

Service Delivery Indicators: Pilot in Education and Health Care in Tanzania

April, 2013



Service Delivery Indicators:

Pilot in Education and Health Care in Tanzania

ABSTRACT

The Service Delivery Indicators ("the Indicators") provide a set of metrics for benchmarking service delivery performance in education and health in Africa to track progress across and within countries over time. The Indicators seek to enhance active monitoring of service delivery by policymakers and citizens, as well as to increase accountability and good governance. The perspective adopted by the Indicators is that of citizens accessing services and facing shortcomings. This report outlines the analytical underpinnings of the proposed indicators and reports on the results from two pilots carried out in the education and health sectors in Tanzania. The report concludes with a discussion of lessons learned and trade-offs, while ultimately proposing that the project be scaled up.

Acknowledgements

This report was prepared for the African Economic Research Consortium (AERC) in Nairobi, in partnership with the World Bank and with generous financial support from the William and Flora Hewlett Foundation. The pilot was implemented under the auspices of the AERC's Institutions and Service Delivery Research Program. The Research for Poverty Alleviation (REPOA) in Tanzania carried out the surveys. The technical team and authors of the report include: Tessa Bold and Jakob Svensson (IIES, Stockholm University), Bernard Gauthier (HEC Montréal), Ottar Maestad (Chr. Michelsen Institute, Bergen), and Waly Wane (The World Bank). Mwangi Kimenyi and Olu Ajakaiye (AERC), Linda Frey (Hewlett Foundation), and Ritva Reinikka (The World Bank) provided strategic guidance during the pilot phase. Philippe Achkar and Cindy Audiguier provided research assistance.

Service Delivery Indicators: Pilot in Education and Health Care in Tanzania

Table of Contents

1. INTRODUCTION	4
2. THE ANALYTICAL UNDERPINNINGS OF THE SERVICE DELIVERY INDICATORS	4
2.1 Service Delivery Outcomes and Perspective of the Indicators.....	4
2.2 Indicator Categories and the Selection Criteria	4
2.3 Indicator Description.....	6
3. IMPLEMENTATION OF PILOT SURVEYS IN TANZANIA	8
3.1 Overview	8
3.2 Sample Size and Design	8
3.3 Survey Instruments and Survey Implementation	9
4. INDICATORS AND PILOT RESULTS	12
4.1 Overview	12
4.2 Education	12
At the School.....	12
Teachers.....	15
Funding	19
4.3 Health	21
At the Clinic.....	21
Medical Personnel	22
Funding	26
5. OUTCOMES: TEST SCORES IN EDUCATION.....	28
6. INDICATOR AGGREGATION PROCESS AND COUNTRY RANKINGS	32
7. LESSONS LEARNED, TRADE-OFFS, AND SCALE-UP	34
7.1 Sample Size and Sample Strategy.....	34
7.2 Defining the Providers.....	35
7.3 Measuring Outcomes	35
7.4 Who are the Audiences?	36
7.5 Costing and institutional arrangement for scale-up	37
References	39

1. INTRODUCTION

Africa faces daunting human development challenges. On current trends, most countries in the region are off-track on most of the Millennium Development Goals. However, a look beneath this aggregate record reveals that much progress has taken place in many countries which started from a low base, and that there have been examples of extraordinary progress in a short time. If successes could be quickly scaled up, and if problems could be ironed out based on evidence of what works and what doesn't, Africa could reach the goals—if not by 2015, then in the not-too-distant future.

To accelerate progress toward the Millennium Development Goals, developing country governments, donors, and NGOs have committed increased resources to improve service delivery. However, budget allocations alone are poor indicators of the true quality of services, or value for money in countries with weak institutions. Moreover, when the service delivery failures are systematic, relying exclusively on the public sector to address them may not be realistic. Empowering citizens and civil society actors is necessary to put pressure on governments to improve performance. For this to work, citizens must have access to information on service delivery performance. The Service Delivery Indicators (hereinafter referred to as "the Indicators") project is an attempt to provide such information to the public in Africa.

To date, there is no robust, standardized set of indicators to measure the quality of services as experienced by the citizen in Africa. Existing indicators tend to be fragmented and focus either on final outcomes or inputs, rather than on the underlying systems that help generate the outcomes or make use of the inputs. In fact, no set of indicators is available for measuring constraints associated with service delivery and the behavior of frontline providers, both of which have a direct impact on the quality of services citizens are able to access. Without consistent and accurate information on the quality of services, it is difficult for citizens or politicians (the principal) to assess how service providers (the agent) are performing and to take corrective action.

The Indicators, which were piloted in Tanzania, provide a set of metrics to benchmark the performance of schools and health clinics in Africa. The Indicators can be used to track progress within and across countries over time, and aim to enhance active monitoring of service delivery to increase public accountability and good governance. Ultimately, the goal of this effort is to help policymakers, citizens, service providers, donors, and other stakeholders enhance the quality of services and improve development outcomes.

The perspective adopted by the Indicators is that of citizens accessing a service. The Indicators can thus be viewed as a service delivery report card on education and health care. However, instead of using citizens' perceptions to assess performance, the Indicators assemble objective and quantitative information from a survey of frontline service delivery units, using modules from the Public Expenditure Tracking Survey (PETS), Quantitative Service Delivery Survey (QSDS), Staff Absence Survey (SAS), and observational studies.

Box 1: PETS, QSDS, and SAS

Over the past decade, micro-level survey instruments, such as public expenditure tracking surveys (PETS), quantitative service delivery surveys (QSDS), staff absence surveys (SAS), and observational studies have proven to be powerful tools for identifying bottlenecks, inefficiencies, and other problems in service delivery.

PETS trace the flow of public resources from the budget to the intended end-users through the administrative structure, as a means of ascertaining the extent to which the actual spending on services is consistent with budget allocations. QSDS examine inputs, outputs, and incentives at the facility level, as well as provider behavior, to assess performance and efficiency of service delivery. SAS focus on the availability of teachers and health practitioners on the frontline and identify problems with their incentives. Observational studies aim to measure the quality of services, proxied for by the level of effort exerted by service providers.

In the Ugandan education sector, for example, Reinikka and Svensson (2004, 2005, 2006) use PETS to study leakage of funds and the impact of a public information campaign on the leakage rates, enrollment levels, and learning outcomes. They find a large reduction in resource leakage, increased enrollments, and some improved test scores in response to the campaign. Using QSDS, the same authors (2010) explore what motivates religious not-for-profit health care providers. They use a change in financing of not-for-profit health care providers in Uganda to test two different theories of organizational behavior (profit-maker versus altruistic). They show that financial aid leads to more laboratory testing, lower user charges, and increased utilization, but to no increase in staff remuneration. The findings are consistent with the view that the not-for-profit health care providers are intrinsically motivated to serve (poor) people and that these preferences matter quantitatively.

Chaudhury and others (2006) use the SAS approach to measure absence rates in education and health services. They report results from surveys in which enumerators made unannounced visits to primary schools and health clinics in Bangladesh, Ecuador, India, Indonesia, Peru, and Uganda, and recorded whether they found teachers and health workers at the facilities. Averaging across the countries, about 19 percent of teachers and 35 percent of health workers were absent. However, since the survey focused only on whether providers were present at the facilities, not whether or not they were actually working, even these low figures may present too favorable a picture. For example, in India, one-quarter of government primary school teachers were absent from school, but only about one-half of the teachers were actually teaching when enumerators arrived at the schools.

The Service Delivery Indicators project takes as its starting point the literature on how to boost education and health outcomes in developing countries. This literature shows robust evidence that the type of individuals attracted to specific tasks at different levels of the service delivery hierarchy, as well as the set of incentives they face to actually exert effort, are positively and significantly related to education and health outcomes. In addition, conditional on providers exerting effort, increased resource flows can have beneficial effects. Therefore, the proposed indicators focus predominantly on measures that capture the outcome of these efforts both by the frontline service providers and by higher level authorities entrusted with the task of ensuring that schools and clinics are receiving proper support. Our choice of indicators avoids the need to make strong structural assumptions about the link between inputs, behavior, and outcomes. While the data collection focuses

on frontline providers, the indicators will mirror not only how the service delivery unit itself is performing, but also indicate the efficacy of the entire health and education system. Importantly, we do not argue that we can directly measure the incentives and constraints that influence performance, but argue that we can, at best, use micro data to measure the outcomes of these incentives and constraints. Because health and education services are largely a government responsibility in most African countries, and quite a lot of public resources have gone into these sectors, the Service Delivery Indicators pilot focused on public providers. However, it would be relatively straightforward to expand the Indicators to include non-governmental service providers.

To evaluate the feasibility of the proposed Indicators, pilot surveys in primary education and health care were implemented in Tanzania in 2010. The results from the pilot studies demonstrate that the Indicators methodology is capable of providing the necessary information to construct harmonized indicators on the quality of service delivery, as experienced by the citizen, using a single set of instruments at a single point of collection (the facility). However, while collecting this information from frontline service providers is feasible, it is also demanding, both financially and logistically. The decision to scale up the project should hence weigh the benefits – having comparable and powerful data on the quality of service delivery – with the costs.

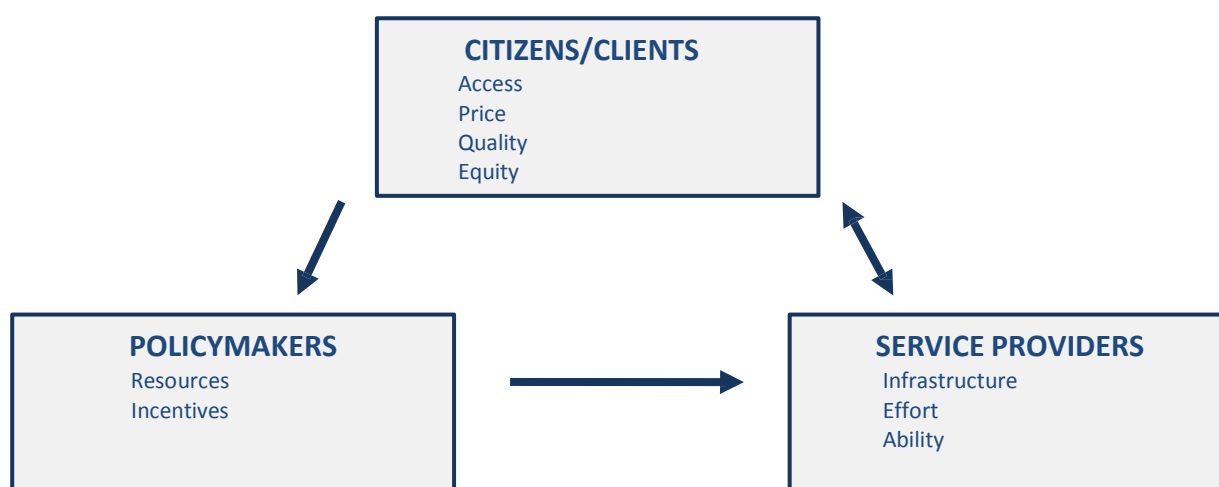
This paper is structured as follows: Section 2 outlines the analytical underpinnings of the indicators and how they are categorized. It also includes a detailed description of the indicators themselves and the justification for their inclusion. Section 3 presents the methodology of the pilot surveys in Tanzania. The results from the pilots are presented and analyzed in section 4. Section 5 presents results on education outcomes, as evidenced by student test scores. Section 6 discusses the advantages and disadvantages of collapsing the indicators into one score or index, and proposes a method for doing so in case such an index is deemed appropriate. Section 7 discusses lessons learned, trade-offs, and options for scaling up the project.

2. THE ANALYTICAL UNDERPINNINGS OF THE SERVICE DELIVERY INDICATORS

2.1 Service Delivery Outcomes and Perspective of the Indicators

Service delivery outcomes are determined by the relationships of accountability between policymakers, service providers, and citizens (Figure 1). Health and education outcomes are the result of the interaction between various actors in the multi-step service delivery system, and depend on the characteristics and behavior of individuals and households. While delivery of quality health care and education is contingent foremost on what happens in clinics and in classrooms, a combination of several basic elements have to be present in order for quality services to be accessible and produced by health personnel and teachers at the frontline, which depend on the overall service delivery system and supply chain. Adequate financing, infrastructure, human resources, material, and equipment need to be made available, while the institutions and governance structure provide incentives for the service providers to perform.

Figure 1: The relationships of accountability between citizens, service providers, and policymakers



2.2 Indicator Categories and the Selection Criteria

There are a host of data sets available in both education and health. To a large extent, these data sets measure inputs and outcomes/outputs in the service delivery process, mostly from a household perspective. While providing a wealth of information, existing data sources (like DHS/LSMS/WMS) cover only a sub-sample of countries and are, in many cases, outdated. (For instance, there have been five standard or interim DHS surveys completed in Africa since 2007). We therefore propose that all the data required for the Service Delivery Indicators be collected through one standard instrument administered in all countries.

Given the quantitative and micro focus, we have essentially two options for collecting the data necessary for the Indicators. We could either take beneficiaries or service providers as the unit of observation. We argue that the most cost-effective option is to focus on service providers. Obviously, this choice will, to some extent, restrict what type of data we can collect and what indicators we can create.

Our proposed choice of indicators takes its starting point from the recent literature on the economics of education and health. Overall, this literature stresses the importance of provider behavior and competence in the delivery of health and education services. Conditional on service providers exerting effort, there is also some evidence that the provision of physical resources and infrastructure – especially in health – has important effects on the quality of service delivery.¹

Box 2: Service delivery production function

Consider a service delivery production function, f , which maps physical inputs, x , the effort put in by the service provider e , as well as his/her type (or knowledge), θ , to deliver quality services into individual level outcomes, y . The effort variable e could be thought of as multidimensional and thus include effort (broadly defined) of other actors in the service delivery system. We can think of type as the characteristic (knowledge) of the individuals who select into specific task. Of course, as noted above, outcomes of this production process are not just affected by the service delivery unit, but also by the actions and behaviors of households, which we denote by ε . We can therefore write

$$y = f(x, e, \theta) + \varepsilon. \quad (1)$$

To assess the quality of services provided, one should ideally measure $f(x, e, \theta)$. Of course, it is notoriously difficult to measure all the arguments that enter the production, and would involve a huge data collection effort. A more feasible approach is therefore to focus instead on proxies of the arguments which, to a first-order approximation, have the largest effects.

The somewhat weak relationship between resources and outcomes documented in the literature has been associated with deficiencies in the incentive structure of school and health systems. Indeed, most service delivery systems in developing countries present frontline providers with a set of incentives that negate the impact of pure resource-based policies. Therefore, while resources alone appear to have a limited impact on the quality of education and health in developing countries, it is possible inputs are complementary to

¹ For an overview, see Hanushek (2003). Case and Deaton (1999) show, using a natural experiment in South Africa, that increases in school resources (as measured by the student-teacher ratio) raises academic achievement among black students. Duflo (2001) finds that a school construction policy in Indonesia was effective in increasing the quantity of education. Banerjee et al (2000) find, using a randomized evaluation in India, that provision of additional teachers in nonformal education centers increases school participation of girls. However, a series of randomized evaluations in Kenya indicate that the only effect of textbooks on outcomes was among the better students (Glewwe and Kremer, 2006; Glewwe, Kremer and Moulin, 2002). More recent evidence from natural experiments and randomized evaluations also indicate some potential positive effect of school resources on outcomes, but not uniformly positive (Duflo 2001; Glewwe and Kremer 2006).

changes in incentives and so coupling improvements in both may have large and significant impacts (see Hanushek, 2007). As noted by Duflo, Dupas, and Kremer (2009), the fact that budgets have not kept pace with enrollment, leading to large student-teacher ratios, overstretched physical infrastructure, and insufficient number of textbooks, etc., is problematic. However, simply increasing the level of resources might not address the quality deficit in education and health without also taking providers' incentives into account.

We propose three sets of indicators: The first attempts to measure availability of key infrastructure and inputs at the frontline service provider level. The second attempts to measure effort and knowledge of service providers at the frontline level. The third attempts to proxy for effort, broadly defined, higher up in the service delivery chain. Providing countries with detailed and comparable data on these important dimensions of service delivery is one of the main innovations of the Service Delivery Indicators.²

In addition, we wanted to select indicators that are (i) quantitative (to avoid problems of perception biases that limit both cross-country and longitudinal comparisons)³, (ii) ordinal in nature (to allow within and cross-country comparisons); (iii) robust (in the sense that the methodology used to construct the indicators can be verified and replicated); (iv) actionable; and (v) cost effective.

2.3 Indicator Description

Table 1 lists, by sector, the indicators that have been identified.

² The suggested indicators for education and health are partly based on an initial list of 50 PETS and QSDS indicators devised part of the project "Harmonization of Public Expenditure Tracking Surveys (PETS) and Quantitative Service delivery Surveys (QSDS) at the World Bank" (Gauthier, 2008). That initial list, which covers a wide range of variables characterizing public expenditure and service delivery, was streamlined using this project's criteria and conceptual framework.

³ See for instance Olken (2009).

Table 1: A service delivery report card

Education	Health
<i>At the school: Inputs and infrastructure</i>	<i>At the clinic: Inputs and infrastructure</i>
Infrastructure (electricity, water, sanitation) Children per classroom Student-teacher ratio Textbooks per student	Infrastructure (electricity, water, sanitation) Medical equipment per clinic Stock-outs of drugs
<i>Teachers: Effort and knowledge</i>	<i>Medical personnel: Effort and knowledge</i>
Absence rate Time children are in school being taught Share of teachers with minimum knowledge	Absence rate Time spent counseling patients per clinician Diagnostic accuracy in outpatient consultations
<i>Funding: Effort in the supply chain</i>	<i>Funding: Effort in the supply chain</i>
Education expenditures reaching primary school Delays in salaries	Health expenditures reaching primary clinics Delays in salaries

The various indicators, and the results from the pilot in Tanzania, are discussed in Section 4. A more detailed description and definition of the indicators are presented in the technical appendix. We will now start by briefly discussing the pilot studies and the data we collected to derive the indicators.

3. IMPLEMENTATION OF PILOT SURVEYS IN TANZANIA

3.1 Overview

The Service Delivery Indicators were piloted in Tanzania in the spring/summer of 2010. The main objective of the pilots was to test the survey instruments in the field and to verify that robust indicators of service delivery quality could be collected with a single facility-level instrument in different settings. To this end, it was decided that the pilots should include an Anglophone country. The selection of Tanzania was also influenced by the presence of strong local research institutes from the AERC network: the Research on Poverty Alleviation (REPOA) in Tanzania. This research institute has extensive facility survey experience and also has grantees of the Hewlett-supported Think Tank Initiative.

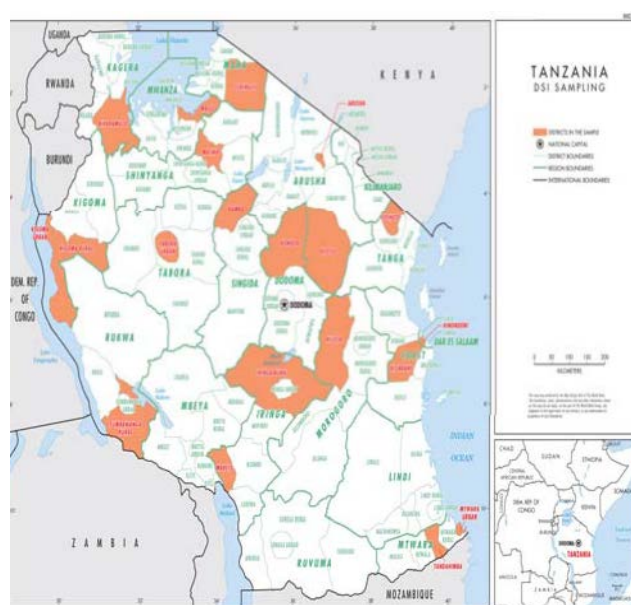
3.2 Sample Size and Design

In Tanzania, the sample was designed to provide estimates for each of the key Indicators, broken down by urban and rural location. To achieve this purpose in a cost-effective manner, a stratified multi-stage random sampling design was employed.⁴ Given the overall resource envelope, it was decided that roughly 180 units would be surveyed in both sectors in Tanzania. The sample frames employed consisted of the most recent list of all public primary schools and public primary health facilities, including information on the size of the population they serve. Table 2 reports summary statistics of the final sample and Figure 1 illustrates the stratification choices.

Table 2: Final sample of facilities by sector in the pilot countries

	Rural	Urban	Total
Health	135	40	175
Education	132	48	180

⁴ Details about the sampling design are provided in the technical appendix.

Figure 1: Map of the sampling areas

3.3 Survey Instruments and Survey Implementation

The survey used a sector-specific questionnaire with several modules (see Table 3), all of which were administered at the facility level. The questionnaires built on previous similar questionnaires based on international good practice for PETS, QSDS, SAS and observational surveys. A pre-test of the instruments was done by the technical team, in collaboration with the in-country research partners, in the early part of 2010. The questionnaires were translated into Swahili for Tanzania.

In collaboration with the in-country research partners, members of the technical team organized a one-week training session, which included three days of testing the instruments in the field. The enumerators and supervisors were university graduates, and in many cases were also trained health and education professionals (teachers, doctors, and health workers) with previous survey experience.

In Tanzania, data collection was carried out by 32 enumerators (16 in each sector) organized into 8 field teams (4 in each sector). Each team consisted of a team leader, 3 enumerators, and a driver. Four senior staff members from REPOA coordinated and supervised the fieldwork. Fieldwork in both education and health started in April 2010 and was completed within a month.

All questionnaires collected during fieldwork were periodically brought from the field to the local partners' headquarters (in Dar es Salaam for REPOA) for verification and processing. In Tanzania, the data were processed by a team of five data entry operators and one data entry supervisor, and entered using CSpro. Data entry lasted 20 days commencing in late May 2010.

Table 3: Instrument modules

Education		Health	
Module	Description	Module	Description
Module 1: Administered to the principal, head teacher or most senior teacher in the school	Self-reported and administrative data on school characteristics, students, teachers and resource flows.	Module 1: Administered to the in-charge or the most senior medical staff at the facility.	Self-reported and administrative data on health facility characteristics, staffing, and resources flows.
Module 2: Administered to (a maximum of) 10 teachers randomly selected from the list of all teachers	Delays in the receipt of wages	Module 2: Administered to (a maximum of) 10 medical staff randomly selected from the list of all medical staff	Delays in the receipt of wages
Module 3: Administered to the same 10 teachers as in module 2	An unannounced visit about a week after the initial survey to measure the absence rates	Module 3: Administered to the same 10 medical staff as in module 2	An unannounced visit about a week after the initial survey to measure the absence rates
Module 4: Classroom observations	Based on 2 observed lessons for grade 4 in either English/French or math. Each observation lasts for 40 minutes	Module 4: Health facility observations	Time use per patient. Based on observations for two hours or at least of 15 patients.
Module 5: Test of teachers	Test of all (a maximum of 10) grade 3-4 teachers in mathematics language and pedagogy to measure teachers' knowledge.	Module 5: Test of health workers. Patient case simulations.	Test of 1-2 medical staff per facility to assess clinical performance.
Module 6: Test of grade 4 children	A test in math and language administered one-on-one to 10 randomly selected grade 4 students to measure learning achievement.		

4. INDICATORS AND PILOT RESULTS

4.1 Overview

This section presents the findings of the pilot surveys in education and health in Tanzania. We report results for each country as a whole, as well as breakdowns by rural and urban locations. While further breakdowns are possible (for example, by geographical area), the Indicators pilot did not seek to generate statistically significant data for these subgroups. As a result, for most indicators, these estimates are not necessarily meaningful.

Sampling weights are taken into account when deriving the estimates (and standard errors), and the standard errors are adjusted for clustering.⁵

4.2 Education

At the School

Infrastructure (electricity, water, sanitation)

Schools often lack basic infrastructure, particularly schools in rural areas. The indicator, *Infrastructure*, accounts for the three basic infrastructure services: availability of electricity (in the classrooms), clean water (in the school) and improved sanitation (in the school). The data are derived from the head teacher questionnaire. While these data are self-reported, our assessment is that the quality of the data is good and the biases are likely to be minimal.

Table 4: Infrastructure (% of schools with electricity, water and sanitation)

All	Rural	Urban
0.03 (0.01)	0.02 (0.01)	0.08 (0.08)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 180 observations, of which 45 are urban schools.

Results are reported in table 4. The infrastructure indicator measures if the school has access to basic infrastructure (= 1); i.e. access to electricity, clean water and improved sanitation, or if they lack one or more of them (= 0). On average, only 3% of the schools in Tanzania have access to basic infrastructure services. Electricity is the key constraint, as just about 20 percent of the schools have access to it.

⁵ Details are provided in the technical appendix.

Looking at the rural-urban breakdown, it is worth noting that the outcome in Tanzania is poor in both urban and rural areas.

Children per Classroom

The indicator, *Children per Classroom*, is measured as the ratio of the number of primary school children to available classrooms. The source for the data is the school enrollment list (for students) and reported classrooms (by the headmaster). Our assessment is that the quality of the data is good, although the enrollment lists may not always be up-to-date.⁶ Table 5 summarizes the results.

Table 5: Children per Classroom

All	Rural	Urban
74.05 (5.29)	70.47 (5.32)	92.51 (12.56)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 180 observations, of which 45 are urban schools.

In Tanzania, urban schools have approximately 92 students per classroom while rural schools have approximately 70.

Student-Teacher Ratio

Teacher shortage is a problem in many developing countries, especially in poor and rural areas. The indicator, *Student-Teacher Ratio*, is measured as the average number of students per teacher. The data on teachers is from the head teacher questionnaire and codes all teachers listed to be teaching. Our assessment is that the quality of the data is good, although the enrollment lists may not always be up-to-date, as noted above. The results are reported in Table 6.

Table 6: Student-Teacher Ratio

All	Rural	Urban
48.71 (2.20)	50.56 (2.47)	39.13 (3.12)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 180 observations, of which 45 are urban schools.

The student-teacher ratio is 48.7 students per teacher in Tanzania. Urban schools in Tanzania have approximately ten students less per teacher than rural schools.

⁶ Enrollment numbers may suffer from over-reporting biases if schools have incentives to report higher enrollment figures in order to attract more funds.

Textbooks per Student

Lack of basic education material may also be an important constraint for learning faced by children and teachers in many developing countries. The indicator, *Textbooks per Student*, is measured as the overall number of textbooks available within primary schools per student. To calculate the indicator, we sum all books per grade and then sum over all grades. Not all schools could report breakdowns of books per grade and subject. In this case, we used data on the reported number of books in total (for a grade).⁷

Measurement errors in the number of books are likely to be an issue, although the enumerators were asked to verify the reports using school records (if available). We do not believe these measurement errors are systematically different in the two countries, thus the cross-country comparison should still be valid.

The results are reported in Table 7.

Table 7: Textbooks per student

All	Rural	Urban
0.94 (0.08)	0.95 (0.09)	0.90 (0.17)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 179 (164 for Language books) observations, of which 44 (43) are urban schools.

Tanzanian children have access to less than a book per child; there are few differences between urban and rural areas.

⁷ As number of subjects (and potentially therefore also the number of books) may differ across countries, it would make sense to (also) report disaggregated estimates for number of mathematics and language books per student. However, records of books per grade and subject were not available for enough schools in the two samples.

Teachers

Absence Rate

In many countries, highly centralized personnel systems, inadequate incentives, and weak local accountability have resulted in high levels of staff absence. The indicator, *Absence Rate*, is measured as the share of teachers not in schools as observed during one unannounced visit.⁸

For cross-country comparisons, we believe the data is of good quality. However, because the information is based on one unannounced visit only, the estimate for each school is likely to be imprecisely measured. By averaging across schools, however, these measurement error problems are likely to be less of a concern. Results are reported in Table 8.

Table 8: Absence Rate

All	Rural	Urban
0.23 (0.02)	0.20 (0.02)	0.36 (0.04)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 180 observations, of which 45 are urban schools.

About one in four teachers in Tanzania, are absent from school on any given school day. Interestingly, the absence rate in urban schools is significantly higher than in rural schools.

Even if at school, however, the teachers may not be in the classroom teaching. As a complementary indicator, we therefore also report absence from the classroom.⁹

Results are reported in Table 9. Findings on absence from the classroom, in Tanzania, are striking. Even when in school, the teacher is absent from the classroom more than half the time. Again, absenteeism is significantly higher in urban schools than in rural schools in Tanzania.

Table 9: Absence rate from classroom

All	Rural	Urban
0.53 (0.03)	0.50 (0.02)	0.68 (0.05)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 179 observations, of which 45 are urban schools.

⁸ In the first (announced) visit we randomly selected 10 teachers from the list of all teachers. We checked the whereabouts of these 10 teachers in the second, unannounced, visit.

⁹ This indicator is also derived using data from the unannounced visit, as the enumerators were also asked to verify if teachers present in the school were actually in the classroom.

Time Children are in School Being Taught

The staff absence survey, together with classroom observation, can also be used to measure the extent to which teachers are in the classroom teaching, broadly defined. In other words, it can be used to measure the indicator, *Time Children are in School Being Taught*. To this end, we start by calculating the scheduled hours of teaching. We then adjust the scheduled time for the time teachers are absent from the classroom on average (this data is reported separately in Table 10). Finally, from the classroom observation sessions we can measure to what extent the teacher is actually teaching when he/she is in the classroom. Here, we use information from the classroom observations done outside of the classroom. Specifically, the enumerator recorded every 5 minutes (for a total of 15 minutes) if the teacher remained in the classroom to teach, broadly defined, or if he/she left the classroom.

As the information is based on one unannounced visit and a short observational period, the estimate for each school is likely to be imprecisely measured. By taking an average across many schools, however, we believe we arrive at an accurate estimate of the mean number of hours children are being taught. We end up with a lower bound of the estimate if, as seems reasonable, the observations done outside the classroom are biased upward due to Hawthorne effects.

The results are reported in Table 10 (for all grades pooled). On average, students in primary schools in Tanzania are taught 2 hours a day, and half an hour less in urban areas. The difference between urban and rural areas is significant. Note that the scheduled time is 5 hours and 12 minutes in Tanzania.

Table 10: Time Children are in School Being Taught (per day)

All	Rural	Urban
2 h and 04 min (10 min)	2 h 11 min (10 min)	1 h 24 min (18 min)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 173 observations, of which 43 are urban schools.

Because the scheduled time differs across grades, a more accurate measure may be to look at the time children in a given grade are in school being taught. These estimates, however, mirror those of the pooled findings reported in Table 10 (results not reported).

Share of Teachers with Minimum Knowledge

Having teachers teaching, however, may not be enough if the teacher's competence (ability and knowledge) is inadequate, a major problem in several developing countries. To assess this issue, up to 10 teachers per school were administered a basic test of knowledge. The teacher test consisted of two parts: mathematics and English for Tanzania.¹⁰ Current teachers of grade 4 students and those teachers who taught the current grade 4 students in the previous year were tested. The test comprised material from both lower and upper primary school in language and mathematics. The test was administered en masse.

The test consisted of a number of different tasks ranging from a simple spelling task (involving 4 questions) to a more challenging vocabulary test (involving 13 questions) in languages and from adding double digits (1 question) to solving a complex logic problem (involving 2 questions) in mathematics.

Table 11: Share of Teachers with Minimum Knowledge and average test score in teacher test

Sample	All	Rural	Urban
Language:			
	0.11 (0.03)	0.13 (0.04)	0.05 (0.04)
Mathematics:			
	0.75 (0.03)	0.75 (0.04)	0.74 (0.06)
Average Share across both Mathematics and Languages:			
	0.42 (0.02)	0.43 (0.02)	0.40 (0.03)

Note: Dependent variable is share of teachers that managed to complete all questions on the primary language and primary mathematics curriculum, respectively. Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 504 observations from 180 schools in Tanzania (260 English teachers and 244 Mathematics teachers), of which 152 (45 schools) are from urban areas. Test scores are averaged at the school level.

While it is a matter for debate what constitutes “‘minimum’ knowledge” for a grade 3 and 4 teacher, a fairly conservative measure is that the teacher demonstrates mastery of the particular curriculum he or she teaches. Our suggested measure for the indicator, *Share of Teachers with Minimum Knowledge*, attempts to capture this. In the basic knowledge test, 14 questions were related to the lower primary curriculum on the language test and 5 questions were related to the primary mathematics curriculum. We define mastery of the primary curriculum as answering all of these questions correctly and derive then the share of teachers that correctly manages to do so. To be precise, for the language section, we derive the share of language teachers who were able to answer all questions correctly. For the mathematics section, we derive the share of mathematics teachers who were able to answer all the questions correctly.¹¹ Of course the content of the lower primary curriculum may vary slightly across countries. We here define lower primary curriculum as all the questions that test basic competencies; i.e. those that were included in the student test.

¹⁰ The test also included a pedagogic section that we do not report on.

As evident from Table 11, only 1 in 10 teachers in Tanzania manage to complete all the questions on the primary language curriculum.¹² For mathematics, the picture is somewhat less bleak, with 3 out of 4 teachers managing to complete all questions on the primary mathematics curriculum. As reported in the last set of rows of Table 11, this implies that on average about 40% of teachers in Tanzania display minimum knowledge. There are no significant differences between urban and rural schools.

Another way to look at the results based on the lower primary curriculum is to assess the results on specific questions. Table 12 reports the findings.

Strikingly, 2 out of 10 teachers in Tanzania struggle to spell simple words; 5 out of 10 could not identify a noun, and 1 in 10 teachers tested failed to correctly subtract double-digit numbers. With the exception of the noun task, there is no significant differences between urban and rural schools here.

Table 12: Scores on particular questions on the tests¹³

Sample	Tanzania
Average score on spelling test	0.82 (0.03)
Share of teachers who could identify a noun	0.51 (0.04)
Share of teachers that could subtract two double-digits numbers	0.90 (0.03)
Share of teachers that could divide two fractions	0.66 (0.04)

Note: Dependent variable is share of teachers that managed to complete all questions on the primary language and primary mathematics curriculum, respectively. Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 504 observations from 180 schools in Tanzania (260 English teachers and 244 Mathematics teachers), of which 152 (45 schools) are from urban areas. Test scores are averaged at the school level.

¹¹ We tested all the teachers in both language and mathematics. However, *all* test statistics we report are based on teachers in the respective subjects only.

¹² With a somewhat more lenient definition of answering 90% or more questions correctly (for language), the numbers jump to 38% in Tanzania.

Funding

Education Expenditures Reaching Primary Schools

The indicator, *Education Expenditures Reaching Primary Schools*, assesses the amount of resources available for services to students at the school. It is measured as the recurrent expenditure (wage and non-wage) reaching the primary schools per primary school age student in US dollars at Purchasing Power Parity (PPP). Unlike the other indicators, this indicator is not a school-specific indicator. Instead, we calculate the amount reached per surveyed school, and then use the sample weights to estimate the population (of all schools) in aggregate.¹⁴

Measuring effective education expenditures reaching primary schools is a challenging task, since resource systems and flows differ across countries. To fully account for the flow of resources reaching the schools from all government sources and programs, schools need to have up-to-date and comprehensive records of inflows. This is not the case in many schools, likely causing us to misinterpret, in some cases, poor records for lack of resources reaching the school. The results are reported in Table.

Table 13: Education expenditures reaching primary schools per primary school age student

All	Rural	Urban
124.54	131.97	99.41

Note: Education expenditures reaching primary per primary school age children in US\$PPP. The estimates are based on data from 180 observation, of which 45 are urban schools.

The amount of recurrent funds (wage and non-wage) reaching primary schools in Tanzania was US\$ 124.54 PPP (per primary school-age student). Rural schools in Tanzania, on average, receive more than their urban counterparts.

The estimates in Table 13 are likely driven both by budget decisions at the central level and the efficiency with which budgeted resources are made available to primary schools. For Tanzania, we can derive an estimate of the latter effect; i.e., the efficiency of the supply chain, by estimating resource leakage in one of the support programs for primary schools (a capitation grant program).¹⁵ The capitation grant is based on the number of pupils attending school and is mainly intended for books and school supplies. As depicted in Table 14, leakage, defined as the share of resources intended for schools, but not received by them, represents 37 percent of the capitation grant budget. Leakage is higher, but not significantly so, in urban areas. Such high levels of resource leakage could potentially have serious consequences for service quality.

¹³ For the spelling question, the teacher had to choose the correct set of letters to fill in the blanks in a list of words. For identifying a noun, the teacher was given a word and asked to identify which parts of speech a particular word belonged to from a given set of options. For the mathematics question, the teacher was asked to subtract two double-digit numbers (i.e. 87-32) and divide two fractions ($3/4 \div 5/8$).

¹⁴ The source for the number of primary school age children, broken down by rural and urban location, is Ministry of Education and Vocational Training (2010) for Tanzania. Quantities and values of in kind items were collected as part of the survey. In cases where values of in kind items were missing, average unit cost was inferred using information from other surveyed schools.

Table 14: Leakage of capitation grant

All	Rural	Urban
0.37 (0.03)	0.36 (0.03)	0.41 (0.02)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 173 observations, of which 41 are urban schools.

Delays in Salaries

The indicator, *Delays in Salaries*, which may have an adverse effect on staff morale and therefore on the quality of service, is measured as the proportion of teachers whose salary has been overdue for more than two months. The data is collected directly from teachers at the school and we believe the data is of good quality. The results are reported in Table 15.

Table 15: Delays in Salaries

All	Rural	Urban
0.02 (.005)	0.02 (.005)	0.006 (.004)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 174 observations, of which 43 are urban schools.

Significant (over two months) delays in salaries do not appear to be a common problem In Tanzania, about 2% of the teaching staff report more than 2 months' delay in salary, and this happens exclusively in rural schools.

¹⁵ Leakage is not included in the Indicators, since we can only measure it for the subset of resources that are allocated by a fixed rule, and not those that are based on bureaucratic discretion.

4.3 Health

At the Clinic

Health clinics often lack basic infrastructure, particularly in rural areas. Access to electricity is important for operating health equipment. Similarly, availability of clean water and sanitation facilities are fundamental for quality services. The indicator, *Infrastructure*, is created in the same way as the parallel indicator for education.

Results for Tanzania are reported in Table 16. On average, only 19 percent of the primary health facilities in Tanzania have access to basic infrastructure.

Table 16: Infrastructure (% facilities with electricity, clean water and improved sanitation)

All	Rural	Urban
0.19 (.07)	0.05 (.02)	0.60 (.13)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 171 observations, of which 40 are urban health facilities.

There are also significant differences in infrastructure availability within countries. While in urban areas, about 60% of facilities in Tanzania have access to electricity, water, and sanitation, this proportion is close to zero for rural areas.

Medical Equipment per Clinic

The lack of basic medical equipment is often a constraint to quality health care. The indicator, *Medical Equipment per Clinic*, is measured as the share of primary care providers that have the following basic equipment available: thermometer, stethoscope, and weighing scale. As with the infrastructure indicator, these data are self-reported. There is a concern that the head of the facility reports availability of medical equipment, even if it may not be fully functional, in which case our results provide an upper bound. Apart from this concern, our assessment is that the quality of the data is good.

Results for Tanzania are reported in Table 17. This indicator measures the health facility's access to all three pieces of equipment (= 1) or lack of one or more of them (= 0). On average, three quarters of the primary health facilities in Tanzania have access to the basic equipment. Or in other words, roughly 2 out of 10 clinics in Tanzania do not have access to the most basic health equipment.

Table 17: Medical equipment per clinic

All	Rural	Urban
0.78 (.04)	0.76 (.05)	0.83 (.04)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 174 observations, of which are 40 urban health facilities.

Rural health clinics in Tanzania are less likely to have access to basic health equipment on average, but the difference is not significant.

Stock-out of drugs

The lack of essential drugs is often a constraint to quality health care. The indicator, *Stock-out of drugs*, is measured as the share of 15 basic drugs which, at the time of the survey, were experiencing stock-out in the primary health facilities. Results for Tanzania are reported in Table 18.

Table 18: Stock-out of drugs

All	Rural	Urban
0.24 (.02)	0.24 (.03)	0.23 (.03)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 175 observations, of which 40 are urban health facilities.

Stock outs of essential drugs is common in Tanzania with about one quarter of the main drugs being out of stock at the moment of the survey.

Medical Personnel

Absence Rate

The indicator, *Absence Rate*, is measured as the share of health staff not in the clinic as observed during one unannounced visit. Our concern with the quality of the data is the same as that for the absence rate indicator in education. The results are reported in Table 19.

Table 19: Absence Rate

All	Rural	Urban
0.21 (.03)	0.17 (.03)	0.33 (.04)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 175 observations, of which 40 are urban health facilities.

We observe that absenteeism is widespread. While one fifth of the health workers are not in the clinic during the random spot check, the ratio reaches one third in urban areas in Tanzania and is significantly higher than in rural areas.

Diagnostic Accuracy in Outpatient Consultations

The indicator, *Diagnostic Accuracy in Outpatient Consultations*, is measured through Patient Case Simulations (PCS, also called “vignettes”). With this methodology, one of the surveyors acts as a case study patient with some specific symptoms. The clinician who is informed of the simulation is asked to proceed as if the enumerator is a real patient, while another enumerator acts as an observer. High quality performance in outpatient consultations entails at least the following: (i) To systematically arrive at a correct diagnosis (or preliminary diagnosis); (ii) To provide an appropriate treatment (or referral); and (iii) To reveal important information to the patient about which actions to take (e.g., how to take the medicine, what to do if the patient does not get better, etc.). The methodology presents several advantages: (a) All clinicians are presented with the same case study patients, thus making it easier to compare performance across clinicians; (b) The method is quick to implement, and does not require waiting for patients with particular diagnoses; (c) We avoid intrusion and ethical issues that would arise if we were studying real patient cases. The method also has its drawbacks. The most important one is that the situation is not a real one and that this may bias the results.¹⁶

The Indicators pilot used five PCSs: (i) Malaria with anemia; (ii) Diarrhea with severe dehydration; (iii) Pneumonia; (iv) Pelvic inflammatory disease; and (v) Pulmonary tuberculosis.¹⁷

There are a number of ways of scoring performance in a PCS and of aggregating the scores across PCSs. The indicator proposed here focus on diagnostic accuracy. Diagnostic accuracy is scored 1 if the correct diagnosis is reached, otherwise zero, and the indicator of diagnostic accuracy is the average score of the five PCSs.

We also report results for process quality, measured based on the share of relevant history taking questions and the share of relevant examinations performed, giving equal weight to both components.¹⁸

The results are reported in tables 20 and 21.

As evident from the last column in Table 20, clinicians in Tanzania reached the correct diagnosis 57% of the cases. Behind these figures is considerable variation across the five different patient cases. In Tanzania, the share of clinicians who made the correct diagnosis for the case of malaria with anemia was 27%; for the case of diarrhea with severe dehydration it was 29%; for the case of pneumonia it was 84%; for the case of pelvic inflammatory disease it was 66%, and for the case of tuberculosis it was 73%. It is particularly worrying that so few clinicians are able to discover the severe and potentially deadly conditions of patients with malaria and diarrhea.

Table 20: Share of clinicians who reached correct diagnosis

Case	Malaria with anemia	Diarrhea with severe dehydration	Pneumonia	Pelvic inflammatory disease	Pulmonary tuberculosis	Diagnostic accuracy (mean)
	0.27 (.046)	0.29 (.047)	0.84 (.034)	0.66 (.059)	0.79 (.037)	0.57 (.030)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 224 observations from 174 health facilities, of which 57 observations from 40 urban health facilities.

¹⁶ Comparisons of Patient Case Simulations with Direct Observation of real patients in low income contexts have revealed that performance scores typically are higher with Patient Case Simulations, but that the correlation between the two measures is substantial (e.g., Das, Hammer, and Leonard, 2008). Some authors have interpreted the score of Patient Case Simulations as a measure of competence or ability rather than actual performance (Das and Hammer, 2005, Leonard et al., 2007). As discussed in the Appendix, there is reason to believe that Patient Case Simulations measure a blend of competence and actual performance, and that the blend depends on the actual design and framing of the tool. The Patient Case Simulations used in the Indicators pilot were framed to resemble actual performance as closely as possible. Nevertheless, one should be aware of a potential upward bias of the *absolute* performance levels. As a measure of *relative* performance, though, we believe that Patient Case Simulations have considerable merit.

¹⁷ These PCS were originally developed by Leonard and Masatu (2007) for Tanzania. We expanded the list of relevant items to be recorded by including items required by the guidelines for Integrated Management of Childhood Illnesses (IMCI) in cases where the patient was a child. These modified PCSs have previously been implemented in Tanzania by Mæstad and Mwisongo (unpublished).

¹⁸ See technical appendix for a more comprehensive discussion on the PCS methodology.

Diagnostic accuracy is higher in urban than in rural areas, but the difference is statistically significant in Tanzania only (see Table 21).

Table 21: Diagnostic accuracy, process quality and the aggregate performance score

	All	Rural	Urban
Diagnostic Accuracy	0.57 (.030)	0.53 (.031)	0.68 (.037)
Process Quality	0.35 (.021)	0.31 (.015)	0.44 (.034)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 224 observations from 174 health facilities in Tanzania, of which 57 observations from 40 urban health facilities.

In Tanzania, clinicians performed on average 35 percent of the questions and examinations relevant for the five PCSs. Process quality is higher in urban than in rural areas. The differences in process quality may be part of the explanation for why there are such large differences in diagnostic accuracy.

Time Spent Counseling Patients per Clinician

The indicator, *Time Spent Counseling Patients per Clinician*, is based on aggregating data from the observational study of medical personnel. In the observational study, the clinician is observed during a two-hour period. By combining data on number of patients treated per day with the observational data on the time spent on each patient, we calculate the total time spent counseling patients per day in the clinic. As the number of clinicians differs across clinics, we normalize the time spent using the number of clinicians, present at the time of the interview, who perform consultations. We then arrive at an estimate of the time spent counseling patients per clinician (at each clinic). Because of the short observational period (two hours), Hawthorne effects may bias the results upward. Poor outpatient records may also affect the precision of the estimate. We do not, however, believe that our estimate is downward-biased.

The results are reported in Table 22.

Table 22: Time Spent Counseling Patients per Clinician (per day)

All	Rural	Urban
29 min (4 min)	26 min (4 min)	36 min (11 min)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 165 observations, of which 39 are urban health facilities.

On average, the time spent counseling patients per clinician in Tanzania is only 29 minutes per day. There is a small and insignificant difference between urban and rural areas.

Funding

Health Expenditure Reaching Primary Clinics

The indicator, *Health Expenditure Reaching Primary Clinics*, captures the resources available to frontline providers. It is measured as the per capita recurrent expenditure (wage and non-wage) reaching the frontline provider in US dollars at Purchasing Power Parity (PPP). As with the education indicator, this indicator is not a clinic-specific indicator. The indicator is created by summing, using the sample weight, the measured amount of resources received per surveyed clinic into a population aggregate.¹⁹

It is important to note that to fully account for the flow of resources reaching the clinics, from all government sources and programs, clinics need to keep adequate records of inflows. This is not the case in many clinics, likely causing us to misinterpret, in some cases, poor records for lack of resources reaching primary clinics. The results are depicted in Table 23.

We observe that the recurrent funds (wage and non-wage) reaching frontline facilities is approximately US\$7.01 PPP per capita in Tanzania. Furthermore, urban clinics in Tanzania receive more per capita resources than rural clinics.

Table 23: Primary Health Expenditure per capita Reaching Primary Clinics

All	Rural	Urban
7.01	5.58	11.15

Note: Health expenditures reaching clinics per capita in US\$PPP. The estimates are based on 175 observations, of which 40 are urban health facilities.

Delays in Salaries

The indicator, *Delays in Salaries*, measures the proportion of health workers whose salary is overdue for more than two months. The data is collected directly from health workers at the clinic, and we believe the data is of good quality. The results are reported in Table 24. We observe that only 2 percent of the health personnel in Tanzania report at least a two-month delay in receiving their salary.

¹⁹ The source for the population data is WDI (2010). Quantities and values of in kind items were collected as part of the survey. In cases where values of in kind items were missing, average unit cost was inferred using information from other surveyed clinics.

Table 24: Delays in Salaries

All	Rural	Urban
0.02 (.01)	0.02 (.02)	0.03 (.02)

Note: Share of health workers whose salary is over 2+ months. Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 172 observations, of which 38 are urban health facilities.

5. OUTCOMES: TEST SCORES IN EDUCATION

To avoid making structural assumptions about the link between inputs, performance, and outcomes, we do not suggest that outcomes should be part of the Service Delivery Indicators survey. However, it may make sense to report separately on outcomes when the various sub-indicators and the potential aggregate index are presented. In health, there are measures for many countries at the national level, such as under-five mortality rates, but no indicator that can be linked directly to the service quality of individual facilities. Quantity outcomes in education are also available (various measures of flows and stock of schooling) for a large subset of countries. However, on quality there are no comparable data available, at least not for multiple countries. Thus, student learning achievement has been collected as part of the survey in education.

Available evidence indicates that the level of learning tends to be very low in Africa. For instance, assessments of the reading capacity among grade 6 students in 12 eastern and Southern African countries indicates that less than 25 percent of the children in 10 of the 12 countries tested reached the desirable level of reading literacy (SACMEQ, 2000-2002). As part of this survey, learning outcomes were measured by student scores on a mathematics and language test.

Table 25: Average score on student test

Sample	All	Rural	Urban
Language			
	0.43 (0.02)	0.41 (0.02)	0.52 (0.03)
Mathematics			
	0.39 (0.02)	0.38 (0.02)	0.48 (0.03)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 1787 observations from 180 schools, of which 449 (45 schools) are from urban areas. Test scores are averaged at the school level.

We test younger cohorts partly because there is very little data on their achievement, partly because SACMEQ already tests students in higher grades, partly because the sample of children in school becomes more and more self-selective as we go higher up due to high drop-out rates, and partly because we know that cognitive ability is most malleable at younger ages (see Heckman and Cunha, 2007).

For the pilots, the student test consisted of two parts: language (English), and mathematics. Students in fourth grade were tested on material for grades 1, 2, 3 and 4. The test was designed as a one-on-one test with enumerators reading out instructions to students in their mother tongue. This was done so as to build up a differentiated picture of students' cognitive skills. Results of the grade 4 student test are presented in Table 25.

The average score on the test was just over 40 percent in Tanzania, for the language section and 45% for the mathematics section.²⁰ In Tanzania English is only introduced as the medium of instruction in grade 3. As expected, rural schools score significantly worse than urban schools.

Table 26: Language: Percentage of student who can read a sentence (in French/English)

All	Rural	Urban
0.06 (0.01)	0.06 (0.01)	0.10 (0.03)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 1787 observations from 180 schools, of which 449 (45 schools) are from urban areas. Test scores are averaged at the school level.

While the mean score is an important statistic, it is also an estimate that by itself is not easy to interpret. Table 26 depicts a breakdown of the results. As is evident, reading ability is low. In fact, only 6 percent of students in Tanzania are able to read a sentence.²¹ In mathematics, 83% of Tanzanian students can add two single digits. Again, as expected, rural schools perform significantly worse than urban ones. For a more detailed description of performance on various tasks, see the technical appendix.

²⁰ The test consisted of a number of different tasks ranging from a simple task testing knowledge of the alphabet (involving 3 questions) to a more challenging reading comprehension test (involving 3 questions) in languages and from adding 2 single digits (1 question) to solving a more difficult sequence problem (1 question) in mathematics. Just as for the teacher test, the average test scores are calculated by first calculating the score on each task (given a score between 0-100%) and then reporting the mean of the score on all tasks in the language section and in the mathematics section respectively. Since more complex tasks in the language section tended to involve more questions, this way of aggregation gives a higher score than simply adding up the score on each question and dividing by the total possible score. Following this latter method of aggregation would lead to a roughly 8-10% lower score in the language section. In the mathematics section the simpler tasks involved more questions, therefore aggregating by task gives a slightly lower score than simply adding up the score on all the questions (roughly 5 %).

Table 27: Mathematics: Percentage of student who can add two single digits

All	Rural	Urban
0.83 (0.02)	0.81 (0.03)	0.93 (0.02)

Note: Weighted mean with standard errors adjusted for weighting and clustering in parenthesis. 1787 observations from 180 schools, of which 449 (45 schools) are from urban areas. Test scores are averaged at the school level.

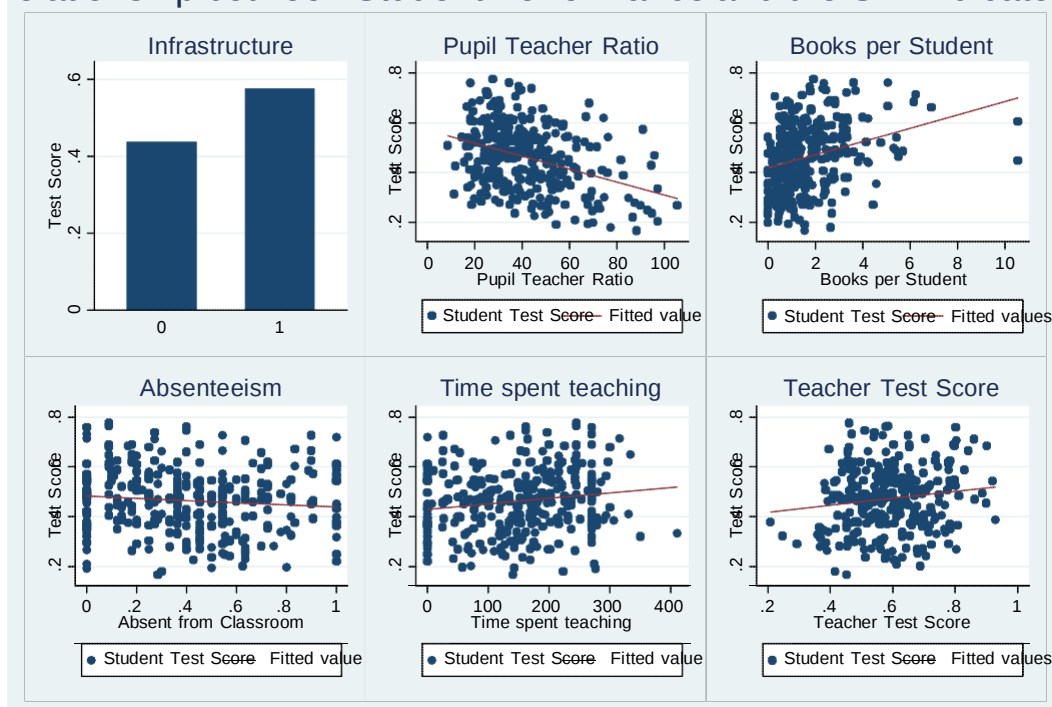
The Service Delivery Indicators are a measure of inputs (including effort), not of final outcomes. Nevertheless, in the final instance, we should be interested in inputs not in and of themselves, but only in as far as they deliver the outcomes we care about. Given that we have collected outcome data in education, we can also check whether our input measures are in some way related to outcomes. Of course, these are mere correlations that cannot be interpreted causally, but we still believe that it is interesting to examine how our Indicators correlate with educational achievement. Figure 21 depicts unconditional correlations between student achievement and the education indicators, where the data from each country is pooled. Interestingly – and across the board – there are fairly strong relationships between the indicators and student knowledge, with all the correlations having the expected sign.²²

²¹ The reading task consisted of reading a sentence with 11 words in Tanzania. We have defined the percentage of students who can read a sentence correctly as those who can read all words correctly. With a somewhat more lenient definition of being able to read all but one word, the numbers rise to 48% and 11%.

²² Results are similar when running a regression of student test score separately on each indicator, a country dummy and a rural/urban dummy.

Figure 21: Relationship between student performance and the education Indicators

Relationship between Student Performance and the SDI indicators



6. INDICATOR AGGREGATION PROCESS AND COUNTRY RANKINGS

The choice of a small set of easily interpretable indicators makes it possible to focus both on a direct comparison of the various sub-indicators as well as an aggregate score. In general, the advantage of collapsing the indicators into just one is clear – it provides an unambiguous and easy-to-understand country ranking. It also potentially reduces concerns about measurement errors in the indicator series. However, as there is no clear theoretical foundation to guide the aggregation, the deeper meaning and interpretation of the ranking is to some extent unclear. Moreover, by appearing precise and certain, the underlying uncertainty is often ignored, even when reported with confidence intervals. In the case of these Service Delivery Indicators, the advantage of aggregation is also less clear because of the short list of indicators that have been identified. Moreover, the Indicators themselves provide direct and interpretable evidence on service performance, further weakening the argument for aggregation.

If a composite index is nonetheless the preferred choice, we propose to use fixed weights when weighting the different indicators. Equal weights are the most commonly used approach for index construction. While there is no analytical reason for this choice, equal weights are a common approach that is often justified on the basis of transparency (see Decanq and Lugo (2008), whose notation we follow here, for a more detailed discussion).

We also suggest, after reversing sign of indicators for which a higher value implies a worse outcome, to use $\Phi(x_k) = [x_k - \min(x_k)] / [\max(x_k) - \min(x_k)]$; i.e., the min-max transformation function to transform each variable into a uniformly distributed variable between 0 and 1.

For each category c , we can then create a sub-indices X_1 to X_3 , where,

$$X_c = \frac{1}{K_c} \sum_1^{K_c} \Phi(x_k | category = c), \quad (3)$$

where K_c is number of dimensions in category c , and $K = K_1 + K_2 + K_3$.

Finally, we can take an equally weighted average of all the dimensions across the sub-indices to get at a Service Delivery Indicator ranking within a sector; i.e.,

$$SDI = \frac{1}{K} \left[\sum_1^{K_1} \Phi(x_k | category = 1) + \sum_1^{K_2} \Phi(x_k | category = 2) + \sum_1^{K_3} \Phi(x_k | category = 3) \right], \quad (4)$$

to get a final index for a given sector.

We can go one step further and derive an overall score and ranking of countries based on the two service delivery sectors.

Table 28 illustrates the methodology for Tanzania. The score for Tanzania, is slightly higher for health. The aggregate Service Delivery Index score is 0.45 for Tanzania.

Table 28: Average scores in total, by sectors and categories

Sample	All	Education				Health services			
		At school	Personnel	Funds	Total	At clinic	Personnel	Funds	Total
	0.45	0.22	0.48	0.61	0.44	0.40	0.44	0.53	0.46

7. LESSONS LEARNED, TRADE-OFFS, AND SCALE-UP

The pilot of the Service Delivery Indicators project in Tanzania demonstrates that this methodology is capable, through a single set of instruments and at a single point of collection, to provide information to construct a set of indices for benchmarking service delivery performance in education and health in Africa. The survey instruments used in the pilots can, with some modifications, be employed when scaling up the project. In what follows, we outline some of the important issues that have arisen during the pilot phase and the trade-offs that would need to be considered when the project is scaled up.

7.1 Sample Size and Sample Strategy

To be credible, the Indicators must be representative of the population in question. In general, being representative is a statement of the type that with 95% certainty, the true value of a parameter is within a distance e of the estimate. This statement can give answers to two questions: (i) What is the required sample size to ensure that with 95% certainty, the true value of a parameter is within a distance e of the estimate; and/or (ii) Given a sample size n , how precisely will the true population mean be estimated (i.e. what is the margin of error). The more precise, the tighter the confidence interval that contains the true mean with 95% probability.

To determine the required sample size, one faces a number of issues. First, the sample size depends on the sampling technique used.²³ Second, when the mean outcome (with specified precision) is not only required for the population as a whole, but also for certain subpopulations, one needs to draw a much larger sample. In the end, the choice one has to make to get precise estimates for the total population, or for subdivisions, be it urban-rural or districts/provinces, comes down to a trade-off between the cost and value of the indicators based on their planned use. (For example, will the Indicators be primarily used for doing within-country comparisons or cross-country comparisons?) Third, the sample size requirement will differ across the Indicators. The easiest way to address this problem is to specify the margin of error for the indicator that is the most vital. The desired standard of precision for the remaining indicators will then probably have to be relaxed. Finally, as the required sample size depends on the variance of the indicator in question, and as that variance is country specific, one cannot determine a precise number for the sample size that holds for all countries. In practice, however, this may be less of a concern. For practical reasons, it is probably better to set a reasonable sample size, adjust the standard of precision for the indicators accordingly, and focus on minimizing non-sampling errors.

²³ Partly to reduce costs, but also to generate statistics for population subgroups, a multi-stage sampling procedure is to be recommended. In general, stratification would tend to increase sampling precision, while clustering will tend to reduce sampling precision.

7.2 Defining the Providers

The aim of the pilot was to measure the quality of primary services. While education is easy, in health care the question of determining the unit of observation is more complex. Primary health services include outpatient consultations, family planning, maternal and child health services, etc. In most health systems, these services are provided both at lower level health facilities (health posts, dispensaries, and community health centers), as well as at the hospital level. A key question is whether all levels of providers should be included, or whether it is sufficient to include those levels that have the provision of primary health services as their main objective. One of the considerations to bear in mind is that it may add considerably to the complexity of the data collection to include the hospital level, unless there is a reasonably clear separation within hospitals between primary health services and other services.

Another factor that is important to bear in mind in this context is that the content of the services for a given type of health facility may vary considerably across countries. For instance, while a health center in some places provides only basic primary health services, it may elsewhere operate like a small hospital. The name of a certain health facility type is therefore not a useful criterion of inclusion.

Our approach was to include all lower level health facilities and then continue to include subsequently higher levels of facilities until we were confident to have covered a significant majority of people's encounters with the primary health services. At least 75 percent coverage was used as a rule of thumb. We were not able to make a quantitative assessment against this criterion but had to rely on the qualitative judgments of our country partners. In Tanzania, we included dispensaries and health centers.

7.3 Measuring Outcomes

The Indicators are designed primarily to measure inputs, including effort, in the service delivery production function. In the final instance, it is of course outcomes that we care about. The question is whether the Service Delivery Indicators pilot should seek to collect data on outcomes where such data do not exist, and where doing so would be feasible and cost-effective.

Our pilot has shown that it is possible to test for learning outcomes in the two countries by administering homogenous instruments. If this project were to be scaled up to all of Africa, the task of creating test instruments for students and teachers becomes more daunting. There are two avenues: Firstly, one could try to find the smallest common denominator of all the curricula in Africa and write a test on the basis of this. This is the strategy currently followed by SACMEQ. This is a process taking several weeks to months in the first instance, involving education experts from every country. An alternative route would be to administer a curriculum-independent test in all the countries. This may be especially appropriate for mathematics, where there is general agreement on what the core skills are.

In terms of which subjects to test, our experience from the pilot has perhaps raised questions about the correct way to test, across countries, for language skills, and this would be even more so the case if testing language skills in many countries. Students start with foreign languages at different ages and this may compromise comparability. Therefore testing in the local vernacular may be more appropriate, but of course this requires the development of different instruments for each country. Alternatively, the test could focus on mathematics only.

In terms of at what level to test, we believe that it is very beneficial to test early. The self-selection in terms of students is much less severe at early ages, so we observe a much more accurate picture of the state of learning in schools than at higher grades and, as demonstrated by the current state of research, there is simply no substitute for early childhood learning.

Finally, one would have to decide how to test. We decided to test children one-to-one and, where possible, give instructions orally in a child's mother tongue precisely to take into account the fact that reading ability may be low. Of course, this is time-intensive and therefore the sample of children we can test is small. Still, we think that the added accuracy is well worth it, and that for testing young children it is surely the optimal method.

Overall then, it may be a worthwhile addition to collect data on outcomes as well as inputs. At the very minimum, this could be done by a simple reading test (in the local language perhaps) and a set of mathematics questions, which would not require a synthesis of a large number of curricula.

7.4 Who are the Audiences?

The purpose of the Indicators is to help policymakers, citizens, service providers, donors, and other stakeholders to enhance the quality of service provision and to ultimately improve development outcomes. However, these different stakeholders may be interested in different types of information. Defining the main audience has implications for both the type of indicators that should be collected and the level at which these indicators should be statistically representative. With regard to the former, we have proposed a set of indicators that are largely context independent. If the purpose was mainly to identify service delivery constraints within a given country, it would make sense to make the indicators more context-specific. For some stakeholders, detailed within-country comparisons would provide the most value-added. For others, it may be primarily cross-country comparisons that are important.

If the aim is to generate statistics for population subgroups, the obvious follow-up question is which subgroups to focus on: Geographical areas? Ethnic affiliation? Standards of living? It is also possible that countries have different preferences over which subgroups are the most important to get specific statistics about. One solution would be to focus on the smallest set of subgroups that will be of general interest (maybe the rural-urban breakdowns) and adjust the sampling strategy and the sample size when there is high demand (within the country) for generating statistics about other subgroups. It is important to keep in mind that while a well designed sampling scheme (i.e. stratification)

can ensure that there will be enough observations to permit useful estimates for each of the groups and keep costs down, generating statistically valid subgroup estimates has potentially large cost implications.

In the final instance, the information provided by the Indicators will only be useful if it is acted on. This, of course, raises another set of questions: Who would have most incentive to act? How will this information reach them? In which format is the data actually useful? For example, if this information would be most useful in the hands of the clients (who have the most to gain from improved service delivery), it is probably also the case that clients care most about information about the actual facilities they use. There is indeed increasing evidence that providing such information leads to better outcomes, as has been shown in Björkman and Svensson (2009) and Andrabi, Das, Khwaja (2010). Of course, doing such an exercise for all countries would be prohibitively expensive. Nevertheless, where these Service Delivery Indicators can be useful is in providing clients with a benchmark that tells them where services in their country or region fall short, and what they can expect their school or clinic to deliver. Disseminating this information is obviously a daunting communications challenge. Nevertheless, in our view, it is paramount that efforts are made to disseminate the information collected to the actual users of the services.

7.5 Costing and institutional arrangement for scale-up

The Service Delivery Indicators pilot focused on public providers in health and education services. A scale-up of the project may want to revisit this choice. It would be relatively straightforward to extend the project to non-governmental service providers and maybe even for-profit ones. Collecting reliable and comparable data from higher level providers, and particularly in health services, however, is a more daunting task. Other sectors, such as water and sanitation, could also be included, although it is questionable if the focus then should be on providers. Obviously, adding additional actors, like non-governmental service providers, will have cost implications.

To provide the right incentives for reform, the Indicators need to be regularly updated. While decisions about the interval between surveys can be taken at a later stage, the incentives to reform may be stronger if governments and service providers know that updated information will be collected in the not too distant future.

Data collection for the Indicators could be organized either through a centralized or decentralized arrangement. A decentralized structure would involve identifying local teams that would take on most of the responsibilities for survey implementation and possibly also data analysis. An alternative approach is instead to rely more heavily on an expert team that would manage the project and supervise data collection activities, relying on local survey partners for the implementation of the survey. A mix between these two arrangements is probably preferable in order to ensure high within-country competence, build local ownership, and ensure that comparable data is collected both over time and across countries

Independent of the chosen approach, the experience of piloting the Indicators points to the importance of assigning a survey supervisor with international experience and local

knowledge to each country. Such a national supervisor should remain in the field for the duration of the survey, including piloting, training, and data entry.

The main cost in constructing the Indicators is related to data collection. As a benchmark, the surveys in Tanzania cost approximately USD 140,000 for both sectors. However, there are a variety of modalities for data collection, all of which have different cost implications. The survey costs incurred during the pilot included all costs incurred by the local partners for survey preparation, implementation and data entry, but does not include the cost for a survey supervisor.²⁴ To these costs, one also needs to add the costs of a core management team, and the costs for data analysis and dissemination.

²⁴ In addition to data collection, steps included a rapid data assessment phase involving a pre-test of the instruments, sample frame and sampling, obtaining necessary governmental authorizations; training and survey pilot; data collection; data entry and initial data cleaning. (See Appendix for details).

References

- Amin, Samia and Nazmul Chaudhury (2008) "An Introduction to Methodologies for Measuring Service Delivery in Education" in Amin, Samia, Das Jishnu and Marcus Goldstein (editors) *Are you Being Served? New Tools for Measuring Service Delivery*, The World Bank, Washington, D.C.
- ANSD (2008) *Estimation de la population en âge de scolarisation : projections démographiques réalisées à partir des résultats du RGPH 2002*, Agence Nationale de la Statistique et de la Démographie, Gouvernement du Sénégal, Dakar, octobre.
- Banerjee, Abhijit, Angus Deaton and Esther Duflo (2004), "Wealth, Health, and Health Service Delivery in Rural Rajasthan", *American Economic Review Papers and Proceedings* 94 (2): 326–30.
- Banerjee, Abhijit, and Esther Duflo (2005), "Addressing Absence", *Journal of Economic Perspectives* 20 (1): 117–32.
- Banerjee, Abhijit, Suraj Jacob, and Michael Kremer with Jenny Lanjouw and Peter Lanjouw (2000) "Promoting School Participation in Rural Rajasthan: Results from Some Prospective Trials," mimeo, MIT.
- Banerjee, Sudeshna, Heather Skilling, Vivien Foster, Cecilia Briceño-Garmendia, Elvira Morella, and Tarik Chfadi (2008), "Africa Infrastructure Country Diagnostic: Ebbing Water, Surging Deficits: Urban Water Supply in Sub-Saharan Africa", Background Paper 12, The World Bank, Washington D.C, June.
- Besley, Timothy and Maitreesh Ghatak (2006) "Reforming Service Delivery", *Journal of African Economies* (16): 127-156.
- Bergeron, Gilles and Joy Miller Del Rosso (2001) "Food and Education Indicator Guide" *Indicator Guides Series*, Food and Nutrition Technical Assistance (FANTA), Academy for Educational Development, Washington, DC.
- Billig, P., Bendahmane, D and A. Swindale (1999) *Water and Sanitation Indicators Measurement Guide*, Indicator Guides Series Title 2, Food and Nutrition Technical Assistance, Academy for Educational Development, USAID, June
- Björkman, Martina, and Jakob Svensson (2009), "Power to the People: Evidence from a Randomized Field Experiment on Community-based Monitoring in Uganda", *Quarterly Journal of Economics* 124 (2).
- Case, Anne and Angus Deaton (1999) "School Inputs and Educational Outcomes in South Africa," *Quarterly Journal of Economics*, 114(3): 1047-1085.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan and Halsey Rogers (2006) "Missing in Action: Teacher and Health Worker Absence in Developing Countries", *Journal of Economic Perspectives*, 20 (1): 91-116.

Cohen, Jessica Pascaline Dupas (2008), "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment", Poverty Action Lab, October.

Das Gupta M, V. Gauri, and S. Khemani (2003), "Primary Health Care in Nigeria: Decentralized Service Delivery in the States of Lagos and Kogi" Africa Region Human Development Working Paper, Series No. 70, The World Bank, Washington D.C., September.

Das, Jishnu, and Jeffrey Hammer, (2005) "Which Doctor? Combining Vignettes and Item-Response to Measure Doctor Quality," *Journal of Development Economics*, 78:348-383.

Das J, Hammer J, and Leonard K (2008). "The Quality of Medical Advice in Low-Income Countries". *Journal of Economic Perspectives*, 22(2):93-114.

Decancq K. and M.A. Lugo (2008) "Setting weights in multidimensional indices of well-being", *OPHI Working Paper No. 18*, August.

Duflo, Esther (2001) "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4): 795-814.

Duflo, Esther, Pascaline Dupas and Michael Kremer (2009) "Additional Resources versus Organizational Changes in Education: Experimental Evidence from Kenya", MIT, mimeo, May

Filmer, Deon and Lant H. Pritchett (1999) "The Impact of Public Spending on Health: Does Money Matter?" *Social Science and Medicine*, 58: 247-258.

Gauthier, Bernard (2008) "Harmonizing and Improving the Efficiency of PETS/QSDS", AFTKL, The World Bank, Washington, D.C. March, mimeo.

Gauthier, Bernard and Ritva Reinikka (2008) "Methodological Approaches to the Study of Institutions and Service Delivery: A Review of PETS, QSDS and CRCS in Africa", African Economic Research Consortium (AERC) Framework paper.

Gauthier, Bernard and Waly Wane (2009) "Leakage of Public Resources in the Health Sector: An Empirical Investigation of Chad", *Journal of African Economies* (18): 52-83

Glewwe, Paul and Michael Kremer, (2006) "Schools, Teachers, and Education Outcomes in Developing Countries," in Hanushek, E and F. Welch (editors) *Handbook on the Economics of Education*, Chap 16, North Holland.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin (2002) "Textbooks and Test Scores: Evidence from a Randomized Evaluation in Kenya," Development Research Group, World Bank, Washington, DC.

-
- Gonzalez de Asis, Maria, Donald O'Leary, Per Ljung, and John Butterworth (2008), "Improving Transparency, Integrity, and Accountability in Water Supply and Sanitation: Action, Learning, and Experiences", The World Bank Institute and Transparency International, Washington D.C., June.
- Hanushek, Eric (2003) "The Failure of Input-Based Schooling Policies," *Economic Journal*, 113(February): F64-F98.
- Hanushek, Eric and Ludger Woessman (2007) "The role of education quality for economic growth," Policy Research Working Paper Series 4122, The World Bank.
- Hoyland, B., K. O. Moene, and F. Willumsen (2010), "The Tyranny of International Index Rankings", Working Paper, University of Oslo.
- Kaufmann D. and H. Kraay (2008) "Governance Indicators: Where Are we, Where we should Be Going?" *World Bank Research Observer* (23):1-30
- Khemani, Stuti (2006), "Can Information Campaigns Overcome Political Obstacles to Serving the Poor," World Bank, Development Research Group. Washington, D.C., mimeo.
- Leonard K., and M.C. Masatu (2007). "Variation In the Quality of Care Accessible to Rural Communities in Tanzania" *Health Affairs*, 26(2).
- Leonard K., M.C. Masatu, and A. Vialou (2007) "Getting doctors to do their best", *The Journal of Human Resources* 42:682-700.
- (2008). "Moving from the lab to the field: Exploring scrutiny and duration effects in lab experiments" *Economics Letters*, 100(2):284-287.
- Maestad, Ottar, Gaute Torsvik and Arild Aakvik (2010) "Overworked? On the Relationship Between Workload and Health Worker Performance", *Journal of Health Economics* 29:686-698.
- Ministry of Education and vocational Training (2010) *Basic Statistics in Education-National*, The United Republic of Tanzania, Dar es Salaam , June
- Morella, Elvira, Vivien Foster, and Sudeshna Ghosh Banerjee, (2008) "Climbing the Ladder: The State of Sanitation in Sub-Saharan Africa," Africa Infrastructure Country Diagnostic The World Bank, Washington, D.C., June.
- OECD (2008) *Handbook on Constructing Composite Indicators: Methodology and User Guide*, Organization for Economic Co-operation and Development, Paris.
- (2009) *Measuring Government Activity*, Organization for Economic Co-operation and Development, JRC European Commission, Paris.
- Olken, Ben (2009) "Corruption Perceptions vs. Corruption Reality", *Journal of Public Economics* 93 (7-8): 950-964.

-
- Ravallion, M., (2010) "Mashup Indices of Development", Policy Research Working Paper 5432, World Bank, Washington DC.
- Reid, Gary J. (2008) "Actionable Governance Indicators: Concept and Measurement" Administrative and Civil Service Reform (ACSR) Thematic Group, The World Bank, Washington DC, February, mimeo
- Reinikka, Ritva and Jakob Svensson (2004) "Local Capture: Evidence from a Central Government Transfer Program in Uganda", *Quarterly Journal of Economics*, 119 (2): 679-705.
- (2005), "Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda." *Journal of the European Economic Association*, 3 (2-3): 259-267.
- (2006) "How corruption affects service delivery and what can be done about it" in Susan Rose Ackerman (ed) *International Handbook on the Economics of Corruption*, 441-446, Edward Elgar Publishing.
- (2010) "Working for God? Evidence from a Change in Financing of Nonprofit Health Care Providers in Uganda." *Journal of the European Economic Association*, 8 (6).
- SACMEQ (2000-2002), Southern and Eastern Africa Consortium for Monitoring Educational Quality, www.sacmeq.org
- Samuel, Paul (2002), Holding the State to Account: Citizen Monitoring in Action, Bangalore: Books for Change.
- Tan, Jee-Peng, Julia Lane, and Paul Coustere (1997) "Putting Inputs to Work in Elementary Schools: What Can Be Done in the Philippines?" *Economic Development and Cultural Change*, 45(4): 857-879.
- UNESCO (2009) Education For All Global Monitoring Report 2009: Overcoming inequality: why governance matters, UNESCO publishing and Oxford University Press.
- WHO (2006), The African Regional Health Report 2006: The Health of the People, The World Health Organization, Washington, D.C.
- WHO (2008), UN Water Global Annual Assessment of Sanitation and Drinking Water, Geneva
- WHO/UNICEF (2008) *Progress on Drinking Water and Sanitation: Special focus on sanitation*, Joint Monitoring Programme for Water Supply and Sanitation (JMP), UNICEF New York, WHO Geneva.
- World Bank (2003) *World Development Report 2004: Making Services Work for Poor People*, The World Bank and Oxford University Press, Washington, DC.
- (2006), Getting Africa on Track to Meet the MDGs on Water and Sanitation: A Status Overview of Sixteen African Countries, Water and Sanitation Program, December.

----- (2008) Global Monitoring Report 2008, MDGs and the environment: agenda for inclusive and sustainable development, The World Bank, Washington DC.

----- (2009), *World Development Indicators*, The World Bank, Washington D.C.

----- (2010), *World Development Indicators*, The World Bank, Washington D.C.