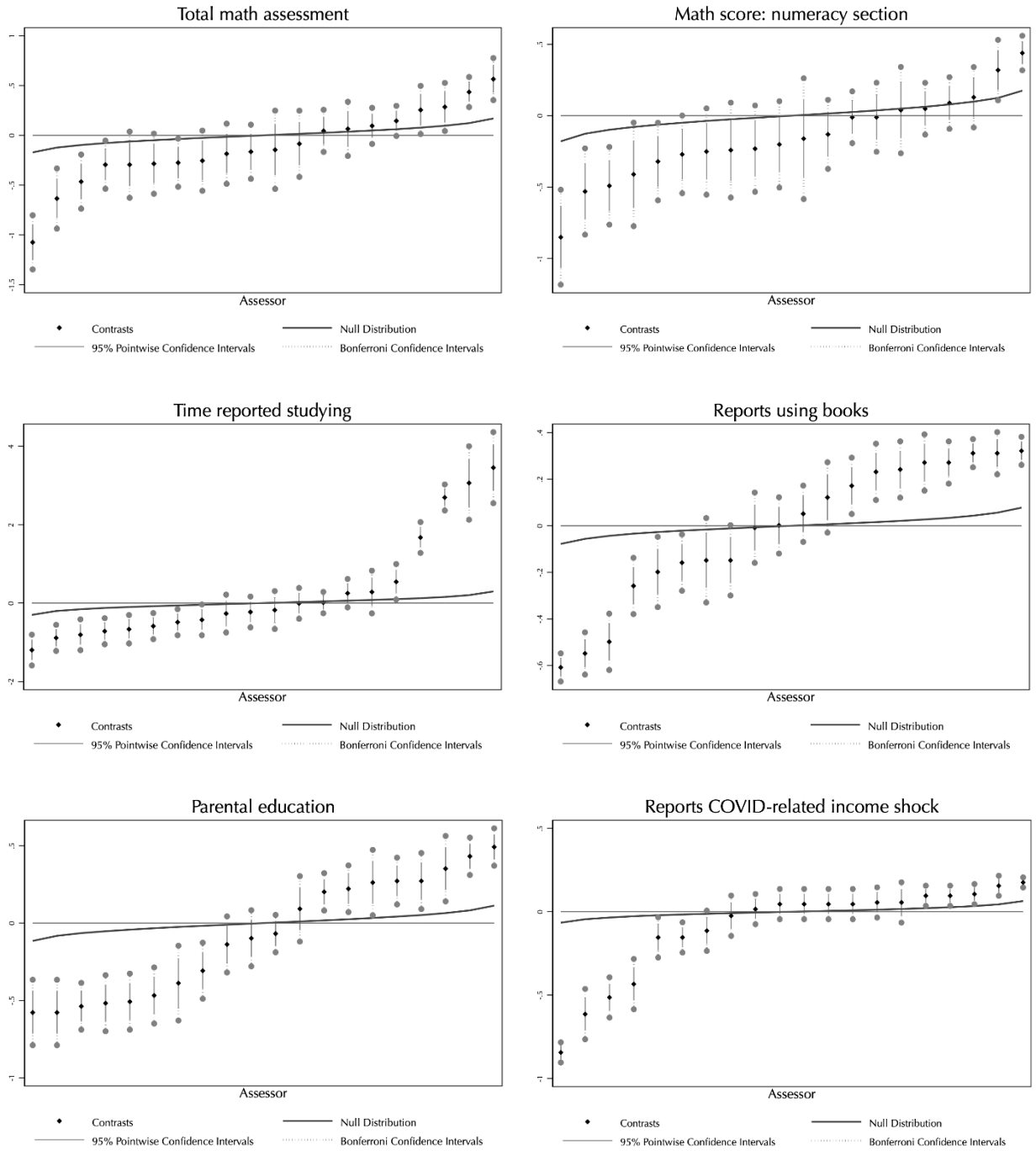
TABLE 3.4

Percentage of Simulation Exercises that Yielded Treatment Effects Different from the Observed Treatment Effects

Level of enumerator assignment	Treatment 1 (T1)	Treatment 2 (T2)	Average T1 and T2	Times higher than student-level
Student	1.4%	0.3%	0.9%	-
Class	11.0%	8.1%	9.6%	11.2x
School	13.0%	12.0%	12.5%	14.7x

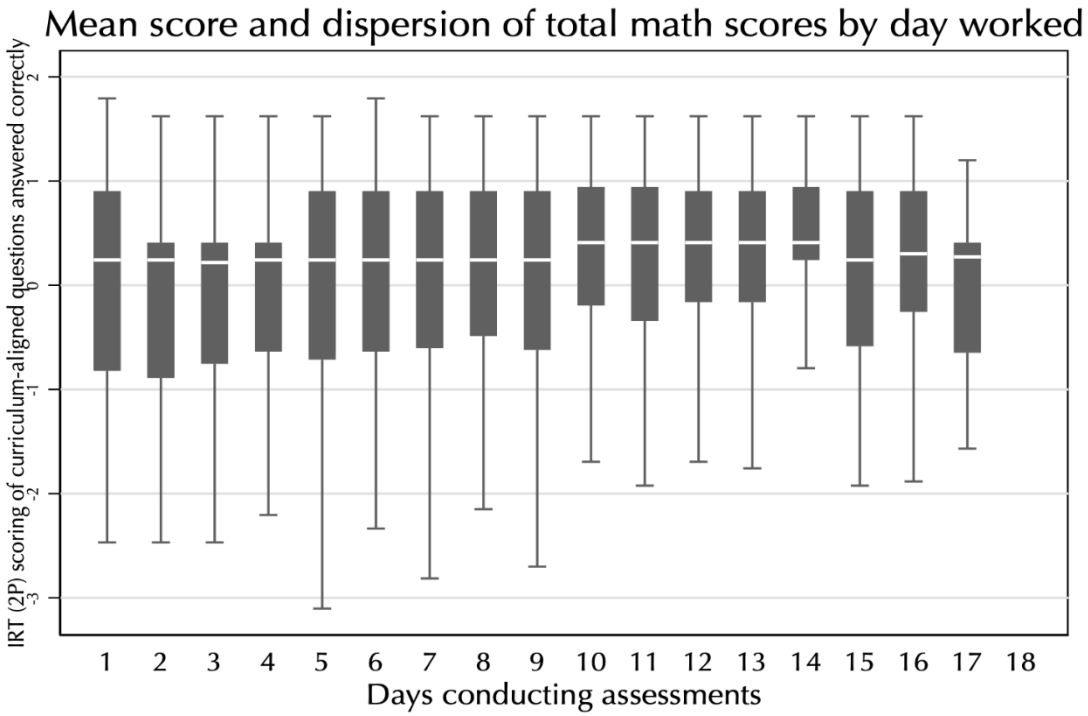
Notes. Numbers obtained from simulating outcomes 10,000 times for each level of enumerator assignment

FIGURE 3.1. *Differences Between Each Assessor and All Other Assessors, by Outcome*



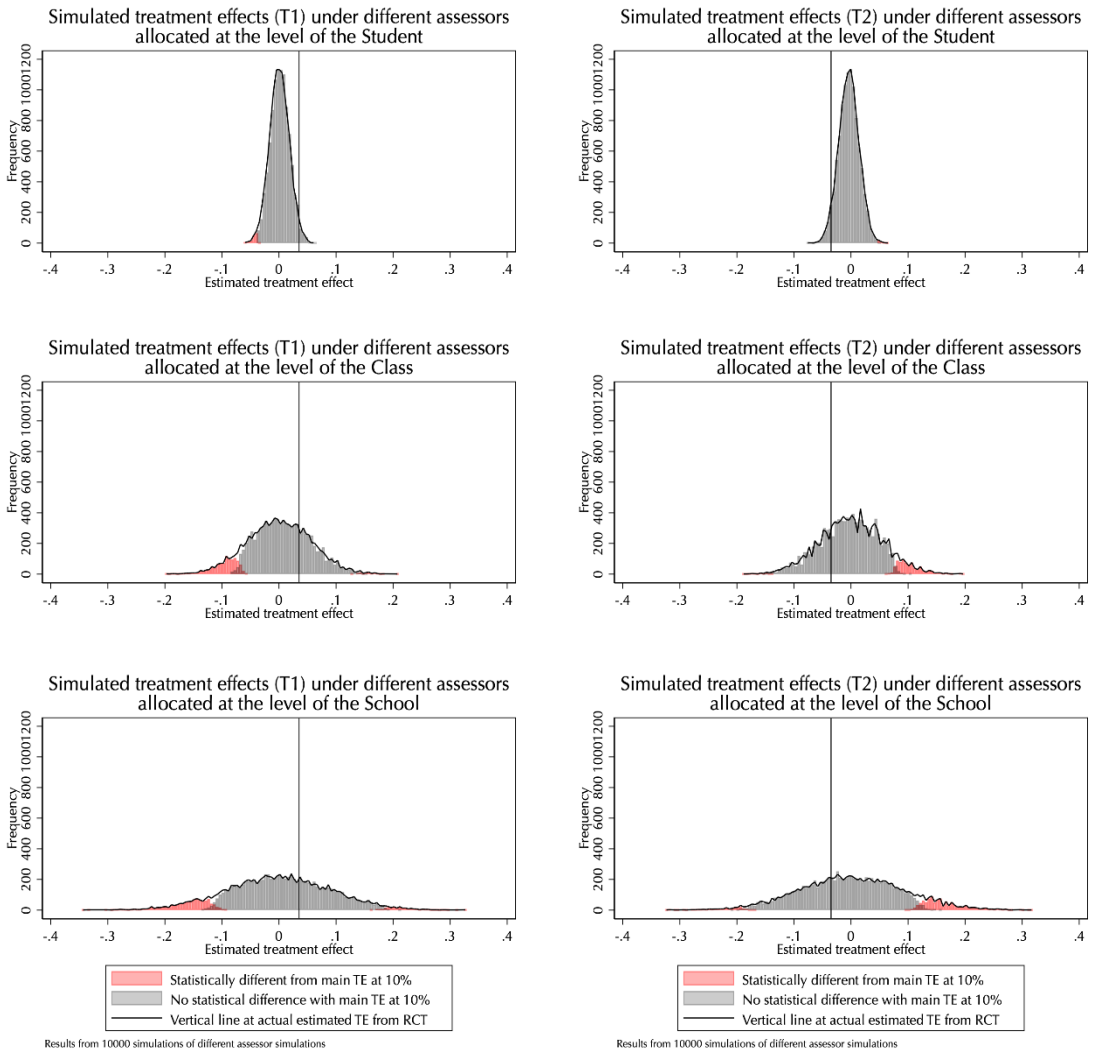
Notes. The difference between each enumerator and the rest is computed by regressing the outcome on an indicator variable, separately for each enumerator. All specifications include a grade- and school-level fixed effects. Coefficients sorted from left to right by plot, meaning that enumerator number does not necessarily match across panels. Standard error bars shown at the 95% level of significance. Standard errors are clustered at the school-level.

FIGURE 3.2. *Distribution of Total Numeracy Scores by the Day Worked for each Assessor*



Notes. Box plot showing the 25th, 50th, and 75th percentile of the total score attained on the phone-based assessment by the day that each child was assessed, according to the days that their assessor worked. Assessment graded using the first component of a principal component analysis.

FIGURE 3.3. *Distribution of Simulated Treatment Effects under Different Allocations of Enumerators, by Treatment Arm and Level of Enumerator Assignment*



REFERENCES

Abuya, B.A., Mutisya, M., Ngware, M., 2015. Association between mothers' education and grade six children numeracy and literacy in Kenya. Education 3–13 43, 653–665. https://doi.org/10.1080/03004279.2013.855250

Adida, C.L., Feree, K.E., Posner, D.N., Robinson, A.L. (2016). Who's asking? Interviewer coethnicity effects in African survey data. Comparative Political Studies. 49: 1630–60

Ahsan, Md N., Banerjee, R., & Hari, S. (2018). Social Promotion and Learning Outcomes: Evidence from India. USC Dornsife Institute for New Economic Thinking. Working Paper no. 18-14.

Akyeampong, K., Delprato, M., Sabates, R., James, Z., Pryor, J., Westbrook, J., Humphreys, S., & Tsegay. A. H. (2018). Tracking the progress of Speed School students 2011-2017. Research Report. University of Sussex: Centre for International Education.

Andrabi, T., Das, J., Khwaja, A.I., 2012. What Did You Do All Day? Maternal Education and Child Outcomes. J. Human Resources 47, 873–912. https://doi.org/10.3368/jhr.47.4.873

André, P. (2009). Is grade repetition one of the causes of early school dropout? Evidence from Senegalese primary schools. MPRA Paper 25665.

Angrist, N., Bergman, P., Evans, D. K., Hares, S., Jukes, M. C. H., & Letsomo, T. (2020). Practical lessons for phone-based assessments of learning. BMJ Global Health, 5(7), e003030. https://doi.org/10.1136/bmjgh-2020-003030

ASER. (2019). Annual Status of Education Report (Rural). http://img.asercentre.org/docs/ASER%202018/Release%20Material/aserreport2018.pdf

ASER. (2021). Annual Status of Education Report (Rural) – Kartanaka (Rural) http://img.asercentre.org/docs/ASER%202018/Release%20Material/aserreport2018.pdf

Asim, M. (2020). Average vs. distributional effects: Evidence from an experiment in Rwanda. International Journal of Educational Development, 79, 102274. https://doi.org/10.1016/j.ijedudev.2020.102274

Atuhurra, J. and Kaffenberger, M. (2020). System (In)Coherence: Quantifying the Alignment of Primary Education Curriculum Standards, Examinations, and Instruction in Two East African Countries. RISE Working Paper Series. 20/057. https://doi.org/10.35489/BSG-RISE-WP_2020/057

Atuhurra, J., Alinda, V. (2018). Basic education curriculum effectiveness in East Africa: A descriptive analysis of primary mathematics in Uganda using the “Surveys of Enacted Curriculum". MPRA Paper No. 87583, University Library of Munich, Germany.

Azevedo, J. P., Goldemberg, D., Montoya, S., Nayar, R., Rogers, H., Saavedra, J., Stacy, B. (2021a). Will every children be able to read by 2030? Defining learning poverty and mapping the dimensions of the challenge. The role of education quality for economic growth. World Bank Policy Research Working Paper No. 9588.

Azevedo, J. P., Goldemberg, D., Montoya, S., Nayar, R., Rogers, H., Saavedra, J., Stacy, B. (2021b). Will every children be able to read by 2030? https://blogs.worldbank.org/developmenttalk/will-every-child-be-able-read-2030

Baird, S., & Özler, B. (2012). Examining the reliability of self-reported data on school participation. Journal of Development Economics, 98(1), 89–93. https://doi.org/10.1016/j.jdeveco.2011.05.006

Banerjee, A. V., & Duflo, E. (2012). *Poor economics: A radical rethinking of the way to fight global poverty*. (Paperback first published). PublicAffairs.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. Journal of Economic Perspectives, 31.(4), 73-102. https://doi.org/10.1257/jep.31.4.73

Banerji, R., Berry, J., Shotland, M., 2017. The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India. American Economic Journal: Applied Economics 9, 303–337. https://doi.org/10.1257/app.20150390

Bau, N., Das, J., & Yi Chang, A. (2021). New evidence on learning trajectories in a low-income setting. International Journal of Educational Development, 84, 102430. https://doi.org/10.1016/j.ijedudev.2021.102430

Bau, N., Das, J., & Yi Chang, A. (2021). New evidence on learning trajectories in a low-income setting. International Journal of Educational Development, 84, 102430. https://doi.org/10.1016/j.ijedudev.2021.102430

Beatty, T. K. M., & Shimshack, J. P. (2011). School buses, diesel emissions, and respiratory health. Journal of Health Economics, 30(5), 987–999. https://doi.org/10.1016/j.jhealeco.2011.05.017

Benstead, L.J. (2014). Does interviewer religious dress affect survey responses? Evidence from Morocco. Politics and Religion 7: 734–60

Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K., & Sheridan, M. (2016). Measuring the measurement error: A method to qualitatively validate survey data. Journal of Development Economics, 120, 99–112. https://doi.org/10.1016/j.jdeveco.2016.01.005

Blaydes, L., Gillum, R.M. (2013). Religiosity-of-interviewer effects: assessing the impact of veiled enumerators on survey response in Egypt. Politics and Religion 6: 459–82

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. Journal of Public Economics, 168, 1–20. https://doi.org/10.1016/j.jpubeco.2018.08.007

Brophy, J. (2006). Grade Repetition. International Academy of Education (IAE) and the International Institute for Educational Planning (IIEP).

Brunette, T., Piper, B., Jordan, R., King, S., & Nabacwa, R. (2019). The impact of mother tongue reading instruction in twelve Ugandan languages and the role of language complexity, socioeconomic factors, and program implementation. Comparative Education Review, 63. (4), 591-612. https://doi.org/10.1086/705426

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location-scale model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(2), 551–568. https://doi.org/10.1111/rssa.12205

Bryan, G., Chowdhury, S., Mobarak, A. M. Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh. (2014). Econometrica, 82(5), 1671–1748. https://doi.org/10.3982/ECTA10489

Carter, E., Rose, P., Sabates, R., Akyeampong, K., 2020. Trapped in low performance? Tracking the learning trajectory of disadvantaged girls and boys in the Complementary Basic Education programme in Ghana. International Journal of Educational Research 100, 101541. https://doi.org/10.1016/j.ijer.2020.101541

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2020). Income Segregation and Intergenerational Mobility Across Colleges in the United States*. The Quarterly Journal of Economics, 135(3), 1567–1633. https://doi.org/10.1093/qje/qjaa005

Chicoine, L. (2019). Schooling with learning: The effect of free primary education and mother tongue instruction reforms in Ethiopia. Economics of Education Review. 69, 94-107. https://doi.org/10.1016/j.econedurev.2019.01.002

Chiplunkar, G., Dhar, Nagesh, R. (2020). Too little, too late: improving post-primary learnin outcomes in India. wORKING PAPER. https://riseprogramme.org/sites/default/files/inline-files/Dhar.pdf

Chisholm, L., & Leyendecker, R. (2008). Curriculum reform in post-1990s sub-Saharan Africa. International Journal of Educational Development, 28. 2, 195-205. https://doi.org/10.1016/j.ijedudev.2007.04.003

Cohen, J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. Educational Researcher, 45(6), 378–387. https://doi.org/10.3102/0013189X16659442

Crépon, B., Ferracci, M., & Fougère, D. (2012). Training the Unemployed in France: How Does it Affect Unemployment Duration and Recurrence? Annals of Economics and Statistics, 107/108, 175. https://doi.org/10.2307/23646576

Crouch, L., Rolleston, C., & Gustafsson, M. (2021). Eliminating global learning poverty: The importance of equalities and equity. International Journal of Educational Development, 82, 102250. https://doi.org/10.1016/j.ijedudev.2020.102250

Cunha, F., & Heckman, J. (2007). The Technology of Skill Formation. American Economic Review, 97(2), 31–47. https://doi.org/10.1257/aer.97.2.31

Datzberger, S. (2018). Why education is not helping the poor. Findings from Uganda. World Development. 110, 124-139. https://doi.org/10.1016/j.worlddev.2018.05.022

DHS: Kenya National Bureau of Statistics and ICF International. Kenya Demographic and Health Survey 2014 [Dataset]. Data Extract from KEHR72DT.dta DHS Household survey.

DHS: Ministry of Health, Community Development, Gender, Elderly and Children [Tanzania], Ministry of Health [Zanzibar], National Bureau of Statistics [Tanzania], Office of the Chief Government Statistician, and ICF. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-16 [Dataset]. Data Extract from TZHR7BDT.dta, DHS and ICF [Distributors]. Accessed from https://dhsprogram.com/ on September 1, 2021.

Di Maio, M., & Fiala, N. (2020). Be Wary of Those Who Ask: A Randomized Experiment on the Size and Determinants of the Enumerator Effect. The World Bank Economic Review, 34(3), 654–669. https://doi.org/10.1093/wber/lhy024

Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S., & Duflo, E. (2017). Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. Science, 357 (6346), 47-55. https://doi.org/10.1126/science.aal4724

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. International Journal of Educational Development, 40, 315–322. https://doi.org/10.1016/j.ijedudev.2014.11.004

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. International Journal of Educational Development, 40, 315–322. https://doi.org/10.1016/j.ijedudev.2014.11.004

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. American Economic Review. *101*(5), 1739-74. https://doi.org/10.1257/aer.101.5.1739

Ejdemyr, S., Kramon, E., & Robinson, A. L. (2018). Segregation, Ethnic Favoritism, and the Strategic Targeting of Local Public Goods. Comparative Political Studies, 51(9), 1111–1143. https://doi.org/10.1177/0010414017730079

Ejdemyr, S., Kramon, E., & Robinson, A. L. (2018). Segregation, ethnic favoritism, and the strategic targeting of local public goods. Comparative Political Studies. 51(9), 1111–1143. https://doi.org/10.1177/0010414017730079

Erling, E., Adinolfi, L., & Hultgren, A. (2017). Multilingual classrooms: Opportunities and challenges for English medium instruction in low and middle income contexts.

Evans, D. K., & Mendez Acosta, A. (2021). Education in Africa: What Are We Learning? Journal of African Economies, 30(1), 13–54. https://doi.org/10.1093/jae/ejaa009

Evans, D. K., & Mendez Acosta, A. (2021). Education in Africa: What Are We Learning? Journal of African Economies, 30(1), 13–54. https://doi.org/10.1093/jae/ejaa009

Evans, D., & Yuan, F. (2019). Equivalent years of schooling: A metric to communicate learning gains in concrete terms. World Bank Policy Research Working Paper, 8752. https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-8752

Evans, D., & Yuan, F. (2020). How big are effect sizes in international education studies? CGD Working Paper 545.

Evans. D. K., Hares, S. (2021). Should governments and donors prioritize investments in foundational literacy and numeracy? Center for Global Development Working Paper 579.

Ganimian, A. J., & Murnane, R. J. (2016). Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations. Review of Educational Research, 86(3), 719–755. https://doi.org/10.3102/0034654315627499

Gibbs, B. G., & Heaton, T. B. (2014). Drop out from primary to secondary school in Mexico: A life course perspective. International Journal of Educational Development, *36*, 63-71. https://doi.org/10.1016/j.ijedudev.2013.11.005

Glewwe, P., & Muralidharan, K. (2016). Improving education outcomes in developing countries. In Handbook of the Economics of Education. (Vol. 5, pp. 653-743). Elsevier. https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Glewwe, P., & Muralidharan, K. (2016). Improving education outcomes in developing countries. In Handbook of the Economics of Education. (Vol. 5, pp. 653-743). Elsevier. https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. American Economic Journal: Applied Economics. 1(1), 112-135. https://doi.org/10.1257/app.1.1.112

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. American Economic Journal: Applied Economics. 1(1), 112-135. https://doi.org/10.1257/app.1.1.112

Glick, P., & Sahn, D. E. (2010). Early academic performance, grade repetition, and school attainment in Senegal: panel data analysis (English). The World Bank Economic Review. *24*(1), 93-120.

Gomes-Neto, J. B., & Hanushek, E. A. (1994). Causes and Consequences of Grade Repetition: Evidence from Brazil. Economic Development and Cultural Change. *43*(1), 117-148. https://doi.org/10.1086/452138

Government of Tanzania. "Big Results Now! Annual report 2013/14," Presidential Delivery Bureau 2015.

Hanushek, E. A., Woessmann, L. (2007). The role of education quality for economic growth. World Bank Policy Research Working Paper No. 4122.

Hanushek, E.A., Schwerdt, G., Wiederhold, S., Woessmann, L., 2015. Returns to skills around the world: Evidence from PIAAC. European Economic Review 73, 103–130. https://doi.org/10.1016/j.euroecorev .2014.10.006

Harding, R., Stasavage, D., 2014. What Democracy Does (and Doesn't Do) for Basic Services: School Fees, School Inputs, and African Elections. The Journal of Politics 76, 229–245. https://doi.org/10.1017/S0022381613001254

Heckman, J. J. (2006). Skill Formation and the Economics of Investing in Disadvantaged Children. Science, 312(5782), 1900–1902. https://doi.org/10.1126/science.1128898

Jones, S., Schipper, Y., Ruto, S., & Rajani, R. (2014). Can Your Child Read and Count? Measuring Learning Outcomes in East Africa. Journal of African Economies, 23(5), 643–672. https://doi.org/10.1093/jae/eju009

Kaffenberger, M. (2019). PISE-D Reveals Exceptionally Low Learning. RISE Programme. Blog post. https://riseprogramme.org/blog/PISA-D_low_learning

Kaffenberger, M., & Pritchett, L. (2020). Aiming higher: Learning profiles and gender equality in 10 low- and middle-income countries. International Journal of Educational Development, 79, 102272. https://doi.org/10.1016/j.ijedudev.2020.102272

Kaffenberger, M., & Pritchett, L. (2021). A structured model of the dynamics of student learning in developing countries, with applications to policy. International Journal of Educational Development, 82, 102371. https://doi.org/10.1016/j.ijedudev.2021.102371

Kaffenberger, M., Sobol, D., Spindelman, D. (2021). The role of low learning in driving dropout: a longitudinal mixed methods study in four countries. RISE Working Paper Series. 21/070. https://doi.org/10.35489/BSG-RISE-WP_2021/070

Kerwin, J. T., & Thornton, R. L. (2020). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. The Review of Economics and Statistics, 1–45. https://doi.org/10.1162/rest_a_00911

Komba, A. and Shukia, R. 2021. Accountability Relationships in 3Rs Curriculum Reform Implementation: Implication for Pupils' Acquisition of Literacy and Numeracy Skills in Tanzania's Primary Schools. RISE Working Paper Series. 21/065. https://doi.org/10.35489/BSG-RISE-WP_2021/065

Laitin, D. D., Ramachandran, R., & Walter, S. L. (2019). The legacy of colonial language policies and their impact on student learning: evidence from an experimental

program in Cameroon. Economic Development and Cultural Change, 68(1), 239-272. https://doi.org/10.1086/700617

Lima, R. C. de A., & Silveira Neto, R. da M. (2018). Secession of municipalities and economies of scale: Evidence from Brazil. Journal of Regional Science, 58(1), 159–180. https://doi.org/10.1111/jors.12348

Luna-Bazaldua, D., Jiberman, J., Levin, V. (2021). Assessing outside of the "classroom box" while schools are closed: the potential of phone-based formative assessments to support learning continuity. Education for Global Development. World Bank Blogs.

Lupu, N., & Michelitch, K. (2018). Advances in Survey Methods for the Developing World. Annual Review of Political Science, 21(1), 195–214. https://doi.org/10.1146/annurev-polisci-052115-021432

Malisa, M., & Missedja, T. Q. (2019). Schooled for servitude: the education of African children in British colonies, 1910-1990. Genealogy, 3(3), 40. https://doi.org/10.3390/genealogy3030040

Manacorda, M. (2012). The cost of grade retention. Review of Economics and Statistics. *94*(2), 596-606. https://doi.org/10.1162/REST_a_00165

Masino, S., Niño-Zarazúa, M. (2016). What works to improve the quality of student learning in developing countries? International Journal of Educational Development 48: 53–65. http: //dx.doi.org/10.1016/j.ijedudev.2015.11.012

Mathew, J.L., 2012. Inequity in childhood immunization in India: A systematic review. Indian Pediatr 49, 203–223. https://doi.org/10.1007/s13312-012-0063-z

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, incentives, and complementarities in education: experimental evidence from Tanzania. The Quarterly Journal of Economics, 134(3), 1627-1673. https://doi.org/10.1093/qje/qjz010

Mbiti, I., Romero, M., Schipper, Y. (2021). Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania. NBER Working Paper No. 25903

Meléndez, L., (2013). Rol del documento de apoyo en el desarrollo de competencias lingüísticas y comunicativas de estudiantes de primer Ciclo: alcances y desafíos 2014-2018. *En Séptimo Informe del Estado de la Educación.* Costa Rica: Programa Estado de la Nación.

Mena, F. (2014, February 18). 45% de los maestros no están capacitados para impartir nuevos programas de Español. CRHoy. https://archivo.crhoy.com/45-de-los-educadores-de-primaria-no-estan-capacitados-para-impartir-nuevos-programas-de-espanol-8244l7l9x/nacionales/

Mensch, B.S., Chuang, E.K., Melnikas, A.J., Psaki, S.R., 2019. Evidence for causal links between education and maternal and child health: systematic review. Tropical Medicine & International Health 24, 504–522. https://doi.org/10.1111/tmi.13218

Minardi, A. L., Rossiter, J., & Hares, S., (2020, February 11). Grade Repetition in Developing Countries: Repeat to Fail or Second Time's a Charm? [Blog post]. Center for Global Development. https://www.cgdev.org/blog/grade-repetition-developing-countries-repeat-fail-or-second-times-charm

Ministry of Education, Science and Technology,. (2016). Curriculum for basic education Standard I and II. ISBN. 978 - 9976 - 61- 436 - 7.

Ministry of Public Education/MEP (2013). Programa de estudio de Español. I Ciclo de la Enseñanza General Básica. Costa Rica: Ministerio de Educación Pública.

Ministry of Public Education/MEP (2020). Indicadores del sistema educativo costarricense. Publication No. 377-17.

Montenegro, C. E., Patrinos, H. A. (2014). Comparable estimates of returns to school around the world. World Bank Policy Research Working Paper No. 7020.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. American Economic Review. *109*(4), 1426-60. https://doi.org/10.1257/aer.20171112

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in India. American Economic Review, 109(4), 1426-1460. https://doi.org/10.1257/aer.20171112

Muralidharan, K., Sundararaman, V. (2011). Teacher performance pay: experimental evidence from India. Journal of Political Economy, 119(1), 39-77. https://doi.org/10.1086/659655

Mwiria, K. (1991). Education for subordination: African education in colonial Kenya. History of Education, 20(3), 261-273. https://doi.org/10.1080/0046760910200306

OECD (2015), PISA mathematics performance by decile of social background. https://www.slideshare.net/OECDEDU/istp-2014-equity-excellence-and-inclusiveness-in-education/13-1616PISA_mathematics_performanceby_decile_of.

OECD (2019), Public spending on education: https://data.oecd.org/eduresource/public-spending-on-education.htm#indicator-chart. Accessed on December 12, 2019.

Okun, A. M. (2015). Equality and efficiency: The big tradeoff. Brookings Institution Press.

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., & West, B. T. (2020). The Past, Present, and Future of Research on Interviewer Effects. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), Interviewer Effects from a Total Survey Error Perspective (1st ed., pp. 3–16). Chapman and Hall/CRC. https://doi.org/10.1201/9781003020219-2

Ozturk, I. (2001). The role of education in economic development: a theoretical perspective. Journal of Rural Development and Administration. 33(1), 39-47

Paltasingh, K.R., Goyari, P., 2018. Impact of farmer education on farm productivity under varying technologies: case of paddy growers in India. Agricultural and Food Economics 6, 7. https://doi.org/10.1186/s40100-018-0101-9

Parket, B. (2019, June 11). The Department of Education's proposed no repeat policy for grade R to 3 will do more harm than good to the state of education in SA. *Parent 24.* https://www.parent24.com/Learn/Learning-difficulties/the-department-of-educations-proposed-no-repeat-policy-for-grade-r-to-3-will-do-more-harm-than-good-on-the-state-of-education-in-sa-20190610

Piper, B., Zuilkowski, S. S., Kwayumba, D., & Oyanga, A. (2018). Examining the secondary effects of mother-tongue literacy instruction in Kenya: Impacts on student learning in English, Kiswahili, and mathematics. International Journal of Educational Development, 59, 110-127. https://doi.org/10.1016/j.ijedudev.2017.10.002

Pritchett, L. (2013). The rebirth of education: Schooling ain't learning. Center for Global Development.

Pritchett, L. (2013). The rebirth of education: Schooling ain't learning. Center for Global Development.

Pritchett, L., & Beatty, A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. International Journal of Educational Development, 40, 276–288. https://doi.org/10.1016/j.ijedudev.2014.11.013

Pritchett, L., & Beatty, A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. International Journal of Educational Development, 40, 276–288. https://doi.org/10.1016/j.ijedudev.2014.11.013

Pritchett, L., Woolcock, M., & Andrews, M. Capability Traps? The mechanisms of persistent implementation failure. Center for Global Development. CDG Working Paper. (232)

Programa Estado de la Educación (2019). Séptimo Informe del Estado de la Educación. Costa Rica: Programa Estado de la Nación.

Randall, S., Coast, E., Compaore, N., & Antoine, P. (2013). The power of the interviewer: A qualitative perspective on African survey data collection. Demographic Research, 28, 763–792. https://doi.org/10.4054/DemRes.2013.28.27

Rodriguez-Segura, D. (2020). Strengthening early literacy skills through social promotion policies? Intended and unintended consequences in Costa Rica. International Journal of Educational Development, 77, 102243. https://doi.org/10.1016/j.ijedudev.2020.102243

Rodriguez-Segura, D., Campton, C., Crouch, L., & Slade, T. S. (2021). Looking beyond changes in averages in evaluating foundational learning: Some inequality measures. International Journal of Educational Development, 84, 102411. https://doi.org/10.1016/j.ijedudev.2021.102411

Rodriguez-Segura, D., Schueler, B. E., (2022). Can learning be measured by phone? Evidence from Kenya. EdWorkingPaper: 22-517. Retrieved from Annenberg Institute at Brown University: https://- doi.org/10.26300/gc6v-qv41

Ross, A. (2014, January 22). Niños de primer grado estrenarán este año plan para leer y escribir. *La Nación*. https://www.nacion.com/el-pais/educacion/ninos-de-primer-grado-estrenaran-este-ano-plan-para-leer-y-escribir/XCIY2RGEURASFEES4XYVVOTBSY/story/

Sa, E. (2007). Language Policy for Education and Development in Tanzania. Working paper. https://www.swarthmore.edu/sites/default/files/assets/documents/linguistics/2007_sa_eleuthera.pdf

Sabates, R., Hossain, A., Lewin, K. (2010) School drop out in Bangladesh: new insights from longitudinal evidence. Consortium for Research on Educational Access, Transitions and Equity. Research Monograph No. 49.

Sandholtz, W.A. (2021). Do voters reward service delivery? Experimental evidence from Liberia. Working paper.

Schueler, B. E., Rodriguez-Segura, D. (2021). A Cautionary Tale of Tutoring Hard-to-Reach Students in Kenya. EdWorkingPaper: 21-432. Retrieved from Annenberg Institute at Brown University: https://- doi.org/10.26300/43qs-cg37

SDI – Service Delivery Indicators. (2010-2014). Education data [Data file]. https://www.sdindicators.org/aboutus#where-are-we-now-

Seid, Y. (2019). The impact of learning first in mother tongue: Evidence from a natural experiment in Ethiopia. Applied Economics, 51(6), 577–593. https://doi.org/10.1080/00036846.2018.1497852

Slavin, Stuart, and Marcel F D'Eon. "Overcrowded curriculum is an impediment to change (Part A)." *Canadian medical education journal* vol. 12,4 1-6. 14 Sep. 2021, doi:10.36834/cmej.73532

Spaull, N., Kotze, J., 2015. Starting behind and staying behind in South Africa: The case of insurmountable learning deficits in mathematics. International Journal of Educational Development 41, 13–24. https://doi.org/10.1016/j.ijedudev.2015.01.002

Strauss, M. E., & Smith, G. T. (2009). Construct Validity: Advances in Theory and Methodology. Annual Review of Clinical Psychology, 5(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Taylor, Y.J., Laditka, S.B., Laditka, J.N., Huber, L.R.B., Racine, E.F., 2016. Associations of Household Wealth and Individual Literacy with Prenatal Care in Ten West African Countries. Matern Child Health J 20, 2402–2410. https://doi.org/10.1007/s10995-016- 2068-z

Teaching at the Right Level: Strengthening Foundational Skills to Accelerate Learning. URL: https://www.teachingattherightlevel.org. Retrieved 25 November 2019

Twaweza. (2014). Uwezo data-Household data https://www.twaweza.org/go/uwezodatasets

Twaweza. (2015). Uwezo data-Household data. [Data file]. https://www.twaweza.org/go/uwezodatasets

UNESCO. (2012). Opportunities lost: The impact of grade repetition and early school leaving. Global Education Digest, 2012. ISBN: 978-92-9189-120-7.

United Nations Educational, Scientific and Cultural Organization Institute for Statistics. (2019). New Methodology Shows that 258 Million Children, Adolescents and Youth Are Out of School (UIS/2019/ED/FS/56; Fact Sheet). http://uis.unesco.org/sites/default/files/documents/new-methodology-shows-258-million-children-adolescents-and-youth-are-out-school.pdf

United Nations. (2021). Sustainable Development Goals (SGDs). https://sustainabledevelopment.un.org/topics/sustainabledevelopmentgoals

USAID. (2015). Testing and reading, writing and arithmetic (3Rs) reform in Tanzania. https://wwww.usaid.gov/sites/default/files/documents/1865/Bruns.pdf

Uwezo. (2017). Are Our Children Learning? Uwezo Tanzania Sixth Learning Assessment Report, Dar es Salaam: Twaweza East Africa.

Valerio, A., Sanchez Puerta, M.L., Tognatta, N.R., Monroy Taborda, S., 2016. Are there skills payoffs in low- and middle-income countries ? empirical evidence using STEP data (No.

Varly, Pierre. (2020). Learning assessments in Sub-Saharan Africa. SDG4-Education 2030 in Sub-Saharan Africa. Analytic Report N°1. https://learningportal.iiep.unesco.org/en/library/learning-assessments-in-sub-saharan-africa

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. Economics of Education Review, 64, 298–312. https://doi.org/10.1016/j.econedurev.2018.01.005

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? Economics of Education Review, 53, 31–45. https://doi.org/10.1016/j.econedurev.2016.05.002

West, B.T., Blom, A.G. (2017). Explaining interviewer effects: a research synthesis. Journal of Survey Statistics and Methodology. 5: 175–211

West, M. (2012). Is Retaining Students in the Early Grades Self-Defeating?. Center on Children and Families at Brookings. CCF Brief No. 49.

Williams, T. P. (2017). The Political Economy of Primary Education: Lessons from Rwanda. World Development, 96, 550–561. https://doi.org/10.1016/j.worlddev.2017.03.037

Williams, T. P. (2017). The political economy of primary education: lessons from Rwanda. World Development, 96, 550-561. https://doi.org/10.1016/j.worlddev.2017.03.037

World Bank DIME (2022). Enumerator Training. https://dimewiki.worldbank.org/Enumerator_Training

World Bank, "Tanzania Education Program for Results," Implementation Status and Results Report P147486 No. 4, World Bank, Washington D.C. 2016.

World Bank, "Tanzania Education Program for Results," Implementation Status and Results Report P147486 No. 5, World Bank, Washington D.C. 2016.

World Bank, "Tanzania Education Program for Results," Implementation Status and Results Report P147486 No. 2, World Bank, Washington D.C. 2015.

World Bank, World Development Indicators. (2021a). Population ages 0-14, total. [Data file]. Retrieved from https://data.worldbank.org/indicator/SP.POP.0014.TO

World Bank, World Development Indicators. (2021b). School enrollment, primary (% net). [Data file]. Retrieved from https://data.worldbank.org/indicator/SE.PRM.NENR

World Bank, World Development Indicators. (2021c). Population growth (annual %) [Data file]. Retrieved from https://data.worldbank.org/indicator/SP.POP.GROW

World Bank, World Development Indicators. (2021d). School enrollment, primary (% gross) – India, 1971 and 2019.[Data file]. Retrieved from https://data.worldbank.org/indicator/SE.PRM.ENRR?locations=INWPS7879). The World Bank.

World Bank. (2017). World Development Report 2018: Learning to Realize Education's Promise. The World Bank.

World Bank. (2017). World Development Report 2018: Learning to Realize Education's Promise. The World Bank. https://doi.org/10.1596/978-1-4648-1096-1

World Bank. (2019). "Tanzania Mainland Poverty Assessment: Tanzania's path to poverty reduction and pro-poor growth. Part I". http://documents.worldbank.org/curated/en/254411585030305188/pdf/Part-1-Path-to-Poverty-Reduction-and-Pro-Poor-Growth.pdf

World Bank. (2020). Cost-effective approaches to improve global learning. Recommendations of the Global Education Evidence Advisory Panel. https://documents1.worldbank.org/curated/en/719211603835247448/pdf/Cost-Effective-Approaches-to-Improve-Global-Learning-What-Does-Recent-Evidence-Tell-Us-Are-Smart-Buys-for-Improving-Learning-in-Low-and-Middle-Income-Countries.pdf

World Bank. World Development Indicators, The World Bank Group, 2019, data.worldbank.org/indicator/

APPENDIX A

a. Additional Figures

FIGURE A.2.1: *Summary Statistics for School Characteristics*

Variable	Mean	SD
How many times did the committee meet last year	5.0	2.8
How many times did the Ministry of Education visit the school last year	2.4	2.8
How many teachers are there in this school	15.3	12.2
Is there an unfulfilled request for teachers	0.8	0.4
How many teachers are absent from the school at the moment	3.1	5.0
Is this an urban school	0.2	0.4
How many classrooms does this school have	9.0	3.8
Are there any unused classrooms or vacant rooms in the school	0.2	0.4
Is overcrowding a problem at this school	0.9	0.3
Does this school provide breakfast	0.1	0.3
Does this school provide Lunch	0.2	0.4
Observations	60	

Notes. Averages calculated at the school level for 2013

FIGURE A. 2.2: *Structure of Group Cells for Difference-in-differences*

			Year	
			Pre period (2014)	Post period (2015)
Grade	1 2	Treatment grades (T)	T-pre	T-post
	3	Comparison grade (C)	C-pre	C-post

Notes. We highlight in green the group of students for which the interaction term between Treatment*Post would equal 1.

FIGURE A.2.3: *Regression Estimates of the Causal Effect of the Curriculum Reform on Learning Using Uwezo Data (2010-2017)*

	Math	English	Kiswahili
Main estimates	0.11*** (0.01) [0.08] 391,813	-0.04*** (0.01) [0.06] 389,974	0.01 (0.01) [0.92] 396,666
Estimates excluding grades 3 and 4	0.17*** (0.01) [0.00] 275,193	-0.06*** (0.01) [0.00] 273,965	0.10*** (0.01) [0.44] 278,760

Notes. These specifications exclude 2016, as Uwezo data was not collected for Tanzania in 2016. By 2017, grades 3 and 4 had been treated, so excluding these groups from the specification is the cleanest approach to avoid contamination of the comparison group. coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

FIGURE A. 2.4: Lee Bounds for Long-term Effects on SFNA Passing Rates in 2018

	Main estimates		Main estimates, no 2017	
	Lower estimate	Upper estimate	Lower estimate	Upper estimate
Kiswahili	-0.13*** (0.00) [0.16] N=7,494,870	-0.12*** (0.00) [0.16] N=7,960,261	-0.15*** (0.00) [0.00] N=5,541,256	-0.14*** (0.00) [0.16] N=7,494,870
English	-0.06*** (0.00) [0.16] N=7,491,304	-0.06*** (0.00) [0.16] N=7,961,820	-0.04*** (0.00) [0.21] N=5,535,941	-0.04*** (0.00) [0.05] N=5,872,576
Math	-0.04*** (0.00) [0.16] N=7,493,808	-0.03*** (0.00) [0.16] N=7,962,156	-0.03*** (0.00) [0.27] N=5,540,742	-0.02*** (0.00) [0.16] N=5,872,439

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

b. What changed within classrooms and schools?

Beyond the change in curriculum, we also explore whether the reform was correlated with other behavioral responses at the school and classroom level. To do so, we group several variables which are available for all years either from head teacher, principal, and teacher surveys, and explore the differences before and after the policy reform. In the aggregation of these variables, we picked all the variables for which we have consistent and reliable data across pre- and post- years. We classified all these variables into these four, admittedly arbitrary, categories. Then, each variable was indexed from 0-1, where 1 was the ”most positive” outcome of the variable. Each category consists of the geometric mean of the indexed version of each variable within it. Note that results hold whether we subset only to control schools or all 350 schools. The current results as displayed are just for control schools. The variables within each category were:

Effort and planning:

- Hours spent teaching in a week
- Hours spent planning lessons
- Hours spent managing and supporting teachers
- Personally taught remedial classes?

Instructional methods:

- Tried a new method of strategically assigning students in groups (tracking)
- Tried a new method of strategically assigning teachers to grades
- Tried a new method of more strictly enforcing student attendance

- Tried a new method of having more teacher supports (volunteers or trainee teachers)

Inputs and monitoring:

- Number of parent-teacher meetings this year
- Number of times Ministry visited the school this year
- (Inverse of) Whether the school holds any classes outside
- Amount of inputs compared to previous years

Teacher training:

- Amount of training compared to previous years
- Training of members of school committee

It is worth noting that current data limitations for these specific surveys only allow us to explore pre- and post- changes for this specific analysis of classroom and school changes, without a clear causal framework. Appendix Figure 5 shows the changes in the post period for the aggregated categories. The only two statistically significant categories are the one that reflects whether teachers are trying out new instruction methods, and the one showing the amount of training that teachers and head teachers got. This could be consistent with a story that beyond the implementation of the curricular change, teachers were better trained and hence did not need to try new pedagogical methods.

FIGURE A.2.5: *Regression Results of Changes in School and Teacher Characteristics After 2014*

	Effort and planning	Instructional methods	Inputs and monitoring	Teacher and training
After 2014	0.045	-0.068	0.057	0.095**
	(0.03)	(0.03)	(0.09)	(0.04)
Observations	164	217	126	127

Notes. Coefficients standardized as z-scores. Robust standard errors in parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

Contrary to the data limitations using the previous outcomes, we can analyze the self-reported teacher satisfaction on different issues using a difference-in-differences framework. Specifically, the treated group are teachers of grades 1-2, and the post- period consists of years 2015. The outcome variable is a discrete variable from 0 to 4, where 4 is the highest level of satisfaction reported on each issue. Appendix Figure 6 displays the coefficient of interest of each outcome. Although none of these coefficients emerges as statistically significant, the two largest coefficients by far are the satisfaction with self-perceived prospects for promotion, and school support. Both of these agree with the story that teachers realize their performance is improving, and that there are external factors beyond teachers that are facilitating this change.

FIGURE A. 2. 6: *Regression Results Using DiD Estimator of Curriculum Change on Satisfaction Ratings by Issue*

Satisfaction with…	DiD estimate
Current job	0.02
	(0.09)
Government support	-0.01
	(0.11)
Job security	-0.03
	(0.10)
Parental support	-0.02
	(0.10)
Promotion prospects	0.11
	(0.11)
Salary	0.03
	(0.10)
School support	0.18
	(0.11)
Observations	1371

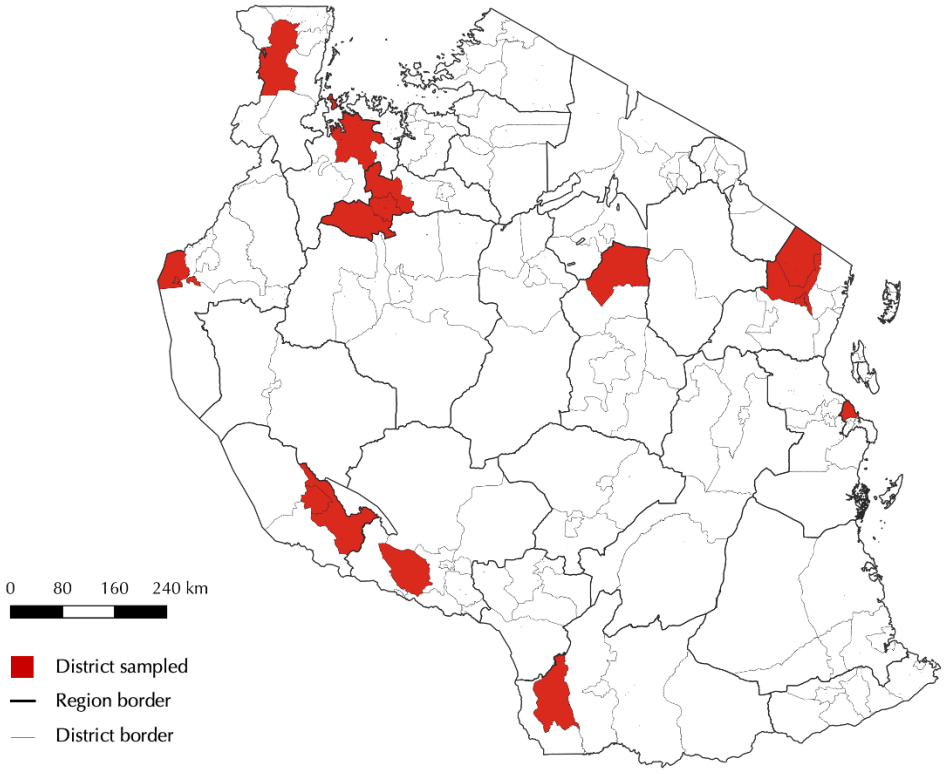
Notes. Coefficients standardized as z-scores. Robust standard errors in parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

c. More contextual details

The map below shows in red the districts where the 60 schools in the sample were drawn from. Specifically, the districts are Geita, Kahama, Karagwe, Kinondoni, Kondoa, Korogwe, Lushoto, Mbinga, Mbozi, Sumbawanga, Kigoma, Kigoma, and Korogwe. According to World Bank poverty estimates at the district level (World Bank, 2019), the poverty rate at the district-level for the selected units is 27.3% with a standard deviation of 9.6%, comparable with the national poverty rate averaged at the district-level of 29.5% with a standard deviation of 13.9%. The average 2013 rank on the Primary School Leaving Examination, a standardized test taken in grade 7, for the districts in the sample is 71.4 out of 151, with a standard deviation of 47.6, a range that covers generally the

national median and mean, and also represents districts in both ends of the achievement distribution.

Figure A.2.7: *Location of the Ten Sampled Districts*



d. Description of other contemporary reforms

FIGURE A. 2. 8: Description of other Big Results Policies in Education Passed Around the Same Time as the Curricular Reform

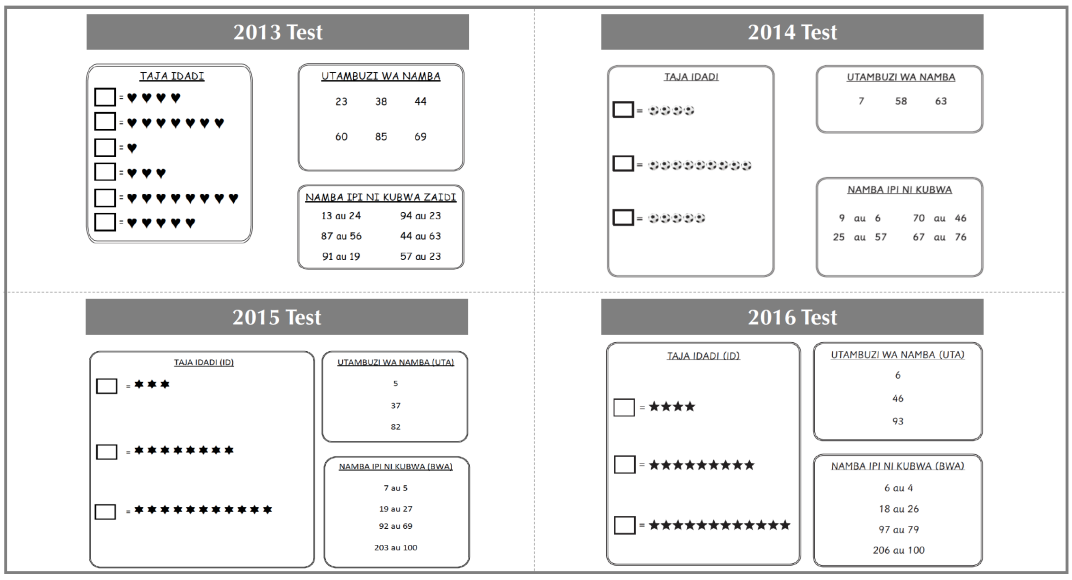
	I. Pressure to perform	
Official school ranking	Ranks all government primary (and secondary schools) by pass rates in PSLE (and CSEE). Each exam has 10 performance bands, which are classified as green, yellow, or red. Results publicly posted and widely disseminated. Both national rankings and district rankings were distributed.	Fully implemented through 2016.
School incentive scheme	Annual monetary and non-monetary incentives for primary and secondary schools that have most improved their performance in the national exams (PSLE & CSEE).	Partially implemented: 60 primary schools received financial awards in 2015. Almost 4000 non-monetary awards (certificates) distributed to primary and secondary schools starting in 2016. No awards prior to 2015.
	II. Teacher motivation	
Teacher motivation	Providing both non-monetary incentives (certificates) to high performing teachers, as well as clearing all outstanding payment arrears for teachers.	Partially implemented: outstanding claims reduced by a third by 2016.
	III. Back to basics	
National 3R assessment	Early learning assessments (Grade 2) under the 3R (reading, writing, and arithmetic) assessment program on a set of randomly selected schools.	Implemented: assessment conducted in 2016.
3R Teaching Training	Teacher training program for Grade 1 and 2 teachers on how to teach reading, writing and arithmetic most effectively to	Implemented: Almost 60,000 teachers were trained by 2016. Training started in 2014.

	this age group. Through a cascade model 37.5% of schools in 40 low-performing districts (out of 136) will be trained.	
Student Teacher English Programme (STEP)	STEP trained primary and secondary school teachers on how to identify and support low performing students. Teachers were trained on how to conduct diagnostic tests to determine which students need extra coaching, as well as how to develop curriculum and conduct classes for low performing students.	Partially implemented: teachers from 5500 primary schools were trained. Training started in 2014.
	IV. School management and finance	
School improvement toolkit	The programs aims to train head-teachers of primary and secondary schools on best practices in the management of schools. A practical toolkit of these practices was distributed to head-teachers.	Implemented: More than 16,000 primary schools received the materials. Distribution started in 2014.
Capitation grants	Timely disbursement of sufficient capitation grants for primary and secondary schools; as well as equalization of funding per student per district (about USD 4.6 per primary student and USD 11.6 per secondary student).	Late implementation: prior to 2016, 31% of schools received funds on time. In 2016 about 90% of schools received funds on time.

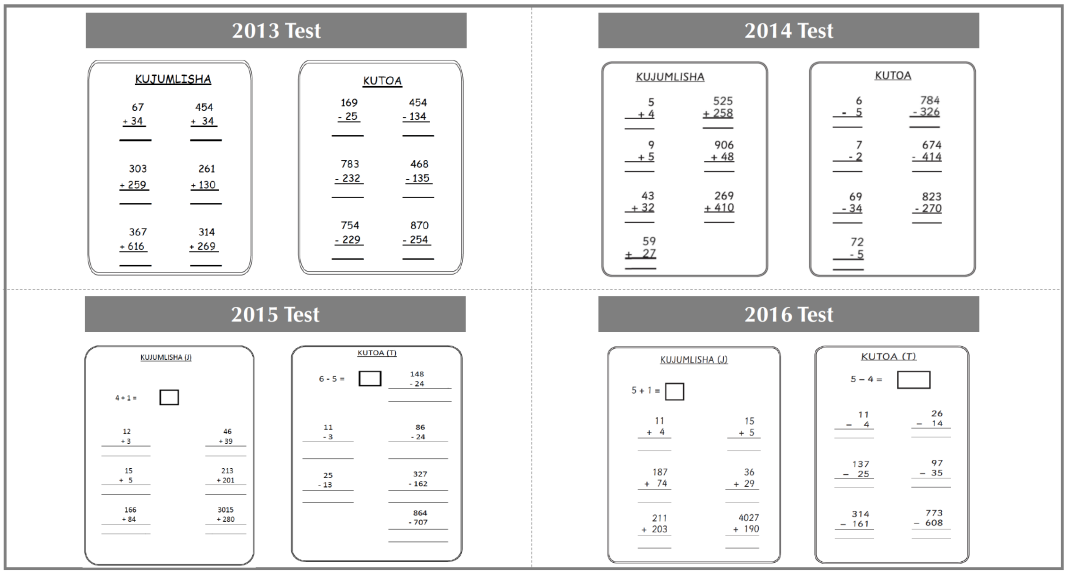
Notes. The delays in implementation were due to lack of funding. BRN was not adequately funded until donors provided funding in 2015. The information displayed in this table is compiled from a series of World Bank Project Implementation Status Reports and Tanzania Government reports.

e. Sample test booklets

Math Test – Grade 1



Math Test – Grade 3



Swahili Test – Grade 1

2013 Test

SILABI

ba he
fu da
sa li
me no

MANENO

panga soma
kula dada
pika safi
pete picha

SENTENSI

1. Damasi anasoma kitabu.
2. Paka anawinda panya.
3. Baba anakata mti.
4. Mama anapika wali.
5. Taa inawaka usiku.
6. Kaka anacheza.

KUANDIKA MANENO - IMLA

kalamu, meza, papai, dawa, redio

2014 Test

SILABI

mi se
ra po
fo

MANENO

kamati duka
sifuri chuma
kiazi chandarua

SENTENSI

1. Mama amepika chakula kitamu.
2. Kabati imejaa vitu.
3. Watoto wameingia darasani.
4. Wanafunzi wameimba vizuri.
5. Nitakuja sasa hivi.
6. Ukuta wa shule umejengwa.

2015 Test

SILABI

da mwi njwa

MANENO

rula dagaa
bomba paka
kamasi chenza

SENTENSI

1. Mtoto anasoma kitabu.
2. Juma analima shamba kubwa.
3. Shule yetu imepata walimu wapya.
4. Dereva anapiga honi ya gari.
5. Baba mdogo ametuletea zawadi.
6. Mwanafunzi amepoteza kitabu na kalamu.

2016 Test

SILABI

mo chu nywe

MANENO

paka sifuri
lengo kuni
sadaka ngedere

SENTENSI

1. Bustani ina maua.
2. Kitabu hiki kina hadithi nzuri.
3. Wazazi wamenunua makochi mapya.
4. Kaka anakata mti.
5. Kuta za nyumba zina nyufa nyingi.
6. Tembo ni mnyama mkubwa wa porini.

Swahili Test – Grade 3

2013 Test

AYA

Zahanati ya kijiji chetu ina jengo moja. Daktari wetu anaitwa Juma. Yeye anawapenda watu na ana huruma. Juzi ametoa ushauri kuwa tutumie vyandarua ili kuzuia tusiumwe na mbu.

HADITHI

Mwaka huu wanakijiji wameamua kutumia utaratibu wa kilimo cha pamoja. Wakati wa masika waliandaa mashamba. Mvua za kwanza ziliponyesha walipanda mbegu. Baada ya siku chache mimea ilichipua.

Ilipofika wakati wa palizi waliweza kufanya kazi hiyo kwa ufanisi. Walipalilia kwa kushirikiana bila kutegeana miongoni mwao. Jitihada hizi ziliwafanya wanakijiji waweze kulima eneo kubwa na wakapata mavuno mengi.

Maswali

1. Wanakijiji waliamua kutumia utaratibu gani ya kilimo?
2. Ni wakati gani wanakijiji huandaa mashamba yao?
3. Kwa nini wanakijiji waliamua kushirikiana?

2014 Test

AYA

Dada ameenda kwa shangazi. Shangazi anaishi Mnazini. Dada amefurahi kufika huko. Ameonana na binamu zake. Wote wamefurahi kumpokea. Wameandaa chakula na vinywaji.

HADITHI

Miaka mingi iliyopita kijiji chetu hakikuwa na maji ya bomba. Mwaka huu kijiji kimepata maji. Wanakijiji walishirikiana kuchimba mfereji wa kupitisha mabomba. Mafundi walifanya kazi ya kuunganisha mabomba ndipo kijiji kikapata maji.

Siku ya kuzindua mradi wa maji wanakijiji walikusanyika. Kila mmoja alitaka kuona mabomba yakitoa maji. Mgeni rasmi alitoa hotuba fupi na kufungua bomba la maji. Maji yalitiririka kwa wingi toka bombani. Wanakijiji walisherehekea kwa kuimba nyimbo na kucheza ngoma.

Maswali ya ufahamu

1. Wanakijiji walishirikiana kufanya nini?
2. Wanakijiji walisherehekea nini?
3. Kwa nini wanakijiji walifurahi?

2015 Test

AYA

Leo shule yetu wamefika wageni. Wageni hao ni wanafunzi kutoka shule iliyo jirani na shule yetu. Tuliwapokea kwa shangwe na walifurahia mapokezi yetu. Ilipofika jioni, timu za shule hizi zilicheza mpira wa miguu. Shule yetu ilishinda kwa kufunga magoli mawili kwa goli moja. Juma lijalo kutakuwa na mechi ya marudiano.

UFAHAMU

1. Shuleni kwetu walifika wageni gani?
2. Juma lijalo kutakuwa na nini?

HADITHI

Tumeelezwa kuwa misitu ni mali. Hivyo lazima tuitunze. Misitu husaidia kuhifadhi udongo. Tunashauriwa tusichome misitu ovyo. Wanyama na ndege wanaoishi msituni watakosa mahali pa kuishi. Kusipokuwa na msitu nchi itageuka kuwa jangwa.

UFAHAMU

1. Misitu ina faida gani kwa wanyama?
2. Misitu ina faida gani kwa udongo?
3. Kuna hasara gani kama tukichoma msitu?

2016 Test

AYA

Leo shule yetu wamefika wageni. Wageni hao ni wanafunzi kutoka shule iliyo jirani na shule yetu. Tuliwapokea kwa shangwe na walifurahia mapokezi yetu. Ilipofika jioni, timu za shule hizi zilicheza mpira wa miguu. Shule yetu ilishinda kwa kufunga magoli mawili kwa goli moja. Juma lijalo kutakuwa na mechi ya marudiano.

UFAHAMU

1. Shule ipi ilishinda?

HADITHI

Siku moja, Amiri aliamka mapema na akaenda kutembea. Siku hiyo kulikuwa na mnada wa mifugo kijiji cha jirani. Akaamua kuwa ni vema akaangalie jinsi mnada unavyoendeshwa. Kwa kuwa ilikuwa mara yake ya kwanza kwenda mnadani, alijifunza mambo mengi pamoja na jinsi biashara ya mifugo inavyoendeshwa.

UFAHAMU

1. Kwa nini Amiri aliamka mapema?
2. Katika kijiji cha jirani kulikuwa na tukio gani?
3. Amiri alijifunza nini?

APPENDIX B

a. Instrument used

Numeracy – Grades 3, 5, and 6

Section	Item number	Questions only for Grade 3	Questions only for Grade 5 and 6
Instructions		Assessor's name: _____ *[Anything written in this font is a note to assessor that should not be read aloud]* **Step 1: Introduction.** Hello. My name is _____ and I am calling on behalf of Bridge International Academies. I am hoping to speak first with your pupil __<insert name>___ about some math problems and then at the end to speak with you again to get a bit of information about how the term has been going for your family. Does that sound okay? **Step 2: Instructions.** First, I would like your child to work on a couple of maths problems. I ask that you put the phone on speaker or repeat out the questions to the pupil to answer. Please have your child answer the problems on their own on a scrap paper. After they are done with a problem, you or they can read out their answer to me. The answers will not count toward grades in school, so it's okay if your child does not get all of the answers correct. Is your pupil ready? **Step 3: Assessment.** ***Core Numeracy Questions*** *[Please ask students the following questions in order. If they get three questions in a row wrong, please do not ask any more of the "core numeracy questions" and move on to the "grade level questions" section below.]*	

Learning assessment – Core numeracy	1	Can you count from 20-30? *[Mark the highest number the student reached consecutively, so if they get to 27 but skipped 25, mark (e) 24]* o No answer o 23 o 27 o 20 o 24 o 28 o 21 o 25 o 29 o 22 o 26 o 30
	2	Which is greater? 64 or 38? o Correct (64) o Incorrect o No answer
	3	What is 62+18? o Correct (80) o Incorrect o No answer
	4	What is 33+49? o Correct (82) o Incorrect o No answer
	5	What is 43-20? o Correct (23) o Incorrect o No answer
	6	What is 81-43? o Correct (38) o Incorrect o No answer
	7	What is 3x4?

		○ Correct (12) ○ Incorrect ○ No answer	
	8	What is the result of 8 divided by 2? ○ Correct (4) ○ Incorrect ○ No answer	
	9	Oil is 200 shillings per liter and rice is 100 shillings a kilogram. How much should I pay for 3 liters of oil and 4 kilograms of rice? ○ Correct (1000 shillings) ○ Incorrect ○ No answer	
Instructions		***Grade Level Questions*** *[Please ask students the following questions in order. If they get three questions in a row wrong, please do not ask any more of the "grade level questions" and move on to the survey question.]*	
Learning assessment – Curriculum-aligned items	10	Complete the following number pattern: 13, 19, ___, 31 ○ Correct (25) ○ Incorrect ○ No answer	What is 3487+2325? ○ Correct (5812) ○ Incorrect ○ No answer
	11	What is 145+213? ○ Correct (358) ○ Incorrect ○ No answer	What is 4756-2149? ○ Correct (2607) ○ Incorrect ○ No answer
	12	What is 278-124?	What is 42x26?

		○ Correct (154) ○ Incorrect ○ No answer	○ Correct (1092) ○ Incorrect ○ No answer
	13	What is 8x5? ○ Correct (40) ○ Incorrect ○ No answer	What is the result of 96 divided by 12? ○ Correct (8) ○ Incorrect ○ No answer
	14	What is the result of 35 divided by 7? ○ Correct (5) ○ Incorrect ○ No answer	What is the result of 3/7 (three sevenths) + 1/4 (one fourth) ○ Correct (19/28 (nineteenth twenty-eighths) ○ Incorrect ○ No answer
Instructions		**Step 4: Student survey.** Nice work. Next I am going to ask you a general question about school. There is no right or wrong answer, please just give your best response.	
Survey – Students	15	How much do you feel your teacher cares about your learning during the remote learning period? *[Read out each answer choice]* ○ Not at all ○ A little bit ○ Some ○ Quite a bit ○ A lot	
Instructions		*[To child]:* Thank you very much. Now, I would like to ask your parent a few questions, could you put them back on? **Step 5: Parent survey.**	

		[To parent]: Thank you. Now I would like to ask you a bit about your child and household during this period of school shutdowns. Your participation is totally voluntary and you are welcome to skip any questions that you do not feel comfortable answering.
Survey – Parents	16	On average over the past week, how many hours a day has your child spent on education? *[This is in reference to the child who completed the test]*
	17	How many times has your child's teacher called you or your child by phone in the past 7 weeks? o 0 o 4 o 8 o 1 o 5 o 9 o 2 o 6 o 10 or more o 3 o 7 o
	18	What are children in your household currently doing to learn?" *[Read out each option and mark all that apply]* o Educational TV programs or radio o Bridge@home o Receiving calls from child's teacher/academy manager/academy o Educational content on the internet o Books we have in the household o Government educational content - courses, audiobooks, or lessons o I/Others in my household are teaching or reading with them o I/Others encourage children to do distance learning (radio, television, phone, etc.) but do not help ourselves o We are paying for in-person tutoring o Other [*Please specify*]:_____ o Nothing

	19	What is the highest level of school that you or someone in your household has completed? o Some primary school o Primary school completion o Some secondary school o Secondary school completion o Certificate or other post-secondary o Some university o University completion o Post-graduate degree
	20	Finally, this has been a hard time for many families due to the coronavirus pandemic. In order to understand how this disruption has influenced children's learning outcomes, it would be helpful to know whether you have experienced any of the following since March when schools were closed: *[Read out options, pausing after each option for a yes/no, and mark all that apply]* o Moved to a different home o Had someone in your home experience health challenges o Had changes to your job or income
Instructions		**Step 6: Closing.** Many thanks for your help with this.