ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 24-940

The Prevalence and Policy Implications of Between-School Heterogeneity in Learning Outcomes: Evidence from Six Public Education Systems

Daniel Rodriguez-Segura
NewGlobe

Savannah Tierney
NewGlobe

While learning outcomes in low- and middle-income countries are generally at low levels, the degree to which students and schools more broadly within education systems lag behind grade-level proficiency can vary significantly. A substantial portion of existing literature advocates for aligning curricula closer to the proficiency level of the "median child" within each system. Yet, amidst considerable between-school heterogeneity in learning outcomes, choosing a single instructional level for the entire system may still leave behind those students in schools far from this level. Hence, establishing system-wide curriculum expectations in the presence of significant between-school heterogeneity poses a significant challenge for policymakers — especially as the issue of between-school heterogeneity has been relatively unexplored by researchers so far. This paper addresses the gap by leveraging a unique dataset on foundational literacy and numeracy outcomes, representative of six public educational systems encompassing over 900,000 enrolled children in South Asia and West Africa. With this dataset, we examine the current extent of between-school heterogeneity in learning outcomes, the potential predictors of this heterogeneity, and explore its potential implications for setting national curricula for different grade levels and subjects. Our findings reveal that between-school heterogeneity can indeed present both a severe pedagogical hindrance and challenges for policymakers, particularly in contexts with relatively higher levels of performance and in the higher grades. In response to meaningful between-system heterogeneity, we also demonstrate through simulation that a more nuanced, data-driven targeting of curricular expectations for different schools within a system could empower policymakers to effectively reach a broader spectrum of students through classroom instruction.

VERSION: April 2024

Suggested citation: Rodriguez-Segura, Daniel, and Savannah Tierney. (2024). The Prevalence and Policy Implications of Between-School Heterogeneity in Learning Outcomes: Evidence from Six Public Education Systems. (EdWorkingPaper: 24-940). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/3psa-e628

The Prevalence and Policy Implications of Between-School Heterogeneity in Learning Outcomes: Evidence from Six Public Education Systems

Daniel Rodriguez-Segura Savannah Tierney

March 2024

Abstract

While learning outcomes in low- and middle-income countries are generally at low levels, the degree to which students and schools more broadly within education systems lag behind grade-level proficiency can vary significantly. A substantial portion of existing literature advocates for aligning curricula closer to the proficiency level of the "median child" within each system. Yet, amidst considerable between-school heterogeneity in learning outcomes, choosing a single instructional level for the entire system may still leave behind those students in schools far from this level. Hence, establishing system-wide curriculum expectations in the presence of significant between-school heterogeneity poses a significant challenge for policymakers — especially as the issue of between-school heterogeneity has been relatively unexplored by researchers so far. This paper addresses the gap by leveraging a unique dataset on foundational literacy and numeracy outcomes, representative of six public educational systems encompassing over 900,000 enrolled children in South Asia and West Africa. With this dataset, we examine the current extent of between-school heterogeneity in learning outcomes, the potential predictors of this heterogeneity, and explore its potential implications for setting national curricula for different grade levels and subjects. Our findings reveal that between-school heterogeneity can indeed present both a severe pedagogical hindrance and challenges for policymakers, particularly in contexts with relatively higher levels of performance and in the higher grades. In response to meaningful between-system heterogeneity, we also demonstrate through simulation that a more nuanced, data-driven targeting of curricular expectations for different schools within a system could empower policymakers to effectively reach a broader spectrum of students through classroom instruction.

Key words: foundational learning, heterogeneity, targeted instruction, South Asia, West Africa, curricular expectations, variance decomposition

Authors' note: we would like to express our gratitude to Alejandro Ganimian, Lant Pritchett, Andreas de Barros, Steven Glazerman, Raissa Fabregas, Doug Johnson, Shannon May, Tim Sullivan, Keuna Cho, and Sean Geraghty for their insightful comments. Additionally, we extend our thanks to Anchal Khandelwal, Michael Kang, and Melanie Gaudet for their excellent research support. Geri Mezzoni, Kate Montgomery, and Sandeep Kumar provided outstanding field coordination for the data collection processes. We also acknowledge the invaluable contribution of the field team across all six contexts, without whom access to these datasets would not have been possible. Rodriguez-Segura and Tierney are both employees of NewGlobe, the organization responsible for directly collecting the data. However, the views expressed here are solely those of the authors and do not necessarily reflect the views of any other organization. Rodriguez-Segura, the corresponding author, can be reached at daniel.rodriguez@newglobe.education. All errors remain our own.

I. Introduction

The prevalence of low learning outcomes in low- and middle-income countries (LMICs) is one of the best-documented facts in education research of the past decades (Pritchett, 2013; World Bank, 2018). Large swaths of children within these educational systems do not meet grade-level expectations for some of the most foundational literacy and numeracy skills (Azevedo et al., 2021), which, in turn, hinders the ability of educational systems to deliver on the economic and social benefits of education (Hanushek & Woessmann, 2007; Montenegro & Patrinos, 2014). In this sense, the vast majority of children in education systems around the world are behind pedagogical expectations —although perhaps at different degrees of severity within and between each system— given what research has shown to be the most fruitful academic path towards strong human capital accumulation.

Among the challenges that LMICs face to achieve higher learning outcomes, there are two related issues that have been documented by researchers. First, one commonly discussed driver of the low learning outcomes in these contexts is the mismatch between the typically "overambitious curricula" that aim higher and move faster than the current level of most children in education systems (Pritchett & Beatty, 2015) and which tend to cater only to top performers in these contexts (Glewwe et al., 2009), and the typical learning outcomes of children in those systems. In fact, when a curriculum reform in Tanzania allowed the foundational literacy and numeracy curriculum to move at a slower pace to meet the needs of more of its children, learning outcomes around foundational skills increased by 0.2 standard deviations nationwide (Rodriguez-Segura & Mbiti, 2022). In this sense, the policy prescription that has generally been advocated for has been to bring the typically high curricular expectations, and in turn, classroom instruction, closer to the —quasi-mythical— performance of the "median child" in these contexts (Hwa et al., 2020; Rodriguez-Segura et al., 2021). As evidenced by the Tanzanian reform, this practice might allow classroom instruction to reach more students as, by definition, there are more students around the median level of a distribution than anywhere else — especially if the curriculum is misaligned and currently targeting the higher levels of the distribution.

The second well-documented challenge that previous research has highlighted repeatedly is that the increase in enrolment in recent decades has brought many children into classrooms who would have otherwise not enrolled in school, in turn widening the within-class disparities in these

systems (Ganimian & Djaker, 2023). Pedagogically, within-class heterogeneity can pose a challenge for teachers, as they have a more difficult task catering instruction to classes with a wider range of proficiency levels in them. In fact, an intervention in Kenya displayed significant learning gains in the order of 0.18 SD when classes were set up to be more homogenous in terms of baseline performance — allowing teachers to reach a larger share of their students only through classroom instruction (Duflo et al. 2011, Cummings, 2017). Similarly, some programs in the vein of "Teach at the Right Level" have aimed to reduce the dispersion of classes through ability grouping, while other work has instead focused on either standardizing instructional materials at scale, or targeting the lowest-performing students to bring them closer to the level of their peers, indirectly decreasing classroom dispersion (Ganimian & Djaker. 2023).

At a more macro-level, when policymakers aim to set curricular expectations for a given grade —typically in a one-size-fits-all fashion within a given jurisdiction— they face a similar challenge as teachers do at a micro level: they observe a systemwide distribution of performance at the school level and, ideally, choose the curricular level that serves the most children and schools. However, in the presence of significant school-level heterogeneity, it is likely that schools that deviate significantly from the curricular level chosen by central planners will not benefit from this instructional level, leaving them either following a curriculum that does not cater to the needs of their children or potentially innovating locally to meet the needs of their students with potentially heterogeneous impacts on student learning outcomes. Hence, together, the facts that the average child in a low- or middle-income country tends to be several grades behind grade-level expectations (e.g., Azevedo et al., 2021), and that there can be within-class heterogeneity in certain contexts that hinders classroom instruction are also still fully compatible with the claim that, in certain contexts, there might also be enough between-school heterogeneity in learning outcomes—even if most schools are, on average, "behind grade-level"— such that the central policy prescription for the curriculum might need to vary by school to cater to their local needs.

In this paper, we explore the extent of and potential pedagogical implications of *between-school* heterogeneity in foundational learning outcomes in public education systems in LMICs. To examine the extent to which between-school heterogeneity persists in public education systems in LMICs, we leverage six unique datasets representative of six public educational systems spanning over 900,000 children enrolled in government schools across two Indian states, a Pakistani

territory, two Nigerian states, and an entire West African nation. Students were assessed with the same foundational literacy and numeracy tools across contexts, enabling a more in-depth mapping of school-by-school learning outcomes to students' proficiency levels and to the potential pedagogical needs different of schools within the same educational system and grade.

We document five key facts. First, although students in even the highest-performing regions of our data generally fall below proficiency expectations, between-school differences in learning levels within the education systems of these six regions can show pedagogically meaningful degrees of heterogeneity, which, in turn, has implications for instruction and how to target a one-size-fits-all set of curricular expectations. Second, while we document that —as previous literature has suggested—raw within-class heterogeneity increases with each subsequent grade level, we also show that this is an incomplete narrative, as raw between-school heterogeneity and the dispersion in pedagogical needs of schools also increase with grade. Third, we find that, on average across all grades and territories, the share of the variance explained by between-school differences in literacy outcomes is 45%, while for numeracy, this figure is only 13% — highlighting potential implications for policy actions that target heterogeneity differentially across subjects, depending on the degree to which within- or between-school heterogeneity is most prevalent. Fourth, we find that the share of the variance explained by between-school differences tends to increase with baseline performance, although this tapers off somewhat at the higher end of the performance spectrum. Conversely, we find that in almost all cases, the share of the variance explained by between-school differences is larger than that explained by between-administrative regions or between urban and rural schools — highlighting that identifying pedagogical needs of schools may not be as simple as using easily-observable school characteristics (stereo)typically associated with performance. Finally, we show that, if the mandated curriculum could be flexibly adapted to the specific needs of each school, a more data-driven catering of the mandated curriculum could allow governments to reach more students within their system, particularly when between-school heterogeneity is high. Yet, these efforts could be dampened by logistical and methodological challenges such as misplacing of schools due to measurement error in the diagnosis of their needs.

Despite the potentially high policy relevance of better understanding and quantifying between-school heterogeneity in LMICs, this exercise has been —to the best of our knowledge— scant in the literature, particularly regarding foundational literacy and numeracy. We believe that there

might be at least two reasons why this is the case. First, while data that are representative of learning outcomes at a large scale in high-income contexts, and to a lesser extent, upper-middle-income countries, are available through either national assessments, which exist as administrative data, or through international assessments like PISA, this type of data has been significantly more scarce in LMICs — particularly for foundational literacy and numeracy outcomes in primary grades. Even the canonical papers that aim to explore the issue of within-class heterogeneity tend to use data that are not representative of larger regions (for instance, Muralidharan et al., 2021), or on selected samples where certain international agencies deployed interventions, which might in turn not be representative of the rest of the educational system due to site selection bias (like in Crouch et al., 2020; Rodriguez-Segura et al., 2021). Even some of the large-scale datasets on foundational skills in LMICs, like Uwezo in East Africa and ASER in South Asia—which are designed with some degree of statistical representativeness in mind—are not at the school-level but at the household-level. Importantly, these are also usually aimed at understanding whether children can reach a certain grade-level performance (e.g., Grade 2 skills). If used to quantify between-school heterogeneity, this might introduce ceiling effects into the analysis in terms of pedagogical prescriptions for higher grades beyond the targeted grade (e.g., beyond Grade 2). Secondly, while still challenging, interventions targeting within-classroom heterogeneity are likely more feasible to enact with the current policy and technological tools (e.g., through edtech solutions within classrooms/schools or ability grouping) than modifying and tailoring curricula differentially for a large number of schools within a system. Therefore, the current gap in the literature might respond to a scarcer set of potential policy solutions —or perhaps one that is currently just harder to implement— even if the issue of between-school heterogeneity were to be well-documented within a system.

In all, our paper makes two key contributions. First, for policymakers in low- and middle-income countries, it highlights a pedagogical challenge which, we believe, has received little policy and research attention, relative to the issue of within-class heterogeneity. Even when policymakers aim to cater to the "median child" within a system, the median child in certain schools might be grade levels ahead of the median child in other schools. Especially when a significant share of the variance in learning outcomes happens between schools, choosing a single level of instruction for national or state-wide curricula might still be a challenging policy decision that leaves out a large

share of children and schools. Second, for researchers, this paper contributes to the broader literature on learning inequality and heterogeneity that has been building in recent years. While our data agree with the literature on the prevalence of within-class heterogeneity, we expand on this literature by showing some of the first estimates of system-wide, between-school heterogeneity in foundational skills in low- and middle-income countries.

The rest of the paper proceeds as follows. Section II explains the data collection process, the instruments used, and the main statistical tool used to decompose the variance in learning outcomes. Section III presents the five key results of the paper, and Section IV discusses the implications of these findings for policymakers in low- and middle-income countries, and the potential policy alternatives (and their risks) that could be leveraged in the face of meaningful levels of between-school heterogeneity in foundational learning outcomes.

II. Data

1. Data collection and sample

For this study, we used representative samples from six different regions across LMICs in South Asia and West Africa. These regions included two states from Northeast India — Meghalaya and Mizoram, the Pakistani capital territory of Islamabad, the African nation of The Gambia, and two Nigerian states — Anambra State and Bayelsa State. We selected these regions primarily because they face a shared challenge of low learning outcomes, but also because they demonstrate considerable geographic and socioeconomic diversity in comparison to one another, enabling a deeper and more externally valid analysis of the extent of heterogeneity in these regions.

The six regions encompass a variety of economic and demographic landscapes. To illustrate, Meghalaya is considered one of the least economically developed states in India (Raghavan & Lodick, 2024; Reserve Bank of India; 2023), with a predominantly rural population —approximately 80%— and an agricultural sector engaging roughly 70% of the workforce (Government of India, 2011). Conversely, although the state of Mizoram shares a similar reliance on agriculture, with approximately 60% of its workforce engaged in the sector, the state currently has one of the fastest-growing economies in India, and a majority urban population (Government of Mizoram, 2016). In further contrast, Islamabad Capital Territory is the capital of the fifth most populous country in the world, and is characterized by rapid urbanization and a burgeoning

information and technology sector (Liu et al., 2020). With regard to the Nigerian states in the sample, Anambra and Bayelsa are both prominent contributors to the nation's oil and gas industry (Audu & Arikawei, 2013; Bello & Nwaeke, 2023), yet they face vastly different demographic profiles; Anambra is the second-most densely populated state in Nigeria, while Bayelsa is the least populated state in the country (National Bureau of Statistics, 2014). In addition to this, these six regions also display income levels that place them in a relatively wide range of average socioeconomic levels. For example, according to the World Bank income classification, The Gambia, with a GDP per capita of USD 808, would be considered a low-income country. In contrast, Bayelsa's GDP per capita of USD 4355 would place it on the same classification scheme as an upper-middle-income region, whereas everywhere else is somewhere in between, classified as a lower-middle-income region (Hamadeh et al., 2023).[1] Importantly, the diversity represented across these regions provides valuable insight into the unique educational challenges and opportunities faced within each context.

For the sampling approach in each of the six regions, the goal was to collect a sample of schools that, collectively, provided an accurate representation of the broader educational landscape in each region. To ensure a comprehensive representation of the state of education within each of the six regions, we utilized a proportional stratified random sampling method in the selection of schools. First, we obtained a list of schools from the respective local or national governments with a roster of all publicly funded primary schools in each region. These lists were then stratified by the next administrative level below each region (e.g., districts in Mizoram, sectors in Islamabad). Then, schools were randomly selected within each sub-region such that the number of schools sampled from a given sub-region was proportional to the total number of schools in that sub-regional area.[2] Within each assessed school and grade, a random sample of students, varying in size across the six regions, was drawn. Hence, this sampling approach ensures that the overall results are broadly representative of the larger educational landscape in each region.

[1] In 2022, the GDP per capita in Meghalaya was 1,360 USD ($4,410 PPP) and 2,619 USD ($8,492 PPP) in Mizoram, according to the Reserve Bank of India (2023). The GDP per capita in The Gambia in 2022 was 808 USD ($2,497 PPP), according to the World Bank (Hamadeh et al., 2023). In Islamabad, the GDP per capita in 2021 was 2,500 USD ($10,466 PPP), according to the Khyber Pakhtunkhwa Bureau of Statistics (Hasan et al., 2021). In 2021, the GDP per capita in Anambra State was 2,002 USD ($5,231 PPP), and 4,355 USD ($11,379 PPP) in Bayelsa State, according to BudgIT (Okeowo & Fatoba, 2022).

[2] The original intent of these studies was for the sampling approach to reflect the number of students in each sub-region, but we found that in the majority of instances, there was no reliable or comprehensive enrollment data.

The data collection process in all regions was conducted by individuals belonging to the same team within the organization — NewGlobe, as part of a broader, common initiative to better understand these educational systems. Therefore, the recruitment of enumerators, training materials, and data collection forms was standardized across all six regions, as the data collection team utilized the same materials. This standardization serves to mitigate potentially differential measurement errors from one place to another, which might have arisen from differences in the data collection process between locations.

In total, the sample comprises 7,413 students distributed across 276 schools and six grade levels (Grades 1-6). On average, 1,236 students were assessed from 46 schools per region, resulting in an average sample size of 27 students per school (refer to Table 1 below for more information regarding sampling within each region). Given the selection procedure described above, the findings in this study can be generalized beyond this sample to represent the entirety of these six public primary education systems, spanning over 900,000 students. In turn, the diversity of this sample, both regionally and within a broader global context, provides a unique opportunity for a more comprehensive examination of the heterogeneity in learning outcomes across a variety of cultural and educational landscapes in South Asia and West Africa.

Table 1: Sample description of each region

Region	Sample Size	Sample Characteristics	Sampling Approach	Representativeness	Date of Assessment
Anambra State (Nigeria)	1,592 students from Grades 1-6 across 44 schools	- 21 local government areas. [3] - Schools 82% rural. - Students 51% female.	Proportional stratified random sampling at the local government area-level, by number of schools	Representative of 353,155 public primary students	June 2023, end of last academic term in 2022-23 school year
Bayelsa State (Nigeria)	966 students from Grades 1-6 across 30 schools	- 8 local government areas. - Schools 73% rural. - Students 49% female.	Proportional stratified random sampling at the local government area-level, by number of schools	Representative of 84,856 public primary students	December 2022, end of first academic term of 2022-23 school year
The Gambia	1,346 students from Grades 1-6 across 40 schools	- 6 regions. - Schools 83% rural. - Students 49% female.	Proportional stratified random sampling at the regional-level, by number of schools	Representative of 276,074 public primary students	October 2023, during first academic term of 2023-24 school year
Islamabad (Pakistan)	1,145 students from Grades 1-5 across 40 schools	- 6 sectors. - Schools 63% rural. - Students 53% female.	Proportional stratified random sampling at the sector-level, by number of schools	Representative of 88,116 public primary students	March 2023, end of last academic term of 2022-23 school year
Meghalaya (India)	1,178 students from Grades 1-5 across 60 schools	- 11 districts. - Schools 92% rural. - Students 49% female.	Proportional stratified random sampling at the district-level, by number of schools	Representative of 86,466 public primary students	May 2023, end of first academic term of the 2023-24 school year
Mizoram (India)	1,186 students from Grades 1-4 across 62 schools	- 9 districts. - Schools 76% rural. - Students 50% female.	Proportional stratified random sampling at the district-level, by number of schools	Representative of 35,267 public primary students	September 2022, during last academic term of the 2022-23 school year

[3] While the original empirical plan was to visit schools from every sub-region, this was not possible in Anambra State due to issues regarding school accessibility. To account for this, the sample from this region was weighted based on the original sampling plan to ensure that the inclusion or exclusion of these schools in our analytical sample did not affect the conclusions of this study. Due to minimal differences in the results based on whether the sample is weighted or not, the unweighted results are used throughout.

2. Assessments and main outcomes

This analysis aims to better understand students' foundational literacy and numeracy capabilities, both in absolute terms and relative to their peers in other schools, grades, and regions. To do so, the same core set of foundational literacy and numeracy assessments was used in each region —although students in different regions sometimes took additional assessments, as required by constraints of a set of companion studies (see Appendix 3 for more information regarding the full set of assessments used). For the purposes of this research, we focus on two key outcomes stemming from this set of common assessments: oral reading fluency, assessed using correct words per minute (cwpm) of a set of Grade 2 passages, and numeracy, measured by the total score on the International Common Assessment of Numeracy (ICAN). These measures provide insight into how students in different grades between schools and systems perform on the same assessment, allowing us to reliably compare metrics of heterogeneity in each system. Below, we elaborate on these two key constructs used in the analysis.

i. Oral reading fluency

To assess students' oral reading fluency, this evaluation relies on Grade 2 English passages from Dynamic Indicators of Basic Early Literacy Skills (DIBELS). The full DIBELS assessment measures the acquisition of early literacy skills, including phonemic awareness, phonics, fluency, vocabulary and comprehension (Good & Kaminski, 2002), although we only used the oral reading fluency subsection. DIBELS serves as a valid assessment of early literacy development, as it is widely used in evaluation studies of educational interventions around the world (Bratsch-Hines et al., 2020; Good & Kaminski, 2002; Petscher & Kim, 2011). In each of the six contexts, students took different Grade 2 passages — although within each context, students of all grades were assessed with the same passage. To measure reading fluency, we use the common unit of "correct words per minute" (cwpm) for all passages and contexts, as this is the most commonly used unit measure of oral reading fluency, given its comparability across languages, assessments, and educational contexts. Therefore, using calibrated Grade 2 passages from DIBELS and a common unit of measurement across all regions enable the comparison of students' fluency levels within their system and across systems in the sample.

There are also some limitations in using oral reading fluency as typically measured in cwpm units as the key outcome to quantify literacy outcomes. First, cwpm units do not take all

components of oral reading into consideration. Reading fluency is comprised of three main components — accuracy, the ability to precisely decode words; automaticity, the ability to recognize and decode words effortlessly; and prosody, the ability to read a text with appropriate expression and intonation (Aldhanhani & Abu-Ayyash, 2020; Raskinski, 2004). Assessing fluency through cwpm does not account for one's automaticity or prosody capabilities, both of which are key determinants of a child's ability to read with accuracy (Valencia et al., 2010). Additionally, one might argue that reading comprehension, as opposed to reading fluency, is a more holistic measure of literacy; reading with comprehension is often considered the "ultimate goal of literacy", and thus, is a consistently used metric to measure literacy across various global assessments (Aldhanhani & Abu-Ayyash, 2020; Abadzi, 2011). While we do have data on reading comprehension for most of these contexts, reading comprehension is a complex skill which first requires the development of many prerequisite skills, including but not limited to decoding (Gough and Turner, 1986; Hoover and Gough, 1990). Given the low average learning outcomes within the sample, and the complexity of reading with comprehension, assessing literacy capabilities using reading comprehension may introduce floor effects, wherein the majority of students score a zero, which would hinder the ability to examine the extent of heterogeneity in learning outcomes — despite children at, for example, 0 cwpm and 35 cwpm requiring very different pedagogical approaches, even if they both score 0% on a reading comprehension assessment. In practice, we do observe that the reading comprehension scores in this sample experience serious floor effects, and as such, we choose oral reading fluency as the most suitable measure for the context of this study.

ii. Numeracy score

Students' numeracy skills are assessed using the ICAN assessment. ICAN, developed by the People's Action for Learning (PAL) Network, is a tool designed to measure performance across a range of core numeracy competencies, all of which are relevant for the age group in this evaluation. ICAN assesses numeracy skills across five domains: number recognition, addition, subtraction, multiplication, and division. Within each domain, there are two subtasks: a simple application of the concept, and a more challenging application of the concept. Two of the domains, subtraction and division, also include a separate word problem, which provides additional insight into the

extent to which students can apply their knowledge of arithmetic operations in real-world situations.

We chose the ICAN for three reasons. First, ICAN is context-agnostic, enabling meaningful comparisons to be made across a variety of contexts (PAL Network, 2020a). Second, compared to other assessments, ICAN incorporates a larger range of numeracy subskills for a wider range of ages, enhancing its validity in reporting on children's basic numeracy competencies (PAL Network, 2020b). Other global assessments, like the Programme for International Students Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMMS), target older students and assess more complex mathematical topics, while assessments such as the Early Grade Mathematics Assessment (EGMA) assess foundational numeracy in lower grades but are not designed for students older than Grade 3 (PAL Network, 2020a). Third, ICAN is not designed to be adaptive, per guidance from PAL Network (2020b). Adaptive assessments, such as India's Annual Status of Education Report (ASER) or Uwezo in East Africa, begin by assigning an assessment task tailored to each child's expected academic level. Subsequent tasks —and their level of difficulty— are then adjusted based on the child's performance on the prior task (ASER Centre, 2023). In contrast, ICAN's administration approach allows for progressive examination of students' performance across all subskills, i.e., not relying on an adaptive process that might vary from pupil to pupil. In fact, we administer the full assessment to each child, starting from the most foundational question, and asking all questions regardless of whether they could answer the simpler subtask first for each skill, which allows us to obtain the same item-level data for the full assessment for all pupils. This, in turn, ensures uniformity in the number of items from which an overall score is calculated. These factors facilitate a more comprehensive analysis of students' learning levels, enabling comparisons across a broader range of grades and international contexts.

As all students took all items in the assessment, we choose to analyze numeracy outcomes as students' total scores on this assessment (i.e., the ratio of correct questions over the total number of questions in the assessment), which allows us to gain a broader understanding of overall numeracy proficiency across all of the domains assessed, without being subject to potential volatility and differences in curricula if we were to choose a single skill (e.g. "simple addition") as

the main indicator.[4] Similarly, continuous variables display preferable empirical properties to estimate certain measures of "learning inequality" than dichotomous variables (Crouch et al., 2021), which is also an additional advantage of using the aggregate ICAN score as our main outcome of interest for numeracy.

3. Mapping of learning outcomes onto proficiency levels

The two outcomes previously discussed exist in the datasets as "raw" variables in units of "correct words per minute" for literacy and "percent correct" for numeracy. In any distribution of learning outcomes at scale, it would be expected to indeed find differences in terms of "between-school" variance of the raw units. Yet, for policymakers, the pressing question is whether once these learning levels are translated into grade-level proficiencies —and therefore, into potential instructional prescriptions— if the "raw variance" also translates into tangible variance in the pedagogical needs of children within the same grade across schools in the same system. Hence, below we discuss how we also translate the literacy and numeracy data in these datasets into a discrete measure of "grade-level proficiency" that allows us to pressure test whether any numeric variance recorded in the key raw outcomes also has concrete implications for the issue of setting curricular levels based on the range of grade-level proficiencies that may exist within a given grade.

i. Literacy

To further analyze literacy outcomes in this regard, we leverage two sources. First, we use a study conducted by Abadzi (2011), in which reading fluency measurements and outcomes from 17 LMICs are examined. Abadzi's study provides average reading fluency levels for each grade across these countries —as data were available for each grade— allowing us to use these averages as proficiency thresholds for the results of our study. In addition to Abadzi's study, we also draw from the Hasbrouck-Tindal oral reading fluency norms — a widely used benchmark developed from a few different assessments, including DIBELS, with data collected primarily in high-income, English-speaking countries. The Hasbrouck-Tindal norms provide a broader scope of analysis of students' learning levels compared to their peers in a high-income context.

[4] Nonetheless, the item-level data for numeracy were collected and are available for all contexts.

These two sets of thresholds have respective advantages and disadvantages for the purposes of this study. First, while the Abadzi thresholds are from comparable settings to the ones studied in this paper —at least in terms of average incomes and the level of development of their education system— they were not intended to be "norms", just averages of existing datasets. These averages do not appear to be farfetched as norms: for instance, it has been suggested that ideally students would learn how to read with comprehension by Grade 3 (World Bank, 2018). Using 45-60 cwpm as a likely proxy to reach this level of mastery (Abadzi, 2012), the Abadzi thresholds fall roughly in this range by Grade 3-4 (see Appendix 4). On the other hand, the Hasbrouck-Tindal thresholds were indeed intended to be norms, but were developed for much more mature educational systems where English is likely a more latent language in society than in each of the six contexts in this study, which means that the levels in the norms are significantly higher than the vast majority of children in English-speaking LMICs. For the purposes of the main text, we display results aligned with Abadzi's thresholds, but wherever relevant, we include the results mapped against the Hasbrouck-Tindal norms in Appendix 1 as a robustness check.

ii. Numeracy

To understand how the numeracy proficiency of students in this study compares to global grade-level expectations, we apply the Global Proficiency Framework (GPF) to map learning outcomes. The GPF is a context-agnostic compilation of numeracy proficiency descriptors developed by the UNESCO Institute for Statistics and myriad contributing organizations. Incorporated within it are the "Global Proficiency Descriptors" (GPD), which leverage mathematics performance data collated from 50 countries to establish a standardized definition of grade-appropriate numeracy skills. Mathematical competencies that may be demonstrated by students at a particular grade level, but exceed expectations for that grade level, are categorized as such, and underperformance is likewise attributed accordingly (UNESCO Institute for Statistics et al., 2023). Given the prominence of the GPF in understanding global numeracy standards, this study established a crosswalk between each skill assessed through the ICAN, and the grade in which children are expected to master that skill according to the GPF. We first carefully identify the mathematical benchmarks in the GPF that most closely correspond with assessment items, considering both the exact problem and its assessed skill. We then use the item-level ICAN scores to determine the grade level at which assessed students should be reaching these benchmarks by

referencing the grade level(s) described under the framework's "Meets Global Minimum Proficiency"[5] threshold (see Appendix 4 for the outcomes of the mapping approach).

III. Results

1. Although generally below proficiency, learning levels within these six education systems show meaningful degrees of school-level heterogeneity with implications for curriculum setting.

Children in the six regions of study exhibit generally low learning outcomes, similar to data for most other LMICs. For example, on average, over a third of Grade 3 students in the sample cannot read a single word from the Grade 2 passage or perform simple two-digit addition without carrying. Even in the two highest-performing regions in this sample in terms of fluency and numeracy outcomes, Islamabad and Anambra, a large share of children are not meeting grade-level expectations. For instance, only 1 in 6 children in this grade in Islamabad meet the Hasbrouck-Tindal median winter norm (97 cwpm), and in Anambra, only 10% of children meet this threshold.[6] Even the significantly lower Abadzi threshold (38 cwpm) is met by only 57% of Grade 3 students in Islamabad and 36% in Anambra. In the most dire case, Bayelsa, only 3% of Grade 3 students reach grade-level proficiency, even by this lower benchmark. Therefore, in absolute terms, the vast majority of children in these contexts, much like in the rest of the LMICs (Azevedo et al., 2021), do not possess the expected foundational literacy and numeracy skills for their grade level.

Despite the overall low learning outcomes, our results show that the severity of these gaps varies significantly by context and by school within each context. To illustrate this, we present the dispersion of literacy and numeracy outcomes for Grade 3 in a series of box plots by subject and context (Figure 1), where each box in a given graph represents a school in each representative sample. The data on numeracy and literacy outcomes were collected and analyzed in a way such

[5] By design, this threshold is formed from a lenient definition of the level of proficiency students need to demonstrate the skill. Therefore, if an ICAN skill is assessed by a problem that is marginally more advanced than the corresponding GPD on the GPF, it is still reasonable to state that students would achieve this skill by the grade level designated by the GPD. Since the GPD describes the minimum level of skill a student can demonstrate that is still considered sufficient, it is likely that a significant proportion of students at this grade level would have stronger proficiency.

[6] For half of the sample (three regions), data were collected in the first term of the school year, and for the other half of the sample data were collected in the last term of the school year. Thus, Hasbrouck-Tindal's median winter norm — the second benchmark out of three for a given grade — served as the most suitable average norm to ensure consistency in interpreting the results.

that they are comparable across all six regions; therefore, each graph allows for a visual examination of dispersion within each system individually (within a row), and relative to the other educational systems (across rows). In Appendix Figures 1-5, we include the equivalent figure for all other grades.

Figure 1 demonstrates that the extent to which the school-level median levels of performance differ within each system actually varies significantly. In Grade 3 literacy, Bayelsa, The Gambia, and Meghalaya show a "flat" profile, where most schools display median levels that are very similar to each other and with few individual outliers within each school beyond 50 cwpm. Conversely, Anambra and Islamabad have much "steeper" profiles, displaying significantly higher within-system and between-school heterogeneity. For example, in Islamabad, 18% of schools have a median level below 10 cwpm, and 43% of schools have a median level at least meeting the Abadzi threshold for Grade 3. Similarly, in Anambra, 27% of schools have a median level of a non-reader (0 cwpm), and 18% of schools have a median level meeting the Abadzi threshold.

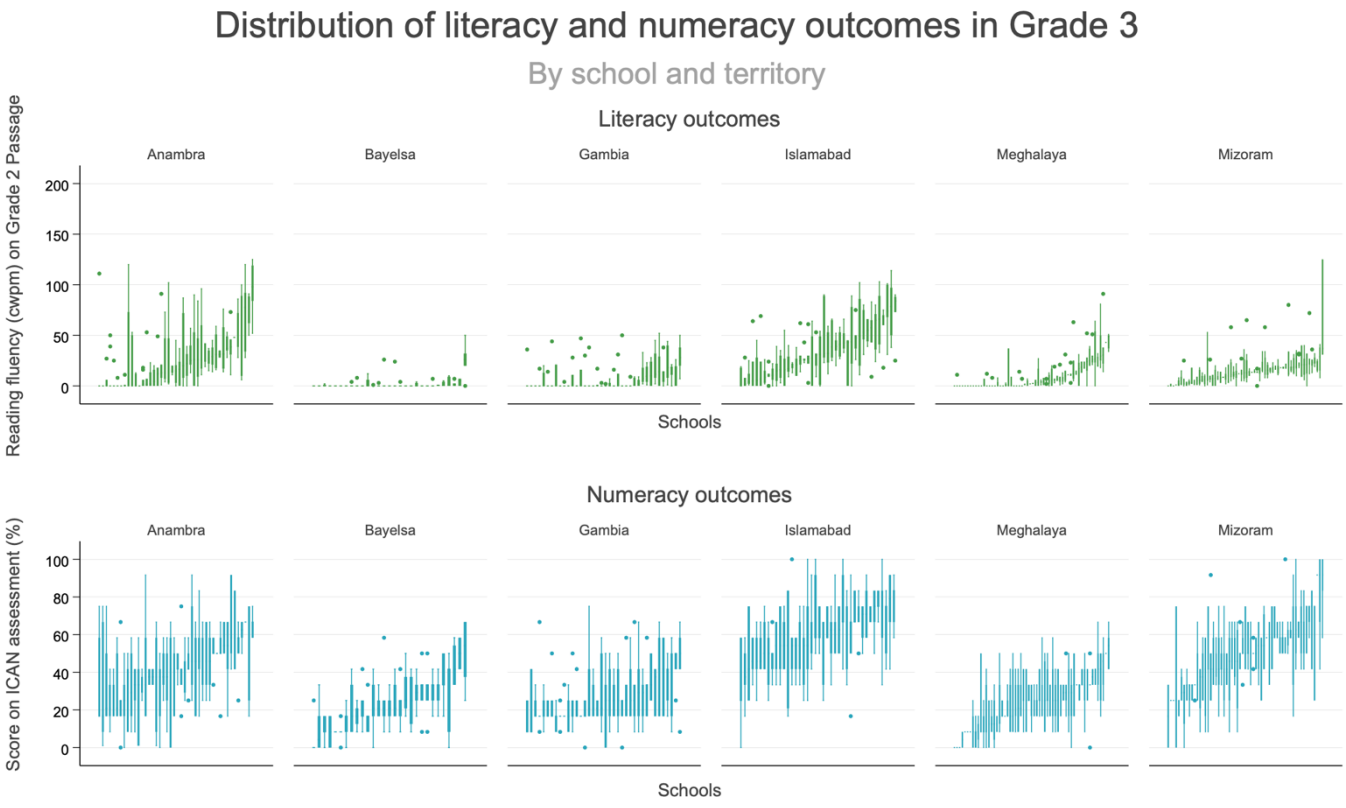


Figure 1

To more systematically explore the differences in school-level median levels within and across systems, by subject and grade, Figure 2 traces the school-by-school median of the two main outcomes of interest for all regions and grades. In other words, the lines for Grade 3, for instance, correspond to the trace of the median levels for each school on Figure 1. Intuitively, the flatter the line for a given region-subject-grade, the more similar the school-level median levels for a given place are. Conversely, the steeper the line for a given region-subject-grade, the more between-school heterogeneity that region has.

First, we find that —for the most part— the steepness of the lines within a subject and grade does seem to vary substantially by region. For instance, Grade 5 for literacy (dark blue) is significantly steeper in Islamabad than in the relatively flatter Bayelsa or Meghalaya. More specifically, each additional 10 percentiles in the school median performance in Islamabad is associated with 10.3 additional cwpm. In Meghalaya, the equivalent figures are only 7.5 additional cwpm, and in Bayelsa, this amounts to 3.3 additional cwpm. Part of these differences are driven by high-performing schools: the school with a median level at the 90^{th} percentile in Islamabad is performing at 122 cwpm, which is over four times the level of the respective school in Bayelsa and almost twice as much as in Meghalaya. Yet, differences at the top of the distribution are not fully driving the overall shape of the curve: while the school with a median level at the 10^{th} percentile is reading at 48 cwpm in Islamabad, the respective schools in Bayelsa and Meghalaya perform at 0 cwpm. In numeracy, the cross-regional differences are not as stark[7] — in fact, the average Grade 5 percentile-on-percentile increase in Islamabad is smaller than that in Meghalaya or Bayelsa — but within each region and grade, the slope is rarely flat, indicating some degree of between-system heterogeneity. For example, the ratio of the average numeracy score among the 25% of schools with the highest median numeracy scores to the equivalent score in the 25% of schools with the lowest numeracy scores is 3.7 in Grade 2 and 2.2 in Grade 4 and 5. Even in the region with the lowest value for this ratio —Islamabad— the average value is 1.9, that is, the top 25% of schools with the highest median scores are performing almost twice as high as the bottom 25%.

[7] Although we cannot rule out that part of this might be due to potential ceiling effects in Islamabad Grade 5 — the highest performing sub-group in the entire sample.

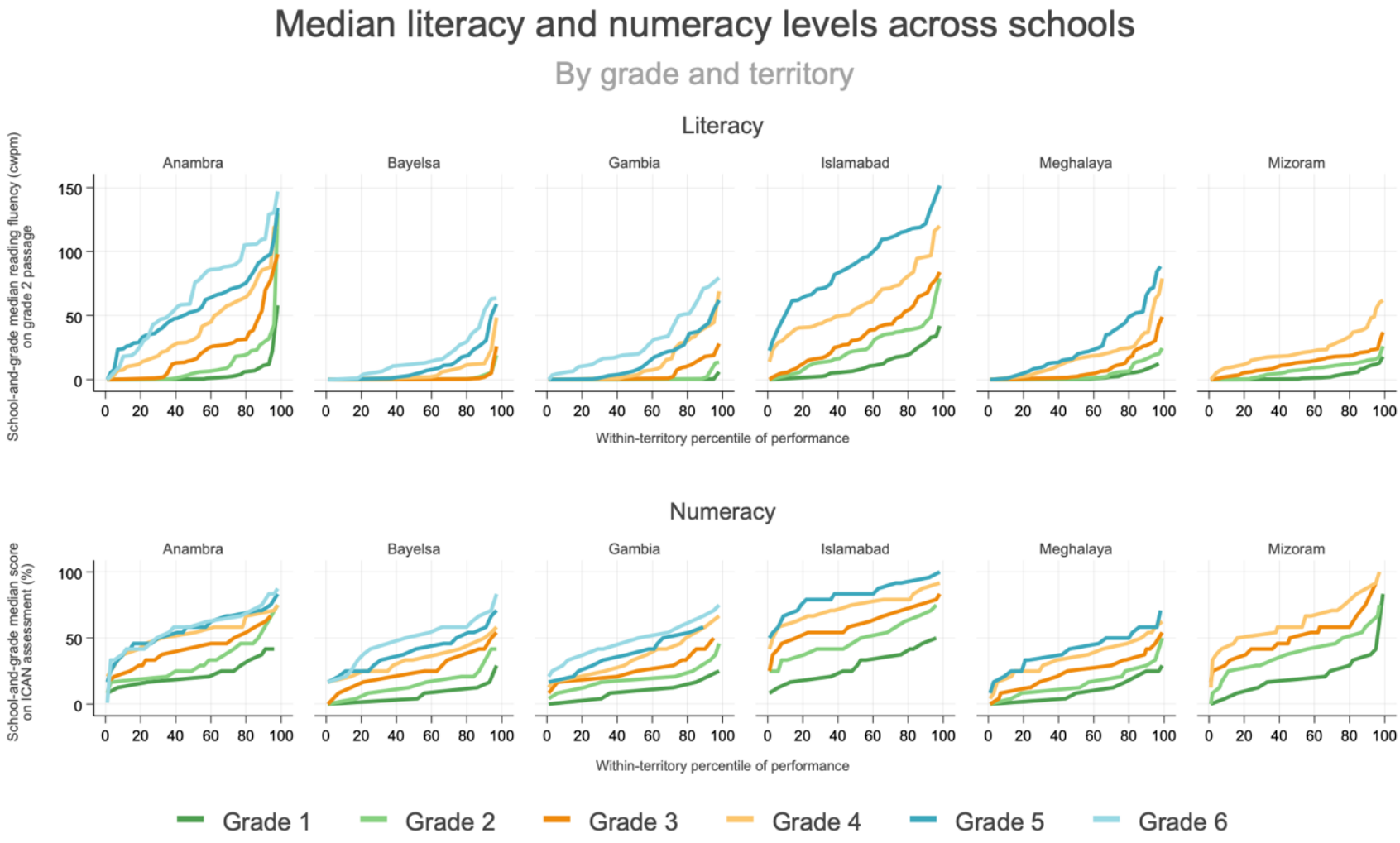

Median literacy and numeracy levels across schools
By grade and territory
Literacy
Anambra
Bayelsa
Gambia
Islamabad
Meghalaya
Mizoram
School-and-grade median reading fluency (cwpm) on grade 2 passage
0
50
100
150
0 20 40 60 80 100
Within-territory percentile of performance
Numeracy
School-and-grade median score on ICAN assessment (%)
Grade 1
Grade 2
Grade 3
Grade 4
Grade 5
Grade 6

Figure 2

Do the variations in learning outcomes between schools within these education systems call for differentiated instruction or curricula at the level of the school? In other words, are these differences "pedagogically meaningful," or are they simply "numeric differences" resulting from natural variation among students who all require the same instructional, likely remedial, level? If these differences do not truly signify distinct pedagogical needs for schools, then policymakers aiming to optimize classroom instruction within their systems might simply adjust the level of their one-size-fits-all curriculum for a given grade to better align with the learning levels of students across the system. However, if these numeric differences also indicate variations in the pedagogical needs of different schools, policymakers might be interested in exploring ways to tailor mandated instruction and curriculum more precisely to the diverse needs of schools, going beyond the current one-size-fits-all approach. To investigate this question, we utilize the mapping of grade-level proficiency discussed in the previous section to the learning outcomes presented in Figure 2. The results for Grade 3 are illustrated in Figure 3, and for the sake of simplicity, outcomes for other grades are provided in Appendix Figures 6-9.[8]

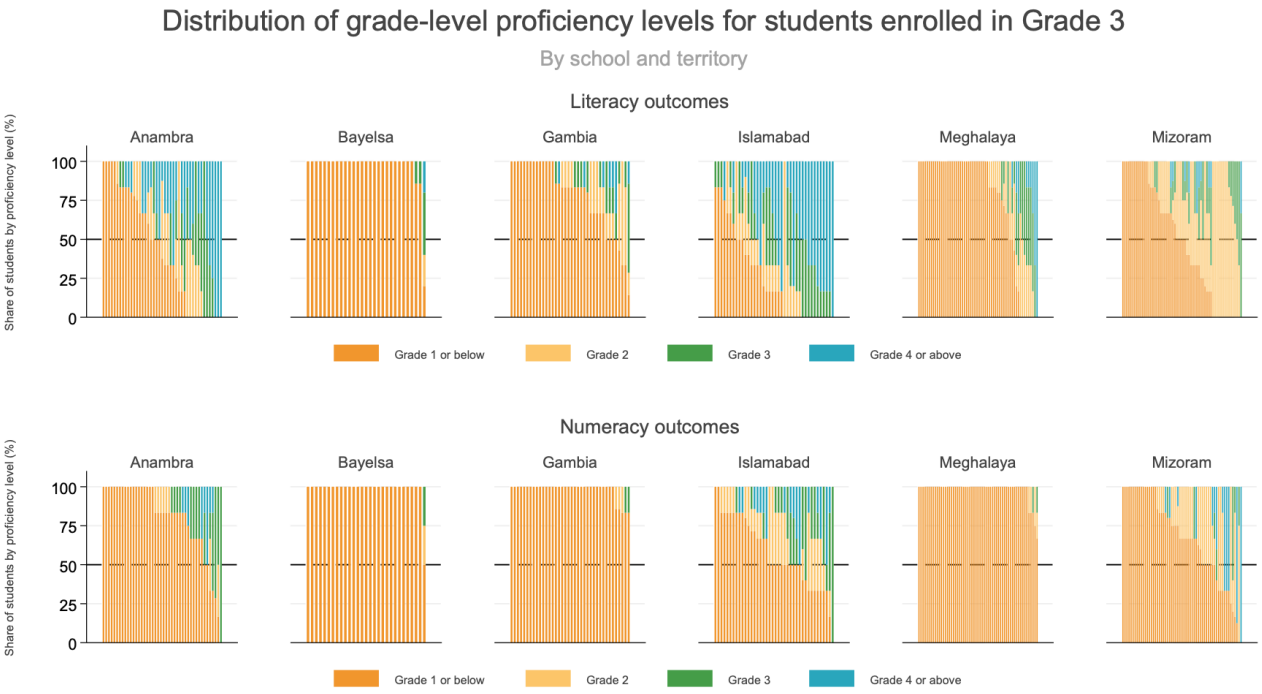


Figure 3

[8] Appendix Figures 10-11 also show the equivalent graphs for Grades 3-4 but using the median Hasbrouck-Tindal winter norms for literacy as opposed to the Abadzi thresholds.

Figure 3 illustrates that there are differences across systems in the extent to which variation in learning outcomes translates into differences in the pedagogical needs of schools. In other words, for certain systems, grades, and subjects, there are not large differences in the pedagogical needs of schools, while in other cases, there are ample differences between schools that merit a closer look into potential differentiation approaches. For instance, in Bayelsa, The Gambia, and Meghalaya, 97%, 79%, and 76% of Grade 3 classes respectively have a median literacy level aligned with Grade 1 proficiency. For numeracy in these same places, the figures are nearly universally aligned with Grade 1 proficiency. In other words, in these cases, between-school heterogeneity is not so large as to warrant exploration of between-school differentiation—especially if this type of potential intervention presents high logistical costs that require them to be offset by similarly large gains in the share of children receiving instruction at their level.

Yet, in other regions, the extent of the between-school heterogeneity in the median grade-level proficiency of each school is significant and might prompt policymakers to consider school-level differentiation of the mandated classroom instruction.[9] For instance, in literacy, 56% of schools in Anambra display a Grade 1 or 2 level, while 44% of them display a Grade 3 level or higher. In Islamabad, 18% of schools display a Grade 1 level, 18% display a Grade 2 level, and 43% of schools are at a Grade 4 level or above. In numeracy, there is a similar pattern, although more schools skew towards the lower levels of proficiency than in literacy. For example, in Islamabad, 45% of Grade 3 classes display a proficiency level akin to a Grade 1 level, while 55% of them are at a Grade 2 level or above. Even in Mizoram, which, in Figure 2, visually appears to be in the middle in terms of the extent of its between-school heterogeneity, 66% of schools are at a Grade 1 level, and all other schools are at a Grade 2 or above, with 8% of them even at Grade 4 or above. In these cases, policymakers might want to consider approaches to differentiate the mandated instruction between schools, as a one-size-fits-all curriculum will not address the needs of a large share of children, regardless of the level at which it is pitched.

[9] These meaningful differences in the pedagogical needs of children within systems and grades are also present when the data are analyzed at the pupil level within each region, grade, and subject, as shown in Appendix Figure 12. This also provides some evidence that the differences between schools displayed here are not solely driven by measurement error in the between-school differences. Yet, while Appendix Figure 12 shows that there are meaningful differences in the proficiency levels within the systems, it obscures whether these are distributed in a similar way across schools (i.e., most of the variation comes from within-school differences) or in different ways between schools (i.e., a large portion of the variation comes from significant differences in the overall learning levels of different schools).

These findings highlight two key takeaways for policymakers. First, the lack (or presence) of meaningful pedagogical needs within a given grade should not be taken for granted. For instance, assuming that there is a large degree of between-school heterogeneity in Grade 3 numeracy in Meghalaya might lead policymakers to solve a non-existent problem, as all schools have a median level equivalent to a Grade 1 level. In this instance, an overall curriculum realignment, if needed, might be enough to reach a larger share of students through classroom instruction. Instead, the lack of concrete data mapped onto literacy proficiency levels in Islamabad might obscure the large extent to which Grade 3 children in different schools have vastly different pedagogical needs. While 1 in 6 students need instruction aligned with levels appropriate for non-readers or close-to-non-readers, 23% of schools are meeting the Abadzi grade-level expectations, and 43% are even exceeding these thresholds. Therefore, high-quality data on learning outcomes are needed across a system to better understand the extent of the policy challenge that between-school heterogeneity might pose.

Second, these analyses highlight the importance of having clear mappings of learning outcomes to proficiency levels and pedagogical needs. Even in assuming that policymakers have access to reliable data on learning outcomes, the "raw" outcomes might not be enough to determine whether a system has meaningful differences in the pedagogical needs of schools, as it is also required that they are mapped onto curricular milestones that place children and schools at different proficiency levels.

2. Within-class dispersion increases with grade — similar to what the literature has suggested so far—but so does between-school dispersion across entire systems.

The literature to date has shown how, across different settings in LMICs, large within-class differences in proficiency levels exist and pose a pedagogical challenge for teachers. Our findings align with this insight. Yet, our findings also highlight that this is only a partial answer to the challenge posed by heterogeneity in learning outcomes, as we also find that systemwide between-school heterogeneity tends to increase with grade too — one of the key contributions of this paper.

To illustrate this, Figure 4 shows two measures of dispersion of learning outcomes by region, grade, and subject. The yellow bar shows the average within-school standard deviation (i.e.,

calculating the standard deviation in the two main outcomes within each school and then taking an average for each region, grade, and subject), and the average between-school standard deviation (i.e., calculating the average of the two main outcomes within each school and then taking the standard deviation of these averages for each region, grade, and subject).[10] One caveat to note in the presentation of these findings is the use of different scales in the measuring literacy and numeracy outcomes, based on the different constructs assessed. Therefore, the scales of the outcomes presented, and any cross-subject comparisons made, should be interpreted with caution as suggestive of relative differences, but also as potentially emerging to some extent due to the different underlying units for each outcome.

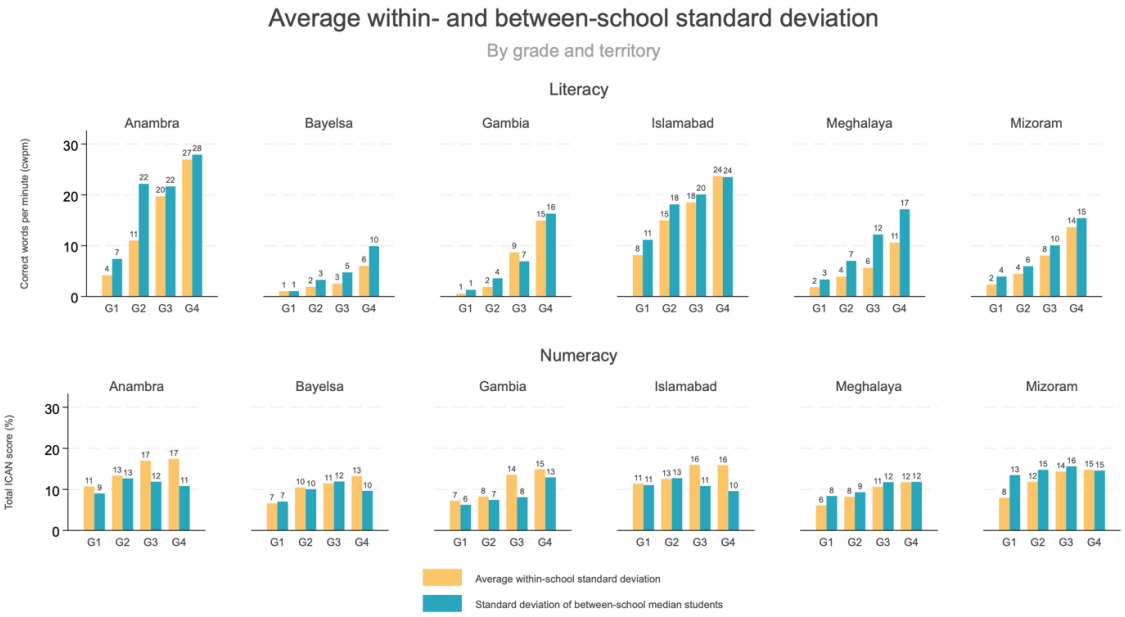


Figure 4

In most cases, we find that the within-class standard deviation (the yellow bar) tends to consistently increase with grade—akin to what previous research has shown. Yet, between-school dispersion (the blue bar) also increases with grade at a similar pace, especially for literacy. In other words, in raw units of standard deviations, the issue of between-school heterogeneity might

[10] For completeness, we also display a similar graph in Appendix Figure 13 but featuring the "coefficient of variation" ("CV"; standard deviation divided by the mean for each region, grade, and subject unit). While this figure shows that the CV decreases with grade, we believe that this only indicates that the average levels tend to increase faster in absolute units than the standard deviation. That is, this does not necessarily mean that dispersion is less meaningful as grade increases, as this is only in "raw units" and not mapped onto proficiency levels, as we do in the main text of the paper.

also pose a challenge in most settings and might require at least some policy and research attention akin to the issue of within-school heterogeneity so far.

3. In these six education systems, the proportion of the variance concentrated between schools is greater for literacy outcomes than for numeracy outcomes—emphasizing the need for different policy responses by subject depending on how the variance is distributed within each system.

So far, we have documented that in many cases, there is substantial between-school heterogeneity that might also have curricular implications for the pedagogical expectations within each school, and that in certain cases, a large portion of the variation might exist at the between-school level, and not just at the within-school level. Yet, it is valuable to more precisely quantify the extent to which the variation in a given region, grade, and subject unit might be due to between- or within-school differences, as there might be different policy recommendations under each scenario.

For example, if a system-wide analysis (akin to Appendix Figure 12) shows that there are meaningful differences within a region, grade, and subject in the proficiency levels of children, and it is also found that a large share of the variation was concentrated at the within-school level, policymakers might consider alternatives like within-school tracking, cross-school remedial tutoring or gifted education programs, or ability grouping interventions for a portion of the day akin to Teach at the Right Level. If, instead, a large portion of the variation was concentrated between schools, policymakers might also consider policy alternatives where children in the same grade across different schools get class content that is better tailored to the level that is most aligned with each school's current proficiency level.

Therefore, to further quantify how much of the raw variance in the two key outcomes of interest is due to between-school variation in each context and grade, we follow a basic variance decomposition of the kind that is typically applied to international assessments like PISA, as proposed by Foy (2005). In particular, we quantify the intraclass correlation (ICC) as the ratio of the between-school variance (i.e., the squared standard deviation of school means) over the total variance — the sum of the between-school variance and the "within-school" variance (i.e., the

squared mean standard deviation within schools) — as described in the following equation (Foy, 2005):

$$\text{Share of total variance concentrated between schools} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

Where σ^2_B is the between-school variance, and σ^2_W is the within-school variance. This method is appealing due to its clear interpretation of the outcome: this ratio always falls within the range of 0 to 1,[11] indicating the proportion of the total variance in the outcome attributable to variations between schools. Therefore, the higher the value of the ratio, the larger the role between-school heterogeneity plays in that context. For all regions, grades, and subjects, Figure 5 shows the share of the total variance in outcomes that is explained by between- and within-school differences, where the yellow portion of each bar represents the share of the variance explained by differences between schools — which in this case, corresponds to the intraclass correlation (ICC). In other words, a greater proportion of the yellow components of each bar signifies a higher level of heterogeneity between schools relative to within schools.

Figure 5 shows at least two key insights into how the variance in learning outcomes is distributed within these education systems. First, between-school variation in learning outcomes explains a larger share of the total variance in literacy outcomes than in numeracy outcomes. Across all regions and grades, between-school variance explains, on average, 13% of the total variance in numeracy outcomes, but 45% of the total variance in literacy outcomes. For context, this same figure was 32% on average across all OECD countries in the 2021 PISA assessment, with low- and middle-income countries typically below this average (e.g., the Philippines at 13%, Cambodia at 14%, Indonesia at 16%, Malaysia at 22%, and Vietnam at 30%). In other words, on average, foundational literacy outcomes in our sample display significantly more between-school variance than OECD countries or most low- and middle-income countries in the PISA assessment, but less than the same benchmark on numeracy assessments. This suggests that there is a meaningful difference in the concentration of learning outcomes between- and within-schools by subject (numeracy vs. literacy), and the likely policy alternatives that may be considered in response to these differences. A larger role of between-school heterogeneity might merit more

[11] Which, for visualization purposes in Figure 5, we present as a percentage from 0-100%.

interventions that are differentiated at the school level and less at the within-class level. In this case, literacy outcomes are in greater need of such types of intervention relative to numeracy outcomes, which would likely benefit more from within-school differentiation.

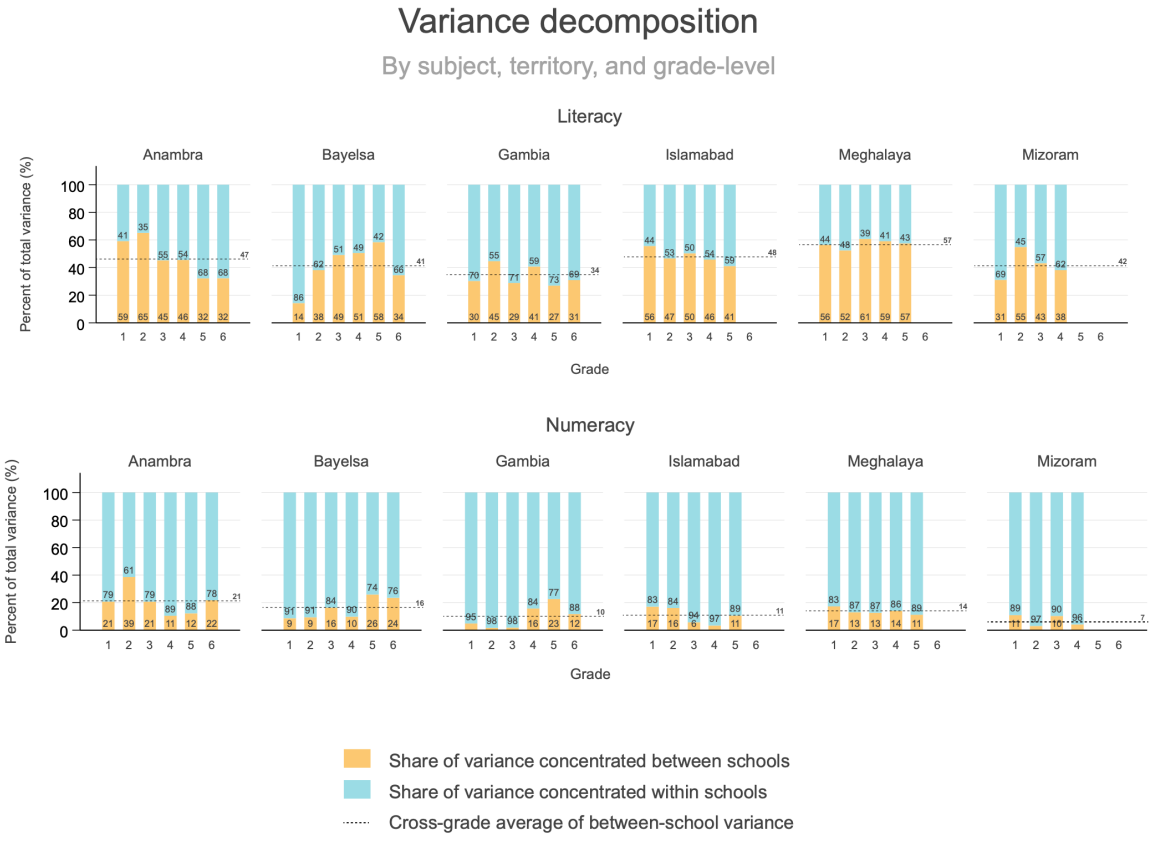


Figure 5

The second finding from Figure 5 is the need for these figures to be understood in conjunction with another absolute measure of heterogeneity in proficiency levels, as shown previously. For example, in Grade 3, the share of the total variance in the "raw outcome" for literacy (cwpm) explained by between-school heterogeneity is larger in Bayelsa than in Anambra. However, when these outcomes are mapped onto proficiency levels, there is more diversity in school-level proficiency levels in Anambra than in Bayelsa. Therefore, these relative shares and potential patterns within a system (or across systems) require additional nuance before delving too deeply into subgroup differences, and it is likely good research practice to avoid over-interpreting smaller

differences between subgroups (e.g., assuming that Grade 5 in Islamabad needs less school-level differentiation than Grade 1, based solely on the yellow bars in Figure 5). This type of variance decomposition is helpful for understanding broader patterns, such as school-level heterogeneity being a significantly larger challenge for literacy than numeracy, but we warn against using the procedure to over-interpret smaller differences between subgroups without the help of other tools like those used elsewhere in this paper.

4. The between-school variance increases with performance, but this increase is not linear and is not directly predicted by regional or urban/rural cuts of the data.

Given that the results have shown so far that school-level variance heterogeneity might pose a significant pedagogical challenge at the system-level, it is also valuable to understand what factors may be helpful to determine the extent of the challenge. If this information were available, then policymakers might be able to respond to system-wide heterogeneity through other observable characteristics of schools without the need to collect high-quality learning outcomes for all schools within a system.

First, we explore in Figure 6 the extent to which heterogeneity is correlated with baseline levels of performance for literacy and numeracy. Intuitively, a higher average level of performance might be expected to lead to more heterogeneity as it is less centered on the left portion of the distribution through fewer children achieving scores of 0, at least from a pedagogical stand point.[12] The orange dots on Figure 6 represent the average within-school standard deviation for each region, grade, and subject unit in the school, and the green dots represent the standard deviation of the average school-level scores for each region, grade, and subject unit.

[12] Mechanically though, several measures of dispersion like the coefficient of variation or even the Gini coefficient decrease when the average level score increases — holding the standard deviation constant.

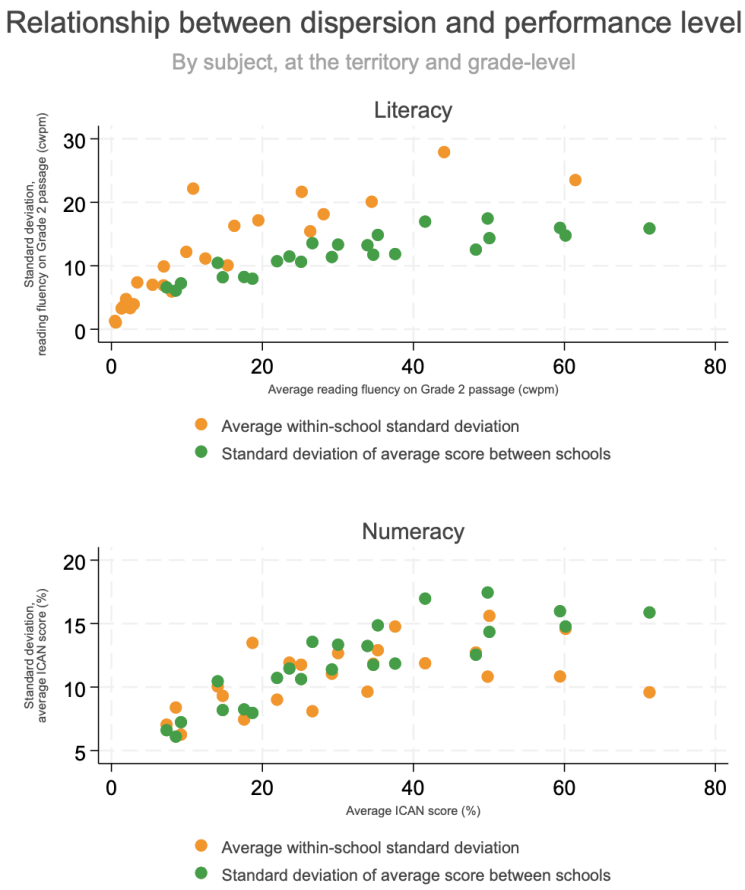


Figure 6

For literacy and numeracy, both metrics of dispersion are positively correlated with baseline performance (statistically significant at the <1% level). Yet, the strength of this correlation varies. For literacy, for example, within-school dispersion is more highly correlated with baseline performance than between-school dispersion, although this pattern is not as clear for numeracy. Similarly, for both literacy and numeracy, the strength of the relationship between both metrics of dispersion and baseline performance tapers off at higher levels of performance. In other words, given a relatively higher-performing sample, it will be harder to predict the extent of within- or between-school heterogeneity than if it were performing at lower levels. This is important because a relatively higher-performing sample, on average, is likely a necessary (but not sufficient) condition to observe meaningful school-level heterogeneity that might have pedagogical implications for policymakers. However, policymakers faced with a relatively higher-performing

sample should not immediately assume a high or low degree of between- and/or within-school level heterogeneity without at least a representative sample with high-quality data from their own context.

Next, we explore whether regional characteristics, such as the subregional unit a school is located in or the rural/urban status of a school, are strong predictors of between-school heterogeneity. In turn, this could serve as a useful proxy for potential curriculum customization in the presence of meaningful heterogeneity. To investigate this issue, we compare, for each region, grade, and subject unit, the share of the variance explained between schools to the share of the variance explained between rural/urban designations or subregional units. In Figure 7, we present the results for literacy, comparing the share of the variance explained between urban/rural designations (vertical axis) and the share of the variance explained by between-school differences (horizontal axis). The patterns are largely the same for numeracy or for sub-regional units (see Appendix Figure 14). Dots above the line of equality would signify a region, grade, and subject unit where rural/urban designations explain a larger share of the variation than between-school differences. Instead, Figure 7 shows that, in most cases, rural/urban designations play a small role in explaining between-school heterogeneity.

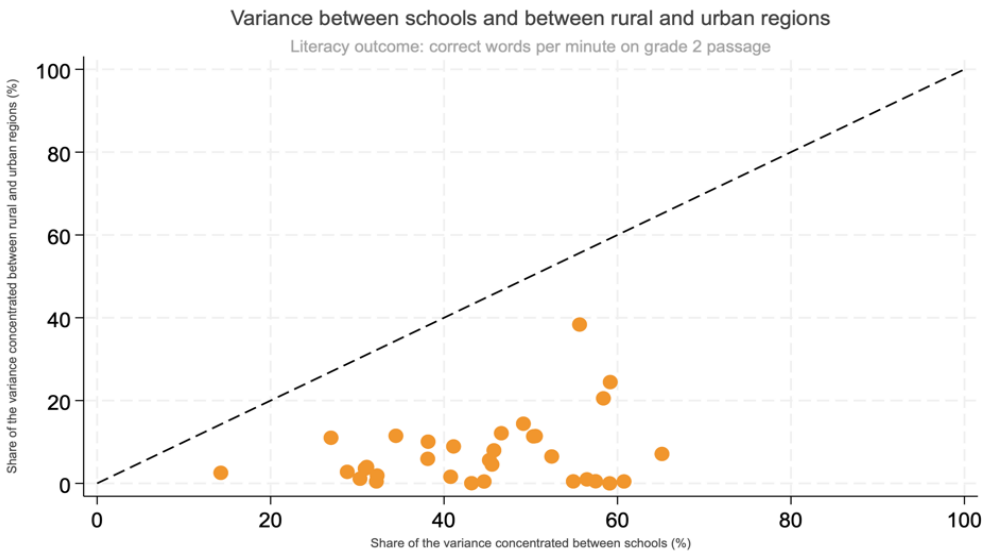


Figure 7

Taken together, the findings here suggest that predicting the extent to which school-level heterogeneity might pose a challenge in determining curricular requirements within a system might not be as straightforward as proxying this through the performance level of the system or observable school-level characteristics usually thought to be related to overall performance, like rural status. While a relatively higher average level of performance might be a prerequisite for meaningful heterogeneity, it is not a guarantee of a pedagogically large extent. Similarly, our results show that, in most cases, there are enough urban and rural schools that are high and low performing (or even by sub-regional unit, like districts) such that, if policymakers target curricular expectations by these characteristics rather than using student-level data, it might lead to significant misleveling of schools. Therefore, these findings leave policymakers facing meaningful school-level heterogeneity with two potential alternatives. First, they may attempt to find an additional observable school- and/or grade-level variable already in their administrative data that may be a stronger predictor of performance than those explored here. The second, and more likely, scenario is that any attempt to customize pedagogical requirements at the school level as a response to meaningful between-school heterogeneity will likely need to be underpinned by reliable, system-wide learning outcomes data.

5. Faced with meaningful between-school heterogeneity, data-driven targeting of the mandated curriculum for different schools can help policymakers reach more students through classroom instruction.

Next, we would like to better understand the potential gains in the share of children that would benefit from targeted classroom instruction under different scenarios, given that governments could level instruction across schools in their purview. To do so, we model four different instructional leveling scenarios using the "overambitious curricula" scenario —where instruction is pitched higher than the proficiency level of most students— as the baseline scenario against which the other three scenarios are compared. More specifically, we consider the potential increase in the share of students that benefit from classroom instruction if the system-wide curriculum were pitched at the level of the median child in a given region and grade, relative to the misaligned curriculum scenario. Additionally, we explore two options for differentiating

instruction at the school level depending on baseline performance as a response to between-school heterogeneity at different levels of fine-grain targeting.

More formally, to quantify the share of children reached by classroom instruction under different curricular leveling scenarios, consider teacher i in school s, region r, and who teaches only one class in grade g with N students. In this class, the teacher faces a median level of performance p, which comes from a distribution of performance P. This teacher pitches their instruction at a level L (proxied in units of the outcome, e.g., correct words per minute), which is set by the central planner, either fixed at $\bar{L}$ for all students in the region r and grade g as $\underline{L}_{\text{rg}}$, or is dependent on the baseline level of performance of each school at L_{rsgp} — depending on the instructional leveling scenario being considered. Similarly, the teacher i is a "curriculum-taker," i.e., they cannot decide the level of instruction that they teach and they take this from the central planner. For the sake of simplicity of the model, assume that at any given level of instructional leveling L, each student is dichotomously either "reached" by instruction or not, and that the share of students reached by instruction is given by $c(L,\ d,\ P)$, which is a function of the instructional leveling L, d — the "span" of children at levels above or below L that can still benefit from instruction at L,[13] and the overall distribution of outcomes P. Therefore, the policymaker's problem is to take the given P and d for their context — which they take as given — and optimally select L such that they maximize $N \cdot \sum^{s,i} c(L,\ d,\ P)$, the systemwide sum of children reached by classroom instruction in region r and in grade g.[14] With this in mind, we consider four different scenarios on how to select L, i.e., how to target instructional leveling within schools, and for each, we quantify the systemwide share of children that would be reached by classroom instruction under each L in each region r, given the differences in their actual distributions of performance P. The four scenarios that we consider are:

[13] To simplify the model, we will assume that d is fixed at d — that is, when instruction is pitched at a level L, children at $L \pm$ d would benefit from it, and we will set d at 10 cwpm, although we do verify that these arbitrary and simplifying decision do not affect the ultimate qualitative conclusions reached through this analysis.

[14] Clearly, without incorporating costs for each level (L) into the model, the central planner should select the most fine-grain L that fits the data at the greatest level of disaggregation. This choice might involve higher costs, and it may not be selected as the optimal response if the model considers costs. However, we maintain the model in its current state for two reasons. First, estimating costs can be highly context-dependent and prone to errors, introducing greater uncertainty into the model. Second, the primary goal of this exercise is to quantify the marginal benefit of different leveling approaches given various performance distributions. This allows us to derive broader insights that policymakers can use to weigh costs in their specific context, which they are likely more familiar with than us as researchers.

- "Curricular misalignment": the canonical case of overambitious curricula, where instruction is pitched at a much higher level than what the majority of children can do. In our model, we make the arbitrary decision of setting instruction in this scenario one full standard deviation above the level of the median child across the system. In other words, the instructional level that everyone in this system receives is $\overline{L}_{rg(p+\sigma)}$.
- "Targeting the median child system-wide": we simulate a scenario where instruction is pitched at the level of the median child within the region, grade, and subject. This represents the basic case where a central government revises its one-size-fits-all curriculum for a given grade to better align with the lower learning levels in their schools. The instructional level that everyone in this system receives is $\overline{L}_{rg(p)}$.
- "Targeting the median child across two distinct performance groups": this scenario represents the case where policymakers have accurate learning outcomes data for all schools at baseline through which they can create two sub-groups of schools. These groups, in turn, receive instruction tailored to the median child in each group. The two instructional levels within this scenario are $\overline{L}_{rg(p1)}$ and $\overline{L}_{rg(p2)}$, where p_1 and p_2 represent the median level of students in group 1 — the lowest-performing half of schools — and in group 2 — the highest-performing half of schools.
- "Targeting the median child across four distinct performance groups": a similar scenario to the previously described targeting of instruction based on baseline levels but with four groups instead of two, where the four targeted levels are $\overline{L}_{rg(p1)}$ through $\overline{L}_{rg(p4)}$. This scenario serves to understand the extent to which even more tailored instructional leveling might lead to additional benefits in the share of students reached — especially in light of the potential additional logistical challenges that this might pose for policymakers, that would need to be offset by meaningful instructional gains. This assumes that policymakers do, in fact, have the necessary resources and materials to implement multiple streams of instruction within each grade.

To showcase how we map this model to the data, we first visually represent the four instructional leveling scenarios for literacy outcomes in Grade 5 Meghalaya in Figure 8. We choose this grade and region as it is an advanced grade that might display a higher level of between-

school heterogeneity but does not stand out in previous analyses either as an extreme outlier in the presence of school-level heterogeneity.

The four base panels of Figure 8 show the same data, similar to the literacy row in Figure 1: a box-and-whiskers representation of reading fluency scores for Grade 5 students in Meghalaya, school by school. What varies from panel to panel in Figure 8 is positioning of the light blue box, which represents the leveling approach being modelled in each panel, centered in every case at the respective L — the given instructional level for each school and scenario — and spans a range of $\bar{d}$, 10 cwpm up and down from L. Therefore, the share of children reached by classroom instruction under the different leveling scenarios *c(L, d)* is displayed as the portions of the individual box-and-whiskers within each graph under this light blue box.

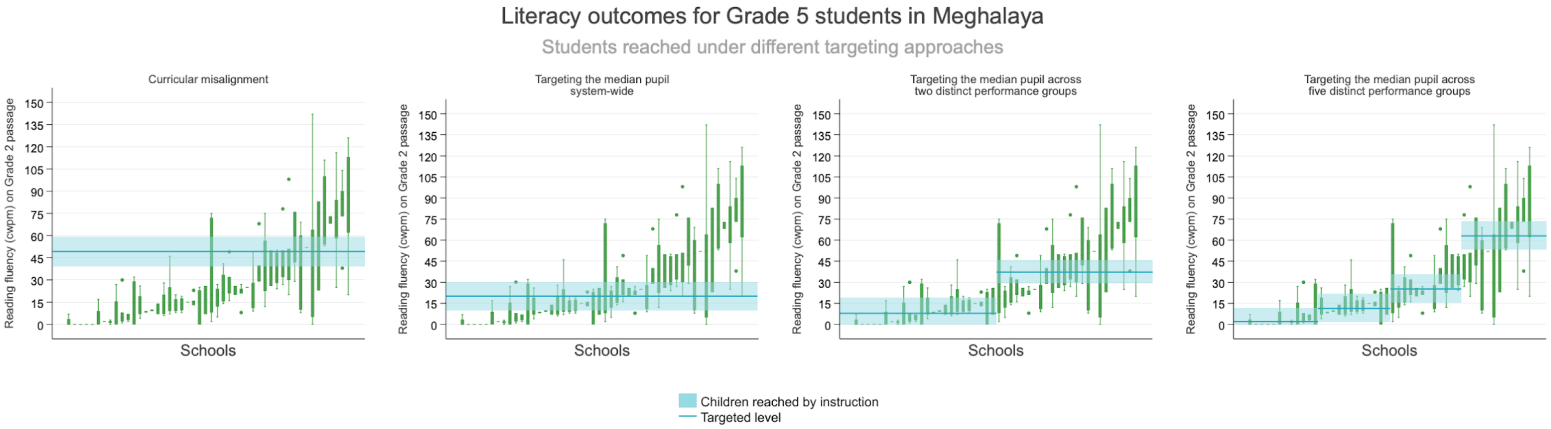


Figure 8

In the case of literacy outcomes for Grade 5 students in Meghalaya, only 11% of the sample is reached under the "curricular misalignment" scenario, as shown in the leftmost panel. Once the curriculum is modified to target the median child system-wide — the next panel on the right — 37% of the sample is reached. Further customizing the instructional requirements across two distinct groups will allow policymakers to reach 58% of Grade 5 children in this context, and differentiating across four different groups (i.e., the right-most panel) will enable policymakers to reach 81% of students. It's important to note that, although each subsequent panel from left to right leads to average improvements, this is not necessarily the case for all students at the individual level. For instance, moving from "curricular misalignment" to "targeting the median

child system-wide" leads to more children being reached on average, but there is a portion of children at the top of the distribution that are actually better served by the first scenario than the second scenario, despite the net gains systemwide.

We now quantify the potential gains in the share of children that would be reached by classroom instruction by repeating this exercise for all regions and grades, displaying the results in Figure 9. To understand how Figure 9 is related to Figure 8, note that the baseline case (curriculum misalignment, in dark green) for Meghalaya Grade 5 has a value of 11%—this is the share of children under the light blue box in the leftmost panel of Figure 8. The next bar over on Figure 9 for Meghalaya Grade 5 displays 26% —this is the additional gain in the share of children reached from the leftmost panel in Figure 8 (11% of children under the light blue box) to the next panel over targeting the system-wide median child (37% of children under the light blue box, for a gain of 26%—as shown in Figure 9).

As we analyze Figure 9, also note that we are not displaying these numbers as ultimate estimates for each region and grade; the precise numbers, as with any other model, depend on the assumptions behind them and parameters used (e.g., setting d at $\bar{d}$=10 for all grades and modeling scenarios, or the extent of the misalignment in the baseline scenario). Instead, these estimates should be taken as a pattern-seeing exercise to help policymakers better understand broader trends in when school-level customization of instructional requirements might be most beneficial, and to what extent and level of differentiation curricula need to be customized.

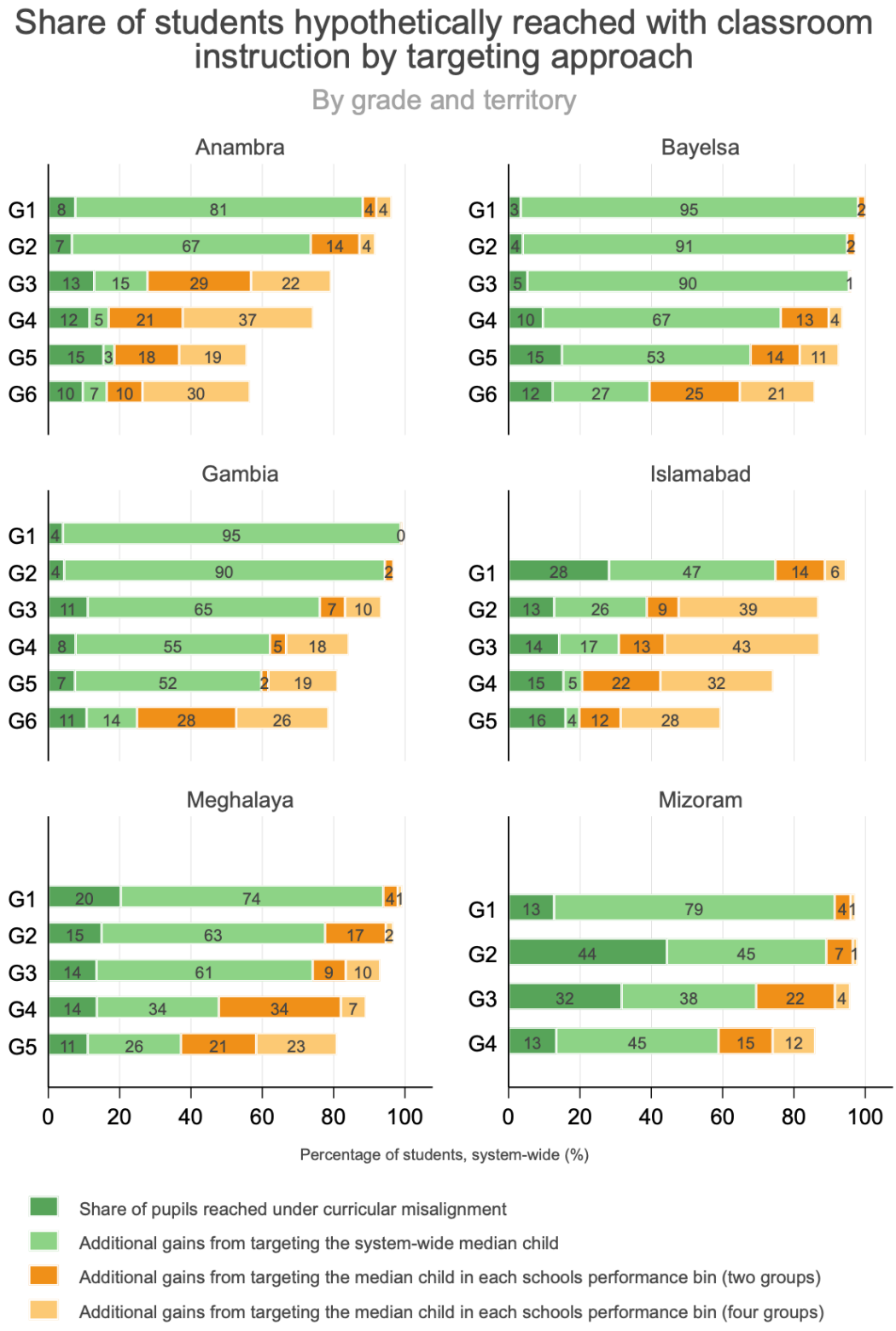

Share of students hypothetically reached with classroom instruction by targeting approach
By grade and territory
Anambra
Bayelsa
Gambia
Islamabad
Meghalaya
Mizoram
G1
G2
G3
G4
G5
G6
Percentage of students, system-wide (%)
Share of pupils reached under curricular misalignment
Additional gains from targeting the system-wide median child
Additional gains from targeting the median child in each schools performance bin (two groups)
Additional gains from targeting the median child in each schools performance bin (four groups)

Figure 9

There are at least three key insights emerging from Figure 9. First, we observe that for the lower grades like Grade 1 or 2, the vast majority of the gains come from a system-wide curriculum alignment with the median level in the whole system. On average across these six systems, the boost from transitioning from the curriculum misalignment scenario to alignment with the median systemwide level is 71 percentage points, relative to an additional 7 percentage points from differentiating schools into two groups or four additional groups based on baseline performance. If this is indeed the case, then policymakers willing to pass curriculum reforms that align systemwide instruction with a more realistic level for their system already have at their disposal a powerful tool to do so, without the need to worry too much about the logistical and political complexities of school-level differentiation.

Secondly, perhaps unsurprisingly, higher grades and regions with more school-level heterogeneity per previous results benefit the most from customized instructional leveling approaches. For example, in Islamabad, from Grade 2 onwards, the most fine-grain approach to curriculum targeting represents the single largest gain in the share of students reached. Interestingly, in Grade 5 in The Gambia, instructional customization in two groups has virtually no added benefit relative to the systemwide realignment of the curriculum towards the median child. Instead, it is not until further differentiation into additional performance groups is added that a significantly larger share of children is reached. These cases exemplify the fact that in the presence of school-level heterogeneity and a one-size-fits-all curriculum for a given grade in the whole region, regardless of where the instructional level is set, a large portion of children will not be served by it. In other words, it is true that in these cases, there are levels of the whole-system curriculum that may maximize the share of children reached (e.g., the median systemwide level), but even at these levels, many children and schools might not benefit from classroom instruction as it may be pitched too high or too low for their specific needs.

Thirdly, Figure 9 shows that at any given level of differentiation (i.e., for any of the instructional leveling approaches shown), the share of children reached by classroom instruction decreases with grade — likely because of the average increase in heterogeneity with grade and/or baseline performance both within-schools and between-schools. In other words, as grades progress, most types of classroom instruction become less able to cater to the needs of all children within those classes, as a larger proportion of them will either need much stronger remediation than the

median level amongst their peers, or because they are several grade levels ahead. In this sense, policymakers might also consider alternative interventions that can cater to those students who, despite more thoughtful instructional tailoring at the level of their school, might still not be reached by this initial intervention. Examples of this might include "pull-out" remedial tutoring for low-performing students or a cross-school gifted education program for high-performing children needing additional stimulation.

Finally, given the potential logistical costs of fine-grain school-level differentiation, we would also like to understand how the gains from targeting four groups—relative to a simpler systemwide curriculum misalignment—correlate with the baseline level of performance and dispersion in the data, as shown in Figure 10.

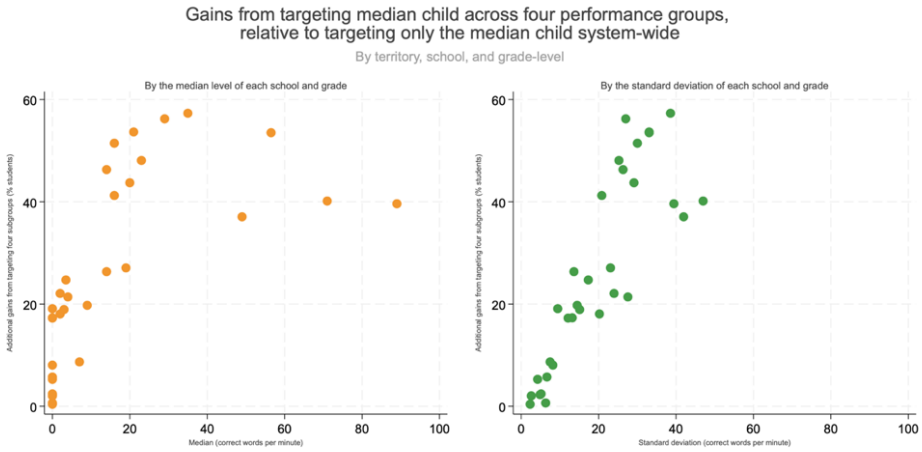


Figure 10

Figure 10 first indicates that the correlation between the gains from fine-grain targeting of school-level instruction is correlated with baseline levels of performance and baseline dispersion, consistent with what Figure 6 also shows. Yet, similarly to before, the correlation between baseline levels and the pedagogical benefits of additional differentiation at the school level tapers off at higher levels of performance. After a baseline level of approximately 20 cwpm, the relationship between these two factors becomes significantly flatter. As such, like before, baseline average levels do not appear to be a strong predictor of what type of curriculum alignment or differentiation would have the largest impact on the accessibility of classroom instruction relative to their

logistical costs. The relationship between the baseline standard deviation and these potential gains is stronger across the distribution, but even then, there is a "fanning out" pattern, where at any given level of baseline standard deviation, there can be a gap as wide as 20 percentage points (or approximately 50%) in the potential gains from differentiation. In this sense, a recurring theme in these findings is that neither the extent of school-level heterogeneity nor the potential benefits of fine-grain differentiation can be easily quantifiable without, at the very least, a representative sample of students and schools with high-quality data on learning outcomes.

IV. Discussion

In this paper, we have explored the issue of between-school heterogeneity in foundational learning outcomes across six public education systems in LMICs and the potential implications that this issue might have for curriculum setting at a large scale. We have shown that in certain regions, grades, and subjects, there are pedagogically meaningful levels of school-level heterogeneity, which might call for a rethinking of setting the same curriculum expectations for all schools within a given system. We have presented evidence that, in the face of school-level heterogeneity across an education system, customizing the instructional level of the curriculum for the needs of different schools given their baseline levels of performance might allow policymakers to reach a significantly higher share of children through classroom instruction. We have presented cases too, such as in the earlier primary grades of low-performing regions, where such an adjustment is not as necessary as simply aligning the national or regional curriculum to better reflect the level of the median child in that grade across the territory.

We also reflect on the fact that the production of the current paper, and one of the reasons why similar exercises are scant, is the relative lack of available, high-quality, at-scale learning data for all grades within education systems in LMICs. Without this type of student-level data, either in the form of census data of learning outcomes or representative samples with similar assessments, policymakers not only are effectively blind to the issue of between- (or within-) school heterogeneity but also lack insight into the broader state of learning outcomes within their educational system. Moreover, while a representative sample allows policymakers to quantify the extent of learning outcomes and/or heterogeneity within their system, census data from every school is needed in order to accurately differentiate instruction. Therefore, the existence of

comprehensive, high-quality learning outcome data for education systems is a critical prerequisite to understand potential policy alternatives to raise learning outcomes, among which could be interventions that address school-level heterogeneity if the data shows that this is indeed a challenge in that context.

Furthermore, the quality of the learning outcomes data is as important as its existence for the quantification of between-school heterogeneity for at least three reasons. First, our data shows that school-level characteristics typically associated with performance levels like urban/rural status do not explain a large portion of between-school heterogeneity. In other words, policymakers wanting to tackle between-school heterogeneity by differentiating curricula by rural status, for example, will actually incur a large degree of measurement error and misleveling of schools. Secondly, accurate school-level data is needed to minimize measurement error at the school level. That is, if a government were to differentiate curricular expectations by school, a poor-quality assessment or one with large degrees of measurement error (i.e., misclassification of grade-level proficiencies) would lead to schools receiving instruction not suitable for them. One can even imagine a scenario in which an assessment is relatively accurate at the system level but inaccurate at the individual level (i.e., the measurement error is a "classical error," or with an expected mean of 0); in this case, the benefits of school-level tailoring wash away across the system despite instructional tailoring for schools, due to poor school-level classification. Hence, high-quality learning data that is accurate at least at the level of the school is a requirement to address school-level heterogeneity. Finally, high-quality data is required to coexist with pedagogically sound mappings of this data to grade-level proficiencies and instructional mandates so that policymakers have a way to translate between the raw outcomes in their data and the potential policy prescriptions at the curricular level.

In all, our paper makes a dual contribution to the field. Firstly, it adds to the literature on learning inequality and heterogeneity, particularly focusing on between-school heterogeneity at a system-wide level. Secondly, it equips policymakers with a set of frameworks to better understand the potential extent of between-school heterogeneity and its relative policy urgency, as well as ways to quantify the potential benefits of different policy alternatives in response to meaningful between-school heterogeneity.

For future research, we identify at least two potential areas that require further exploration. Firstly, this paper does not provide a clear framework for policymakers facing both between- and within-school heterogeneity on how to weigh these two potential challenges and direct relative portions of their efforts to one and/or the other. To begin to think about such a framework, obtaining full census data — at least from each sampled school — might be advisable to minimize potential measurement error in within-school heterogeneity due to small sample sizes within each school and grade combination. Secondly, even if, in theory, there might be gains in learning outcomes if the curriculum or instruction were better tailored to the level of each class or school in the presence of between-school heterogeneity, the magnitude —if at all— of these gains is still uncertain. As such, building a body of research on interventions that address between-school heterogeneity, akin to the budding body of research addressing within-system heterogeneity, will be valuable for policymakers considering potential alternatives for their education systems and specific challenges with heterogeneity within them.

V. References

Abadzi, H. (2011). Reading fluency measurements in EFA FTI partner countries: Outcomes and improvement prospects. World Bank. https://documents1.worldbank.org/curated/en/925221468179361979/pdf/797780WP0readi0Box0379789B00PUBLIC0.pdf

Aldhanhani, Z. R., & Abu-Ayyash, E. A. S. (2020). Theories and research on oral reading fluency: What is needed? Theory and Practice in Language Studies, 10(4), 379. https://doi.org/10.17507/tpls.1004.05

ASER Centre. (2023). Annual Status of Education Report (Rural) 2022. https://img.asercentre.org/docs/ASER 2022 report pdfs/All India documents/aserreport2022.pdf

Audu, N. P., & Arikawei, A. R. (2013). Oil and gas exploration in the Niger Delta: Assessment of its impact on rural development in Bayelsa State. Research on Humanities and Social Sciences, 3(17), 47-57.

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

Bello, A. T., & Nwaeke, T. (2023). Impacts of oil exploration (Oil and gas conflicts; Niger Delta as a case study). Journal of Geoscience and Environment Protection, 11(03), 189–200. https://doi.org/10.4236/gep.2023.113013

Bratsch-Hines, M., Vernon-Feagans, L., Pedonti, S., & Varghese, C. (2020). Differential Effects of the Targeted Reading Intervention for Students With Low Phonological Awareness and/or Vocabulary. Learning Disability Quarterly, 43(4), 214-226. https://doi.org/10.1177/0731948719858683

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

Cummins, J. R. (2017). Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment. Economics of Education Review, 56, 40–51. https://doi.org/10.1016/j.econedurev.2016.11.006

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. American Economic Review, 101(5), 1739-74. https://doi.org/10.1257/aer.101.5.1739

Foy, P. (2005). Estimating and interpreting variance components in international comparative studies in Education. Studies in Educational Evaluation, 31(2–3), 173–191. https://doi.org/10.1016/j.stueduc.2005.05.009

Ganimian, A., Djaker, S. (2023). How Can Developing Countries Address Heterogeneity in Students' Preparation for School? A Review of the Challenge and Potential Solutions. Working Paper.

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

Good, R. H., & Kaminski, R. A. (Eds.) (2002). Dynamic indicators of basic early literacy skills (6th ed.). Institute for the Development of Education Achievement, University of Oregon. http://dibels.uoregon.edu

Government of India (2011). Census tables. Office of the Registrar General and Census Commissioner. https://censusindia.gov.in/census.website/data/census-tables

Government of Mizoram. (2016). Mizoram Vision 2030. Sustainable Development Goals Mizoram. https://sdg.mizoram.gov.in/uploads/attachments/8ffc1e44ea430770e384e2b61d4ffdfb/8-mizoram-economy.pdf

Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. Remedial and Special Education, 7(1), 6–10. https://doi.org/10.1177/074193258600700104

Hamadeh, N., Van Rompaey, C., & Metreau, E. (2023). World Bank Group country classifications by income level for FY24 (July 1, 2023- June 30, 2024). World Bank Blogs. https://blogs.worldbank.org/opendata/new-world-bank-group-country-classifications-income-level-fy24

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

Hasan, S. M., Beyer, R. C. M., & Hassan, K. (2021). GDP of Khyber Pukhtunkhwa's districts: Measuring economic activity using nightlights. Khyber Pakhtunkhwa Bureau of Statistics. https://kpbos.gov.pk/assets/docs/reports/NTL-PolicyBrief-Aug-1.pdf

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. Reading and Writing, 2(2), 127–160. https://doi.org/10.1007/BF00401799

Hwa, Y., Kaffenberger, M., Silberstein, J. (2020). Aligning Levels of Instruction with Goals and the Needs of Students (ALIGNS): Varied Approaches, Common Principles. https://riseprogramme.org/publications/aligning-levels-instruction-goals-and-needs-students-aligns-varied-approaches-common

Liu, Y., Shaker ul din, & Jiang, Y. (2021). Urban growth sustainability of Islamabad, Pakistan, over the last 3 decades: A perspective based on object-based backdating change detection. GeoJournal, 86(5), 2035–2055. https://doi.org/10.1007/s10708-020-10172-w

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

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

National Bureau of Statistics (2014). Population (States). Nigeria Data Portal. https://nigeria.opendataforafrica.org/wytkbxb/population-states

OECD. (2023). PISA 2022 Results (Volume I): The State of Learning and Equity in Education. PISA. https://doi.org/10.1787/53f23881-en

Okeowo, G. & Fatoba, I. (Eds). (2022). State of states: 2022 edition. BudgIT. https://yourbudgit.com/wp-content/uploads/2022/10/2022-State-of-states_Official.pdf

PAL Network (2020a). ICAN: International Common Assessment of Numeracy. Background, f eatures and l arge-scale i mplementatio n. Nairobi: People's Action for Learning Network.

PAL Network (2020b). ICAN: International Common Assessment of Numeracy. Frequently asked questions. Nairobi: People's Action for Learning Network. https://palnetwork.org/wpfd_file/english-faqs/

Petscher, Y., & Kim, Y.-S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. Journal of School Psychology, 49(1), 107–129. https://doi.org/10.1016/j.jsp.2010.09.004

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

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

Raghavan, C., & Lodrick, D. O. (2024). Meghalaya. Encyclopedia Britannica. https://www.britannica.com/place/Meghalaya

Rasinski, T. V. (2004). Assessing Reading Fluency. Pacific Resources for Education and Learning (PREL). https://eric.ed.gov/?id=ED483166

Reserve Bank of India. (2023). Handbook of statistics on the Indian economy. https://rbidocs.rbi.org.in/rdocs/Publications/PDFs/HBS20222023FULLDOCUMENT2FB950EDD2A34FE2BAE3308256EAE587.PDF

Rodriguez-Segura, D., & Mbiti, I. (2022). Back to the Basics: Curriculum reform and student learning in Tanzania [Working paper]. RISE. https://www.povertyactionlab.org/sites/default/files/research-paper/Back%20to%20the%20Basics-%20Curriculum%20Reform%20and%20Student%20Learning%20in%20Tanzania.pdf

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

Rodriguez-Segura, D. (2021). EdTech in developing countries: A review of the evidence. The World Bank Research Observer, 37(2), 171-203. https://doi.org/10.1093/wbro/lkab011

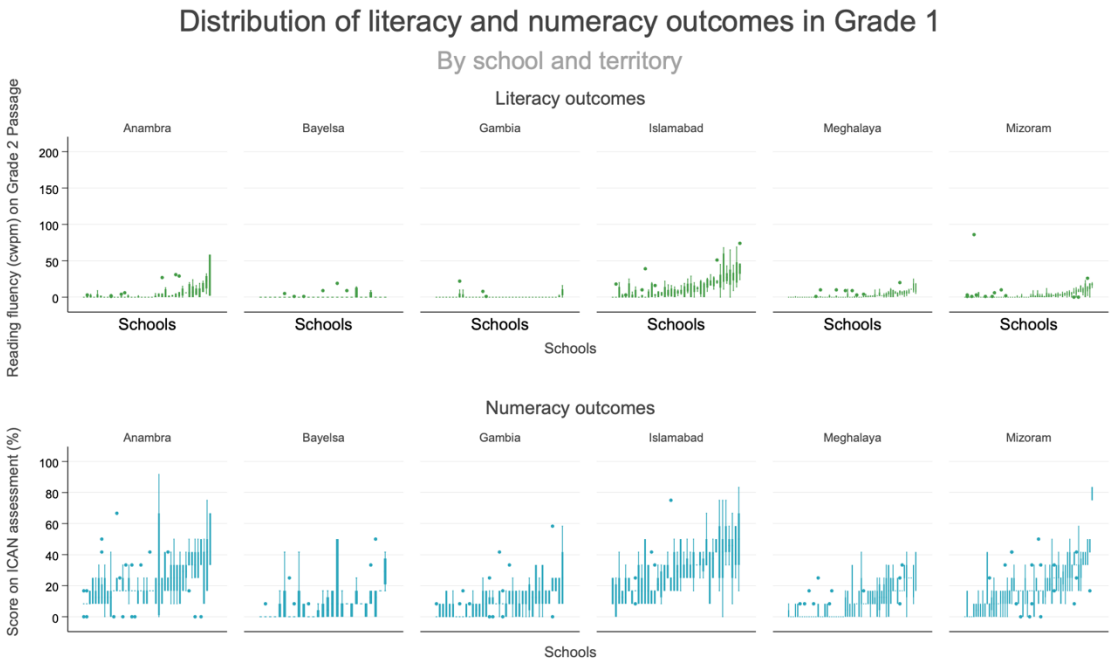
UNESCO Institute for Statistics, Global Alliance to Monitor Learning, UKAID, World Bank. (2023). Policy Linking for Measuring Global Learning Outcomes Toolkit: Linking assessments to the Global Proficiency Framework. UNESCO. https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/Policy_Linking_for_Measuring_Global_Learning_Outcomes_Dec-2020.pdf

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. Reading Research Quarterly, 45(3), 270–291. https://doi.org/10.1598/rrq.45.3.1

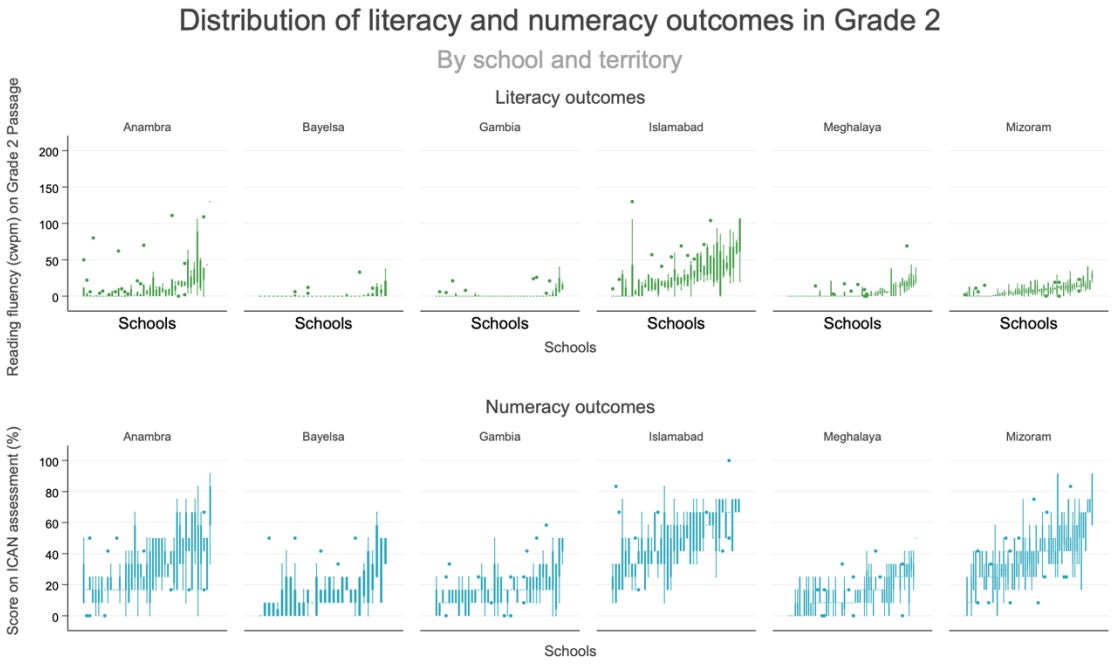
World Bank. (2018). World Development Report 2018: Learning to Realize Education's Promise. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1096-1

VI. Appendix

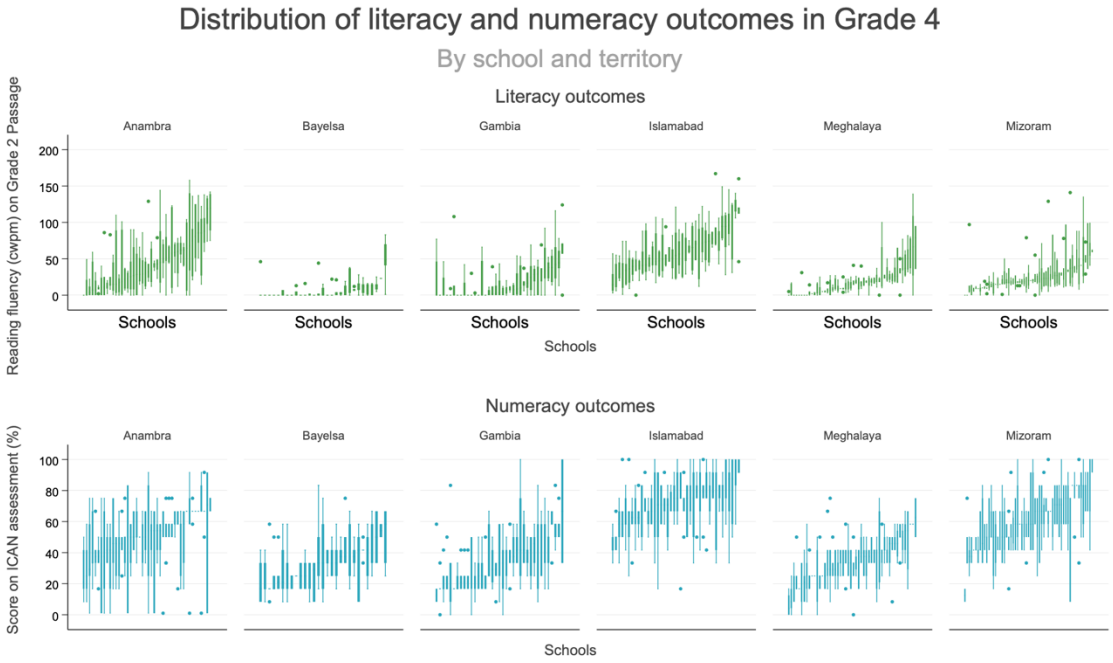
1. *Additional figures*



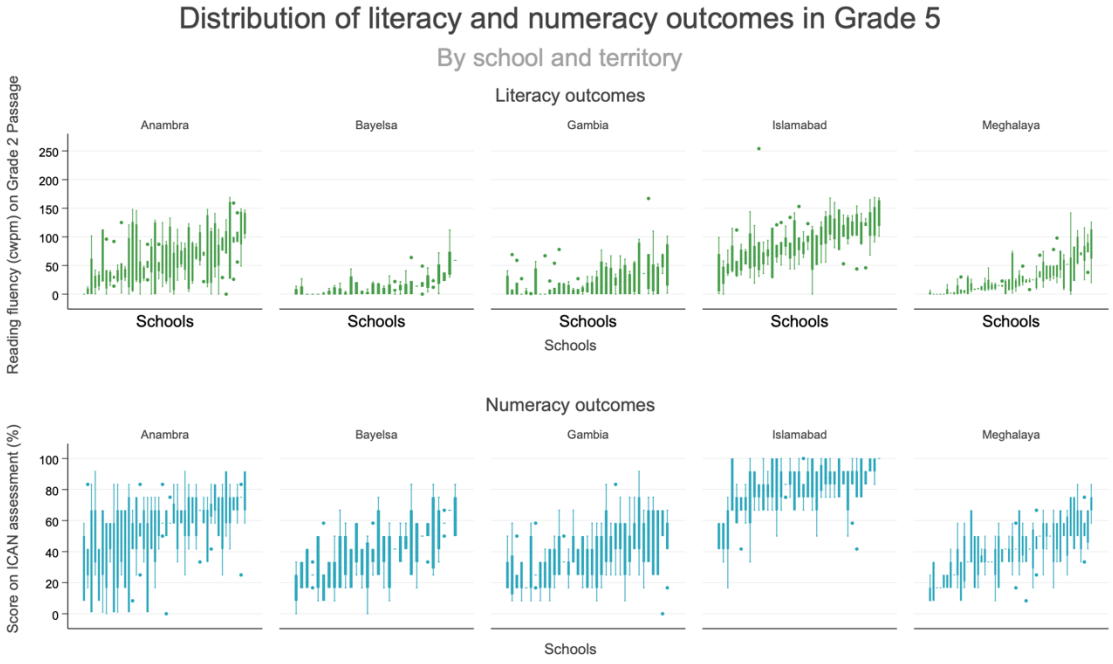
Appendix Figure 1



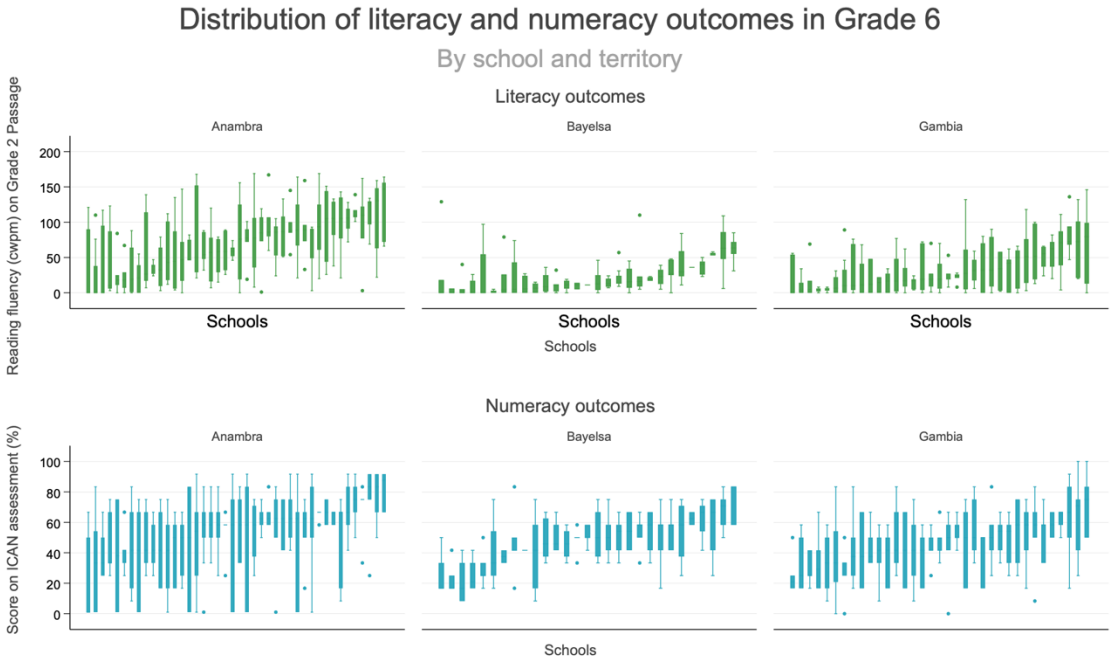
Appendix Figure 2



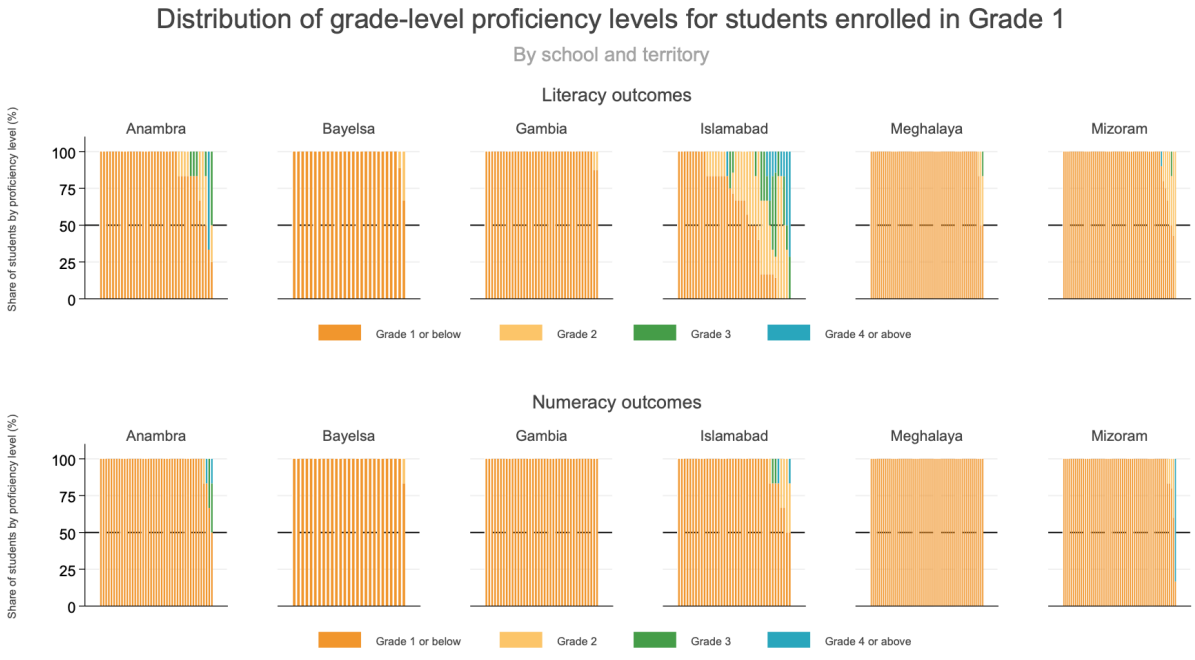
Appendix Figure 3



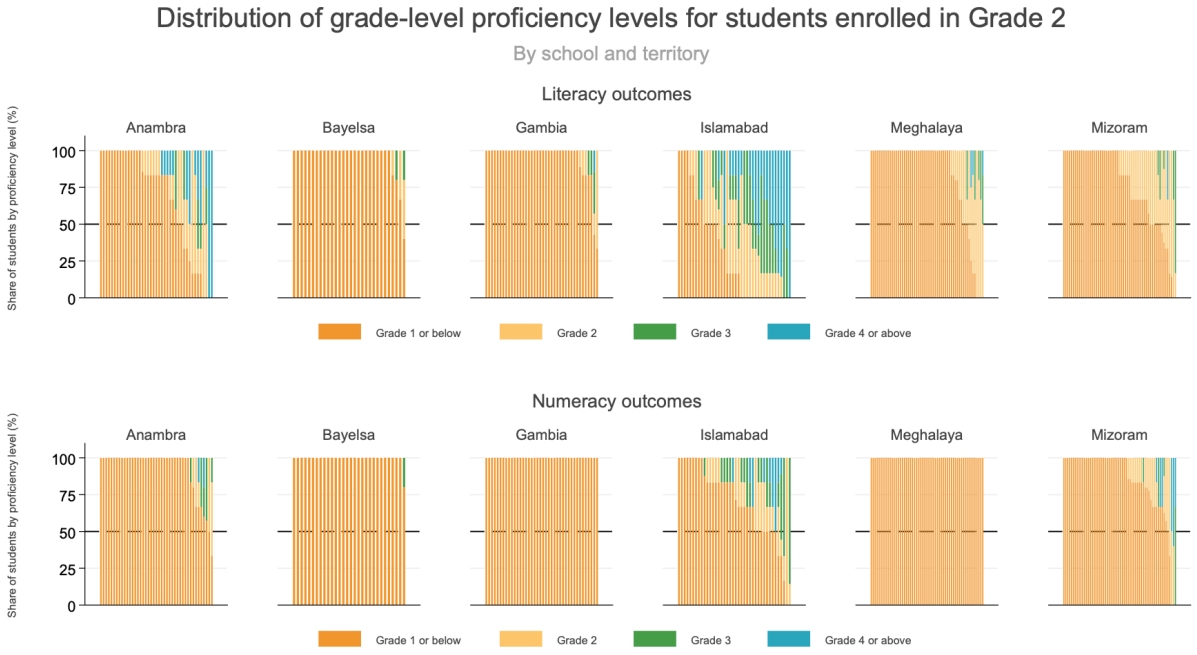
Appendix Figure 4



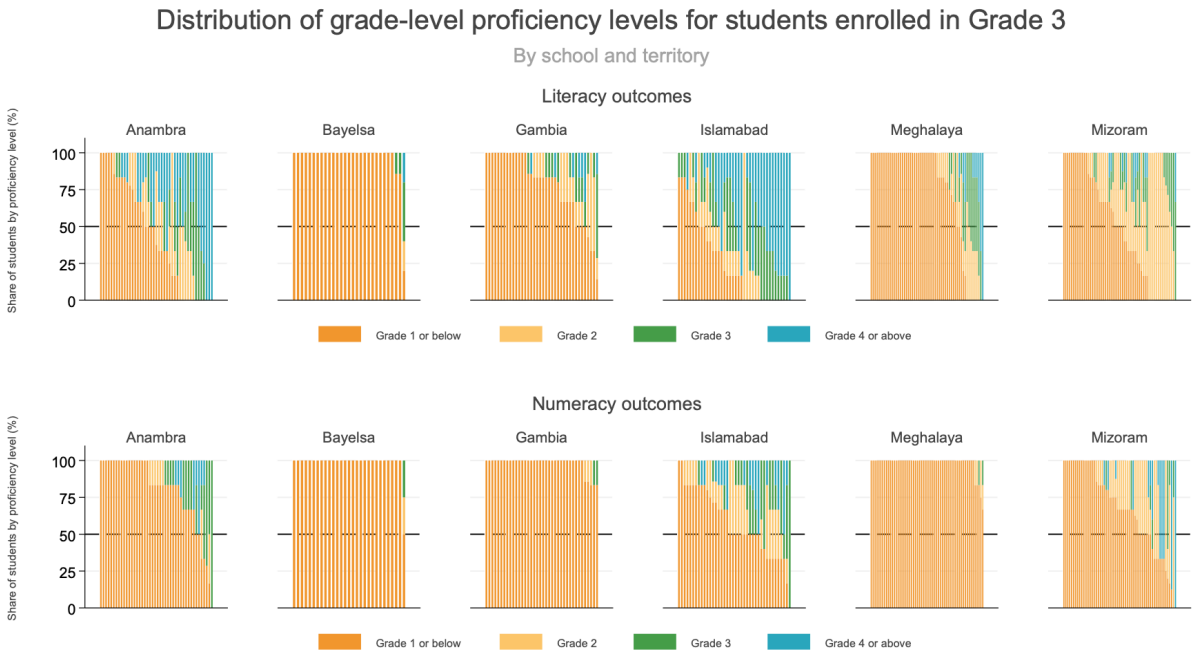
Appendix Figure 5



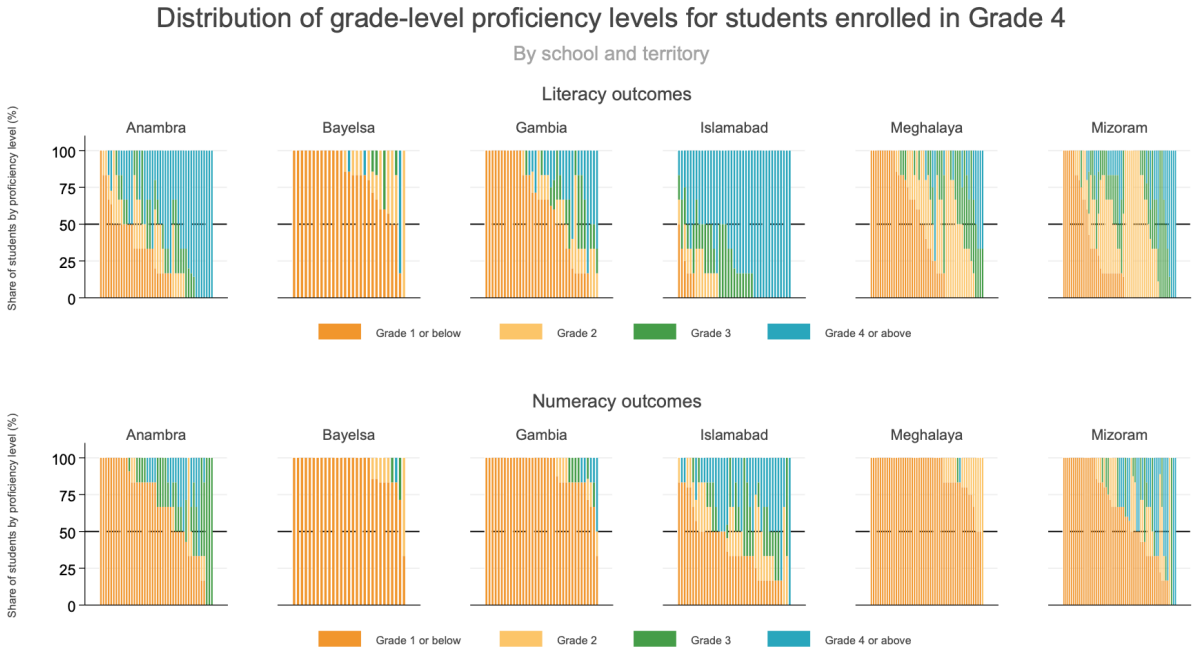
Appendix Figure 6



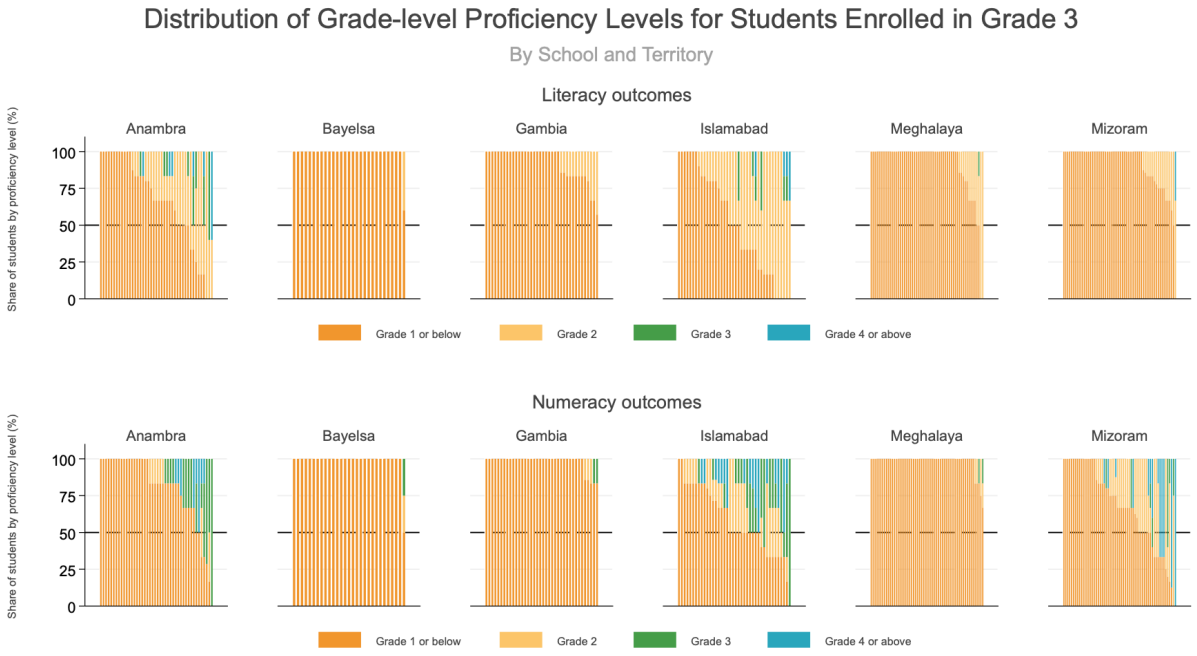
Appendix Figure 7



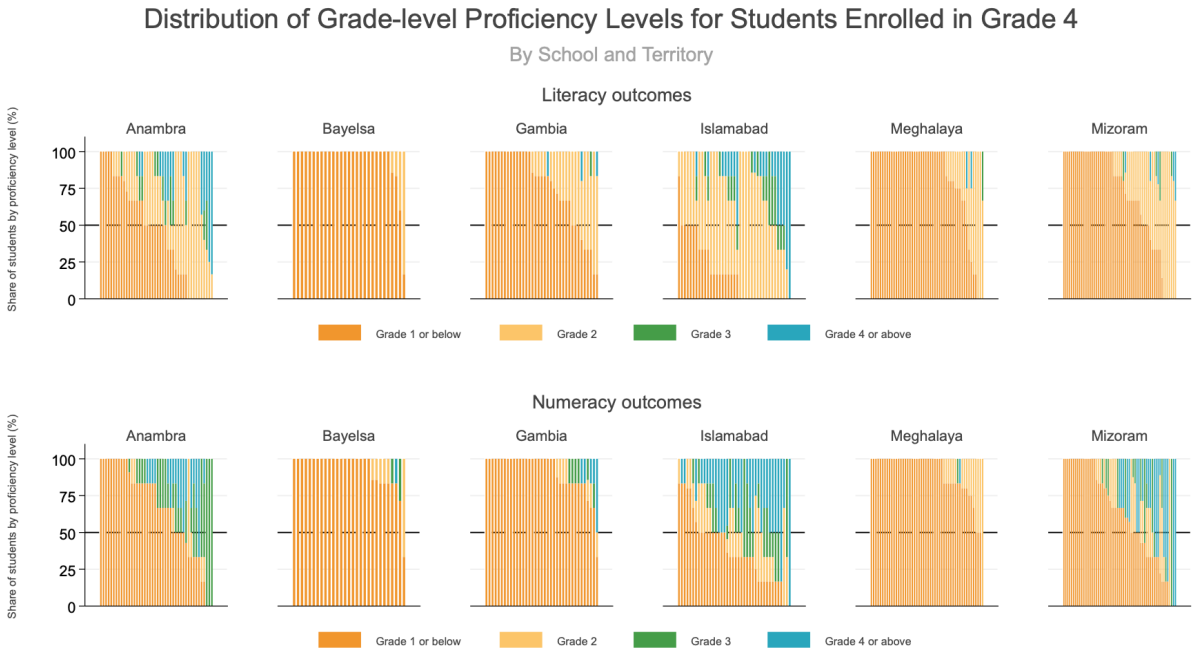
Appendix Figure 8



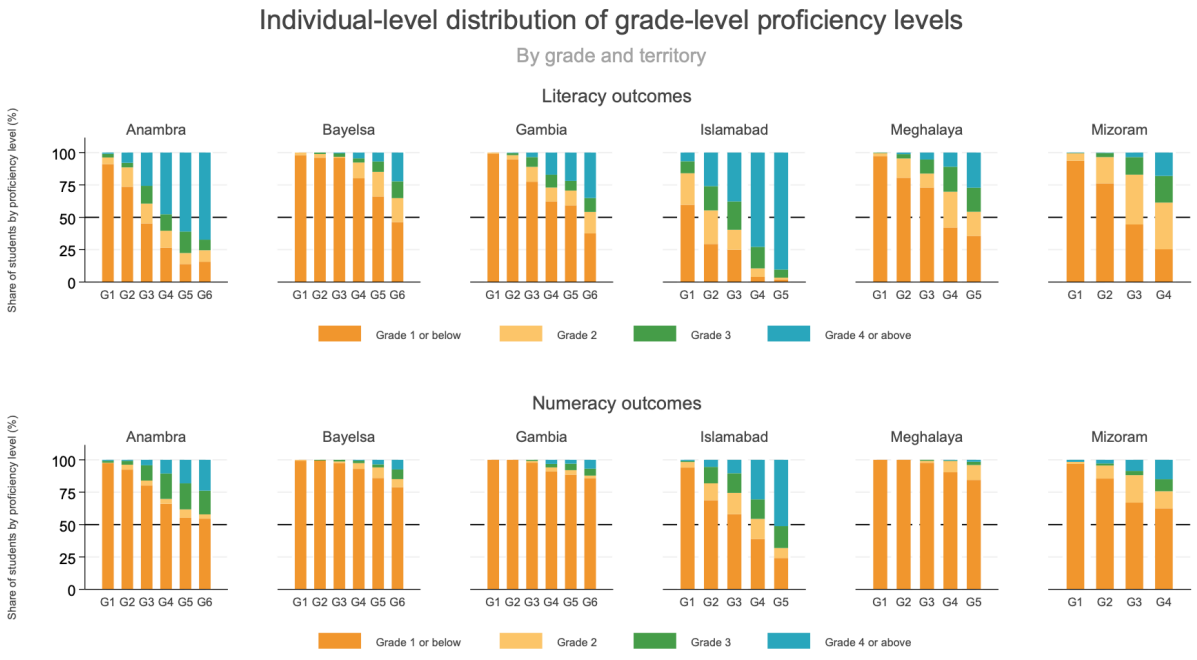
Appendix Figure 9



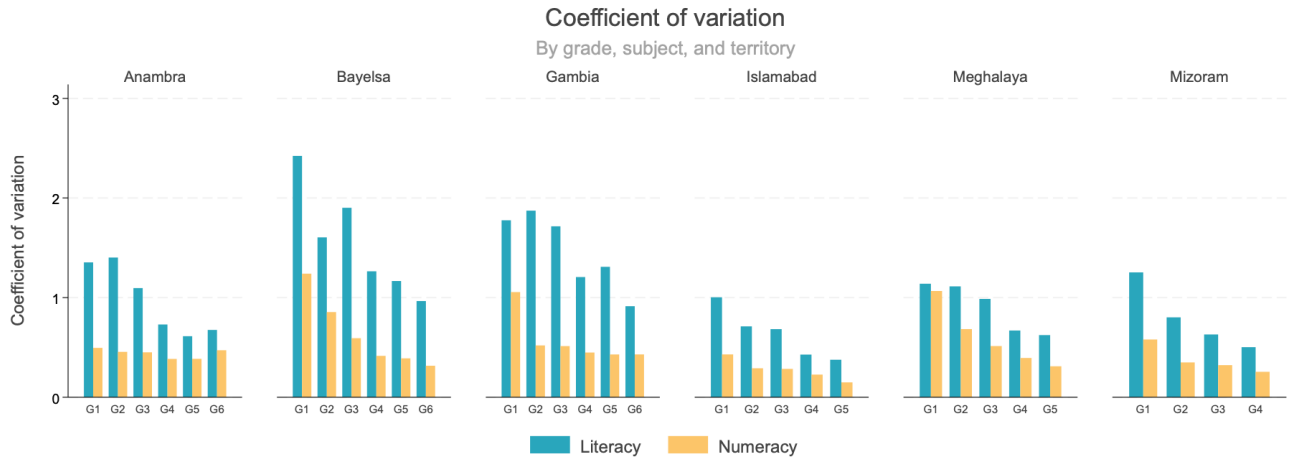
Appendix Figure 10



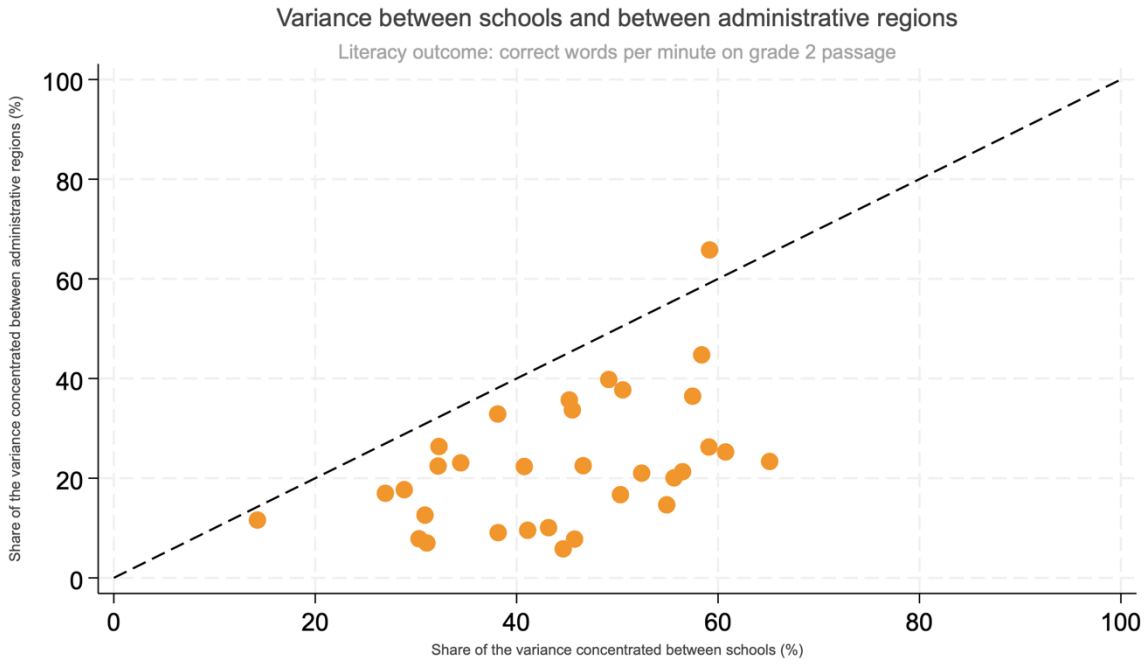
Appendix Figure 11



Appendix Figure 12



Appendix Figure 13

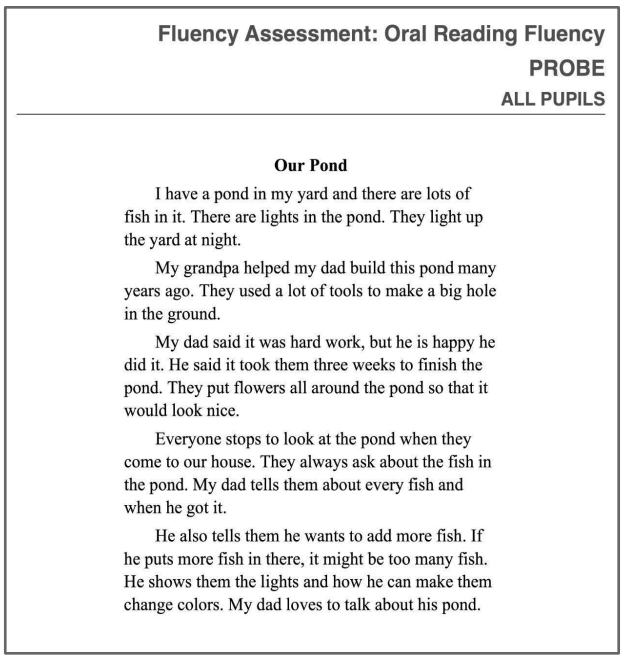


Appendix Figure 14

2. Assessment instruments used for the two main outcomes of interest

i. Oral reading fluency

Grade 2 passage: Anambra State, Islamabad, Meghalaya



Fluency Assessment: Oral Reading Fluency
PROBE
ALL PUPILS

Our Pond

I have a pond in my yard and there are lots of fish in it. There are lights in the pond. They light up the yard at night.

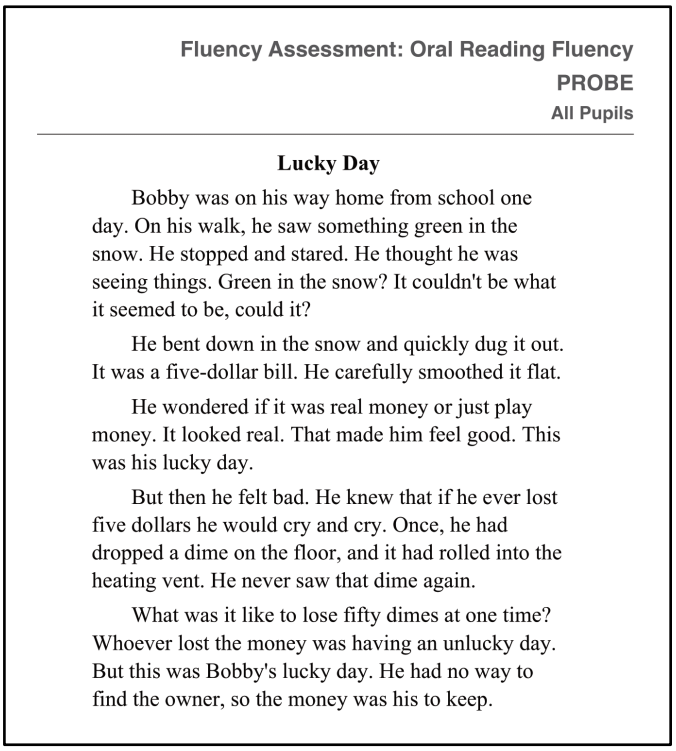
My grandpa helped my dad build this pond many years ago. They used a lot of tools to make a big hole in the ground.

My dad said it was hard work, but he is happy he did it. He said it took them three weeks to finish the pond. They put flowers all around the pond so that it would look nice.

Everyone stops to look at the pond when they come to our house. They always ask about the fish in the pond. My dad tells them about every fish and when he got it.

He also tells them he wants to add more fish. If he puts more fish in there, it might be too many fish. He shows them the lights and how he can make them change colors. My dad loves to talk about his pond.

Grade 2 passage: Bayelsa State, Mizoram



Fluency Assessment: Oral Reading Fluency
PROBE
All Pupils

Lucky Day

Bobby was on his way home from school one day. On his walk, he saw something green in the snow. He stopped and stared. He thought he was seeing things. Green in the snow? It couldn't be what it seemed to be, could it?

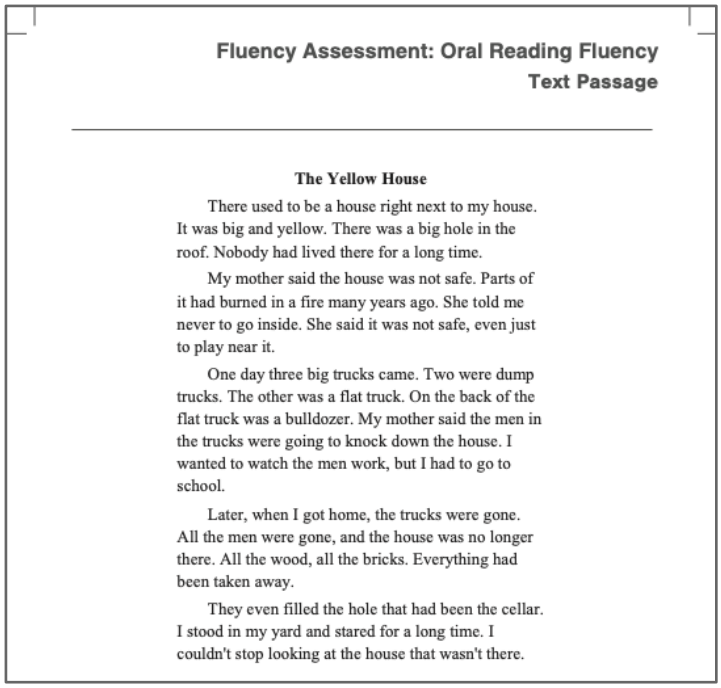
He bent down in the snow and quickly dug it out. It was a five-dollar bill. He carefully smoothed it flat.

He wondered if it was real money or just play money. It looked real. That made him feel good. This was his lucky day.

But then he felt bad. He knew that if he ever lost five dollars he would cry and cry. Once, he had dropped a dime on the floor, and it had rolled into the heating vent. He never saw that dime again.

What was it like to lose fifty dimes at one time? Whoever lost the money was having an unlucky day. But this was Bobby's lucky day. He had no way to find the owner, so the money was his to keep.

Grade 2 passage: The Gambia



Fluency Assessment: Oral Reading Fluency
Text Passage

The Yellow House

There used to be a house right next to my house. It was big and yellow. There was a big hole in the roof. Nobody had lived there for a long time.

My mother said the house was not safe. Parts of it had burned in a fire many years ago. She told me never to go inside. She said it was not safe, even just to play near it.

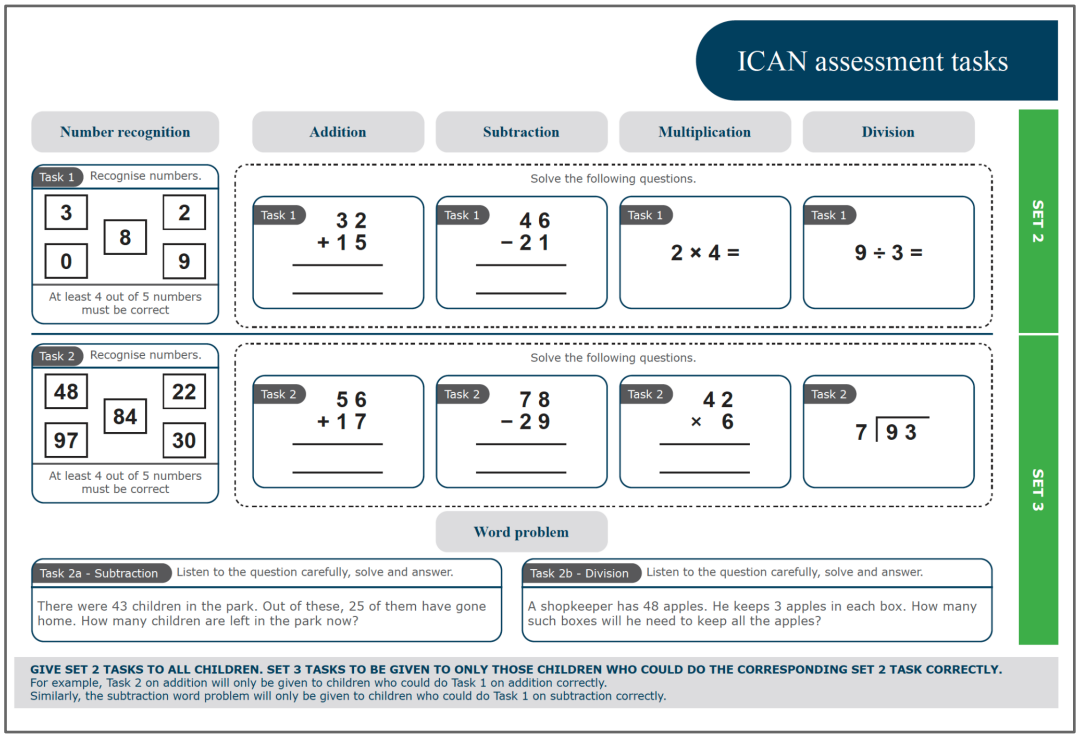
One day three big trucks came. Two were dump trucks. The other was a flat truck. On the back of the flat truck was a bulldozer. My mother said the men in the trucks were going to knock down the house. I wanted to watch the men work, but I had to go to school.

Later, when I got home, the trucks were gone. All the men were gone, and the house was no longer there. All the wood, all the bricks. Everything had been taken away.

They even filled the hole that had been the cellar. I stood in my yard and stared for a long time. I couldn't stop looking at the house that wasn't there.

ii. Numeracy

International Common Assessment of Numeracy (ICAN)



3. Full list of assessments given in each region

Appendix Table 1: Detailed overview of all assessments given in each of the six regions

Region	Common assessments across grades	Grade-level passage	Reading comprehension assessment	Additional assessments
Anambra State	Grade 2 passage from DIBELS	Grade-level passage from DIBELS	2 additional questions after each passage: 1 "direct thinking" and 1 "inferential thinking" question	English oral language assessment
	ICAN			Internally-developed diagnostic numeracy assessment
Bayelsa State	Grade 2 passage from DIBELS	Grade-level passage from DIBELS	2 additional questions after each passage: 1 "direct thinking" and 1 "inferential thinking" question	Internally-developed diagnostic numeracy assessment
	ICAN			
The Gambia	Grade 2 passage from DIBELS	Grade-level passage from English textbooks approved by the government	4 additional questions after each passage: 3 "direct thinking" and 1 "inferential thinking" question	English oral language assessment
	ICAN			Internally-developed diagnostic numeracy assessment
Islamabad	Grade 2 passage from DIBELS	Grade-level passage from DIBELS	2 additional questions after each passage: 1 "direct thinking" and 1 "inferential thinking" question on both an English and Urdu passage	English oral language assessment
				Grade 2 Urdu passage from a government-approved textbook
	ICAN			Grade-level Urdu passage from a government-approved textbook
				Internally-developed diagnostic numeracy assessment

Meghalaya	Grade 2 passage from DIBELS	Grade-level passage from DIBELS	2 additional questions after each passage: 1 "direct thinking" and 1 "inferential thinking" on both an English and native language passage	Grade 2 Garo passage from a government-approved textbook
				Grade 2 Khasi passage from a government-approved textbook
				Internally-developed diagnostic numeracy assessment
Mizoram	Grade 2 passage from DIBELS	Grade-level passage from DIBELS	2 additional questions after each passage: 1 "direct thinking" and 1 "inferential thinking" question	Internally-developed diagnostic numeracy assessment

4. Additional information on mapping proficiency levels

Appendix Table 2: Grade-level averages abstracted from reading fluency studies

Grade	Average words per minute	Number of countries where these grades were tested
1	12	8
2	23	13
3	38	13
4	62	5
5	70	3
6	56	2

Source: Abadzi, H. (2011). Reading fluency measurements in EFA FTI partner countries: Outcomes and improvement prospects. World Bank. https://documents1.worldbank.org/curated/en/925221468179361979/pdf/797780WP0readi0Box0379789B00PUBLIC0.pdf

Hasbrouck-Tindal Oral Reading Fluency Norms

The Hasbrouck-Tindal Oral Reading Fluency Norms are widely used as a tool to benchmark appropriate student progress in English oral reading fluency, given their developmental stage at different points of their primary school experience. These benchmarks are developed based on data from a few different assessments, including DIBELS, collected primarily in high-income, English-speaking countries. The chart below contains the Hasbrouck-Tindal grade-level benchmarks for students in the 25th, 50th, and 75th percentiles during the Spring term, the last term of the school year. Furthermore, the chart also includes the average expected growth per week from a student in the 50th percentile at this point of the school year.

Appendix Table 3: Oral Reading Fluency Norms (Correct words per minute)

	25th percentile	50th percentile	75th percentile	Median average weekly improvement
Class I	34	60	91	2.0
Class II	72	100	124	1.6
Class III	91	112	139	0.9
Class IV	105	133	160	1.2
Class V	119	146	169	0.8

Appendix Table 4: Mapping ICAN Results onto Global Performance Standards

ICAN skill	Sample problem	Grade-level expectation, according to GPF	Rationale[15]
Simple number recognition: One-digit number recognition	3, 0, 8, 2, 9	KG	G1: N1.1.1_M Count in whole numbers up to 30.
Complex number recognition: Two-digit number recognition	48, 97, 84, 22, 30	G1–2	G1: N1.1.1_M Count in whole numbers up to 30. G2: N1.1.1_M Count in whole numbers up to 100.
Simple addition: Two-digit addition without carrying	32 + 15 = ____	G2	G2: N1.3.1_M Add and subtract within 20 (i.e., where the sum or minuend does not surpass 20), and represent these operations with objects, pictures, or symbols. G3: N1.3.1_M Demonstrate fluency with addition and subtraction within 20 and add and subtract within 100 (i.e., where the sum or minuend does not surpass 100), with and without regrouping, and represent these operations with objects, pictures, or symbols (e.g., 32 + 59; solve an addition or subtraction problem presented by images of bundles of tens and ones; use number lines or skips on a hundreds grid to reason through or solve addition and subtraction problems).
Complex addition: Two-digit addition with carrying	56 + 17 = ____	G3	G3: N1.3.1_M Demonstrate fluency with addition and subtraction within 20 and add and subtract within 100 (i.e., where the sum or minuend does not surpass 100), with and without regrouping, and represent these operations with objects, pictures, or symbols (e.g., 32 + 59; solve an addition or subtraction problem presented by images of bundles of tens and ones; use number lines or skips on a hundreds grid to reason through or solve addition and subtraction problems).

[15] In this column, the Global Proficiency Descriptors for each grade level are coded in accordance with which domain and where in the GPF they are located ("N" stands for "Number Operations" and "A" stands for "Algebra"). The "M" at the end of each descriptor's label indicates that this is the expectation for the "Meets Minimum Proficiency" level. To access the GPF firsthand, please follow this link: https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/Global-Proficiency-Framework-Math.pdf

Simple subtraction: Two-digit subtraction without borrowing	46 - 21 =____	G2	G2: N1.3.1_M Add and subtract within 20 (i.e., where the sum or minuend does not surpass 20), and represent these operations with objects, pictures, or symbols.
Complex subtraction: Two-digit subtraction with borrowing	78 - 29 = ____	G3	G3: N1.3.1_M Demonstrate fluency with addition and subtraction within 20 and add and subtract within 100 (i.e., where the sum or minuend does not surpass 100), with and without regrouping, and represent these operations with objects, pictures, or symbols.
Simple multiplication: One-digit multiplication without regrouping (exact multiplication)	2 x 4 = ____	G3	G3: N1.3.2_M Multiply and divide within 100 (i.e., up to 10 x 10 and 100 ÷ 10, without a remainder), and represent these operations with objects, pictures, or symbols.
Complex multiplication: Two-digit multiplication with regrouping	42 x 6 = ____	G5	G5: N1.3.2_M Multiply, with and without regrouping, and divide, with no remainder, any number by a one-digit number and multiply two, 2-digit numbers, with and without regrouping (e.g., 342 x 4 = __; 42 x 34 = __; 1380 ÷ 5 = __).
Simple division: Exact, one-digit short division with no remnant	9 ÷ 3 = ____	G3	G3: N1.3.2_M Multiply and divide within 100 (i.e., up to 10 x 10 and 100 ÷ 10, without a remainder), and represent these operations with objects, pictures, or symbols.
Complex division: Short division of a two-digit dividend by a one-digit divisor with a remnant	93 ÷ 7 = ____	G6	G6: N1.3.2_M Multiply any number by a 2-digit number, with and without regrouping, and divide any number by a 1-digit number, with and without a remainder (e.g., 3427 x 68; 1380 ÷ 6 = __).
Simple fractions: Recognition of the magnitude of fractions	Which is greater: 4/5 or 3/15	G5	G5: N2.1.3_M Compare and order fractions with different but related denominators up to 12. G6: N2.1.3_M Compare and order proper and improper fractions with different, unrelated denominators.
Complex fractions: Addition of a fraction and a mixed number	1 1/6 + 1/3 =____	G6	G6: N2.2.1_M Add and subtract improper fractions or mixed numbers with different but related denominators.
Simple algebraic equations: Solving for a variable	17x = 68 x = ____	G6	G6: A3.2.1_M Find a missing value in a number sentence using any one of the four operations.

requiring one step			
Complex algebraic equations: Solving for a variable requiring two steps	-5y - 3 = 12 y = ____	G7	G7: A3.3.1_M Represent and solve problems, including real-world problems, using a two-step equation with any of the four operations.
Subtraction word problem	There were 43 children in the park. Out of these, 25 of them have gone home. How many children are in the park now?	G4	G4: N1.4.1_M Solve simple real-world problems involving addition and subtraction of whole numbers within 100 (i.e., where the sum or minuend does not surpass 100) with and without regrouping, including problems involving measurement and currency units.
Division word problem	A shopkeeper has 48 apples. He keeps 3 apples in each box. How many such boxes will he need to keep all the apples?	G5	G5: N1.4.2_M Solve simple real-world problems involving the multiplication of two whole numbers to 10, and associated division facts.
Fractions word problem	There were 108 goats in the pen. 1/6 of them were black. How many goats were NOT black?	G5	G5: N2.3.2_M Solve real-world problems involving the multiplication and division of a proper fraction and a whole number.
Algebraic equations word problem	A number plus 8 equals √144. What is the number?	G7	G7: A2.1.1_M Use linear expressions to represent problem situations with a single variable (e.g., The cost of buying cinema tickets online is £12 per ticket plus a £2 booking fee. Write this as an expression where x is the number of tickets purchased). G7: A3.3.1_M Represent and solve problems, including real-world problems, using a two-step equation with any of the four operations (e.g., solve 3x + 4 = 22; Some people got on a bus, doubling the number of passengers. At the next stop, 8 people got off, leaving 16 people on the bus. Represent the situation as an equation, and solve to find the number of people on the bus originally).