CSAE Working Paper WPS/2024-05

Does Effective School Leadership Improve Student Progression and Test Scores? Evidence from a Field Experiment in Malawi

Salman Asim*, Ravinder Casley Gera†, Donna Harris‡, and Stefan Dercon§

Evidence from high-income countries suggests that the quality of school leadership has measurable impacts on teacher behaviors and student learning achievement. However, there is a lack of rigorous evidence in low-income contexts, particularly in Sub-Saharan Africa. This study tests the impact on student progression and test scores of a two-year, multi-phase intervention to strengthen leadership skills for head teachers, deputy head teachers, and sub-district education officials. The intervention consists of two phases of classroom training along with follow-up visits, implemented over two years. It focuses on skills related to making more efficient use of resources; motivating and incentivizing teachers to improve performance; and curating a culture in which students and teachers are all motivated to strengthen learning. A randomized controlled trial was conducted in 1,198 schools in all districts of Malawi, providing evidence of the impact of the intervention at scale. The findings show that the intervention improved student test scores by 0.1 standard deviations, equivalent to around eight weeks of additional learning, as well as improving progression rates. The outcomes were achieved primarily as a result of improvements in the provision of remedial classes.

JEL Classification: I21; I28; C93.

Key words: Education Quality; Primary School; Education Policy; Field Experiment

Registration: Pre-registered with World Bank's Strategic Impact Evaluation Fund (SIEF) in May 2019. Awarded the World Bank Human Development Award for Excellence in using Evidence to build Human Capital in June 2019.

Timeline: Baseline data collection was conducted prior to the beginning of implementation in April 2019. The intervention was conducted over two school years, 2018/2019 and 2019/2020. Endline data collection was conducted from October 2021 to February 2022.

*Senior Economist, Education Global Practice, World Bank: sasim@worldbank.org.
†Education Specialist, Education Global Practice, World Bank
‡Research Fellow, Department of Economics, University of Oxford
§Professor of Economic Policy, Blavatnik School of Government, University of Oxford

1 Introduction

Evidence from high- and middle-income countries suggests that the quality of school leadership has measurable impacts on teacher behaviors and on student learning achievement (Leithwood and Sun, 2012; Branch et al., 2012; Bloom et al., 2015; Fryer Jr, 2017; Leaver et al., 2019). In a meta-analysis of six rigorous studies based on data from the United States, Grissom et al. (2021) find the impact of improvements in principals' practices on student achievement to be nearly as large as those of a comparable improvement in teacher's practices identified in similar studies.[1] The authors identify three key areas through which school leaders achieve these impacts: organizing resources strategically around people and instruction; overseeing and providing feedback on teaching; and managing and effectively communicating with teachers, parents and the community. Adopting these three broad areas as a framework, we note that the literature finds that school leaders can improve the performance of teachers by (i) motivating them to attend school regularly and fulfil their assigned duties (Leithwood and Sun, 2012; Leaver et al., 2019), and (ii) directly supporting improvement in pedagogical practices, for example by observing lessons and providing feedback (Grissom et al., 2021; Hallinger, 2005; Ingersoll, Dougherty, and Sirinides, Ingersoll et al.; Grissom et al., 2013). Finally, school leaders can inculcate school cultures that are conducive to learning and support strong identification and communication between teachers, students and parents (Akerlof and Kranton, 2002; for Economic Co-operation and Development, 2005). However, there is limited evidence on the impacts of school leadership quality on educational outcomes in low-income countries.[2]

The apparent importance of school leadership has led to a growing body of experimental evidence on interventions to improve school leadership skills, with mixed findings. Fryer Jr (2017), de Hoyos et al. (2020) and Tavares (2015), reporting on interventions in the United States, Argentina and Brazil respectively, find significant impacts from such interventions on student outcomes, although impacts were frequently heterogenous and took 1-2 years to emerge. Studies from Mexico (Romero et al., 2022) and India (Muralidharan and Singh,

[1] The authors identify an impact of 0.13 standard deviations (S.D.) in mathematics and 0.09 in reading from a 1 S.D. improvement in principal effectiveness, as compared to a 0.17 S.D. gain in mathematics and 0.13 S.D. in reading from a 1 S.D. improvement in teacher effectiveness.

[2] There are reasonable grounds to expect a relatively high marginal return from improved leadership at the school level in these countries. In particular, constraints of inputs unsurprisingly tend to be higher in low-income countries, raising the importance of efficient allocation and utilization of these resources; however, misallocation and poor utilization tend to be more common in low-income countries, particularly those in Sub-Saharan Africa (Bashir et al., 2018), suggesting a greater potential gain from efforts to build the skills of principals in resource management and utilization. For example, in Malawi, the context of this study, shortages of teachers are exacerbated by widespread misallocation of teachers between grades (see section 2.3). Similarly, challenges of teacher behavior, including absenteeism and inappropriate behavior towards children, are on average more severe in low-income countries and particularly Sub-Saharan Africa (ibid.); this potentially increases the benefits to learning and progression improvements in the skills and practices of school leaders in these areas.

2020) observe null effects on student outcomes. Reviewing studies from 20 predominantly middle-income countries, Anand et al. (2023) finds an average impact on learning of 0.04 S.D.. However, there is a lack of experimental evidence on the impact of interventions to improve school leadership skills in low-income countries, particularly in Sub-Saharan Africa. Three studies included in Anand et al. (2023) are conducted in low-income countries. However, two of these do not report impacts on student learning or progression (Devries et al., 2015; Lassibille, 2016). The study which does report learning impacts, Blimpo et al. (2015), finds null effects .

There are reasonable grounds to expect such training to be effective in low-income contexts. In particular, while high-income countries typically have formalized merit-based promotions for head teachers, and middle-income countries including Colombia, Peru, Ecuador and Mexico have established similar systems in recent years (Commission for Quality Education for All, 2016), in most low-income countries, including Malawi, promotions are made on a discretionary basis and reflect seniority more than merit ; furthermore, the overwhelming majority of head teachers and deputy head teachers receive no training on entering their roles (Mulkeen, 2010). This suggests that there are substantial potential benefits for a well designed and implemented program of leadership training to raise the skills of head teachers, and in so doing, improve school outcomes. There is therefore a strong need for rigorous experimental evidence on the impact of such interventions in low-income settings.

This study aims to fill this lacuna in the literature by testing the impact on student progression and test scores of a two-year, multi-phase intervention in Malawi to strengthen leadership skills for head teachers, deputy head teachers, and sub-district education officials (known as Primary Education Advisers, PEAs). We implement an intervention specifically tailored to the school environment in Malawi. The intervention consists of two phases of classroom training along with follow-up visits, implemented over two years. It focuses on skills relating to school organization in a resource-constrained environment, particularly: making more efficient use of resources available at the school; motivating and incentivizing teachers to improve performance; and curating inclusive school cultures which meet the needs of all students including overage students, students with special needs, low-performing students, and girls. We conduct a randomized controlled trial, with 1,198 schools in the treatment group and a further 1,198 in the control group. The interventions are designed by a partnership of the Government of Malawi, University of Oxford, University of Malawi, and an international consulting organization, along with the World Bank; and implemented by the government and University of Malawi with support from the consulting organization.

For data, we employ a large-scale, nationally representative, independent survey of conditions, practices and learning outcomes in schools, administered by an independent research

firm. The survey and test instruments were designed by the World Bank's research team in partnership with the government and country level psychometricians and research consultants. Our endline data collection took place two years after the completion of the first (main) phase of the intervention, enabling us to observe medium-term impacts on test scores.

Our study makes two distinct contributions. This study is the only the second of which we are aware to generate rigorous experimental evidence in a low-income setting on the impact of interventions to improve school leadership on learning and progression. If the interventions can positively influence student progression and test scores in Malawi, it will have important policy implications for governments in low-income countries. Second, we evaluate a government program at national scale, with leaders from 1,198 schools (more than one-fifth of all public primary schools) from all districts in Malawi participating, in contrast to pilot interventions generally evaluated at a small scale and in particular districts or regions[3]. This implies the results of the study are relevant to all public primary schools in Malawi and the evaluated approach, if successful, is likely to have similar impact on student achievement if scaled up.

2 Design

2.1 Research Question and General Hypothesis

Our research question is: Can student progression and students' test scores improve as a result of a custom-designed multi-phase school leadership training program?

Our general hypothesis is:

> The intervention is expected to improve utilization of resources and the behavior of teachers, and create school cultures which meet the needs of over-age students, students with special needs, low-performing students, and girls; and as a result improve student progression and test scores (compared to control).

2.2 Theoretical Framework

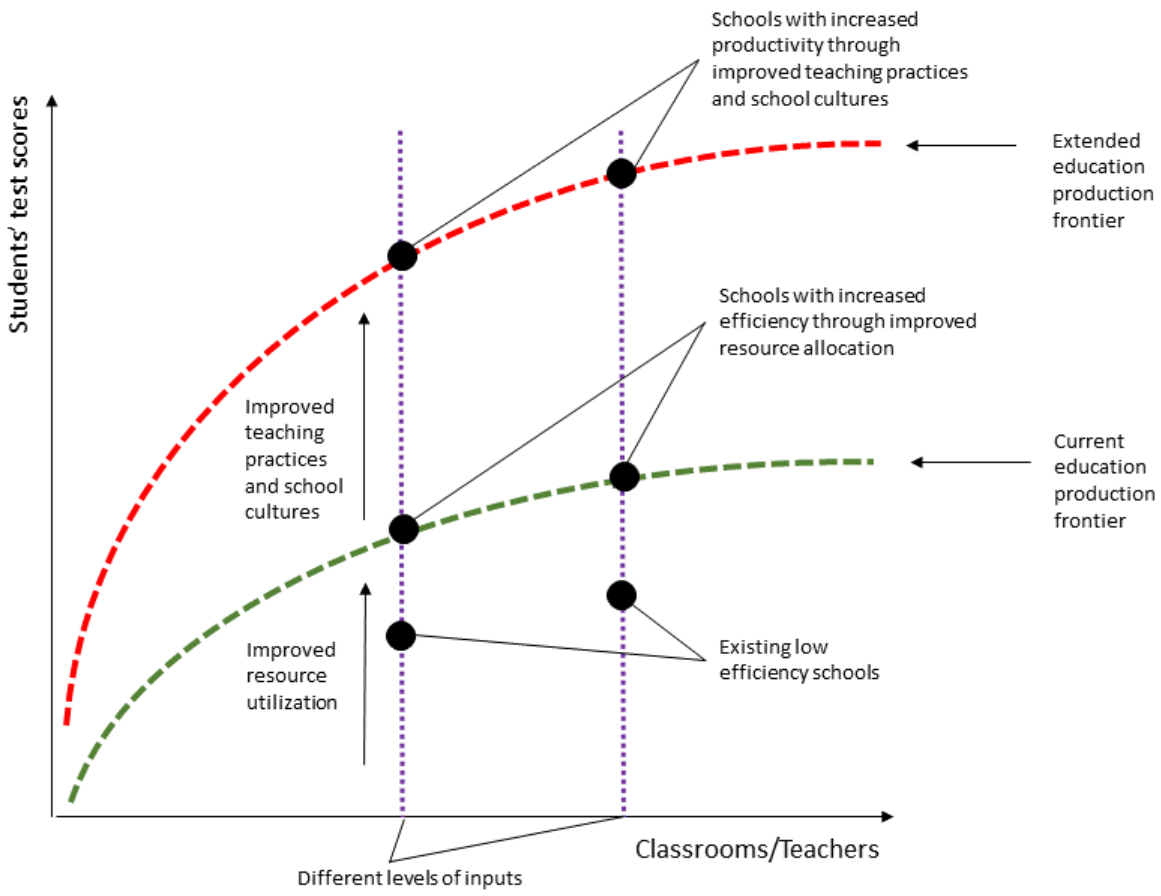
Poor outcomes in education in low-income countries reflect systemic resource limitations, poor utilization of resources, and weaknesses in education production (Hanushek, 1987; Delprato and Antequera, 2021). We employ the term *education production function* (Bowles, 1970;

[3] For example, the previous experimental study which reports learning impacts, Blimpo et al. (2015), employs a total of 273 schools across all treatment and control groups.

Hanushek, 2020) to describe the capacity of schools to convert inputs – such as teachers and classrooms – into the output of learning. In low-income countries, there exists a low level of inputs in comparison to higher-income countries, as a result of limitations of the quantity and quality of available resources such as teachers, classrooms, and textbooks. In the context of low-income countries, there may be limited scope to raise the level of inputs, as a result of overall resource constraints. In Malawi, for example, it is estimated that an additional 75,000 classrooms are required at primary level alone to bring all schools to an acceptable ratio of 60 students per permanent classroom (MoEST, 2020); even adopting a low-cost model of construction, it is estimated that this would cost US$562 million, more than one year's entire public education budget (Government of Malawi, 2022). With such large barriers to achieving significant increases in inputs, there is a need for concerted efforts to increase the education production function to achieve improvements in outcomes.

In addition, we employ the term *education production frontier* to describe the level of education production a school can produce, assuming efficient utilization of the available inputs (Jimenez and Gonzalez, 2011). Many schools operate within their education production frontier as a result of misallocation or other inefficiencies in the utilization of available inputs. School leadership has the potential to move a school's education production function to the frontier by maximizing the efficiency of allocation of resources. Furthermore, even where schools operate on the education production frontier, there may be potential to extend the frontier by increasing the productivity with which the school converts efficiently allocated resources into learning. School leadership has the potential to move a school's education production frontier outward by improving teaching practices and school cultures (Figure 1).

Figure 1: Education Production Function



School leadership can improve a school's education production function through a number of channels, including improved teacher satisfaction, effort and performance; improved allocation of resources; and improved alignment of policies, practices, and curricula (Leithwood et al., 2004; Day et al., 2009; Day and Sammons, 2013). In this study, we focus on a number of key channels of relevance to the Malawi context. Our intervention (see section 2.4) is designed to utilize each of these channels to achieve improvements in learning.

Utilization of resources. Material and human resources, including teachers, teaching and learning materials such as textbooks, and student learning materials such as notebooks and pencils, are necessary inputs for learning (Verstegen and King, 1998). The availability of re-

sources is substantially constrained in low-income contexts, and addressal of these constraints is a necessary (but not sufficient) step to improving learning outcomes (Mbiti et al., 2019; Das et al., 2013). Inefficient allocation and utilization of available resources can exacerbate these constraints. In particular, inequitable allocation of teachers to classes can exacerbate school-level staff shortages by leading to larger class sizes in relatively understaffed classes, with negative impacts for learning (Angrist and Lavy, 1999; Grissom et al., 2017), and weaknesses in utilization of textbooks can exacerbate limited availability (Milligan et al., 2017). School leaders can improve allocation and utilization of resources by allocating teachers equitably to classes; encouraging or mandating proper usage of textbooks; and otherwise monitoring and directing full use of available resources in school (Grissom et al., 2021).

Teacher behavior. Teacher behaviors are a key determinant of learning outcomes (Chetty et al., 2014). In particular, it can negatively affect students' learning outcomes if teachers do not attend school regularly, or when in school are not in the classroom. By contrast, teachers can improve learning outcomes by employing effective pedagogical techniques including assigning, collecting and marking homework and using teaching aids such as posters and models (Eren and Henderson, 2004; Asokhia, 2009). School leaders can improve teachers' motivation to attend school regularly and apply effort to achieve students' learning. Although some sources of low teacher motivation, such as poor pay or conditions, may be out of the locus of control of school leaders, school leaders may be able to exercise some influence on teachers' motivation through simple low-cost measures such as verbally praising effective or high-effort teachers. Second, school leaders can improve teachers' pedagogical skills by observing lessons, providing feedback, and encouraging and mandating specific pedagogical techniques that are underutilized (Grissom et al., 2021).

School cultures. School culture may be defined as the guiding beliefs and expectations evident in the way a school operates (Fullan, 2007), including the disciplinary climate, how well students and teachers get along, and how strongly students identify with their school (for Economic Co-operation and Development, 2005). Employing evidence from the OECD's Programme for International Student Assessment (PISA) examinations in 2000, for Economic Co-operation and Development (2005) find that an average of 8 percent of variation in PISA outcomes between schools is explained by school climates, a greater amount than school policies or resource availability. Akerlof and Kranton (2002), conducting a systematic review of case studies and other non-economic literature on successful schools, conclude that issues of culture and identity, particularly students' sense of identification and alignment with the perceived values of a school, explain a significant portion of the success of certain schools in the face of challenges of resource availability or difficult social contexts. They note that head teachers in schools with successful cultures typically invest a non-negligible share of school resources into the creation of community and minimization of differences between students.

The three channels of impact of school leadership outlined above – resource utilization, teacher behavior, and school cultures – are intrinsically connected. For example, school leaders' ability to control the allocation of resources may extend to teachers, the largest single area of expenditure in most schools – the selection and management of whom is a key driver of school culture.

2.3 Context

Malawi has made impressive achievements in access to schooling since the introduction of free primary education in 1994, but its education system has failed to keep pace with increased enrolments, resulting in overcrowded classrooms and understaffed schools.[4] These extremely poor conditions lead to high rates of grade repetition[5] and dropout. In 2017/18, at the time of the baseline for this study, repetition rates nationwide averaged 32 percent in Grade 1 and 24 percent in all grades.[6] (Ministry of Education, Science and Technology of Malawi (MoEST), 2018) Only 59 percent of students enrolling in Grade 1 remain in school to Grade 5 (ibid.). Only one in three students who enter primary education complete all eight grades (Ravishankar et al., 2016).

Even students who remain in primary school experience poor learning outcomes. At Grade 2 level, fewer than 25 percent of students achieve minimum proficiency levels in the Early Grade Reading Assessment (EGRA) conducted by the United Stages Agency for International Development (USAID). At Grade 6, fewer than 25 percent of students achieve minimum proficiency levels in the Southern African Consortium for Monitoring Educational Quality (SACMEQ) assessment in mathematics, placing Malawi near the bottom in the region. Learning outcomes also vary widely between students, even within a single class: in 86 percent of schools, according to the baseline for this study (MLSS, 2018) the gap in knowledge scores was at least one S.D., 115 points. This is equivalent to more than two years of schooling, effectively creating multi-grade classrooms. Scores are lower for girls, over-age students, and students from minority language groups (ibid.).[7]

Malawi's challenges are not unique: analyses of school systems in Sub-Saharan Africa have identified a set of challenges common to countries that have rapidly expanded their primary

[4] The average class in Grade 1 has about 150 students; and in Grade 2, about 125 students (MLSS, 2018).
[5] Most Malawian schools require students to repeat grades if they fail to pass an end-of-year examination.
[6] Grades are known as standards in Malawi, but will be referenced as grades throughout this paper. Grades range from Grade 1-12. Primary Education is from Grade 1-8, and Secondary Education from Grade 9-12.
[7] Percentage scores in mathematics, Chichewa, and English adjusted using Item Response Theory and converted to knowledge score mean-centered at 500 points with an S.D. of 100 points. See section 2.7 for details.

enrollments, including high rates of repetition and dropout in lower grades (Majgaard and Mingat, 2012; Bashir et al., 2018). However, Malawi appears to be facing greater challenges than most comparable countries. According to a recent World Bank report, among 12 sampled countries in Sub-Saharan Africa, Malawi's students are the most likely to repeat early grades, and the least likely to remain enrolled to Grade 8 (Bashir et al., 2018), and learning levels are low even by regional standards as described above. We hypothesize that, in addition to system-wide shortages of inputs, these particularly poor outcomes also reflect (i) inefficiencies in the utilization of resources in schools, (ii) weaknesses in teaching behaviors, and (iii) school cultures which are not conducive to learning for all students (Asim and Casley Gera, 2024).

We then hypothesize that these problems reflect weaknesses in school leadership. The majority of newly appointed head teachers receive no training in school leadership when appointed and consider the position as a continuation of regular teaching duties; an official program of training, which some head teachers have attended in the past, is outdated and overly focused on administrative responsibilities (World Bank, 2015).

We further hypothesize that improvements in the skills of school leaders, engendered through a targeted and high-quality program of up-skilling, could enable improvements in practices in each of these areas and, as a result, in outcomes.

Utilization of resources: Malawi's teachers and classrooms are disproportionately targeted to upper grades, reflecting poor allocation decisions by school leaders. This contributes to extremely large class sizes in lower grades: the median school has an average class size of 139 in Grade 1, but just 59 in Grade 7 (Asim and Casley Gera, 2024). Although teachers are allocated to schools by Government, their allocation to specific grades is the responsibility of head teachers, who may need to overcome pressure to over-allocate teachers to upper grades, both from teachers themselves, because of a perceived lack of prestige or promotion opportunity from working in lower grades; and from parents of students in upper grades, who prefer small class sizes to maximize the potential scores of upper grade students in the Primary School Leaving Certificate Examination (PSLCE), a high-stakes examination completed in Grade 8 which determines transition to secondary school.

Lower grades' classrooms are also typically in worse condition than those of upper grades, with fewer chairs, textbooks and other classroom materials (ibid.). Management of materials is subject to other inefficiencies, for example textbooks being stored centrally, using lesson time for distribution and preventing students from using them for home study. Although the

overall supply of these inputs in schools is largely outside of the control of school leaders,[8] head teachers can set and enforce policies on allocation and utilization, for example by allocating classrooms and teaching and learning materials to grades and by setting and enforcing policies ensuring the use of materials in the classroom and the right of students to take them home.

Problems of resource utilization extend to the most expensive and important education input, teachers. Malawi's teachers demonstrate extremely low levels of motivation, attendance, and time on task. At baseline (MLSS, 2018), 13 percent of teachers at sampled schools were absent on the day of data collection, and in 50 percent of observed classrooms the teacher was out of the room when enumerators entered during scheduled lesson time. Teachers' own estimations suggest that they spend less than 2.5 hours per day teaching. In 10 percent of observed lessons, the teacher was engaged in actual teaching activities for half of the allotted lesson time or less. Only 10 percent of teachers in observed lessons set homework for students, and only 5 percent collected or reviewed homework previously assigned to students, suggesting that thousands of teachers may be routinely giving homework that is never marked.

Teaching practices: Malawi's teachers are as likely as those in neighboring countries to employ a range of positive teaching practices, including correcting students' mistakes and providing positive reinforcement. However, Malawi's teachers lag behind on more sophisticated pedagogical techniques: only 24 percent of observed teachers at baseline used teaching aids such as posters and models to support learning; only 22 percent used local information and materials to make lessons more relevant to students; and only 56 percent of classrooms at rural schools have instructional materials such as maps and charts displayed on the walls.

School cultures: Many of Malawi's schools have not successfully developed school cultures in which all students are consistently treated with respect by teachers. Seven percent of head teachers say that teachers in their schools verbally mistreat students at least once per month, and 9 percent of students say that they have been embarrassed by a teacher in front of other students. Issues of respect extend to interactions between students: one-fifth of head teachers say that students physically harm other students at least once per month. In addition, girls achieve significantly lower learning outcomes than boys as early as Grade 4, and overage students also underperform their peers (Asim and Casley Gera, 2024), suggesting that school cultures do not equitably meet the need of all students.

[8] Teachers and classrooms are allocated to schools by Government, although schools can hire or construct more from available financial resources. Teaching and learning materials are both supplied directly to schools in response to requests by Government and donors, and obtained through grants to schools dedicated to their purchase.

2.4 Intervention

The intervention consisted of a multi-phase program to strengthen leadership skills for head teachers, deputy head teachers, and PEAs. The intervention employed both classroom training and follow-up visits to schools to measure changes in practices and behaviors and to inform a refresher phase of training conducted a year after the main training. The intervention was part of the Malawi Education Sector Improvement Project (MESIP), a major program of education investment and reforms implemented by the government from 2016 to 2021.[9]

The intervention, known as the School Leadership Program, aimed to improve the capacity of school leaders to conduct key activities in each of the three channels identified above: utilization of resources, managing teacher behavior, and building inclusive school cultures for all students, particularly overaged students, low-performing students, students with special needs, and girls. The skills and practices encouraged by the intervention were selected based on the literature on effective school leadership practices for improvements in learning, and through consultations with Government and school leaders on particular constraints in leadership practices in Malawi's schools.

To address inefficiencies in allocation and utilization of resources, the training included a module dedicated to *Improving school management in a resource constrained environment*. This was intended to provide participants with skills to, among others, equitably allocate teachers to classes and improve utilization of textbooks (Grissom et al., 2021) and classrooms. We anticipate that this module would lead to improvements in the allocation of teachers between grades, leading to improvements in pupil-teacher ratios (PTRs) in lower primary (Grades 1-4), and improved textbook utilization practices, leading to improvements in observed pupil-textbook ratios in class.

To address weaknesses in teacher behavior, including absenteeism and low time-on-task, the training included a module on *Improving teachers' motivation and rewarding performance*. This was intended to provide participants with skills to support and reward teachers' pedagogical performance (Ingersoll, Dougherty, and Sirinides, Ingersoll et al.; Grissom et al., 2013); in particular to, among others, identify issues affecting teacher morale and motivation and

[9] MESIP was financed by the Global Partnership for Education and was complemented by MESIP-Extended, a partner project financed by the Royal Norwegian Embassy which extended the MESIP interventions to additional schools. Our first cohort of schools represents schools treated under MESIP and our second, schools treated under MESIP-Extended. The implementation of the intervention was conducted concurrently for the two cohorts with the same tools, staff, and procedures.

provide support; assess teacher performance and identify strong performers and problem areas; reward strong performance using social awards, recognition in school assembly and other low-cost methods; and address problems through model lessons, in-service training and other methods. We anticipate that this module would contribute to improvements in the frequency with which head teachers observe teachers, provide feedback, and reward teachers for strong performance, leading to improvements in teacher attendance, effort and teaching practices.

To address problems of school culture, the training included a module on *Creating a positive and inclusive culture towards over-age children and girls*. This was intended to provide participants with skills to, among others: understand, monitor and prevent inappropriate treatment by teachers and students of over-age children, learners with special needs and girls, and its impact on well-being and learning outcomes; train teachers to identify, encourage and support low-performing students; and monitor and address bullying. We anticipate that this module would lead to increased Head Teacher effort to prevent bullying, leading to reduced incidence of bullying.

Finally, the training included a module on *Maintaining and promoting utilization of financial and management records*. This supported improved resource utilization, by training participants to maintain full and up-to-date records of school income and expenditures; utilization of textbooks and other non-financial inputs; and teacher attendance and time on task. It also supported improvements in teaching practices and school cultures by training participants to maintain full and up-to-date academic records, including on student attendance and learning, measured through continuous assessment; and to use these records to monitor and address poor teaching practices and chronic student absenteeism and low teacher effort.

We anticipate that, if the intervention were successful, the improvements in resource utilization, teaching practices, and school cultures engendered by the intervention would ultimately lead to improvements in student progression and dropout. See section 2.5.

The intervention was developed by a consortium of The Ministry of Education of Malawi (MoE, particularly the Department of Teacher Education and Development and Directorate of Basic Education; Asian Institute of Developmental Studies, Inc. (AIDSI), a Malaysia-based consulting organization selected by MoE through tender; The School of Education of the University of Malawi (SEUM), acting as local partner to AIDSI; the University of Oxford; and The World Bank, which provided technical support and guidance to the development of content. Following initial development of the materials and protocols, implementation was carried out by the MoE in partnership with AIDSI and SEUM.
The intervention consisted of three main phases:

Main training: This consisted of 10 days' intensive classroom training, conducted in batches of 100 participants (divided into smaller groups for particular activities). The training included a mix of direct instruction, focused on building awareness of appropriate practice and behavior, and interactive practical exercises, focused on developing goals for improvement. A pre-test and post-test were administered to participants at the start and end of training to measure the short-term gain in knowledge; the average score increased from 48 percent to 72 percent, suggesting gains in knowledge at least in the short term.

Follow-up visits: This consisted of three follow-up visits around two to six months after initial training. The first visit was conducted by one of the consortium's trainers, the second and third by the PEA of the zone who had also undergone the main training. The visits observed behavior change by participants and provided instruction and guidance on correct implementation of the strategies and approaches learned during classroom training, as well as informing the design of the refresher training (see below).

Refresher training: This consisted of three days' classroom training, focused on areas identified as having seen limited improvement in follow-up visits and by a third-party monitoring firm.

2.4.1 Impact of COVID-19 on intervention

The main training and follow-up visits were completed in April -December 2019, prior to the onset of the COVID-19 pandemic. The refresher training took place in August-September 2020 during a period when Malawi was enacting restrictions (including the closure of schools) as a result of the pandemic. Batches were divided into classes of 50 or fewer to comply with regulations on gatherings. No changes were made to the content or schedule of training.

2.5 Theory of Change

If the training is effective, we anticipate that the intervention will have direct impacts on the attitudes and behaviors of participating school leaders. In particular, we expect to observe increased incidence of the practices specifically targeted by the training, including: improved distribution of teachers between grades, measured through the teacher utilization ratio between lower and upper primary; more frequent observation of teachers by head teachers and more use of rewards for teacher performance, supporting improved teacher attendance, effort and teaching practices; and improved efforts by head teachers to prevent bullying.

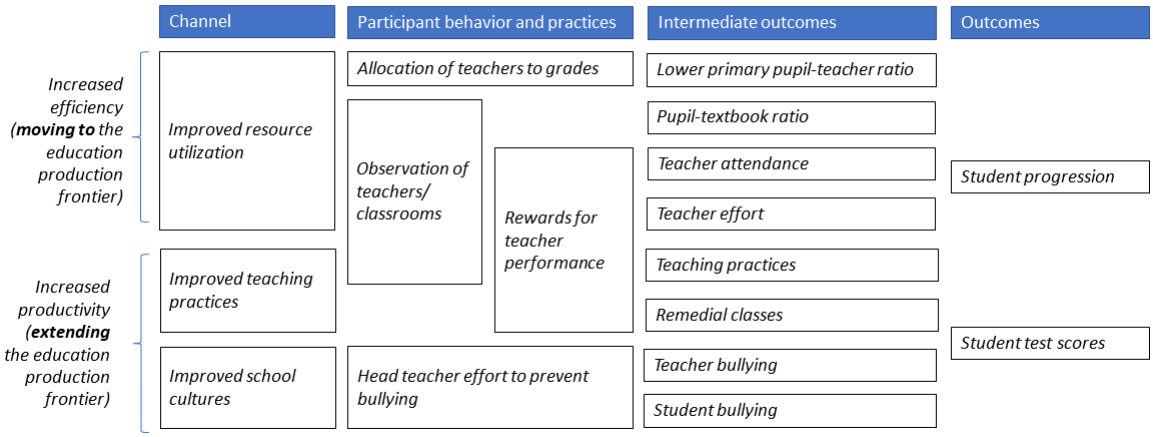
We then anticipate intermediate impacts as changes in school leaders' behaviors and practices

influence and cascade down to teachers and students' behaviors. In terms of *resource utilization*, we anticipate improvements in lower primary PTRs, as a result of improved allocations of teachers to grades; improvements in pupil-textbook ratios, as a result of improved utilization practices; and improved teacher attendance and effort in response to increased observation and reward. In terms of *teaching practices*, we anticipate improvements in the use of more sophisticated pedagogical techniques such as use of local information and materials, and the provision of remedial classes. In terms of *school cultures*, we anticipate reduced incidence of bullying of students, both by teachers and students.

We then expect these intermediate impacts to ultimately lead to improvements in student progression and test scores.

This is detailed in Figure 2.

Figure 2: Theory of Change



2.6 Identification strategy

We employ a randomized controlled trial design. Schools are assigned randomly, at the school level, to treatment (1,198 schools) or control (1,198 schools) from a population frame of all public primary schools in Malawi.[10]

We estimate the intent-to-treat (ITT) effect by fitting an Analysis of Covariance Model (ANCOVA) model. ANCOVA is the preferred estimator, as the standard difference-in-difference

[10] Excluding 'junior primary' schools which teach only Grades 1-4; see section 2.9.1

(DiD) estimator would require twice as much sample to yield the same power when baseline is taken, and autocorrelations are low (McKenzie, 2012).

For the empirical analysis, we expect two additional assumptions to hold in our ANCOVA specification. (1) Independence of the covariate and treatment effect: as schools are randomly assigned to treatment groups, we find independence between our covariates and treatment (see **Tables 1 and 2**); (2) Homogeneity of regression slopes: we expect a positive slope between the covariate and the outcome across all treatment groups. We construct two separate specifications, the first (Equation 1), measures the student-level effects, and the second (Equation 2) measures the school-level effects.

$$Y_{is} = \alpha + \beta \cdot T_i + \eta \cdot Y_{is,0} + \alpha \cdot X_{is,0} + \epsilon_{isz} \tag{1}$$

For the hypotheses that relate to individual, pupil-level outcomes, we report treatment-effect estimates based on a simple comparison of post-treatment outcomes for treatment and control individuals.

In Equation 1, Y is is the outcome of interest for student i in school s at endline. β is the coefficient of interest, capturing the effect of the intention to treat (ITT). T is an indicator for whether school s was randomly assigned to Treatment, Y_0 captures the outcomes of interest at baseline. X_0 is a vector of baseline covariates. ϵ is a mean zero error term. We cluster our errors at the school level, the unit of treatment assignment.

$$Y_s = \alpha + \beta \cdot T_i + \eta \cdot Y_{s,0} + \alpha \cdot X_{s,0} + \epsilon_s \tag{2}$$

To investigate the school-level effects, we report treatment-effect estimates based on a simple comparison of post-treatment outcomes for treatment and control schools. In Equation 2 where Y is the outcome of interest for school s and T indicates that the school was assigned to treatment, Y_0 captures the outcomes of interest at baseline. X_0 is a vector of baseline covariates. ϵ is a mean zero error term.

2.7 Outcomes

Our primary outcomes of interest are student progression and test scores.

2.7.1 Student progression

Rates of student progression are low in Malawi, as a result of high rates of dropout and repetition, driven in part by large class sizes and limited teacher effort. We anticipate improvements in class sizes, as a result of improved allocation of teachers to lower grades. We also anticipate improvements in teacher effort. As a result of these, we anticipate improvements in student progression, particularly in lower grades.

Our primary measures of student progression are dropout and repetition rates.

Dropout Rate: This is measured by the number of students in each grade dropping out in the most recently completed school year, divided by the total number of students enrolled in that grade for that year.

Repetition Rate: This is measured by the number of students in each grade repeating a grade in the most recently completed school year, divided by the total number of students enrolled in that grade for that year.

As these measures are drawn from school records, we measure them at the school level and report them overall across grades, and for lower and upper primary (Grades 1-4 and 5-8 respectively).[11]

2.7.2 Student test scores

As a result of improvements in class sizes, teacher effort, and school learning environments as described above, we anticipate improvements in the test performance of sampled students in learning assessments in mathematics and English. We utilize data from norm referenced assessments that cover competencies from Grade 1 through Grade 6 (see section 2.8. We measure these outcomes at the student level through a longitudinal sample (see section 2.9.4). Raw scores are converted to *knowledge scores* on a mean-centred scale using Item Response Theory (see section A).

Mathematics Knowledge Score: This is measured by the knowledge score of each student in the mathematics test for each school.

[11] Owing to the fact that our data is not collected annually (see section 2.8), we are unable to use standard definitions which divide the number of repeaters by the previous years' enrollment.

English Knowledge Score: This is measured the knowledge score of each student in the English test for each school.

Total Knowledge Score: This is measured by the knowledge score of each student averaged across both subjects for each school.

We also measure intermediate impacts variables, combined into Anderson indices: *resource utilization*, including the ratio of PTRs between lower and upper primary, lower primary PTRs, and lower primary pupil-textbook ratios; *teacher management*, including frequency of observation of teachers by head teachers and of use of rewards for teacher performance; *teacher effort*, including teacher attendance and the share of teachers at a school who were found in the classroom on MLSS visit; assigned or marked homework during observed lessons; reported satisfaction with their job; or were above the median in terms of the amount of time they spent teaching on a typical day; *teaching practices*, including whether teachers asked students questions, employed teaching aids, and employed local materials to help deliver lessons; *remedial classes*, including the frequency of such classes and the share of grades in which they are offered; and *school cultures*, including the existence of clear rules on bullying and the incidence of bullying by teachers and students. For more detail on the construction of these indexes and the underlying variables, see section A.0.2.

2.8 Data

We draw on data from the Malawi Longitudinal School Survey (MLSS). The MLSS is an independent, nationally representative survey that provides data on conditions, practices and outcomes in Malawi's primary schools. The MLSS was designed to align with the MESIP project cycle to both provide information to the government on implementation and outcomes of the project and to support a number of impact evaluations of pilot interventions under the project, following rigorous experimental designs; as well as to inform general education policymaking. Visits are unannounced.[12] Data is collected through observations and interviews – for example, observations of school and classroom facilities; observation of lessons and teaching practices; interviews with head teachers, teachers, and members of community committees; and interviews and testing of Grade 4 students, as well as a longitudinal sample of students in later rounds.

The MLSS has its own randomly selected, nationally representative sample of 924 schools

[12] The first visit to each school is unannounced to help capture the real situation of the school in terms of infrastructure, school performance, school and classroom management practices, student and teacher absenteeism and student learning outcomes. If required to complete all instruments, a second visit is made on an announced or pre-scheduled basis.

which overlaps with, but does not encompass, the treatment and control samples for this intervention. MLSS baseline data collection was conducted in 733 of these schools across multiple rounds between May 2016 and December 2018 (Asim and Casley Gera, 2024).[13] All these rounds were conducted prior to the intervention. 659 schools underwent endline during September 2021 - February 2022. This endline data is used for measurement of outcomes of interest. For more details of the MLSS instruments and procedures, see section A.

2.9 Treatment Allocation and Sampling

2.9.1 Defining the population frame

To select school communities for treatment and control, we begin by drawing schools from the total population of public primary schools in Malawi.[14] We use administrative data, the 2017 Education Management Information System (EMIS) dataset (MoEST, 2017), which includes 5,552 public primary schools within 34 education districts.[15] Due to its pilot nature, the government requested the intervention to be implemented in schools with both lower and upper grades at primary level. Therefore, we exclude 'junior primary' schools which include only Grades 1-4. This leaves 5,405 schools remaining across all 34 districts.

2.9.2 Randomized allocation to treatment

A total of 1,198 schools were randomly assigned to treatment and 1,198 to control, across two cohorts of schools. In the first cohort, 800 schools were randomly assigned to treatment and 800 to control. In the second cohort, 398 were assigned to treatment and 398 to control. In both cohorts, the treatment group receives the intervention and the control group receives no intervention. There are no significant differences between the treatment and control groups and population (see Appendix B), and **Table 1** demonstrates there are no significant differences in characteristics between the pooled treatment and control groups.[16]

[13] Baseline could only be completed in 765 of the 924 schools owing to budgetary and logistical constraints. The authors do not find any significant differences for key school-level variables in the reduced 765 school sample using administrative school census data. Results available on request.

[14] This population excludes private schools, but includes schools owned by religious institutions which are accredited and financed by Government and considered part of the public school system.

[15] Malawi has 28 local government authorities, including districts and municipalities. Some of these are subdivided into two or three components for the purposes of education management, producing a total of 34 education districts. In this paper, we use 'district' to refer to these education districts.

[16] There are similarly no significant differences between the treatment and control groups within either cohort (see Appendix B)

2.9.3 Impact Evaluation Sample

Our impact evaluation sample is comprised of those intervention and control schools which are also part of the sample for the MLSS.

MLSS sample

The intervention and MLSS are run independent of each another. The MLSS sampling was conducted prior to the randomization using the same population frame.[17] The MLSS sampling uses probability proportional to size (PPS) to keep the sample representative not only at a national but at a sub-regional (division) level. Malawi has six education divisions across its three regions;[18] our strata are defined by division, with an additional stratum for all schools within urban locations. The number of schools selected within each stratum depends on the size of that stratum (section A.0.2). The final MLSS sample consists of 924 schools across all districts of Malawi and is representative at division level (see Appendix B).[19]

Overlap between treatment/control sample and MLSS

The scale of our intervention, targeting 1,198 schools across Malawi, enables us to exploit the overlap with the nationally representative MLSS. Of the 924 schools in the MLSS sample, 386 were among the 2396 randomly assigned to treatment or control, of which 202 are treatment schools and 184 are control schools.[20] These schools are used for the evaluation.[21] The full overlap of 386 schools was included in the endline data collection and is exploited for the ANCOVA specification.[22] This overlap is adequately powered to detect an impact on test scores of approximately 0.2 S.D. (see section 2.9.5). This impact evaluation sample is balanced between treatment and control on key variables (**Table 2**).[23]

[17] Like the intervention sampling, the MLSS sampling uses a population frame of all Malawi public primary schools. However, the MLSS sampling does not exclude junior primary schools.

[18] Divisions are sub-regional administration bodies established for management of education.

[19] MLSS baseline data (Asim and Casley Gera, 2024) shows that variability in inputs and outcomes is substantial at sub-district level (the lowest unit) and shows limited differences between three regions and six divisions (overall).

[20] Of these, 258 are from the first cohort, of which 133 are treatment schools and 125 are control schools; and 128 are from the second cohort, of which 69 are treatment schools and 59 are control schools.

[21] Owing to an error in implementation planning, the initial batch of training included 51 schools which were not part of the treatment group. Eight of these additional schools are in the control group. However, none of these are in the MLSS overlap and therefore not in the IE sample.

[22] Owing to financial constraints, 337 of these were visited for the baseline round. As we employ a longitudinal sample of students, only this subset of schools is used for analyses including student test scores. There are no significant differences between this subset of schools and the rest of the impact evaluation sample (see Appendix B).

[23] The impact evaluation sample is also balanced within each cohort (see Appendix B).

2.9.4 Student and teacher sampling

A number of outcomes of interest, including test scores, are drawn from a sub-sample of students and teachers that are interviewed and tested as part of MLSS. For teachers, the MLSS samples 10 teachers per school at baseline, all of which are re-surveyed at endline if eligible, with additional teachers added to replace those which become ineligible for interview and testing (for example, having left teaching). See section A.0.2 for details.

For students, the study uses a longitudinal sample. At baseline, a gender-balanced random sample of 25 Grade 4 students was sampled in each school. At endline, a subset of 15 students were randomly selected for re-survey at endline.[24] Students who had transferred to new schools were tracked to the new schools, and students who had dropped out were tracked to their homes where possible. The longitudinal sample enables us to measure the impact of the intervention on individual students' learning trajectories, measured through endline test scores. The median student from the longitudinal sample was in Grade 7 at endline.[25]

2.9.5 Power Calculations

We follow convention in the social sciences for power calculations, using a significance level (probability of Type I error) of 0.05 and power (probability of avoiding a Type II error) of 0.8.

We are interested in calculating the endline sample size necessary to detect an effect of $\sim$0.2 S.D. in student learning, our main outcome of interest, in each of the two specifications.

Our randomization takes place at the school level. We use the mean score for the baseline control group of 500, S.D. of 115, cluster size of 15 students per school. The intra-cluster correlation is 0.28. The required sample size to achieve this effect is 206 (103 in the treatment group and 103 in the control group).

For the identification of the effects of the intervention, our MLSS overlap exceeds the necessary sample size for both treatment and control. We are reasonably powered for subgroup analysis.

[24] A reserve sample of two additional students from the baseline sample was used to replace students who were ineligible for the interview and testing (for example, who had died) or were untraceable.

[25] At endline, in addition to the tracking and re-testing of the longitudinal sample, a new gender-balanced random sample of 15 Grade 4 students was sampled. Findings for this Grade 4 sample are presented in Appendix B.

2.10 Study Timeline

The study was conducted over six years. Baseline data collection took place in two rounds between April 2016 and October 2018. Development and pre-piloting of the intervention was conducted between October 2018-March 2019, with the intervention delivered in two rounds between April 2019 and September 2020 along with midline data collection. Endline data collection was conducted in October 2021 - February 2022.

2.11 Compliance

Noncompliance is feasible as a result of non-attendance by participating head teachers and/or deputy head teachers in one or more aspect of the training.[26] For the first cohort of schools, full records of attendance are available to the impact evaluation team for review and identification. We find that 743 schools attended fully (meaning that both Head Teacher and deputy Head Teacher attended both the main and refresher training). Employing a looser standard, we find that 798 of 800 schools in the treatment group participated at least substantially (meaning that at least one of the Head Teacher and deputy Head Teacher attended both the main and refresher training). In the second cohort, 385 of 398 schools in the treatment group attended at least partially. These high rates of attendance reflect a number of factors in the design and implementation of the intervention, including the context of a large-scale investment project, MESIP, which ensured a significant level of inputs in terms of staffing and management; provision of allowances for travel, accommodation and food, ensuring no cost to participants; and intensive effort by the implementing team to ensure attendance. Balance checks confirm there are no significant differences between the participating group and the entire sample for either the treatment/control or IE samples, in either cohort (see Appendix B).

2.12 Attrition

The potential sources of attrition at endline are students transferring to other schools, or dropping out; and teachers transferring to other schools, retiring, dying, or leaving teaching. To minimize attrition, students and teachers who have transferred are tracked to their new school to complete interview and learning assessment; students who have dropped out are tracked to their homes. In total, we reached 85 percent of the longitudinal sample of students, and 80 percent of teachers. There are no systematic differences in attrition for either students or teachers between treatment and control group (**Table 3**).

[26] As the evaluation is conducted at the school level and PEAs operate at subdistrict level, we do not consider PEAs in the compliance analysis.

3 Results

Tables 4 to 13 present ITT findings using ANCOVA. All student-level findings use Equation 1, while all school-level findings use Equation 2.

3.1 Test Scores

Student level results on test scores, are reported in **Table 4**. We include controls for students' baseline scores. We observe significant impacts from the treatment. Students in (or tracked from) treated schools achieved scores an average 13 points higher in mathematics scores, equivalent to 0.1 S.D. English scores also appear to have been increased by the treatment, but the result is not statistically significant. A statistically significant impact of 10.4 points in total scores is observed, but we cannot reject the hypothesis that this is driven by the impact on mathematics scores. We present findings for the same sample of students, without controlling for baseline scores, in Appendix B; we observe similar impacts with a slightly stronger treatment effect.

3.2 Student Progression

Table 5 presents the school-level results on dropout and repetition rates. We find evidence that the intervention significantly reduced repetition rates in lower primary by 2.2 points compared to control, approximately a nine percent reduction. Upper primary repetition rates also appear to have been reduced by the intervention but the result is not statistically significant. We do not observe significant impacts on dropout rates.

3.3 Heterogeneity

Malawi's schools operate in a diverse range of conditions, with wide variation in remoteness, staffing, and learning outcomes, while Malawi's students also possess a wide range of different characteristics with significant impacts on learning (Asim and Casley Gera, 2024). In order to investigate the differential impact of the intervention on schools in different contexts, we conduct heterogeneity analysis. Tables 6 to 9 show student-level estimates on test scores interacted with dummies for the relevant characteristics. As an additional check, we conduct sub-group analysis presenting ITT findings for particular sub-groups of the school and/or student sample (see Appendix B).

PTR. **Table 6** shows estimates on test scores interacted with a school-level dummy for whether a schools' PTR in lower primary was below the median at baseline. We observe significantly different impacts from the treatment for students in these schools, with English scores in

treated schools 23 points above similar control schools. We conclude that the treatment was most effective in schools with smaller PTRs.[27]

Gender. **Table 7** shows estimates on test scores interacted with a student-level dummy for female students. We do not observe significantly different impacts from the treatment for female students.

Age. **Table 8** shows estimates on test scores interacted with a student-level dummy for students who were below the median age for the sample at baseline. We do not observe significantly different impacts from the treatment for these students.[28]

Score at baseline. **Table 9** shows estimates on test scores interacted with a student-level dummy for students whose scores were below the median at baseline. We observe increased treatment effects for these students, with mathematics scores increasing by 17 points as a result of the intervention. This suggests that the intervention supported students with lower levels of learning to catch up to their peers as they progressed through upper grades.[29]

3.4 Robustness Checks

3.4.1 Alternative explanations for learning impact.

In order to confirm that the observed impacts on test scores in the longitudinal sample are real, we test a number of potential alternative explanations.

First, we explore the possibility that the treatment reduced dropout among the longitudinal sample. As dropped-out students are expected to learn less than those still enrolled, a reduction in dropout could potentially explain the effects of the treatment on learning in the longitudinal sample. However, as noted above, we do not observe impacts from the treatment on school-wide dropout rates (**Table 5**).

[27] We observe similar results in our sub-group analysis (Appendix B); the treatment impact on total scores was 22 points in schools with low PTR at baseline, versus 11 points in the entire longitudinal sample.

[28] In the sub-group analysis (see Appendix B), we do observe lower impacts from the treatment for younger students which do not attain statistical significance, which may suggest that the impacts of the treatment were smaller for younger students in the longitudinal sample but that our heterogeneity analysis is underpowered to detect this effect.

[29] We observe a similar effect in the sub-group analysis (Appendix B) in which stronger impacts are observed on mathematics scores (21 points) and significant impacts on English scores (15 points) for students with below median scores at baseline. This suggests that the overall effects on the intervention on the longitudinal sample were driven in large part by gains among those students whose performance was low at baseline.

As a further check, we test for differences in the share of students from the longitudinal sample who were found to have dropped out at endline (**Table 3**). We find no significant difference in this share as a result of the treatment. We then explore the possibility of selection bias: if the rate of attrition was higher in the control group than in the treatment group, or vice versa, this could lead to the observed impacts if higher-performing students were more or less likely to remain in the sample. However, we find no difference in attrition between the treatment and control groups.

3.4.2 Spillovers

Spillovers from the treatment to control group are feasible as a result of informal communication and collaboration between head teachers, deputy head teachers, and PEAs, and learnings by PEAs applied to other non-trained schools in their zones. To address this, we conduct subgroup analysis excluding control schools within the same zone as treated schools. We find no substantial differences from the main ITT estimates (see Appendix B).

3.5 Mechanisms

Table 10 presents school-level results on our exploratory intermediate indicators, grouped into indices. We observe significant impacts on the index combining measures of the provision of remedial classes. We do not observe significant impacts on other mechanisms from the treatment.

To disaggregate the impacts on remedial classes, **Table 11** presents school-level results for the individual indicators that make up this index. We find significant impacts from the treatment on the average frequency of remedial classes, the share of grades in which they are offered, and the share of grades in which they are offered for free. We find the impacts on the share of grades in which remedial classes are offered to be focused on the upper grades.

Additional tables in Appendix B present the breakdown of the mechanism indicators for the other indices. We observe indicative significant impacts from the treatment on pupil-textbook ratios in lower primary and on the likelihood that head teachers reward teachers for strong performance.

4 Discussion

The results suggest that an intervention targeted to improve leadership skills in a low-income context can improve students' test scores and progression in a short time period. We find

significant impacts on mathematics test scores from the intervention, and significant impacts on repetition rates in lower primary. The coefficient on mathematics scores, 13 points on our 500-centred IRT scale, is equivalent to 0.1 S.D., which compares favorably with the average impact of 0.04 S.D. observed in high- and middle-income countries in Anand et al. (2023). This impact is equivalent to approximately eight weeks' learning.

In terms of the mechanisms of impact, the most significant observed impacts on intermediate indicators relate to remedial classes, specifically the incidence, frequency, and share of grades in which remedial classes are offered. This is commensurate with the finding from our heterogeneity analysis, that the benefits to learning accrued to those students in the longitudinal sample which had learning scores below the median at baseline, who would be the most likely to benefit from remedial classes.

Recalling our theoretical framework, we conclude that the intervention helped improve schools' education production functions by increasing the amount of learning produced with the existing level of inputs. More specifically, the intervention helped schools move towards the education production frontier by maximizing the effort made by teachers to support low-performing students. However, we do not observe impacts on teaching practices or school cultures, suggesting that the intervention did not enable schools to extend their education production frontier outward.

4.1 Cost Effectiveness

To estimate the cost effectiveness of the intervention, we consider the marginal cost of delivery of the main training package following initial design activities, including printing of materials, allowances, and staff costs; and the average number of students in upper primary in treated schools who are exposed to the potential benefits. We estimate the per-student cost of the intervention at US$5.23. This can be considered a conservative estimate as it does not reflect the cost of the follow-up visits or refresher training. As noted above, the learning impacts observed from the intervention on the longitudinal sample of students are equivalent to approximately 0.1 S.D. Using the J-Pal recommended approach to cost-effectiveness analysis (Bhula et al., 2022), this equates to a cost of US$53 per S.D. of additional learning, or 1.88 S.D. per US$100 invested. This compares favorably with other recent experimental trials in low-income countries including conditional cash transfers in Malawi; provision of textbooks and contract teachers in Kenya; and remedial education in India (ibid.).[30]

[30] Using the entire cost of the intervention, rather than the marginal cost, the estimated impact falls to 0.97 S.D. per US$100, which remains competitive with a number of similar recent experimental interventions.

4.2 Sustainability

The intervention benefited from a significant amount of national and international expertise being brought to bear at the design stage, through the partnership between the government, Universities of Oxford and Malawi, and AIDSI. This public-private partnership enabled the development of concise, innovative content which was customized to the Malawi context while embodying international best practice. Crucially, while the entire partnership contributed to the design of the intervention, implementation was led by the government and SEUM with AIDSI providing support to logistics and administration. Prior to the completion of the intervention, a handover process was conducted to enable the continued implementation of the training by MoE in partnership with SEUM.

In 2022, a Memorandum of Understanding was signed between the local partners, MoE and SEUM, to enable SEUM to continue to support the implementation of the intervention. Following the completion of the evaluation of the pilot and the positive impacts observed, the intervention is now being scaled up to all school leaders who have not yet received the training, with financial support from the Global Partnership for Education and World Bank as part of the successor project to MESIP, the Malawi Education Reform Program (MERP). It is expected that the continuing participation of SEUM will help to ensure that the impacts are maintained at the scale-up stage.

5 Conclusion

We test the effect of training school leaders in 1,198 schools to improve use of school resources, motivate teachers and create an inclusive school culture for students to attain their maximum learning potential. We find that exposure to the intervention produced significant improvements in mathematics test scores for a longitudinal sample of students, equivalent to approximately 0.1 S.D. or eight weeks' learning. We also find that the treatment significantly reduced repetition in lower grades.

The findings suggest that improvements in school leadership can be achieved in a low-income context, with Government implementation and at scale, through a multi-phase intervention combining classroom training with follow-up visits and refresher training; and that these improvements can ultimately lead to improvements in student test scores and progression.

Acknowledgments

We thank the Ministry of Education, in particular Honorable Minister Madalitso Kambauwa Wirima, Secretary of Education Dr Mangani Katundu, and Director of Basic Education Grace Milner; former Directors of Education Planning Rodwell Mzonde and Francis Zhuwao, and former Directors of Basic Education Joseph Chimombo and Gossam Mafuta, for support in data collection; Muna Salih Meky, Safaa El-Kogali, Hugh Riddell, Greg Toulmin, Inam Ul Haq, Sajitha Bashir, Marina Bassi, and other World Bank colleagues in Malawi and elsewhere for encouragement, support and advice; Martin Moreno, Chris Burningham, Xu Wang, Miranda Lambert, Radhika Kapoor, Stefan Nippes, Gift Banda, Linda Matumbi, Leah Mziya, and Archit Singal, for invaluable research assistance; Vigdis Aaslund Cristofoli and Elin Ruud, Royal Norwegian Embassy, and Arianna Zanolini and Sabina Morley, UK Foreign, Commonwealth and Development Office, for comments and suggestions, and funding for the preparation of this paper. The findings, interpretations, and conclusions expressed herein are our own and do not necessarily represent the views of the World Bank, its Board of Directors, or any of its member countries. All remaining errors are our own.

Administrative information

Ethics approval

The Malawi Longitudinal Schools Survey has been subject to ethics review and clearance by the National Commission of Science and Technology for Malawi. The Principal Investigator, Salman Asim, has completed the training "Protecting Human Research Participants" provided by the National Institutes of Health (Certification Number: 921497; Date of completion: 05/16/2012)

Funding

Principal financing for data collection was provided by the Royal Norwegian Embasssy, Malawi, and the United Kingdom Foreign, Commonwealth and Development Office (FCDO). The interventions were financed by the Global Partnership of Education through the MESIP program and the Royal Norwegian Embassy through MESIP-Extended.

Tables

Table 1: Balance: Impact Evaluation Sample: Treatment vs Control

		(1) Control		(2) Treatment	T-test Difference
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)
Number of Toilets	1197	14.335	1194	14.091	0.244
		(0.284)		(0.262)	
Pupil Classroom Ratio	1188	108.807	1188	108.229	0.578
		(1.710)		(1.683)	
Pupil Teacher Ratio	1196	80.684	1193	79.708	0.976
		(1.037)		(1.122)	
Enrollment Grades 1-4	1197	608.030	1194	613.517	−5.487
		(12.275)		(13.909)	
Enrollment Grades 5-8	1197	295.145	1194	307.861	−12.716
		(8.354)		(9.957)	
Repetition Rate	1185	0.259	1181	0.253	0.006
		(0.003)		(0.003)	
Dropout Rate	1185	0.048	1181	0.044	0.004
		(0.002)		(0.002)	

Notes: Indicators extracted from EMIS 2017. The value displayed for t-tests are the differences in the means across the groups. Robust standard errors are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 2: Balance: Impact Evaluation Sample: Treatment vs Control

		(1) Control		(2) Treatment	T-test Difference
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)
Number of Toilets	184	14.973	202	13.267	1.705∗
		(0.734)		(0.530)	
Pupil Classroom Ratio	182	107.964	201	105.360	2.604
		(4.526)		(4.364)	
Pupil Teacher Ratio	184	75.928	201	77.714	−1.786
		(1.984)		(2.343)	
Enrollment Grades 1-4	184	618.212	202	577.119	41.093
		(31.354)		(27.025)	
Enrollment Grades 5-8	184	316.685	202	285.287	31.398
		(21.046)		(17.675)	
Repetition Rate	184	0.258	200	0.248	0.011
		(0.008)		(0.008)	
Dropout Rate	184	0.048	200	0.049	−0.001
		(0.005)		(0.005)	

Notes: Indicators extracted from EMIS 2017. The value displayed for t-tests are the differences in the means across the groups. Robust standard errors are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 3: Differential Dropout and Attrition

	(1) % Students Dropped Out b/se	(2) % Students Non-Attritted b/se	(3) % Teachers Non-Attritted b/se
Treatment	−2.841	0.176	−0.943
	(2.14)	(2.00)	(2.11)
Constant	34.264 * **	84.833 * **	80.866 * **
	(1.52)	(1.45)	(1.53)
Observations	337	337	337
R-sqr	0.005	0.000	0.001

Data Source: MLSS 2021.

Shares of longitudinal sample. Analytical sample restricted to MLSS schools observed in baseline and endline.

Robust standard errors are reported in the second row (in parentheses).

***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 4: Impact of the Intervention on Test Scores

	(1) Math Score	(2) English Score	(3) Total Score
Treatment	12.979 * *	8.566	10.443*
	(6.094)	(6.355)	(5.940)
Constant	339.098 * **	325.052 * **	305.597 * **
	(9.468)	(10.726)	(10.788)
Observations	4363	4363	4363
R-sqr	0.113	0.123	0.151

Notes: This table looks at the impact of the school leadership on Test Scores (columns 1-3) at the student level. The outcome variables are: Math Knowledge Score (column 1); English Knowledge Score (column 2); Total Knowledge Score (column 3). Analytical sample restricted to MLSS schools observed in baseline and endline. All regressions are adjusted by the value of the outcome variable observed at MLSS baseline. First row of the table shows point estimate coefficients and statistical significance. Robust standard errors (clustered at the school level) are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 5: Impact of the Intervention on Student Progression

	Repetition Rate			Dropout Rate		
	(1)	(2)	(3)	(4)	(5)	(6)
	Lower Primary	Upper Primary	Overall Primary	Lower Primary	Upper Primary	Overall Primary
Treatment	−2.249∗	−0.798	−1.773	0.308	0.181	0.087
	(1.339)	(1.393)	(1.170)	(0.826)	(0.823)	(0.744)
Constant	25.108 ∗ ∗∗	18.627 ∗ ∗∗	21.039 ∗ ∗∗	5.210 ∗ ∗∗	6.614 ∗ ∗∗	5.922 ∗ ∗∗
	(1.494)	(1.466)	(1.370)	(0.628)	(0.739)	(0.648)
Observations	385	380	385	385	380	385
R-sqr	0.041	0.051	0.072	0.018	0.028	0.024

Notes: This table looks at the impact of the school leadership on Repetition Rate (columns 1-3) and Dropout Rate (columns 4-6) at the school level. All regressions are adjusted by the value of the outcome variable observed at MLSS baseline. First row of the table shows point estimate coefficients and statistical significance. Robust standard errors are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 6: Impact of the Intervention on Test Scores (Heterogeneity by PTR in lower primary at baseline)

	(1)	(2)	(3)
	Math Score	English Score	Total Score
Treatment	7.739	−0.813	3.585
	(9.012)	(9.236)	(8.757)
Below Median PTR in Lower Primary	11.592	4.262	7.970
	(8.737)	(9.288)	(8.566)
Heterogeneous Treatment Impact	12.759	23.384∗	17.700
	(12.955)	(13.850)	(12.788)
Constant	489.581 ∗ ∗∗	493.805 ∗ ∗∗	491.809 ∗ ∗∗
	(6.152)	(6.375)	(6.014)
Observations	4346	4346	4346
R-sqr	0.008	0.008	0.009

Notes: This table looks at the heterogeneous impact of the school leadership on student level Test scores (columns 1-3) with respect to school-level Lower Primary (Grades 1-4) Pupil-Teacher Ratio (PTR) at baseline. The heterogeneous treatment impact is the estimated coefficient for the value of the treatment effect interacted with the variable of interest. Analytical sample restricted to MLSS schools observed in baseline and endline. All regressions are adjusted by the value of the outcome variable observed at MLSS baseline. The outcome variables are: Math Knowledge Score (column 1); English Knowledge Score (column 2); Total Knowledge Score (column 3). All regressions include clustered (at school level) standard errors. Robust standard errors (clustered at the school level) are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 7: Impact of the Intervention on Test Scores (Heterogeneity by gender)

	(1) Math Score	(2) English Score	(3) Total Score
Treatment	14.361∗	10.796	12.670∗
	(7.620)	(8.244)	(7.520)
Female	−9.922	1.668	−3.772
	(6.093)	(6.010)	(5.606)
Heterogeneous Treatment Impact	−0.900	−1.166	−1.316
	(8.322)	(8.304)	(7.719)
Constant	500.340 ∗ ∗∗	495.023 ∗ ∗∗	497.635 ∗ ∗∗
	(5.129)	(5.556)	(5.023)
Observations	4363	4363	4363
R-sqr	0.004	0.002	0.003

Notes: This table looks at the heterogeneous impact of the school leadership on student level Test scores (columns 1-3) with respect to student gender. The heterogeneous treatment impact is the estimated coefficient for the value of the treatment effect interacted with the variable of interest. The outcome variables are: Math Knowledge Score (column 1); English Knowledge Score (column 2); Total Knowledge Score (column 3). Analytical sample restricted to MLSS schools observed in baseline and endline. All regressions are adjusted by the value of the outcome variable observed at MLSS baseline. Robust standard errors (clustered at the school level) are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 8: Impact of the Intervention on Test Scores (Heterogeneity by student age at baseline)

	(1) Math Score	(2) English Score	(3) Total Score
Treatment	13.355 ∗ ∗	9.174	11.226∗
	(6.652)	(6.860)	(6.396)
Below Median BL Student age	64.247 ∗ ∗∗	72.024 ∗ ∗∗	67.681 ∗ ∗∗
	(6.300)	(7.138)	(6.123)
Heterogeneous Treatment Impact	−1.390	0.911	−0.343
	(8.830)	(10.129)	(8.743)
Constant	477.676 ∗ ∗∗	476.505 ∗ ∗∗	477.379 ∗ ∗∗
	(4.319)	(4.644)	(4.229)
Observations	4183	4183	4183
R-sqr	0.055	0.070	0.073

Notes: This table looks at the heterogeneous impact of the school leadership on student level Test scores (columns 1-3) with respect to student age at baseline. The heterogeneous treatment impact is the estimated coefficient for the value of the treatment effect interacted with the variable of interest. The outcome variables are: Math Knowledge Score (column 1); English Knowledge Score (column 2); Total Knowledge Score (column 3). Analytical sample restricted to MLSS schools observed in baseline and endline. All regressions are adjusted by the value of the outcome variable observed at MLSS baseline. Robust standard errors (clustered at the school level) are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 9: Impact of the Intervention on Test Scores (Heterogeneity by total score at baseline)

	(1) Math Score	(2) English Score	(3) Total Score
Treatment	4.242	4.126	4.064
	(7.645)	(7.954)	(7.219)
Below Median BL Total Knowledge Score	−82.778 * **	−82.831 * **	−82.374 * **
	(5.827)	(5.884)	(5.223)
Heterogeneous Treatment Impact	16.327*	10.494	13.547
	(8.938)	(8.988)	(8.258)
Constant	539.553 * **	540.607 * **	540.001 * **
	(4.912)	(5.170)	(4.585)
Observations	4203	4203	4203
R-sqr	0.085	0.091	0.103

Notes: This table looks at the heterogeneous impact of the school leadership on student level Test scores (columns 1-3) with respect to total score at baseline. The heterogeneous treatment impact is the estimated coefficient for the value of the treatment effect interacted with the variable of interest. The outcome variables are: Math Knowledge Score (column 1); English Knowledge Score (column 2); Total Knowledge Score (column 3). Analytical sample restricted to MLSS schools observed in baseline and endline. All regressions are adjusted by the value of the outcome variable observed at MLSS baseline. Robust standard errors (clustered at the school level) are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 10: Impact of the Intervention on several mechanism indices

	(1) Resource utilization	(2) Teacher management	(3) Teacher effort	(4) Teacher practices	(5) Remedial classes	(6) School culture
Treatment	0.088	0.148	0.031	−0.003	0.193∗	−0.116
	(0.094)	(0.101)	(0.103)	(0.102)	(0.102)	(0.102)
Constant	−0.046	−0.099	−0.009	−0.057	−0.085	0.007
	(0.074)	(0.080)	(0.082)	(0.076)	(0.082)	(0.113)
Observations	386	386	386	386	386	386
R-sqr	0.144	0.039	0.007	0.024	0.022	0.004

Notes: This table looks at the impact of the school leadership on summary indices of intermediate outcomes at school level. Indices constructed using the generalized least-square method proposed by Anderson (2018: Journal of the American Statistical Association 103: 1481–1495). All regressions are adjusted by the value of the outcome variable observed at MLSS baseline. Significant findings are robust to multiple hypothesis testing based on FWER Romano-Wolf p-values from 2000 bootstrapped replications. Robust standard errors are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.
Resource utilization includes indicators relating to lower primary PTRs, the ratio of PTRs between lower and upper primary, and lower primary pupil-textbook ratios.
Teacher management, includes frequency of observation of teachers by Head Teachers and of use of rewards for teacher performance.
Teacher effort includes teacher attendance and presence in the classroom, assigning and marking of homework, teacher job satisfaction, and time on task.
Teaching practices includes whether teachers asked students questions, employed teaching aids or local materials to help deliver lessons, or displayed instructional materials in the classroom.
Remedial classes includes the frequency of remedial classes and share of grades in which they are offered.
School cultures includes measures of the incidence of physical and verbal bullying by teachers and students.
See Appendix A for full details of intermediate indicators.

Table 11: Impact of the Intervention on the components of the Remedial Classes mechanism

	(1) School offers remedial classes	(2) Remedial Classes Average Frequency	(3) Remedial Classes (Share of Grades)	(4) Free Remedial Classes (Share of Grades)	(5) Lower Primary Remedial Classes (Share of Grades)	(6) Upper Primary Remedial Classes (Share of Grades)
Treatment	0.042	0.655 ∗ ∗	5.325 ∗ ∗	5.220 ∗ ∗	3.884	6.602∗
	(0.026)	(0.328)	(2.550)	(2.519)	(2.970)	(3.678)
Constant	0.900 ∗ ∗∗	4.518 ∗ ∗∗	54.397 ∗ ∗∗	53.614 ∗ ∗∗	87.990 ∗ ∗∗	15.364 ∗ ∗∗
	(0.046)	(0.241)	(2.003)	(1.989)	(2.412)	(2.638)
Observations	386	386	386	386	386	386
R-sqr	0.009	0.022	0.015	0.015	0.006	0.015

Notes: This table looks at the impact of the school leadership on prevalence of remedial classes at the school level. All regressions are adjusted by the value of the outcome variable observed at MLSS baseline. Significant findings are robust to multiple hypothesis testing based on FWER Romano-Wolf p-values from 2000 bootstrapped replications. Robust standard errors are reported in the second row (in parentheses). ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

References

Akerlof, G. and R. Kranton (2005). "Identity and the economics of organizations". Journal of Economic Perspectives 19(1), 9–32.

Akerlof, G. and R. Kranton (2008). "Identity, supervision, and work groups". American Economic Review 98(2), 212–217.

Akerlof, G. and D. Snower (2016). "Bread and bullets". Journal of Economic Behavior & Organization 126, 58–71.

Akerlof, G. A. and R. E. Kranton (2002). "Identity and schooling: Some lessons for the economics of education". Journal of economic literature 40(4), 1167–1201.

Akerlof, R. (2016). "'We thinking' and its consequences". American Economic Review 106(5), 415–419.

Anand, G., A. Atluri, L. Crawfurd, T. Pugatch, and K. Sheth (2023). "Improving School Management in Low and Middle Income Countries: A Systematic Review". Technical report, IZA Institute of Labor Economics.

Angrist, J. D. and V. Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". The Quarterly Journal of Economics 114(2), 533–75.

Asim, S. and R. Casley Gera (2024). "What Matters for Learning in Malawi? Evidence from the Malawi Longitudinal School Survey". World Bank. Forthcoming.

Asokhia, M. (2009). "Improvisation/Teaching Aids: Aid to Effective Teaching of English Language". International Journal of Educational Sciences 1(2), 79–85.

Bashir, S., M. Lockheed, E. Ninan, and J.-P. Tan (2018). Facing Forward: Schooling for Learning in Africa. World Bank.

Batson, C. D. (2009). "Two forms of perspective taking: Imagining how another feels and imagining how you would feel". In K. Markman, W. Klein, and J. Suhr (Eds.), Handbook of imagination and mental simulation, pp. 267–279. Psychology Press.

Bernard, T., S. Dercon, K. Orkin, and A. S. Taffesse (2019). "Parental Aspirations for Children's Education: Is There a "Girl Effect"? Experimental Evidence from Rural Ethiopia". AEA Papers and Proceedings 109, 127–132.

Bhula, R., M. Mahoney, and K. Murphy (2022). "Conducting Cost-effectiveness Analysis". https://www.povertyactionlab.org/resource/conducting-cost-effectiveness-analysis-cea. Accessed 2022-07-21.

Blimpo, M. P., D. K. Evans, and N. Lahire (2015). "Parental human capital and effective school management: evidence from The Gambia".

Bloom, N., R. Lemos, R. Sadun, and J. Van Reenen (2015). "Does management matter in schools?". The Economic Journal 125(584), 647–674.

Bowles, S. (1970). "Towards an Educational Production Function". In W. L. Hansen (Ed.), Education, Income, and Human Capital, pp. 11 – 70. NBER.

Branch, G. F., E. A. Hanushek, and S. G. Rivkin (2012). "Estimating the effect of leaders on public sector productivity: The case of school principals". Technical report, National Bureau of Economic Research.

Benabou, R. and J. Tirole (2006). "Incentives and prosocial behavior". American Economic Review 96(5), 1652–1678.

Chen, Y. and S. Li (2009). "Group identity and social preferences". American Economic Review 99(1), 431–457.

Chetty, R., J. Friedman, and J. Rockoff (2014). "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood". American Economic Review 104(9), 2633–2679.

Commission for Quality Education for All, T. (2016). Building Quality Education: A Pact with the Future of Latin America.

Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan, and V. Sundararaman (2013). "School Inputs, Household Substitution, and Test Scores". American Economic Journal: Applied Economics 5(2), 29–57.

Day, C. and P. Sammons (2013). Successful leadership: A review of the international literature. ERIC.

Day, C., P. Sammons, D. Hopkins, A. Harris, K. Leithwood, Q. Gu, E. Brown, E. Ahtaridou, and A. Kington (2009). "The impact of school leadership on pupil outcomes". Department for Children, Schools and Families Research Report DCSF-RR108.

de Hoyos, R., A. J. Ganimian, and P. A. Holland (2020). "Great things come to those who wait: Experimental evidence on performance-management tools and training in public schools in Argentina". Working Paper.

Delprato, M. and G. Antequera (2021). "School efficiency in low and middle income countries: An analysis based on PISA for development learning survey". International Journal of Educational Development 80.

Devries, K. M., L. Knight, J. C. Child, A. Mirembe, J. Nakuti, R. Jones, J. Sturgess, E. Allen, N. Kyegombe, J. Parkes, E. Walakira, D. Elbourne, C. Watts, and D. Naker (2015). "The Good School Toolkit for reducing physical violence from school staff to primary school students: a cluster-randomised controlled trial in Uganda". The Lancet Global Health 3(7), 378–386.

Eren, O. and D. Henderson (2004). "The impact of homework on student achievement". The Econometrics Journal 11(2), 326–348.

for Economic Co-operation, O. and Development (2005). School factors related to quality and equity: Results from PISA 2000. Organisation for Economic Co-operation and Development.

Fryer Jr, R. G. (2017). "Management and student achievement: Evidence from a randomized field experiment". National Bureau of Economic Research Working Paper 23437.

Fullan, M. (2007). The new meaning of educational change. Routledge.

Government of Malawi (2022). 2022-2023 Budget Statement.

Gregory, B., K. Nathan Moates, , and S. Gregory (2011). "An exploration of perspective taking as an antecedent of transformational leadership behavior". Leadership & Organization Development Journal 32(8), 807–816.

Grissom, J., D. Kalogrides, and S. Loeb (2017). "Strategic Staffing? How Performance Pressures Affect the Distribution of Teachers Within Schools and Resulting Student Achievement". American Educational Research Journal 54(6), 1079–1116.

Grissom, J., S. Loeb, and B. Master (2013). "Effective Instructional Time Use for School Leaders: Longitudinal Evidence from Observations of Principals". Educational Researcher 42(8), 433–44.

Grissom, J. A., A. J. Egalite, and C. A. Lindsay (2021). How Principals Affect Students and Schools: A Systematic Synthesis of Two Decades of Research. University of North Carolina at Chapel Hill.

Hallinger, P. (2005). "instructional leadership and the school principal: A passing fancy that refuses to fade away". Leadership and policy in schools 4(3), 221–239.

Hanushek, E. (1987). "Assessing the Effects of School Resources on Student Performance: An Update". Educational Evaluation and Policy Analysis 19(2), 141 – 164.

Hanushek, E. (2020). "Education Production Functions". In S. Bradley and G. C. (Eds.), The Economics of Education (Second Edition), pp. 161 –170. Academic Press.

Ingersoll, R. M., P. Dougherty, and P. Sirinides. School leadership counts. University of Pennsylvania.

Jimenez, J. and D. Gonzalez (2011). "Are You On The Educational Production Frontier? Some Economic Insights On Efficiency From PISA". In M. Pereyra, H. Kotthoff, and R. Cowen (Eds.), PISA Under Examination , pp. 169–182. Sense Publishers.

Kranton, R. E. (2016). "Identity economics 2016: Where do social distinctions and norms come from?". American Economic Review 106(5), 405–409.

Lassibille, G. (2016). "Improving the management style of school principals: results froma randomized trial". Education Economics 24(2), 121–141.

Leaver, C., R. Lemos, and D. Scur (2019). "Measuring and Explaining Management in Schools: New Approaches Using Public Data". World Bank Policy Research Paper 9053.

Leithwood, K., K. S. Louis, S. Anderson, and K. Wahlstrom (2004). "How Leadership Influences Student Learning. Review of Research.". Technical report, Wallace Foundation.

Leithwood, K. and J. Sun (2012). "The nature and effects of transformational school leadership: A meta-analytic review of unpublished research". Educational Administration Quarterly 48(3), 387–423.

Majgaard, K. and A. Mingat (2012). Education in sub-Saharan Africa: A comparative analysis. World Bank.

Mbiti, I., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani (2019). "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania". The Quarterly Journal of Economics 143(3), 1627–1673.

McKenzie, D. (2012). "Beyond baseline and follow-up: The case for more T in experiments". Journal of development Economics 99(2), 210–221.

Milligan, L. O., L. Tikly, T. Williams, J.-M. Vianney, and A. Uworwabayeho (2017). "Textbook availability and use in Rwandan basic education: A mixed-methods study". International Journal of Educational Development 54, 1–7.

Ministry of Education, Science and Technology of Malawi (MoEST) (2018). Malawi Education Statistics 2017/18.

MLSS (2018). Malawi Longitudinal Survey 2018 Baseline. Unpublished data.

MoEST (2015). Education Management Information System 2014-15 [database].

MoEST (2017). Education Management Information System 2016-17 [database].

MoEST (2020). National Education Sector Investment Plan 2020 – 2030.

Mulkeen, A. (2010). Teachers in Anglophone Africa. World Bank.

Muralidharan, K. and A. Singh (2020). "Improving Public Sector Management at Scale? Experimental Evidence on School Governance in India". National Bureau of Economic Research Working Paper 28129.

Ravishankar, V., S. E.-T. El-Kogali, D. Sankar, N. Tanaka, and N. Rakoto-Tiana (2016). Primary Education in Malawi: Expenditures, Service Delivery, and Outcomes. The World Bank.

Romero, M., J. Bedoya, M. Yanez-Pagans, M. Silveyra, and R. de Hoyos (2022). "Direct vs indirect management training: Experimental evidence from schools in Mexico". Journal of Development Economics 154.

Smith, W. (2021). "Consequences of school closure on access to education: Lessons from the 2013–2016 Ebola pandemic". International Review of Education 67, 53–78.

Tavares, P. A. (2015). "The impact of school management practices on educational performance: Evidence from public schools in São Paulo". Economics of Education Review 48, 1–15.

Verstegen, D. A. and R. A. King (1998). "The Relationship Between School Spending and Student Achievement: A Review and Analysis of 35 Years of Production Function Research". Journal of Education Finance 24(2), 243–62.

World Bank (2015). Project Appraisal Document for a Malawi Education Sector Improvement Project.

A Appendix

Malawi Longitudinal School Survey

The Malawi Longitudinal School Survey (MLSS) collects extensive data on school, classrooms, teachers, students, community members and parents. This Appendix provides summary details. For additional details, see Appendix B.

A.0.1 Instruments

The survey contains the following instruments:

1. Observation of school and classroom facilities
2. Lesson observation
3. Head Teacher interview, including details of teachers and committee members; information about their background and procedures; and school information from records
4. Student interview
5. Student learning assessment
6. Teacher interview
7. Teacher knowledge assessment
8. Community interviews with members of the School Management Committee, Parent-Teacher Association Executive Committee, and Mother Group
9. Group Village Headman interview

All instruments were included in all rounds, except the Group Village Headman interview, which was not included in the 2016 phase of baseline. [31]

The MLSS instruments are based on similar tools used as part of the Service Delivery Indicators (SDI) survey implemented by the World Bank. The SDI instruments were adapted with additional indicators which were appropriate for the Malawian context and/or specific to the MESIP program and related impact evaluations.

[31] A Group Village Headman is an intermediary-level official in Malawi's Traditional Authority structure, broadly analogous to a Village Chief

Learning assessments:: MLSS includes learning assessments in English and mathematics. These subjects not only allow for capturing of students' literacy and numeracy skills but also allow to test on wide range of cognitive skills. The curriculum of these subjects is relatively standardized across schools. The assessments are targeted to Grade 4 students but contain items aligned with the curricula for Grades 1-6. The test was designed by a psychometrician in collaboration with experts who had prior experience in designing similar tests and were familiar with the primary school syllabus of Malawi. Teachers in government schools were also consulted during the design and formulation of the test items so as to keep the structure of questions as close as possible to textbooks. Items were developed in reference to international tests and adapted for Malawian context. Student percentage scores are converted to knowledge scores using Item Response Theory, according to the distribution of correct answers across rounds, using a mean-centered scale centered at 500 with an S.D. of 100.

A.0.2 Sampling

The sampling frame for the MLSS was derived from the most up-to-date list of schools available prior to baseline, the 2015 EMIS (MoEST, 2015). [32]

Twenty percent of urban schools that were mainly concentrated in the four major cities of the country (Blantyre, Lilongwe, Mzuzu and Zomba Urban districts) were randomly selected. For rural schools, a stratified probability proportional to size (PPS) sampling was used, with strata defined based on the six educational divisions. From each stratum, a random sample of schools was selected using PPS, using the number of schools in each stratum as measure of size.[33] At the next stage, for districts that had few schools selected using the first round of PPS sampling, random oversampling was conducted to increase the final number of schools in each district to about 24. This oversampling allows district specific analysis. The urban and the rural samples were then combined to form the final survey sample of 924 schools.

As the MLSS is a longitudinal study, it employs both a cohort and longitudinal sample for students. A gender-balanced random sample of 25 Grade 4 students per school (13 girls and 12 boys) is selected at baseline. We then select 15 of these at random (8 girls and 7 boys) for resurvey at endline. Students who have dropped out or transferred to other schools are traced to the new schools or their homes and complete learning assessment and a modified version

[32] The original sample frame contained 5,738 schools with identifier variables such as division, district and zone. Of these, 323 private schools were removed from the sample frame, which leaves the frame with 5,415 primary schools subordinated to government or religious agencies.

[33] Number of schools is used as a measure of stratum size instead of enrollment as the EMIS enrollment data was found unreliable in many instances.

of the student interview. Two students are additionally selected as a reserve sample to replace students who have died, left Malawi, or cannot be traced. We expect to reach 90 percent of the selected students at endline. In addition, a new cohort of 15 Grade 4 students is surveyed at endline.

For teachers, a primarily longitudinal sample is used. Ten teachers per school are selected at baseline, using a protocol which ensures representation of lower and upper primary and of female and male teachers while maintaining random selection. All of the sampled teachers are resurveyed at endline if eligible. Teachers who have transferred to new schools are tracked and administered a modified version of the interview, while those who have died, left Malawi, left teaching, or cannot be traced are replaced with teachers who have more recently joined the school or were not selected at baseline. We expect to reach 90 percent of the originally selected teachers at endline.

Intermediate Indicators

Resource utilization: This Anderson index measures changes in the distribution and utilization of teachers and textbooks within a school. It includes: the ratio of PTRs between lower and upper primary; the ratio of female PTRs (female enrollment divided by female teachers) between lower and upper primary; lower primary PTRs; and lower primary pupil-textbook ratios.

Teacher management: This Anderson index measures participants' active management of teacher performance. It includes: frequency of observation of teachers by head teachers; the share of head teachers who report providing rewards for good teacher performance; and the share of teachers who report having been rewarded for good performance.

Teacher effort: This Anderson index measures the level of effort made by teachers in the school. It includes: teacher attendance; the share of teachers at a school who were found in the classroom on MLSS visit; the share of teachers at a school who assigned homework during observed lessons; the share of teachers at a school who marked homework; the share of teachers at a school who reported satisfaction with their job; and the share of teachers at a school who were above the national median in terms of the amount of time they spent teaching on a typical day.

Teaching practices: This Anderson index measures the extent to which teachers use practices associated with higher learning outcomes. It includes the share of observed teachers who: asked students questions; employed teaching aids; or employed local materials to help deliver lessons; and the share of observed classes where instructional materials were posted on the walls.

Remedial classes: This Anderson index measures the availability of remedial classes at the school. It includes: whether the school offers remedial classes; the frequency with which they are offered; the share of grades in which they are offered; the share of grades in which they are offered free of charge; and the share of grades in which they are offered in lower and upper primary specifically.

School culture: This Anderson index measures the extent to which schools have a culture conducive to learning. It includes: the share of students reporting having not recently experienced verbal bullying by other students; the share of students reporting having not recently experienced physical bullying by other students; and the share of students reporting having not recently experienced bullying by teachers.

Context of COVID-19

The COVID-19 pandemic, the associated closures of schools, and the resulting learning loss, are important pieces of context for the results. As described in section 2.4.1, the refresher training was subject to adjustments in class structure as a result of the COVID-19 pandemic which are not expected to affect the overall impact of the intervention. However, we do anticipate potential effects on the impact of the intervention from the pandemic as a result of the closure of schools. The Government of Malawi closed all schools and universities from late March to early October 2020 and again in January-February 2021. Following the reopening of schools, most underwent various adjustments to their normal operations, including the use of shifts with a reduced length of school day to reduce class sizes.

Learning loss as a result of COVID-19, and the associated closuree of schools, is expected to be substantial and exacerbate existing inequities. Increases in dropout are also likely as a result of the closure of schools (Smith, 2021). Our ANCOVA specification, and the fact that the pandemic and closure affected treatment and control schools equally, should mitigate the effects of these impacts on our ability to ascertain the overall impacts of the intervention. However, there is a chance that marginal gains in test scores or progression could be wiped out by large disruptions from the pandemic and school closures.

In order to explore this context, we estimate the extent of COVID-19-related learning loss for students in the longitudinal sample from schools in the treatment and control groups. We employ data from three rounds of the MLSS (2016-2018; 2018-19; and 2021) and exploit the

fact that the closure of schools took place between the second and endline rounds.[34] In total 2,832 students from our longitudinal survey had data for all three rounds. Excluding dropouts which occurred before the pandemic or were not related to it,[35] we estimate the pace of learning before and after COVID for 676 students and compare learning levels in 2021 with those we would expect if the pandemic had not occurred.

Table 12 shows results. Model 1 provides estimates of the overall loss in learning as a result of COVID-19.[36] We find that, in the control group, students' learning was 86 points below where we would project if the pandemic had not taken place. This is equivalent to around 1.8 years of lost learning. However, in the treatment group, the learning loss was 5.7 points lower at 80 points, equivalent to around 45 days' additional learning. In terms of trajectories of learning, we estimate that between the various rounds of MLSS, students in control schools gained an average 12.8 points in learning for each 100 days of schooling while students in the treatment group gained an average 13.8 points in learning in the same period.

Model 2 enables us to estimate the trajectory of learning after the reopening of schools.[37] We find that, prior to the closure, the pace of learning was higher in treatment schools, with students achieving an average 13.9 points of learning per 100 days versus 12.9 for students in control schools. Following the closure of schools, the slowdown in learning was also lower in treatment schools, with students learning an average 0.8 points less per 100 days than prior to COVID-19 in treatment schools, versus 1.3 points less per 100 days in control schools.

This provides suggestive evidence that the treatment achieved its impacts both by improving the pace of learning prior to COVID-19, and by reducing the slowdown in learning following the reopening of schools.

[34] The MLSS was conducted in three rounds, with the first phase of the second round taking place before the intervention described in this paper and treated as part of the baseline for this paper. However, for this analysis of COVID-19-related learning loss, we employ the three rounds as distinct sources of data.

[35] We identify students who dropped out after the closure of schools whose characteristics are similar to those who dropped out prior to the closure. We use propensity score matching on a range of characteristics including age, gender, parental literacy, and home language. We consider the 300 such students to have dropped out for non-pandemic related reasons and exclude them from the analysis, along with those who dropped out prior to the closure.

[36] Model 1 measures the average impact of the COVID-19 shock on students' test scores after the pandemic, controlling for time spent between two exams for each student.

[37] Model 2 incorporates the interaction between the COVID-19 shock and time spent between two exams. That helps to divide the average impact of the COVID-19 shock from the previous model into two parts: the one-off shock associated with the closure of schools, and any change in learning pace that students experience following the reopening of schools.

Table 12: Total Scores vs School Exposure: Treatment with Interaction

	(1) Model 1 b/se	(2) Model 2 b/se
Covid	-86.026***	-74.409**
	(7.15)	(29.09)
Covid&Treatment	5.725	1.512
	(9.59)	(40.32)
Time	12.791***	12.909***
	(1.18)	(1.26)
Time&Treatment	1.039	0.991
	(1.57)	(1.67)
Time&Covid		-1.307
		(3.38)
Time&Covid&Treatment		0.469
		(4.71)
Observations	1352.000	1352.000
R-sqr	0.261	0.261

Data Source: MLSS 2016/18/19/20/21.

Scores are for longitudinal students and exclude 187 non-COVID-related dropouts.

Robust standard errors are reported in the second row (in parentheses).

***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Augmented Intervention

As noted above, we do not observe significant impacts from the intervention on any outcomes relating to school cultures. We hypothesize that the dosage of the treatment was too small to achieve changes in these areas. Evidence from Fryer Jr (2017), in which the benefits for learning from school leadership training were weakest for black students, female students and students from lower-income households, suggests that such training may struggle to help schools build inclusive cultures. In order to explore the reasons for the intervention's lack of impact on school cultures, we present findings from an augmented intervention which combined the standard intervention with an additional, one-day intervention focused on school cultures.

The augmented intervention was conceived following the initial phase of follow-up visits. Evidence from the follow-up visits suggested that school cultures were a key theme in which change was limited following the main training. In addition, evidence from qualitative research conducted during the design stage of the augmented intervention suggested that bullying and corporal punishment practices in schools, which were subject to limited focus in the standard intervention, are key factors that have a negative impact on students' enthusiasm for school attendance and learning.

The augmented intervention, developed by the authors in collaboration with a team at the University of Oxford; consisted of the standard intervention plus one additional day of intensive training, delivered at the end of the refresher training. This additional training was delivered through interactive group discussions and specially-developed videos and focused exclusively on identifying particular behaviors by school leaders as problematic and presenting alternative (positive) behaviors which build positive and inclusive school cultures, to address the problems of bullying and corporal punishment as well as training the school leadership teams how to communicate effectively to their staff, students, and parents. The design of the augmented intervention built on theoretical and empirical literature in economic and social psychology which emphasizes the role of identity (Akerlof and Kranton, 2005, 2008; Akerlof, 2016; Akerlof and Snower, 2016; Benabou and Tirole, 2006; Chen and Li, 2009; Kranton, 2016), role modelling (Bernard et al., 2019), and perspective-taking capacity in leadership (Batson, 2009; Gregory et al., 2011).

The augmented intervention was tested in the second cohort of 398 schools, who received training under MESIP-Extended (see section 2.4 and section 2.9). Of these, 385 were present at the refresher training at which the assignment to treatment for the augmented intervention was conducted. Assignment to treatment was completed randomly at batch level. 212 schools were randomly selected to receive the augmented intervention. These schools receive both

the standard intervention and the augmented intervention. The remaining 173 schools in the second cohort receive only the standard intervention. There are no significant differences in characteristics between the two groups (Appendix B).

In line with the evaluation of the standard intervention, findings for this additional evaluation are drawn from endline via ANCOVA. Since the augmented intervention was added during implementation, the MLSS overlap was not designed with the second cohort in mind. Therefore, an endline survey was conducted by telephone with all schools in the second cohort. The survey collects data on the beliefs, attitudes, and mindsets of the school leadership team (head teachers, deputy head teachers, and PEAs). These include their understanding of school cultures; factors that could affect learning outcomes; current leadership practices; understanding of challenges faced by students; and attitudes to learning, bullying and corporal punishment. See **Table 13** for details of the index indicators. For more details on the augmented intervention and data, see Appendix B.

A.0.3 Findings

Table 13 shows results. We report naive p-values and those corrected using Romano-Wolf multiple hypothesis testing. We do not observe significant impacts from the augmented intervention on any of the outcomes of interest, compared to the standard intervention. We conclude that, even with this additional dose, the overall dosage of treatment in these areas is not large enough to achieve changes in participant behaviors.

Table 13: Augmented Intervention

VARIABLES	(1) Performance	(2) Student Beliefs	(3) Growth Mindset	(4) Teacher Practices	(5) Leadership	(6) Challenges	(7) Corporal Punishment (Participant)	(8) Corporal Punishment (Student)	(9) Bullying Awareness
{Treatment Effect}	0.011	0.023	-0.004	-0.092	0.002	-0.038	-0.071	-0.043	0.111
Unadjusted p-value	0.832	0.550	0.966	0.016	0.954	0.431	0.052	0.377	0.154
Bonferroni	1.000	1.000	1.000	0.132	1.000	1.000	0.483	1.000	1.000
Holm's	1.000	1.000	0.959	0.132	1.000	1.000	0.429	1.000	1.000
List et al. (2019)	0.995	0.956	0.959	0.117	0.998	0.933	0.340	0.927	0.665
Romano-Wolf	1.000	0.960	1.000	0.120	1.000	0.952	0.379	0.952	0.665
Controls (Female, Age)	Yes	Yes	(F Yes	Yes	Yes	Yes	Yes	Yes	Yes

Results are robust to multiple hypothesis testing based on FWER Romano-Wolf p-values from 2000 bootstrapped replications.
Robust standard errors are reported in the second row (in parentheses).
***, **, and * indicate significance at the 1, 5, and 10 percent critical level.
All indicators are Anderson Indexes drawn from endline telephone survey conducted with participants.
Performance measures beliefs about the drivers of student achievement.
Student Beliefs measures views about the importance to student outcomes of students' own beliefs about their abilities.
Growth mindset measures the extent to which participants believe their school instils a growth mindset in students.
Teacher practices captures how often participants believe their school's teachers engage in practices that uplift and encourage students who are struggling.
Leadership captures how often participants feel their school engages in cooperative leadership activities like involving teachers and staff in decision making.
Challenges measures participants' awareness of the obstacles to learning in their school and efforts to address them.
Corporal Punishment (Participant) captures head teachers' perceptions of corporal punishment.
Corporal Punishment (Student) captures head teacher's perception of how their students feel about corporal punishment.
Bullying Awareness assesses participants' ability to judge if different situations are considered bullying.

B Supplementary Data

Supplementary data associated with this article can be found in the Online Annex at:

https://bit.ly/malawi-school-leadership-supplemental.