Learning to Teach by Learning to Learn*

Vesall Nourani
UChicago

Nava Ashraf
LSE

Abhijit Banerjee
MIT

November 10, 2023

Click here for latest version

Abstract

Massive learning gaps persist in most developing countries, undermining rapid gains in school attendance. While a pedagogy based on rote learning of facts is often cited as a factor for limited school effectiveness, evidence of success of introducing new pedagogies to teachers is scarce. Indeed, most in-service teacher interventions evaluated in the literature have been ineffective. In this paper, we report on the randomized evaluation in Uganda of an intervention in which teachers were trained in a "learning how to learn" approach. The curriculum, called *Preparation for Social Action*, invited teachers to learn how to learn like scientists: posing sharp questions, framing specific hypotheses, using evidence and data gathered from everyday life whenever possible. We find evidence of change in pedagogy, with dramatic effects on learning: The intervention raised the pass rate in the national exam that determines progression from elementary to secondary school from 51% to 75%, which places the program in the top five percentile of all rigorously evaluated education interventions in terms of learning-adjusted years of schooling per USD. Effects persist four years after the first cohort of teachers was trained, including in the learning outcomes of students who continue to secondary school.

* We would like to thank Kimanya Ngeyo Foundation for Science and Education for providing us the opportunity to study their teacher training program in Uganda. Many seminar and conference presentations have contributed to the refinement of ideas in this paper. Their commitment to systematically learning about their own activities has benefited this study immensely. This study is funded by the Abdul Latif Jameel Poverty Action Lab's Post-Primary Education Initiative and, in its early stages, by the International Growth Centre's Small Project Facility in Uganda. The corresponding author acknowledges support from the National Science Foundation under grants numbers DGE-1144153 and 1809427. The RCT in this paper was pre-registered as AEARCTR-0002647. IRB approval was provided by LSE IRB and UNCST #REC REF 0510-2017. Numerous research and field assistants have provided excellent support along the way. Special thanks goes to Azizi Buyinza, Andrew Chemonges, Ed Davenport, Moustafa El-Kashlan, Stefan Faridani, Joshua Obaya Humphreys, Timo Kapelari, David Liedig, Beatrice Montano, Kim Sarnoff, Misha Seong and Nick Swanson.

1 Introduction

While the last few decades have witnessed an unprecedented increase in access to school worldwide, this access has not been accompanied by a commensurate increase in student achievement. In Uganda, the country where our study takes place, close to 100% of primary age students attend primary school - but only 35.5% of students who started primary earlier in the late 2000's qualified for the last grade of primary education by 2016, and of these only 59% of primary-school leavers transitioned to lower secondary school (UNESCO, 2020). Even after reaching grade 7 (the last year of primary), 30% cannot read and understand a story in English, the primary medium of education (UWEZO, 2016). The results from many other developing countries such as India, Pakistan and Nigeria, are similarly concerning.[1]

Dealing with this learning crisis is widely recognized to be a key challenge of the day (World Bank Group, 2018). The public discussion of the strategies to deal with it emphasize teacher education. The main strategy across diverse contexts is to revamp the teacher education framework.[2] However the Global Education Evidence Advisory Panel's recent (2020) report on "Best Buys in Education," lists in-service teacher training as an area where lots of resources are spent "but evidence on how to do it effectively is low", and goes on to say

> "However, there is little evidence showing that the typical stand-alone general-skills in-service training is cost-effective. Indeed, much of the rigorous evidence that is available suggests that it does not improve student learning outcomes."

The Report however contrasts this conclusion about general-skills in-service trainings with trainings that are dedicated to help teachers implement specific pedagogical practices strongly supported by evidence from multiple studies.

This paper stakes out a claim for a specific category of general-skills in-service training. We report evidence from an RCT in Uganda of very large improvements in student out-

[1] In India, for example, only 50% of fifth graders can do second grade level maths. Only 16% of children in rural India attending Grade 1 of primary school were able to read a Grade 1 level text. While every child in Grade 1 in India is expected to recognize 2-digit-numbers (according to the National Council of Educational Research and Training's specification of learning outcomes), only 41.1% of children in Grade 1 were able to do so (ASER Centre, 2019). In rural Pakistan, only 59% of children in Grade 5 were able to read a story in their local language in 2019, while 57% were able to do division. In urban Pakistan, the numbers were 70% and 66% respectively (ASER Pakistan, 2019). After having received over three years of language training, 80 percent of students in Nigeria were not able to read a simple word of English. In Uganda, around half of the students who had received three years of mathematics teaching were not able to order numbers between 0 and 100 (Bold et al., 2019).

[2] For example, India's recent New Education Policy (Ministry of Human Resource Development, 2020) and Uganda's Effective Teacher Policy (MoES, 2014).

comes resulting from a one-year-long in-service teacher training intervention, combining three periods of intense study during term breaks with monthly in-service visits from tutors. Teachers study a subset of a curriculum titled *Preparation for Social Action* (PSA).[3] The teachers are trained to teach in a way that is inspired by the scientific method, emphasizing the use of precise language, posing sharp questions, framing precise hypotheses and using evidence to decide. To make this approach more effective, the teacher is encouraged to use examples from everyday life wherever possible, so that the questions, the hypotheses and the relevant evidence are likely to be concrete. The training takes place through demonstrations, where the trainer takes the teachers through multiple instances of the method, applied to specific contexts.

However this is still very far from a scenario where the trainer sets very specific teaching goals and lays out the exact steps to get there. The teachers in this program still have to figure out how to apply these ideas in the context of what they are teaching, which can be English, Social Studies or some branch of Science, and do so for multiple different topics within each field. In this sense, the intervention we study is very different from Teaching at the Right Level (TaRL) and Structured Lesson Plans, which are the two pedagogy/teacher training interventions that have been widely and successfully replicated in multiple developing countries, and much closer to general-skills training. Both of those interventions narrow down the pedagogy to a relatively small number of steps combined with a clear set of proximate goals — for example, basic skills development for young children. The current intervention, on the other hand, enables a broad-based shift in the capabilities of the teachers, so that the teacher becomes less the source of the answer and more able to assist students to learn to find the answers themselves across diverse applications and contexts. This is important, for example, when teaching young adolescents, aged 11 to 15, who study multiple different fields and are expected to cover much more material, which makes it harder to provide teachers with a minute-by-minute lesson plan.

In 2017, we randomly selected 18 out of 35 primary schools in a subcounty of Jinja, Uganda using a pairwise matching procedure (Bruhn & McKenzie, 2009) to participate in the training.[4] The District Education Office (DEO) invited these schools to send two to three primary school teachers (emphasizing upper primary — grades 4 through 7 — over lower primary) to a training opportunity offered by KN. Therefore, our analysis measures treatment effects for schools where two cohorts of teachers participated in the training

[3]The curriculum was originally developed by Fundación para la Aplicación y Enseñanza de las Ciencias (FUNDAEC) in Colombia in the 1970s and implemented in Uganda by an Ugandan NGO called Kimanya Ngeyo Foundation for Science and Education (KN) and the Jinja District (Government) Education Office. The translation of FUNDAEC into English is "The Foundation for the Application and Teaching of the Sciences."

[4]Between 2017 and mid-line data collection in 2019, three schools became insolvent, causing them to close. Due to pairwise matching, we proceed with the analysis unhindered by dropping each school in the pair from the study, resulting in outcomes from 29 schools analyzed in the present study.

— one in 2018 and the other in 2019.

The intervention had large and positive effects on almost every outcome where we expected one, as recorded in the pre-analysis plan.[5] Student performance in treatment schools increases by 0.5 standard deviations in high stakes (primary leaving) national exams that determine whether or not a student can proceed to secondary school — the pass rate in treated schools increases by 24 percentage points relative to a control school mean of 51%. This result reflects improvements across the entire distribution of student achievement and for boys and girls alike. There were large improvements in performance in all the fields that were featured in the training — English, Social Studies and Science — and these are largely robust to using Randomization Inference to deal with the small number of clusters and Benjamini-Hochberg adjustments for multiple hypothesis testing. The one area where there was no clear evidence of improvement was mathematics, which was explicitly left out of the teacher training due to time constraints.[6]

To understand better the mechanisms at the student level, we were interested in whether the students have a deeper understanding of the conceptual material or whether the main source of improvement was more effective rote memorization, resulting from say greater effort or higher self-confidence. To do this we administered an exam using questions from a commonly utilized assessment booklet in Uganda (SIPRO, 2019) and classified questions in this booklet that reflect higher-order learning using classification categories described in (Burdett, 2017).[7] Student performance in these tests improves by between 0.5 and 0.8 standard deviations.

To test the idea that the students actually absorbed the scientific method, we implemented a novel field test of scientific competencies. We coordinated with the district education office to institute a series of science shows, where student-led experiments tested methods of soil conservation and clean water preparation in practice. We present results of these experiments to assess practical scientific competencies.[8] Specifically, an independent panel of judges hired and trained by IPA-Uganda assessed the performance of students by attending the shows held by each of the schools in the study. The treatment students demonstrate a higher aptitude for scientific thinking as measured by the judges' assessments of the school-based science shows. Students proficiency in designing

[5]The experiment is registered in the AEA RCT Registry under AEARCTR-0002647, where the pre-analysis plan was published on October 2, 2020.

[6]As a consequence of these results, KN is revisiting approaches to including PSA materials focused on mathematical capabilities.

[7]According to this classification, there are, roughly, three broad categories of assessment questions that measure student recall, application of understanding, and critical thinking and reasoning. As Burdett (2017) notes, there is a paucity of questions assessing critical thinking and reasoning in most developing country examinations. We, therefore, augmented the common assessment with additional questions testing critical thinking and reasoning using the framework described in Liu et al. (2014).

[8]The National Curriculum Development Centre (NCDC) strongly encourages the use of science shows, though they are scarcely used in practice. We drew on this encouragement in our conversations with schools and the district education office and tied our measurement of scientific competencies almost exactly to those requested by the NCDC.

experiments, articulating and analyzing hypotheses, measuring observations, and communicating results increase by 31%, 80%, 61% and 60% respectively over control school peers, all significant at conventional levels.

We use several other measures of teacher outcomes to link improvements in student learning to pedagogy. Classroom observations test whether there are observable changes in the conduct of the classroom and we find that students in treated schools are 39 percent more likely to be engaged in the activity conducted by a teacher relative to their peers in control schools. Students in treated schools are 6 percentage points more likely to have asked teachers questions in class when they don't understand a concept that is being presented relative to a control school mean of 22%. We develop a novel measure indicative of the degree of attention they give to their students by quizzing them on details of their students' lives and past attendance. Teachers in treated schools are 10 percentage points more likely to know these details.

In 2022, four years after the initial cohort of teachers were trained, we administered assessments to all primary school students, secondary students who we can trace to the primary schools in our study and primary school teachers. We also carried out a household census with recall data on the previous five years of school enrollment for each household member to develop measures of persistent school enrollment and grade progression. We find that all treatment effects associated with student learning persist in the long run, including in the learning outcomes of secondary students in 2022 who were assigned to treated schools in 2018 but are no longer taught by trained teachers. Furthermore, we do not find differences in teacher test scores when we test them on questions relevant to the subjects they teach. However, we do find that large effects in a series of questions that test whether teachers are more precise and consistent in their "use of language" — exercises where they choose between words such as "science" and "technology" or "understand" and "know" in sentences where an untrained eye might not be able to differentiate word choice precisely.

The student learning results in particular, measured in Learning Adjusted Years of Schooling (LAYS), place the program in the top five percentile of all education interventions across the current list of 150 studies analyzed by Angrist et al. (2020). A conservative estimate suggests that an additional $100 of funding increases learning adjusted years of schooling by 9.62 years.

While we do not have direct observations of what happens during the training itself, there is a large body of writing that can help us understand why the PSA approach has the potential to promote improved learning (Freire, 1970, Ranciére, 1991, Duckworth, 2006). Teachers in these settings, as in most of the world, are accustomed to "chalk and talk," or lecture and blackboard-centered pedagogy (Glewwe & Muralidharan, 2016). To quote Paolo Friere, the Brazilian educationist who partly inspired FUNDAEC's work,

> Narration (with the teacher as narrator) leads the students to memorize me-

> chanically the narrated account. Worse yet, it turns them into "containers," into "receptacles" to be "filled" by the teachers. The more completely she fills the receptacles, the better a teacher she is. The more meekly the receptacles permit themselves to be filled, the better students they are. Education thus becomes an act of depositing, in which the students are the depositories and the teacher is the depositor.

By contrast, the PSA training adopts a premise that, in part, suggests that knowledge is to be discovered by the learner (Freire, 1970) and that one of the teacher's primary roles is to help the student discover strategies for learning. FUNDAEC describes its tutors which would include the trainer who trained the teachers as well as the trained teachers themselves thus:

> Tutors guide the students through the textbooks, raise questions, clarify obscure matters, encourage reflection on real-life experience and supervise experiments and social action. They do not lecture or dictate but nor are they mere facilitators of group discussion (Arbab & Lample, 2005).

In practice what the teacher does differently can be quite simple. For example, science teachers typically introduce plant biology by drawing a plant on the blackboard and asking students to memorize the location of the "roots," "stem," "leaves," "buds," "flowers," etc. Instead, teachers might now bring actual plants into the classroom and ask students to differentiate the parts of a plant according to what they observe. The teacher might ask the students to reflect on the process of a growth of a plant by planting, observing, and measuring key aspects of a plant at distinct stages, thereby facilitating the use of scientific methods. Once the learner has exerted sufficient effort to understand plants through observation, the teacher provides clarity by linking the learner's efforts to what we know about the core characteristics of plants, but the source of knowledge has shifted in a fundamental manner through this exercise. The pedagogical act now disassociates teachers from the origin of knowledge, without compromising the teacher's role as an experienced guide, and students slowly begin to see how their own critical thinking abilities can be a source of insights into the reality around them.

Moreover this process of discovery, experimentation and research spills over to problem-solving within the school as well. Whenever teachers complain about a challenge they face in the school environment, such as student absenteeism, or a confusing directive from school managers, KN tutors ask them "Do you know what the source of this challenge is? Have you done enough research to find out?"[9] If the trained teachers take this lesson

[9]For example, one teacher shared that a pupil of theirs was consistently late to school. Prior to participating in the training, the teacher's main strategy was to berate the pupil and send him back home so that he would learn his lesson and arrive on time the following day. This did not happen and the pupil was consistently sent home day after day. When the teacher brought this challenge up during

to heart and take the lead in solving problems, the culture of the entire school, including teachers not included in the program, may change.

The emphasis on how to learn to learn is closely related to the increasing emphasis on metacognition (thinking about thinking) in the literature on how to improve teaching practices (Perkins, 1992, Kolencik & Hillwig, 2011, Tanner, 2012, Lang, 2012, Swartz & Perkins, 2017). This literature argues that giving students more control over the learning process helps them reach a deeper understanding of the material and become more engaged, but also gain confidence, which helps them make better use of what they know.

To the best of our knowledge, this is the first rigorous impact evaluation of a "learning to learn" intervention. Evidence from multiple RCTs have established the idea that the traditional top-down pedagogy, where the teacher teaches a fixed curriculum and is largely unresponsive to specific gaps in what the students know, is a major source of the learning gap in many developing countries. When the pedagogy is altered to focus on specific child level gaps, they make rapid progress (Piper et al., 2014, Banerjee et al., 2017). While most of these studies focus on basic skills for younger children, Muralidharan et al. (2019) show that a tablet-based TaRL program that coaches children based on where they are making mistakes, can lead to large gains in learning levels for children in Upper Primary or Middle school. Bando et al. (2019) show that another technique, inquiry- and problem-based pedagogy, improve student performance in math and science, primarily for boys, in a way that increases performance over time. We offer that a general-skills training in which the teacher learns how to learn may enable her to draw on the appropriate pedagogy depending on the learning objective on hand.

Moreover, this insight helps us think more broadly about the relationship between human capital accumulation, technology and economic growth (Becker, 2009, Goldin & Katz, 2009). A growing literature argues that the type of human capital required of modern economic growth matters (Kremer, 1993, Acemoglu et al., 2012, Hanushek & Woessmann, 2012, Flabbi & Gatti, 2018). More often than not, labor needs to be flexible to the demands of modern technologies. We propose that metacognitive, "learning to learn," approaches to human capital development, rather than measuring simple years of schooling, may capture vital skills for economic growth and development. We explore this further in our conclusion.

The rest of the paper proceeds in the following manner. In section 2 we provide details of KN's Intervention, articulating our theory of change and its relationship to our research questions and hypotheses in our pre-analysis plan. In section 3, we provide more details

a KN teacher-training, the tutor asked her to investigate why the student was consistently late. After some research, the teacher learned that the student spent his morning hours selling donuts at the local market in order to pay school fees. Feeling sympathy towards this student, the teacher began to allow the student to arrive late to school whenever necessary and engaged more deeply with his family to see if there were other ways of supporting the child so that he did not miss important class time.

on our experimental design and data. In section 4, we describe our identification strategy and share and interpret results before providing a cost-benefit analysis of KN's teacher training program. We conclude with thoughts for next steps in section 7. All appendix materials are available online.

2 The Intervention and Conceptual Framework

2.1 Preparation for Social Action (PSA) Teacher Training

The PSA curriculum has been in development since the mid-1970s and is continuously reviewed and updated through an ongoing process of action-research, often taking up to ten years to produce texts before distribution.[10] KN coordinates the study of PSA materials among groups of tutors in Eastern Uganda, including in Jinja district where our study is based. This section weaves together a description of the content of the training with a conceptual framework that explains how the approach cultivates a general approach to developing teacher capabilities.

2.1.1 Learning How to Learn

The PSA materials are studied by the teachers together with the tutor in a highly participatory manner. The initial experience is intended to enable the teacher to experience a new type of learning environment and is, therefore, jarring for most Ugandan teachers accustomed to training where they are passive recipients of lectures. Instead, the tutor starts by handing out a PSA text that all teachers take turns reading out loud. The text poses questions that the teachers are invited to answer. Initially, teachers tend to be hesitant to share their thoughts lest they embarrass themselves by providing a "wrong" answer.

As an example, here is the first page of the first text in the PSA training, on "Properties,":

> When we look at the world around us, we immediately notice that objects have different shapes and sizes. This enables us to differentiate things and describe them to others. For example, in everyday conversation we speak of round things, square things, long, wide, big, and small things. Let us think about these concepts to find out how much we really understand them.
>
> At one time or another, we have all played in the mud or the sand building things of different shapes. Perhaps we are now too scrupulous to play with

[10] In the early 2000s, FUNDAEC, the organization that develops the PSA materials, began partnering with collaborating institutions around the world which now includes a growing network of organizations across Africa, including KN, which was established in 2007.

mud, so let us use a bit of clay or play-dough to make some interesting shapes and try to describe them to one another.

[Participants spend 15 minutes making shapes out of clay and then attempt to describe the shapes to one another for another 30 minutes with questions interspersed by the tutor. Next, the text asks.]

Was it easy to describe all the shapes?

Teachers initially provide common responses to a spherical object someone may have crafted such as "round," "circular," or it "looks like a ball." The tutor introduces questions that challenge these descriptions such as "What do you mean by circular? I thought circles were flat?" or "But that is not entirely round. I notice a few bumps on that object. How would you describe it in its entirety?" The conversation around the shapes teachers have crafted soon transitions into a conversation about dimensionality as well as one about regular and irregular shape names — all through an activity where they attempt to describe the shapes they themselves made. As teachers struggle to find words that make their description more precise, the tutor helps them realize that their prior descriptions required an underlying understanding of the words "round" or "circle."

While seemingly simple, the novelty of the exercise exposes teachers to new learning experiences that the cohort of teachers can reflect on collectively. In section **??** we describe how later parts of the training help teachers systematically analyze the pedagogical environment that led to these learning experiences. For now, it is worth pointing out that as teachers analyze the exercise of describing shapes they recognize that prior knowledge is a pre-requisite of precise language and is required to generate new knowledge. This underscores the necessity of being precise with one's language, and for a shared understanding of prior beliefs/concepts. If this can matter with something as simple as creating a sphere out of clay, how much more important can it be with more complex shapes and concepts?[11]

In later lessons, for example as teachers carry out activities to learn about the growth of a plant, the participant recognizes the values of distinguishing and describing different parts of a plant: its root system, its branches, its leaves, and so on. As the participant learns to describe each part in greater detail, her description of the process of the growth of a plant is rendered more accurate.

[11] The tutor might subsequently ask, "what other shapes and words are important to know to describe these objects more precisely?" The text then reads "as we deal with more and more complex shapes, the task of describing them becomes increasingly difficult." Teachers consider how they can describe increasingly complex objects by combining regular shapes in their descriptions. For example, a house can be described as having a long rectangular cuboid concrete slab as a foundation with a cuboid structure on top of the foundation. On top of the cuboid is a pyramid or prism-like shape that we call a roof.

2.1.2 Developing Scientific Capabilities

The above exercise illustrates a non-hierarchical mode of teaching related to one aim of science: to describe reality with increasing degrees of precision. However, it builds on this illustration with a series of additional experiences that foster the development of scientific capabilities in teachers.[12] These dimensions include, but are not limited to: 1) the use of precise language, 2) observing and describing the world around oneself with increasing clarity, 3) designing experiments and testing hypotheses, 4) distinguishing objective from subjective statements, 5) identifying sources of data and processes worth studying, and so on. We list below concrete examples of exercises done in the teacher training.

Using precise language. In "Transition to Agriculture," one exercise reads "Depending on how it is used, a word can carry one or another meaning, which is easily seen by the explanations given in a dictionary." Teachers are then asked to consider how the context in the reading determines the meaning of a pre-selected set of words used in the materials. They recognize that precision of language does not only depend on the definition of a word, but also on the context in which it is placed.

Observing and describing the world with increasing clarity. When agriculture is introduced in the module "Planting Crops" teachers are asked to observe the leaf structure of the plants in their environment and classify them according to their biological properties.

Designing experiments and testing hypotheses. Participants set up an experiment that allows them to measure the boiling point of water, testing the hypothesis that the boiling point is 100°C.

Distinguishing between objective and subjective statements. When articulating hypotheses, teachers are careful not to use subjective terms such as "improved soil health," but rather objective measurements of soil health, potentially observable to all, such as soil pH, moisture absorption, precise proportions of clay, sand and loam, and so on.

Identifying processes worth studying and sources of data. When learning about "urban gardens" they measure the amount of labor and financial input as well as both the nutritional and economic value of the crops produced. In a text titled "Nurturing Young Minds," participants carry out interviews with village elders to collect information on past and present forms of child rearing and education, often focusing on the moral, social and intellectual dimensions of child development. This

[12] There are multiple ways in which the PSA materials foster these capabilities; We provide coarse summaries of each course in appendix A.

> allows for further and deeper analysis of the forces that enable or impede development of their learners. In 'environmental issues," teachers visit local markets and investigate waste management practices among small and large scale manufacturers.

These are but a few examples of experiences teachers have as they experience various modalities of learning. Beyond providing them with an illustration of a non-hierarchical mode of teaching the deepening of awareness of scientific approaches to learning generates an orientation of humility, an openness to data and refinement of ideas.

2.1.3 Understanding How Knowledge is Produced

The above examples of learning activities are not experienced in isolation from each other during training. From the very first exercise of describing a shape, teachers are asked to reflect on how any descriptive language or shared understanding relies on assumptions/prior concepts. This is reinforced by the scaffolding of the learning approach within the training — the learning experience in PSA is cumulative and makes clear that knowledge can grow along a path of discovery that starts from initial building blocks. As learners encounter gaps in their understanding, new observations lead to the need to develop more precise language to learn about, and describe, any object of reality. The experience of the training thus sheds light onto how science — and knowledge — advance, by incrementally adding basic shared concepts to describe increasingly complex phenomena.

A passage in "The Heating and Cooling of Matter" helps illustrate this point. In this course, participants develop an appreciation for the atomic theory of matter and how the evolution of this theory helped humanity understand the relationship between heat and matter through an iterative process of observation, refinement of models, and further tests over time. Toward the end of the course, participants reflect on the following questions

> The way we have chosen to introduce the subject of heat may lead you to believe that a theory is somehow the conclusion one reaches after making many observations. But pause a little and ask yourselves if this is really true. Would people actually begin to observe any process in nature if they did not already know something about it? what kind of questions would you ask about the process of heating and cooling if you did not have some notion, right or wrong, of what heat is? Think of very small children. Does what they learn about heat and temperature arise directly from their sense of touch and sight, or does somebody have to teach them about these concepts?

Teachers may spend over an hour discussing these question in their group. The result of this exercise is a deeper appreciation that (scientific) knowledge can emanate from

observations made by anyone, including young children. In order for knowledge to advance, people build on the accumulated insights of past experiences as well as the forward movement of new ideas — including new observations made as language of description becomes more refined. The ability of each individual to contribute meaningful insights and observations beckons teachers towards an orientation of humility to the perspectives of others, including children — in gathering and incorporating new data. Knowledge of the process of knowledge production gives rise to a sense of agency in the ability to generate knowledge oneself.

2.1.4 Creating Conceptual Connections

Learning comes with understanding new *concepts*, including the meaning of the term concept itself, and learning how concepts generalize to broader contexts. For example, continuing in "Properties," teachers begin discussing how the concept of "properties," can be applied to their own understandings of their students. Towards the end of "Properties," teachers read:

> Up to now we have been studying the properties of matter. We have examined certain general properties common to all matter and a few specific ones that make each substance different from others. We would now like to ask whether the concept of property also applies to human beings. Do they have general and specific properties that we can use to describe them?...
>
> It so happens that we do not normally use the word *property* to describe human beings. We use words such as *characteristic, quality,* and *attribute.* The underlying concept, however is not so different. In the same way that it is the property of water to flow, certain virtues such as honesty, courage, and generosity are properties of human beings. So, too, is the capacity to think and to express thoughts through language.

The material, with the help of the tutor, carefully draws a connection between how we observe and describe the properties of nature and our observations and descriptions of human beings and human behavior. Creating this conceptual connection can have profound effects. For example, on several occasions the PSA materials explicitly or implicitly ask teachers to examine what it means for a material substance to change phases — say, from a liquid to a gaseous state. Following the sections that analyze "properties" of human being, the tutor also helps the teachers to reflect on their own students. They might ask: "What are the qualities you desire in your students? How can you facilitate a process that helps the students acquire these qualities? How can we apply our understanding of the "change of phase" of substances to this process?"[13]

[13]This extends to practical exercises as well. In "The Heating and Cooling of Matter," teachers ex-

The fact that material substance can change phase becomes an analogy for how human "properties" (qualities) may also not be fixed and static, but rather subject to change and transformation. This can change how teachers view their students, from having fixed qualities (bad at math, for example) to capable of change - just like other material substances.

A second example from "Properties" further illustrates how scientific concepts can affect teachers' perceptions of their students. In one lesson, teachers explore the concept of "objective" and "subjective" descriptions.[14] For example, as they discuss the scientific imperative of striving to make and test objective statements, they gain insights into the concept by reading a series of statements and discussing whether they are objective or subjective descriptions, e.g., "the school is the largest building in the village" or "he has a big heart." The tutor then asks them to reflect on the various statements they might make about their students and whether it describes an objective or subjective characteristic of their students. Teachers soon realize that statements that describe students may differ from teacher to teacher. They begin to recognize pitfalls in their thinking — that subjective terms such as "smart," "slow," and "difficult" may reflect their own biased perspective more than anything else. They begin to realize that their colleagues might see the very same student in a different light, propelling them to seek the advice of others to reduce the bias in their opinion. The recognition of subjective bias even propels changes in the teacher's own attitudes and behaviors towards students, a process we describe in more detail when describing our conceptual framework and predictions in section 2.2.

The concepts "change in phase," "subjective" and "objective" descriptions are only a few concepts that the teachers unpack in depth. Additional concepts include the difference between "information" and "concept" and the associated ideas of "*assimilating* information" and "*understanding* concepts." This exploration of concepts accompanies the other features of the rich new learning experiences teachers have over the course of the year. Together, these experiences help teachers appreciate how anyone can learn to produce new knowledge, or "learn how to learn." This process can begin with very basic building blocks where each block gives rise to subsequent questions that require more investigation. Teachers recognize that when knowledge is explored and organized in this manner, the educational experience allows the learner to appreciate the role they can play in advancing knowledge generation, granting them with a sense of agency for

periment with measuring the boiling point of water — the point at which molecules break apart and water changes phase from liquid to gas. The concrete nature of this exercise aids tutors to initiate a conversation in which teachers think more concretely about the "transformation" they desire to see in their students. Tutors are not satisfied with vague statements and point out how the teachers' description of this transformation should strive towards being as concrete and precise as how they can now describe the change of phase of a substance.

[14] In a similar vein, teachers discuss the difference between "general" and "specific" properties of substances. A dialogue between a KN tutor and participating teachers, highlighting aspects of KN tutors' pedagogy, is available to read in Appendix A.7.

contributing meaningfully to processes of discovery. Equipped with this experience and realization, the final step of the training involves assisting the teacher to create similar learning experiences for their students, that is, "learning to teach."

2.1.5 Learning How to Teach

The second part of the training pivots to teaching, using the same insights used to learn how to learn. Teachers begin by carrying out an explicit analysis of the PSA curriculum using "Discourse for Social Action (DSA): Unit 2 Education," part of another sequence of courses developed by FUNDAEC. This material is used in the teacher training to help the teacher think about how the curriculum and pedagogy in PSA influenced her own thinking — a meta-cognitive analysis of the pedagogical act. By unpacking the path between pedagogy and learning, this process helps the teacher generalize beyond her own experience and to think how her pedagogy influences her pupil's thinking.

It does so by walking the reader through FUNDAEC's description of the aim and purpose of its own educational programs — "to assist our students in their spiritual and intellectual growth as individuals and in the development of their capacity to contribute to the transformation of society." Teachers reflect more broadly on the aim of education, for example by understanding the difference between "assimilating information" and "understanding concepts," by recognizing the importance of integrating theory with practice, being with doing, moral and academic achievement, all while avoiding the tendency to simplify reality by espousing extreme views and false dichotomies such as "the aim of education is to help students understand concepts rather than assimilating information."

It is important to note that what distinguishes this approach from other approaches to teacher training is that, prior to carrying out the meta-cognitive analysis, teachers in KN's training first experience the PSA curriculum *as learners*. Thus, their analysis is not of the curriculum or pedagogy alone — they also analyze their own experiences together with an analysis of the PSA curriculum and pedagogy, rendering the act of generalizing pedagogy more concrete and practicable. Rather than copying and pasting the PSA curriculum for use in their own classroom, teachers see the value of thinking clearly about the purpose of any curriculum and ensuring that the pedagogy and content are coherent with this purpose.

What does this mean in practice? When reflecting on the difference between assimilating information and understanding concepts, for example, teachers recognize that much of their (former) pedagogy is aimed at helping their students memorize definitions, formulas or procedures depending on the subject. From their own experience with the PSA material as well as reflecting on their aim as an educator, they realize that memorization may be insufficient as an objective of an educational endeavor. They recognize that a different pedagogy can engender forms of learning that allows one to apply understand-

ing to ever-broadening contexts.[15] They analyze the approach taken by the KN tutor to guide participation in the PSA courses and reflect on the underlying aim, purpose, and structure of Uganda's National Curriculum and how it correlates with their own evolving notions of the aim and purpose of education. Crucially, rather than mirroring approaches to teaching offered by their own teachers, they are able to think more critically about pedagogical choices that advance student learning more holistically, beyond the mere assimilation of information.

Each of the pieces of KN's training are brought together through this exercise, and teachers see how the experience of "learning how to learn" will lead them to learn how to teach: they use their newfound scientific capabilities to observe and describe their own practices with increasing clarity; they consider more carefully the concepts underlying the curriculum they are told to teach; and they reflect on how their own pedagogy can fit the learning task at hand, articulating concrete plans and approaches that advance student learning towards diverse learning objectives. This newfound approach to learning has profound implications for how teachers teach. We summarize the implications in the next subsection.

2.2 Summary and testable hypotheses

The traditional approach to teaching in these schools is best described as "explication" — an approach that works under the assumption that knowledge is a static and well-defined object, and the main problem is that students lack this knowledge which teachers possess Ranciére (1991). Under this set of presumptions, even a teacher with the best of intentions is unlikely to explore solutions or answers together with the student.

By contrast, the pedagogical approach teachers experience in KN's training conceives of knowledge as an evolving construct and emphasizes that everyone can participate, to varying degrees, in its generation and application. The most important function of the training may be to make teachers more aware of the implications of this proposition with the goal of helping them develop an appreciation for the freedom of thought required of students who learn through a process of discovery.

The theory behind this pedagogy is that this shift in their perspective on teaching has many implications for the teaching-learning process. We summarize the main potential changes here:

1. The training leads teachers to internalize and adopt a pedagogical approach that is more exploratory in nature, and to develop a taste for active learning that they then try to inculcate in their students.

[15] It is during this moment in the training that one representative teacher commented: "I have been teaching for 30 years and am only now realizing that all I was ever doing was assisting my pupils to assimilate information."

2. It also makes the teachers more inquisitive about the world around them and the material they are teaching, which makes their own learning, and consequently their teaching, less rote-based.

3. Taking a more inquiry-based approach to teaching makes the teachers more interested in the thought-processes of the students, which in turn makes them more sympathetic to the students and less adversarial.

4. Both because of the closer connection with the students and because of the overall emphasis on more open learning, the teachers are more likely to encourage students to explore ways to answer a question on their own (and perhaps explore it with them) than handing them a textbook answer.

5. Being more open about the sources of knowledge also enables teachers to learn more from their colleagues, but also share their own knowledge and insights with their fellow-teachers.

These changes are all multidimensional, and there are surely many other potential changes beyond those that we have listed here. In particular, we expect that these changes will have knock-on effects on the families of these students and their village. Our plan is to pursue many of these questions in our ongoing work on this project. Here, we focus on the changes in the main educational outcomes, two years and four years after the beginning of the intervention. We listed many of these as a set of testable hypotheses for the outcomes we measure in our pre-analysis plan, which we list below:

(Q1) Does the teacher training change teacher pedagogy, attitudes, and effort?

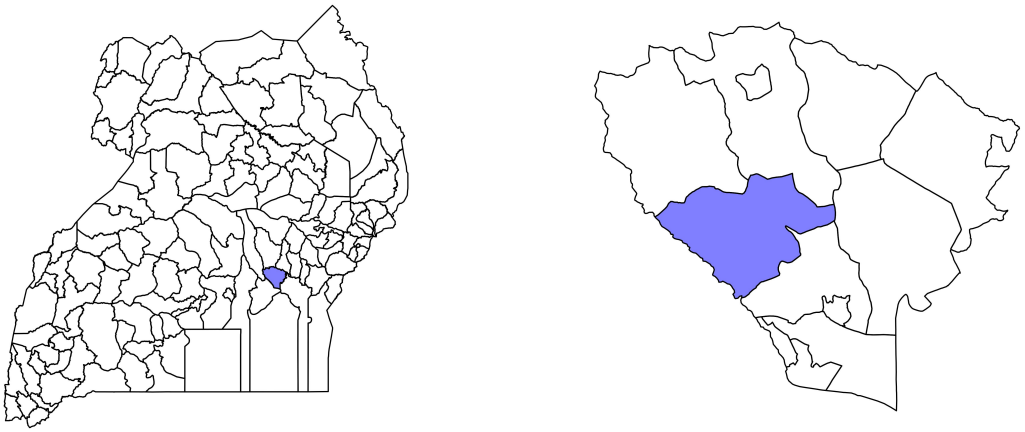
(H1.1) The teacher training will increase the following classrooms outcomes: student engagement,the number of students asking questions in class, pedagogical techniques that engage critical thinking and practical exploration, pedagogical techniques that facilitate understanding of concepts and deeper learning. It will decrease the use of corporal punishment and other harsh responses to poor student behavior.

(H1.2) Trained teachers exhibit higher effort and knowledge of their students.

(H1.3) Teachers will be more willing to learn from others and collaborate with their peers following the teacher training.

(Q2) Does exposure to trained teachers improve student outcomes?

(H2.1) Traditional learning outcomes are higher for students taught by trained teachers via: Pass-through Rates and standardized test scores(Primary Leaving Exam — PLE)



(a) Jinja District in Uganda **(b)** Sub-county in the Study

Figure 1: Study Area

(H2.2) Higher order learning outcomes are higher for students taught by trained teachers.

(H2.3) Scientific capabilities are higher for students taught by trained teachers.

(H2.4) Students taught by trained teachers find more creative uses of local resources.

There are implications of the theory that are not directly addressed in the hypotheses outlined above such as the degree of active learning pedagogy, teacher inquisitiveness, and an increase in teacher sympathy towards students. We provide additional results that explore these aspects of the theory in our discussion in section 4.

3 Experimental Design and Data

Jinja District is located in Uganda's Eastern region and contains the source of the river Nile (Figure 1a). The study takes place in one sub-county in Jinja District (now Jinja City).[16]

[16]We had initially planned to augment this sample with schools from four additional sub-counties and divisions in Jinja District. We collected baseline data from 53 additional schools to this effect in 2019. However we did not anticipate at the time that Covid-19 would effectively cause schools in Uganda to be closed for two full academic years, 2020 and 2021. We released our initial manuscript at the end of 2020 when it seemed schools would be re-opened in 2021 prior to the global concern with the Delta variant of Covid-19. We managed to extend the final data collected as part of the study to the end of 2022, but decided not to collect data in the augmented sample in lieu of collecting more long-run outcomes in the original set of schools in Budondo sub-county, those selected into the study in 2017.

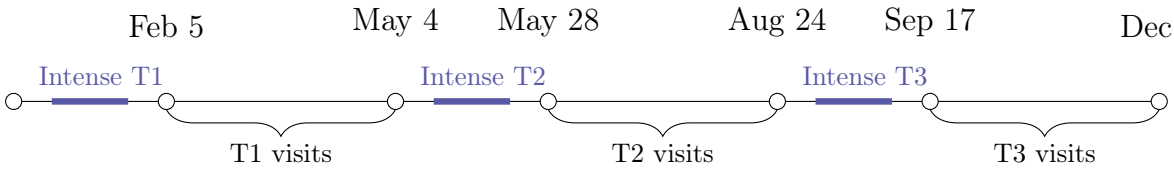


Figure 2: Timing of Teacher Training in 2018

3.1 Experimental Implementation & Timeline

We focus on KN's use of the PSA materials in its approach to train school teachers, an initiative that began in late 2014. KN collaborates with the district education officials (DEO) to invite schools to select 2-5 teachers per school (depending on the size of the school) to participate in the in-service Teacher Training Program over the course of the school year.

The training combines three two-week periods of full-time programming during school holidays with in-service periods and classroom visits during the school year. There are three school terms during the course of a typical school year in Uganda. In 2018, for example, term 1 took place from February 5 to May 4, term 2 from May 28 to August 24 and term 3 from September 17 to December 7, the gaps between terms are utilized by KN for "intense" study of PSA materials as depicted in Figure 2. During these periods, teachers study PSA materials for roughly 11 days together with a KN trained tutor. During school terms, PSA tutors try to visit teachers in their classrooms at least once each month to observe how teachers engage with students and reflect on how to improve pedagogy and other dimensions of practice.

KN worked with the DEO to invite 50 teachers to participate in the training in each year of the intervention that we analyze in this paper, inviting teachers from both primary and upper primary. They expected that some would not be able to attend and were aiming for roughly 40 teachers participating in each training. In total, 42 teachers attended a portion of the intense sections of the training in 2018 and 46 attended a portion of the training in 2019. Each cohort of teachers is dubbed a "unit," and KN employs two PSA tutors to engage with each unit of teachers at any given time.[17] Intense study periods were held at a central location near the schools so that practical exercises could link the PSA materials to local contexts in which lessons can be applied. Notwithstanding that teachers are not paid to participate (only a transport subsidy is provided, which schools are requested to refund), training participation rates are high — 33 and 32 teachers attended the majority of the training in 2018 and 2019 respectively.

[17]We also note that, in conversations with KN, it became clear that the tutors they hire do not necessarily have experience as educators in formal educational settings — indeed, one of the two tutors of the Budondo unit had not completed secondary school before he started engaging with the PSA program in the late 2000s.

3.2 Sampling

3.2.1 School Sampling Frame

To identify the sample in Budondo subcounty, we took a census of all primary schools in the summer of 2017 and collected information on the ownership status of each school, enrollment numbers, the highest class taught, and the number of teachers (among other measures). Given the study's focus on upper primary teachers, we limited schools to those that teach up to the highest primary school level, P7. This left us with 37 remaining schools of which 2 closed doors before the end of the 2017 school year. We included the remaining 35 schools in our sample - 15 of which are government-owned. KN began working with the randomly selected treatment schools in December 2017 after baseline data was collected.

3.2.2 School Assignment to Treatment

We randomly selected 18 out of 35 Budondo primary schools and offered training opportunities only to these schools. Randomization was conducted in November 2017 by pairwise matching (Bruhn & McKenzie, 2009) on 7 school characteristics drawn from administrative data.[18] First, one school was chosen uniformly at random for treatment. Then, the remaining 34 schools were greedily matched into pairs according to minimum Mahalanobis distance among the 7 matching characteristics. Finally, one school from each of the 17 pairs was chosen uniformly at random for treatment. Importantly, this procedure allows us to treat each of the 17 pairs, as well as the odd school, as strata in which one side of the pair was randomly selected into the treatment. Thus, if schools drop out of the study for any reason, we can drop the pair of schools and proceed unhindered without losing identification. Three private schools were no longer solvent in 2019, which caused us to drop them and their paired schools from the study at that point. Thus, the proceeding analysis takes into account the remaining 29 schools in Budondo sub-county.

The top panel in Table 1 presents summary statistics for school characteristics as measured in 2017 for the sample of 29 schools whose outcomes we analyze in this paper. Out of the 29 schools, 15 are government schools and, in 2017, each school employed 13.2 teachers on average for a total of 383 teachers with 248 of them teaching at some level of upper primary. Of these, 43% teach P6 and P7 in addition to other classes, which shows that teachers have responsibilities for many classes across the school. Only 24% of teachers have a post-secondary diploma, as is now mandated of all teachers by MoES in the newly implemented National Teacher Policy, and 20% participated in in-service

[18] Government ownership, student enrollment, number of teachers, fraction of teachers who teach 6th and 7th year primary school, mean population under 15 years old in feeder villages, mean number of years that teachers have worked at the school, and fraction of teachers who completed upper secondary school.

trainings prior to 2017.[19] Roughly 520 students were enrolled in each school on average in 2017 with an average pupil-teacher ratio of 38.2. This table further breaks down averages by control and treatment status.

Teachers often transfer in and out of schools in Uganda, especially in private school settings. Due to poor record keeping around such transfers, we cannot know with precision the percent of trained teachers in a given school at a given time. However, we do know that no trained teacher transferred to a control school from a treated school. We also know that, of the 233 upper primary teachers in the sample in 2019, 127 teach in treated schools and 49 (38.6%) participated in the training at some point during the two years of the intervention.

3.2.3 Teacher, Classroom and Student Sample

Outcomes will be measured at the teacher-, student- and classroom-level to address the research questions above. Due to resource constraints, we limit the analysis specified in the pre-analysis plan and in this paper to measures from the following three sampling frames within our sample schools: upper primary teachers, upper primary classrooms, students in primary six (survey and researcher-administered assessments) and students sitting for primary leaving exams. The measures we use are best thought of as measures collected of multiple cross-sections across time, though in cases where teachers are retained, we have teacher-level panel data. For teacher outcomes, we interview the universe of upper primary teachers in our school sample. Furthermore, we conduct classroom observations of all upper primary classes during the morning hours of operation at schools in the sample (typically between 8:00 AM and 1:00 PM, though this can vary by school and school day). In 2019, we interviewed 230 teachers in this way and observe 238 P4 to P6 class periods and 69 P7 class periods across the 29 schools.

In the student survey, we also asked P6 students to identify which of the upper elementary teachers regularly teach them at their school. When considering teacher behavior measures constructed using student responses, we restrict the teacher sample to those teachers who teach P6 students as indicated by at least 90% of surveyed students in a school — there are 95 such teachers in our sample as indicated by student responses.

Our sampling frame for students who participate in the student survey and assessment is the universe of P6 students in each school.[20] We select 10 to 15 such students (depending on the size of the school) in each school to our sample, balancing across school-specified performance quintiles and student gender.[21] We provide alternative stu-

[19] MoES is now establishing a credit-based continuous professional development (CPD) training system that requires each registered teach to obtain a certain number of CPD credits every two years.

[20] We began collecting student data in 2019, therefore we do not have baseline student survey data for Budondo schools.

[21] We use Term 1 records from 2019, provided to us by the schools in the study, to determine the rank and gender of students in each school. This list serves as our sampling frame as well.

dents within quintile-gender strata in case a student is unavailable for interviews and assessments on the day our team of enumerators visits the school. We select 329 students to participate in the student survey and researcher assessment in this way across the 29 schools.

We had the opportunity to collect long-run data on a set of indicators in 2022 after the nearly two-year-long Covid-19 induced school lockdowns in Uganda. Given the unique circumstances of the intermediary period, we will present and analyze this data in section 5 as long-run follow-up data.

3.3 Instruments

Our study will make use of a data-set compiled of survey data, classroom observations, school administrative data (including exam results), a researcher-administered assessment and a series of science fairs organized by the DEO in collaboration with the co-authors. These data are collected as follows:

Survey data. Survey data is collected by trained enumerators hired by IPA-Uganda. We work with two types of survey data in this paper: teacher and student. The teacher and student surveys are conducted at the schools in our sample during the third term on the same day of a given academic year.[22] Student surveys were collected in 2019 and teacher surveys were collected in both 2017 and 2019.

Classroom observations. Classroom observation data is collected by trained enumerators leveraging the Stallings observation tool. This tool provides information on teacher time-use in class. Observations are in the form of a "snapshot" in which the enumerator observes what the teacher is doing, the degree of student engagement in the activity facilitated by the teacher, what non-engaged students are doing, and the materials used by teachers and students. We calibrate the Stallings tool so that enumerators collected ten such snapshots for each scheduled class in the school's timetable.

School administrative data. We attempted to collect school administrative data from all schools that agreed to participate in the study. These data include information on teacher and student attendance, student registration, and student exam scores in each term. In practice, this data proved difficult to collect and its use was limited to constructing sample frames for student surveys.[23]

Researcher Administered Student Assessment. In 2019, primary six students who participated in student interviews also responded to questions on an assessment we provided them. In 2022, we carried out assessments with all students in the school. We

[22] We intended to collect household and community-leader surveys in March-April 2020 at the homesteads of respondents. However, during the ongoing Covid-19 pandemic, we were forced to pause these activities.

[23] While they are required by law to do so, many schools did not systematically collect data on teacher and student attendance.

describe both assessments in greater detail, including the variables we construct out of it, in appendix sections C and D.

Science Shows Held in Collaboration with DEO. Each school in Budondo sub-county was requested to hold a science show in October 2019. During these shows, student groups in P6 classes, and organized by teachers at the school, presented results of an experiment they conducted on themes related to the P6 curriculum. Schools formed 79 groups in this way, depending on school size. Two independent judges assessed student performance during the course of these science shows. We describe these shows in greater detail in section appendix sections C and E.

Uganda National Examination Bureau Data. Primary 7 "candidates" register for the Primary Leaving Examination (PLE) in June of each school year. They sit for the examinations in November and results are provided in January to assess whether a candidate can attend secondary school. Students from different schools often sit at a centralized school for exam-taking purposes. Therefore, we collect student registration numbers from each school after students have registered but prior to students having taken the PLE. We merge these registration numbers with official records from the Uganda National Examination Bureau (UNEB) which provide student-level results for Math, English, Science and Social Studies. We purchase these results in February of each year relevant to our study.[24] We collected 964 records from schools in this manner in 2019, which were merged with 899 records in the official UNEB data. The remainder (65 students) did not take the PLE for reasons we do not observe. We categorize these students as failing the PLEwhen analyzing pass rates.

Budondo census data.: Finally, we collect detailed census data in 2018 and 2022 in Budondo sub-county which includes household rosters with information on age, gender, highest education attainment, schooling status in 2018, schooling status in 2017 among other variables. We use this information to provide a richer understanding of the context around schooling in Budondo in 3.4 and appendix section B.2.

Table 2 provides a timeline for all data used in this study. We have post-treatment data for instruments and school terms highlighted in blue in this table. These instruments were used to make observations and measurements in Term 3 of 2019. These outcomes are analyzed in section 4. The treatment effects in this section are conditional on two cohorts of teachers participating in the teacher-training from Budondo subcounty. Additionally, we collected some long-run outcomes on students and teachers in 2022 after schools were re-opened. We describe these in section 5.

[24]Given poor practices around record-keeping in a subset of schools, we were unable to compile the list of pre-registered PLE takers at baseline for the full set of schools in our study. However, the records we collect in 2019 contain information for the 29 schools in this paper.

3.3.1 Balance Checks and Summary Statistics

Table 1 further provides balance tests for a selection of variables across the set of instruments we use at baseline.The balance tests are broken up according to the intersection of the dataset used to generate the variable and the analyzed unit of observation. For example, we compare school characteristics at the school-level and teacher characteristics at the teacher-level, clustering by school. We conduct five specifications in this way: 1) school characteristics, 2) teacher survey, 3) teacher dyad outcomes, 4) classroom observation outcomes, 5) student PLE results.[25] For each specification, we estimate the following equation (with indices changing according to the unit of observation):

$$\text{Treated}_{is} = \beta \mathbf{x_{is}} + \epsilon_{is}, \tag{1}$$

where Treated_{is} indicates the treatment status of individual/classroom snapshot i (or dyad ij) in school s, $\mathbf{x_{is}}$ is the vector of covariates tested for the given unit of observation and ϵ_{is} is the error term clustered at the school level. We wish to understand whether covariates are significantly different across treatment and control schools, the joint distribution (according to each specification) of these differences is reflected in β. In addition to presenting p values of each covariate, Table 1 presents joint tests of orthogonality of these covariates against treatment status, which is reflected in the F score (and the p value associated with the F score). None of the p values associated with the F scores are significant at conventional levels.

3.4 Schooling in Budondo Subcounty

The learning crisis is evident in Budondo subcounty. Grade repeat rates are between 12 and 18 percent across every level of the schooling pipeline (as measured between 2017 and 2018 in the household census and summarized in Table B.1). Appendix Figure B.2 shows the distribution of ages enrolled in different levels of schooling in primary school. Here, we see a very large age variance in schooling level and this variation widens as students advance in school levels (as expected with a high repeat rate). For example, in Primary 1, there is significant density throughout ages 6 and 10, suggesting that either students start Primary 1 quite late, or there is a significant tendency of repeating classes — most likely, both factors play a role. The age-variation in Primary 4 (the transition year between primary and middle school) and Primary 7 increases relative to Primary 1. In Primary 7, there is high density throughout ages 13-18. A likely cause of the patterns

[25] For 2017 PLE results, we needed to collect PLE registration numbers ex post. Some schools had lost these numbers, thus they are only available in 2017 for 26 schools. As we were collecting PLE registration numbers for schools we added to the study in early 2020, COVID-19 induced lockdowns impeded our access to schools which prevented us from collecting 2019 PLE registration numbers from each school. We intend to retrieve these as soon as conditions allow.

in these figures is poor teaching quality which we explore further in our analysis below.[26]

[Vesall: Removed Section Measurement of Outcome Variables and placed it in the appendix.]

4 Identification and Results

We estimate the intent to treat effect (ITT) by comparing outcomes across schools invited to participate in the training (treatment) and other (control) schools. We regress endline outcome y_{isp} for teacher or student i (or dyad ij) in school s and matched pair p surveyed by enumerator e on a dummy T_{sp}, indicating treatment status of school s within its pair. We include a vector of fixed effects depending on the characteristics of the variable selected but will always include pair fixed effects, γ_p, and generally include enumerator fixed effects when available, ζ_e.[27] The coefficient of interest β_1 is the intent to treat effect. Since treatment is assigned at the school level, we cluster standard errors at the school level. Formally, the estimating equation for the ITT is equation 2:

$$y_{ispe} = \beta_0 + \beta_1 T_{sp} + \gamma_p + \zeta_e + \epsilon_{ispe} \tag{2}$$

We pre-committed to using estimators that cohere with the distribution of outcome variables of interest. The table of all specifications and estimators, linking variables described in section C to a variant of equation 2, including the unit of observation and each specification's associated fixed effects is available in Table G.3.

To test the robustness of our results, we report two sets of p values associated with the ITT in each specification. First, we report the conventional p value derived from the use of the estimator specified in Table G.3. That is, we will present p values that test the null that the ITT is significantly different from zero for the average school. Second, following Young (2018), we report the p value derived from randomization inference with at least 1,000 permutations of the treatment variable. That is, we will present p values that test the sharp null that the ITT is zero for all schools in the sample. These p values have the added benefit that they allow for tests that are exact — with a distribution that is known no matter the sample size or characteristics of the error terms. Nevertheless, when drawing inference from our results we will take results of all tests into consideration. Since we are testing multiple outcomes for a number of our hypotheses, we control the false discovery rate (FDR) across outcome measures within each hypothesis using the Benjamini-Hochberg method detailed in Benjamini & Hochberg (1995). The set of pre-

[26] A more thorough description of the schooling and socio-economic context in Budondo sub-county is available in appendix section B.2

[27] Enumerator fixed effects are not available, for example, in administrative data. Nor do they make sense to use when we construct teacher outcomes using student responses. Variables stemming from classroom observations of teachers will also include "class" or "grade" fixed effects.

specified results, containing both specification and randomization inference p values are presented in Tables 3 through 5. These tables also present the Benjamini-Hochberg corrected critical p values at the 5% level and furthermore highlight, using the $^{\pm}$ symbol, whether a given test can be rejected at either the 5% or 10% level after accounting for the FDR under the critical p values at the 5% and 10% levels respectively. Given the sensitivity of test statistics to the small sample size, we separately windsorize our data by chopping off the top and bottom 5% of observations within treatment and control schools separately. We replicate analysis with this sample and present results in appendix Tables G.4 through G.6. We note that the statistical significance of all of the results we discuss below are stronger after windsorizing our data.

4.1 Tests of (H2.1): Traditional Learning Outcomes.

Table 3 presents results associated with our analysis of high-stakes primary leaving exam scores and pass-through rates to secondary school and from P6 to P7. In the first four columns, we find that students who attend a school whose teachers were invited to participate in KN's training scored 0.64σ higher in English, 0.56σ higher in Science, and 0.44σ higher in social studies than students in control schools — all statistically significant at the 1% level and robust to multiple hypothesis testing (controlling the FDR across all variables in the table) and randomization inference. Students perform higher in math by 0.12σ, but this result is not statistically significant.[28] The Ugandan National Examination Bureau aggregates scores across all subjects to produce an "aggregate" score. Students in treated schools perform 0.51σ higher than students in control schools. Additionally, student performance along this dimension in treated schools has first-order stochastic dominance over the distribution of performance in control schools as is inferred from Appendix Figure H.1b, suggesting that the program increases outcomes for every student across the distribution of student performance.[29]

The fifth column shows that students in treated schools are 24 percentage points more likely to pass the PLE examination at division three or above relative to a control school mean of 51% — this is in the same range as the number reported by UNICEF of 59% of primary-school leavers transitioning to lower secondary school.[30] In the sixth column, we show that students in P6 in 2018 are 12 percentage points more likely to transition to P7

[28]The result is, however, significant upon windsorizing the data as demonstrated in Table G.4, which shows a treatment effect of 0.19σ after chopping off the top and bottom 5% of the sample within treatment and control schools separately. Nevertheless, as we discuss in the introduction, of the six PSA materials KN studies in the teacher training, they have yet to find a way to integrate a book specific to mathematical capabilities. These books exist in the PSA curriculum and as a result of these findings KN is now considering what it can do to integrate more of the lessons from these materials into its training.

[29]We provide distributions of non-normalized performance across all measures in Table 3 in Figure H.2, distinguishing between the distribution of performance in control and treatment schools.

[30]Students scoring below division 3 are almost never admitted into secondary schools — even students scoring in division 3 may struggle to gain admittance.

than students in control schools. This is relative to a control school mean of 0.81, which is similar to the range reflected in responses to the household census in 2017 and reported in appendix Table B.1. Notice that in this final column we include enumerator fixed effects into the specification because the variable is sourced from the student survey, whereas PLE data are sourced from UNEB. These are large effects. We use the same table to extrapolate total increases in expected years of schooling for a P6 student whose chances of progressing to P7 increase by 12 percentage points and whose chances of progressing from P7 to secondary increase by 25 percentage points. Such a student participates in 0.59 extra years of schooling, a 5.5% increase over the average student in control schools. If we extrapolate pass rates to all grades along the pipeline (except for secondary grades), years of schooling for an average student in treated schools increases by 0.86 years, or 9.6% more than the average student in control schools.

4.2 Tests of (H2.2) through (H2.4): Higher-Order Learning, Science Shows and Creativity Outcomes.

In section 2.2, we describe how the pedagogy applied by teachers will also lead students to seek ways of engaging with knowledge in novel and creative ways — here, we test those hypotheses. The first two columns of Table 4 measure treatment effects associated with higher-order learning outcomes. Column one shows that P6 students who attend a treated school scored 0.73σ higher on questions that measure applied understanding and 0.45σ higher on questions that measure critical thinking, both results of which are robust to multiple-hypothesis testing and randomization inference. In third and fourth columns we explore whether students are better able to their critical thinking skills to concrete practices that solve locally relevant problems. The third column does this by measuring the treatment effect associated with our aggregate measure of performance in science shows. We find that students in treated schools score 0.87 points higher relative to a control school mean of 2.63 — a 33% increase.[31] Finally, students score 0.44 points higher in the creativity measure we constructed relative to their control school counterparts, suggesting that they are better able to consider ways in which common, every-day, objects can be used to serve creative and unique ends. While these two tests are robust to multiple hypothesis testing, they are not, however, robust to randomization inference. Looking at appendix figure H.3, it appears possible that this may be driven more by the influence of outliers than any meaningful pattern in the treatment effect that may stray from our interpretation of the data. We discuss this possibility in further detail in section 4.4 below.

[31] We decompose this measure in section 4.4 below to gain further insights.

4.3 Tests of (H1.1) to (H1.3): Teacher Pedagogy and Effort Outcomes

We attribute these improvements in student outcomes to the improved teacher pedagogy. The first evidence of this is demonstrated in the first column of Table 5. Here, we analyze the classroom observer's assessment of the proportion of students who were engaged in the activity led by the teacher during observed class sessions. Recall, we do not ask the observer to record a precise percentage of engaged students, which calls for an ordered logit estimator. We show that for every snapshot observed across P4 through P6, accounting for class fixed effects, classrooms in treated schools are 39% more likely to have a larger proportion of engaged students in class than control schools — the control school average of 4.41 reflects that enumerators indicated that close to half of the students were engaged in an activity with the teacher on average.

Next, we use teacher measures from the student survey to analyze teacher outcomes. Each measure in column 2 and 3 should be interpreted as reflecting the percent of a given teacher's students who stated that they ask questions in class when they don't understand a concept (column 2) and who stated that the teacher has used a form of corporal punishment (column 3), both reflecting actual incidences spanning the two most recent school terms. To facilitate descriptions of results, we call the variable in column 2 "student inquisitiveness." On average, student inquisitiveness is 22% in control schools as depicted in column two of Table 5. This same number is 6 percentage points higher in treated schools as measured by a tobit estimator that treats zero as a lower bound and 1 as an upper bound, suggesting that students are more inquisitive in treated schools. However, they are no less likely to be be subject to corporal punishment in treated schools. The baseline level of corporal punishment is large — the average teacher was mentioned by 53% of her students as a teacher who has used a cane to beat the students in the past. This outcome does not change on average across treatment and control schools, which we suspect is due in part to how it is measured. Recall, we asked each student to mention teachers who applied corporal punishment at any point in the last two school terms. Thus, teachers in treated school who were in the second cohort may have applied corporal punishment prior to the second training which invites them to think more carefully about their relationship with their students. Thus, questions that measure a reduction of an unwanted behavior may not yet manifest in the data at this time.[32]

Despite this result, column four suggests that teachers put more effort into learning intimate details about their students' lives. We quiz teachers on the knowledge they

[32]There is also a strongly entrenched belief, often reinforced by parents, that suggests learning, discipline and upright conduct is developed through corporal punishment. This belief may be difficult for teachers to shake off in the short-run, however the evolution of this belief will be a subject of future research.

have regarding the school attendance and living circumstances (relationship with the guardian they stay with) of a randomly selected subset of students. When matching student and teacher responses, we find that the average teacher in control schools provides correct responses 62% of the time. Teachers in treatment schools are 10 percentage points more likely to provide a correct response using a Tobit estimator that censors observations below 0 and above 1. The distribution of outcomes along these lines is also worth mentioning. Appendix Figure H.4 provides a distribution of all outcome measures.

Column five reports treatment effects associated with teachers' desire to learn, as measured through a change in interactions with colleagues with the objective of learning from one another. The measure reflects the degree of intensity of teacher learning within dyad with 3 being the most intense relationship and 0 indicating no such relationship. We can divide the mean and coefficient by 3 to interpret results in percentage point terms. Teachers in treatment schools are 9 percentage points more likely to form a learning relationship with their colleague relative to a control school mean of 60%.[33]

Results in columns four and five support the view that the emphasis on metacognition in the training facilitates a desire to learn how to overcome a given challenge rather than take it as given. This is reflected in their willingness to learn more about about their students' lives and their willingness to learn with their colleagues about how to improve their pedagogy, consistent with results in columns 4 and 5.

4.4 Additional Discussion

The previous set of results contain pre-specified analysis outlined in our pre-analysis plan. Below, we explore the data further to provide additional insights into the nature of change created by the intervention.

Science Shows. We go into further detail and analyze each of the outcomes associated with the science shows. Science show judges assessed student groups based on 12 outcome variables that we place into five categories: 1) measures that assess pupils' ability to frame the question they are investigating (Framing); 2) measures that assess the quality of the pupils' experiment (Experiment); 3) Measures that assess whether the experiment and hypotheses are well-aligned (Hypothesis); 4) student groups' ability to systematically measure and log outcome variables associated with the experiments and hypotheses (Measurement); and 5) students' ability to independently and accurately articulate the conclusions they drew from the experiment (Articulating). Each of the twelve measures are outlined in detail in Table E.1.

Appendix Figure H.5 displays the distribution of each of the 12 outcome variables and shows a large degree of bunching of outcome variables on the lowest measure — 1 —

[33]Recall that this is unlikely to be due to experimenter demand effects because we verify the existence of a collaborative relationship between two teachers through the response of both teachers within the pair.

across many of the outcomes. As a result, we use Tobit estimators to analyze treatment effects, setting one as the lower bound and 10 as the upper bound across all specifications. Table 6 estimates the treatment effect of each of theses 12 measures.

Apart from outcome measures reflecting the students' ability to frame their research question, we find significant impacts across all categories of outcome measures.[34] In other words, there is reason to believe that we observe positive treatment effects in students' abilities to design an experiment, describe and analyze hypotheses, measure outcome variables related to the hypothesis, and clearly articulate their arguments. The most robust outcomes are variables in the latter two categories, whose randomization inference p values suggest significant impact at the 10% level, accounting for the false discovery rate within outcome variable category.

Furthermore, these effects are large. Take, for instance, treatment effects of 2.12 and 2.37 in columns (11) and (12) respectively. The control school means for these outcomes is 3.20 and 2.65. Therefore, on average, treated schools perform 66% and 89% higher in these two categories respectively.

We also display a sample of the text summaries articulated by the science show judges for the 10 highest performing student groups and 10 median performing student groups in Appendix Tables E.2 and E.3 to provide qualitative insights demonstrating the clear difference between a science project receiving a high score and low score. It is clear from reading these summaries that our elicited measures had their intended effect of measuring an important dimension of student learning — applied critical thinking and reasoning — through science shows.

Experimenter Demand and Hawthorne Effects. One natural concern of a teacher-training intervention is that teachers in trained schools have been conditioned by the training — therefore, their subjective responses to survey questions will reflect that they know what the researchers are looking for. Notice however that none of the variables constructed in our analysis rely on a single teacher's subjective response. Even our main indicators of teacher behavior are provided primarily an outside observer (column 1 of Table 5) or their own (columns 2 and 3 of Table 5). The two measures constructed using teacher responses in columns 4 and 5 of Table 5 are also conditioned by either their students (column 4) or their colleagues (column 5). As a result, we do not believe that there should be any issue concerning experimenter demand effects.

Hawthorne effects are similarly unlikely since Table 3 provides results from a standardized test score administered by a government institution. Students across treatment and control schools were unlikely to change their effort because their results would be used in the study — passing the PLE is already a high-stakes exercise requiring most of

[34] In retrospect, the letter we crafted together with the Jinja DEO framed the research exercise for the students and teachers, so it is not surprising that there is no treatment effect along the "framing" dimension.

a child's attention.

Active Learning Pedagogy. Among the arguments we make in our theory in Section 2.2, we claim that teachers who go through the training are better able to implement active learning pedagogy. Active learning pedagogy involves students in an experiential learning process similar to the ones that teachers in KN's training participate in. We use teacher survey responses and classroom observations to show that teachers are, in fact, utilizing pedagogical principles that suggest a more active learning environment. We do this in two ways.

First, in the teacher survey we ask teachers whose schools have a garden plot how many days they spent either in the clearing, planting, weeding, or harvesting stages of growing crops. There is a positive, but mostly insignificant difference between treated and control schools along this measure. However, when we further ask teachers to indicate how many days they spend on the garden plot working *with their students*, we see large and significant differences. Using a Tobit estimator censored below by zero, Table G.8 shows that teachers in treated schools spend 1.93, 2.80, 2.52, and 2.14 more days engaged in each of the four activities, respectively, than their control school counterparts.

Second, a teacher adept at implementing active learning pedagogy can engage learners in learning activities even when she is not around. This is often the case in P7, where students are tasked with "revising" course content from previous years of school in preparation for the PLE. We observed classrooms of PLE takers two months prior to the exam in 2019 and noted that 20.1% of the time, teachers were "out of the classroom" during our observations.[35] However, just because teachers are not around does not mean that students are not engaged in a learning activity. To check whether this is the case, We further observe what activity the students are engaged in when the teacher is not conducting an activity with pupils. There are 168 instances of classroom observations where this is the case in P7. In these cases, the classroom observer can indicate whether the students are distracted by indicating that learners are "not engaged" or "socializing," or that learners are engaging in the following learning activity categories "assignment or classwork," "copying," or other.[36] Figure H.6 shows the distribution of activities in P7 classrooms when teachers are not around, where we see that learners are engaged in learning activities 75.2% percent of the time in treatment schools while only engaged in learning activities 52.7% of the time in control schools.

Teachers are more inquisitive. We also argue that the teacher training makes teachers more inquisitive and their teaching becomes less rote-based. The latter is reflected in the improvements in student learning in treatment schools along dimensions of higher-order learning. We present additional evidence here that teachers are more

[35] This as compared to 12.7% of teachers in P4 to P6 who were "out of the classroom."

[36] We compile "other learning activities" using responses that suggest the learners are "Reading out loud," engaged in "Explanation/lecture," "Interactive demonstration," "Question and answer/ discussion," or Practice and drill.

inquisitive. In the teacher survey, we ask teachers whether they have ever been asked a question by their pupils that they don't know the answer to (1 if yes, 0 if no). Table G.9 shows that there is no significant difference across treatment and control in response to this question. However, conditional on responding yes, we further ask teachers to share the most common way they respond to such a question. Teachers respond freely and we record their responses according whether they carry out further research (inquisitive) or if they ignore the question (ignore).[37] We find that teachers in treatment schools are 10 percentage points more likely to state that they are inquisitive and 6 percentage points less likely to ignore the student.

We do not believe these answers to be driven by experimenter demand. In order to suggest this, we interact school-level average teacher inquisitiveness with treatment and regress the interaction effect against science show outcomes in Table G.10. Teacher inquisitiveness is negatively correlated with performance in control schools, but positively correlated in treatment schools across all categories of outcomes, a result we depict graphically in Figure H.7. What this result shows is that teachers who present themselves as inquisitive are more able to improve science show performance in treated schools as opposed to control schools. This suggests that teachers in treated schools know how to use their inquisitiveness to promote student exploration. Inquisitiveness may be a correlated feature of a broader capability that enables teachers to facilitate a process of exploration that leads to learning.

Teachers are more sympathetic and less adverserial. Our pre-specified result on corporal punishment does not show much change across treatment and control school outcomes. However, there is still reason to believe that teachers become more sympathetic and less adverserial towards their students. In the teacher survey, we ask a series of yes and no responses in this regard. First, we ask whether students are free to disagree with their teachers about concepts they are being taught. Second, we ask if students are free to decide how to spend time during a lesson. Third, we ask teachers whether they agree with the statement that they are "learn as much from my pupils as they learn from me" (on a five-point likert scale, we code a response of "Strongly Agree" — the strongest agreement response — as 1 and 0 otherwise). We show that teachers are 50% more likely to agree with the first statement and 76% more likely to agree with the third, both significant at five percent and one percent levels, respectively. An index that averages across all three variables is also significinat and positive at the 1% confidence level.

We also ask teachers how they respond to pupils who mis-behave in class. Teachers can select from a menu of strategies and we code their responses as "gentle" or "harsh." Table G.12 shows that teachers in treatment schools are 56% more likely to provide gentle

[37]For the former, teachers response is either "Tell the student you do not know the answer and will do research to find the right answer" or "suggest to the student how to investigate the answer on his or her own." For the latter, teachers response is either "Ignore the student and do not respond" or "Provide the best response you can provide."

responses and 43% less likely to provide harsh responses.

Again, the above results could be consistent with experimenter demand effects. Anecdotal evidence suggests that as teachers learn more about their students' lives, they often find that student misbehavior is a result of student hunger. To address this, they often ramp up school feeding programs. If this is the case, then students should be less hungry at school in treatment schools, a hypothesis we test in appendix Table G.13. We ask students whether they ate a meal at school yesterday and whether they agree with the statement "I am usually hungry during the school day" (five point likert scale). We show that students in treated schools are 17 percentage points more likely to eat a meal at school the previous school day and are more likely to disagree with the statement about hunger at school.

5 Long-Run Outcomes Measured in 2022

As described earlier, we had originally hoped to augment the sample by 53 additional schools in 2020 and study end-line outcomes in 2021, where we would also study long-term follow up results in the original set of schools in our study. This plan was thwarted by the Covid-19 pandemic, which resulted in a two-year school lockdown in Uganda. Schools were not re-opened until February 2022 (only "candidate classes" were allowed to meet in 2021), one additional school closed entirely (causing us to drop two schools from the study), the approach to training was constrained by an inability to carry out tutor visits to schools, and many teachers chose not to return to their previous posts. Indeed, of the 388 teachers from these schools in 2019, only 236 (61%) remained in 2022. The total number of teachers reduced from 388 to 340 — consistent with anecdotes around the country that seemed to suggest a general drop in the teacher population due to post-covid exit from the profession. This also affected the teachers trained prior to 2020 in similar proportion — such teachers made up 23% of all teachers in treated schools in 2022. Kimanya Ngeyo continued with the study portion of their training, studying in small groups of teachers to ensure adequate social distancing. However, since schools were closed, tutors were unable to follow up with teachers in classrooms, reflecting on teacher practices for the two years of government-imposed school closures. With this caveat in mind, we still note that by the time data was collected in 2022 58% of teachers found in treated schools in 2022 (112 of the 192) had participated in training of some kind — though only 55 of these participated in the training prior to 2020 under the original model.

5.1 Persistence and Enrollment in Secondary Higher in Treated Schools

In 2022, we collected data on all school-going children in our study sample using a household census. We briefly interviewed the most senior member of the household, requested that they list all household members and their school-going status for each year since 2018. This allows us to assign a treatment status to children who attended schools in our sample in 2018 in order to estimate measures on long-run enrollment in primary or secondary school as well as persistence through school. We can study two questions: did the intervention affect student enrollment in any school after four years? did the intervention increase secondary enrollment rates after four years of the intervention? did the intervention increase persistence, the observation of a student passing through successive classes, after four years?

Specifically, we define an outcome variable related to each of these variables in 2022 and assign students to treatment based on the school they were enrolled in in 2018. For the first question, we define an outcome variable using a dummy equal to one if a student is enrolled in any kind of school in 2022. For the second question, we define an outcome variable equal to one if a student is enrolled in secondary school in 2022 (we only define this for students in grades 5 or higher in 2018). For the third question, we take the difference between the student's grade level in 2022 and 2018. If a student was not enrolled in school in 2022, we replace missing values with the most recent grade in which the student was enrolled.

Table 7 presents outcomes in three panels, one for each question. Column one aggregates all observations of students enrolled in sample primary schools in 2018 and the remaining seven collumns disaggregate by the student's 2018 grade-level. We find strong and large positive effects across all relevant specifications. The only meaningful null results are in Panel A, which analyzes enrollment in any school, for students enrolled in P1 or P2 in 2018. Notice that the control-school average of whether a student continues to be enrolled in 2022 for this subset is at least 0.97, consistent with a more general observation that almost all primary-school aged children enroll in universal primary school. Consistent with our PLE results from 2019, we find that students enrolled in treated primary schools are 10 percentage points more likely to be enrolled in a secondary school in 2022. Finally, we find that treated students across all grades, including lower primary, have passed through more grades relative to their control school counter-parts. The effect seems to increase for students enrolled in higher grades in 2018 — from 0.09 additional grades for treated P1 students to 0.25 additional grades for treated P7 students.

It is worth noting that this measure of treatment assignment does not reflect any type of treatment intensity at the student-level. For example, students assigned to treated schools in 2018 could have transferred to control schools or vice versa. Thus, the treatment

effects described above are likely lower bounds for the effect of the treatment on the specified outcomes.

5.2 Secondary School Performance when Treated in Primary

We are interested in understanding whether the ability to "learn how to learn" transfers to the students — one indicator of this is whether treatment effects persist to learning outcomes even after students' exposure to treated teachers ends. In 2022, one of our co-authors initiated a RCT within secondary schools, where he conducted a student learning assessment in lower secondary (S1 to S4, corresponding to grades 8 through 11 in the U.S. context). We use these data to analyze a long-run effect on student learning.

Following a fuzzy match procedure, we identified 528 students in the secondary assessment data who we could identify in the 2022 household census with a high degree of confidence (following manual checks). Of these, 320 attended primary schools from our sample at some point between 2018 and 2021 — 189 students attended a treated school in 2018 and 131 did not (only 9 of the 131 attended a treated school after 2018 before entering secondary school). Furthermore, 165 of the 320 were in S1, 89 in S2, 63 in S3 and 3 in S4. Consistent with our effects on persistence, a larger share of treated students identified in the assessment data are in later grades of secondary (53% in S2 and above) relative to control students (42%), which increases our confidence in the matches we found. Questions for the assessment were compiled from past Trends in International Mathematics and Science Study (TIMSS) assessments and select reading comprehension questions tailored to the local context. Additionally, we tested students on the precise use of words that teachers engage with during the course of the teacher training — activities include juxtaposing "information" with "concept," "science" with "technology," and "knowing" with "understanding." Greater consistency in differentiating these words across sentences leads to higher performance.

The results suggest a fairly large and persistent effect on student learning even after student exposure to treated primary teachers ends. On average, students who attended primary schools in 2018 outperformed students who attended control schools by 0.33 standard deviations on an index that aggregates all assessment questions into a single outcome. Effects are positive across all subjects (ranging from 0.08 to 0.24 SD), including Biology, Chemistry, Physics, Math, Reading and Language. Effects are statistically significant for Math (0.14 SD) and Reading (0.24 SD) at the 5% and 10% level respectively.

5.3 Teacher and Student Assessments in 2022

Finally, in 2022 we were able to administer and digitize written student assessments to all upper primary students in our student sample to better understand the distribution of learning outcomes across grade levels. We also took this opportunity to administer

subject-specific assessments to teachers to see whether the training affected teacher content knowledge. We were unfortunately not able to administer assessments to teachers in all schools since many teachers were too busy to take our exams. We nevertheless present results for all teachers who were available.

We present student results in table 9. We continue to provide results on critical thinking (in Panel D, similar to Table D.1) but break down additional test questions into their respective subject categories of math (Panel A), English (Panel B) and Science (Panel C). Student learning outcomes remain positive, significant and large for all categories of outcomes, including mathematics in grades four through six in Panel A. Treatment effects on math outcomes are largest in P4 (0.59 standard deviations) and equal to 0.43 standard deviations in P5 and P6. We suspect that these effects are driven by improvements in pedagogy in earlier grades of schooling. Consistent with our earlier result in column 4 of Table 5, it is possible that teachers are more aware of their students' mathematical abilities at younger ages. Given the cumulative nature of math skills, and the importance of having a strong mathematical foundation for later math outcomes, we suspect that improved pedagogy in earlier grade levels results in improved outcomes in later grades as well.[38]

Teacher learning outcomes are presented in table 10. We note that columns 1 and 2 reflect teacher responses to questions related to subject proficiency and critical thinking — we include teacher-subject fixed effects for column 1 since the questions differed depending on the subject that the teacher taught. The critical thinking questions are similar to those we give to primary students.

We do not see a significant effect in either teacher content knowledge or critical thinking. This subjects that teachers across treated and control schools have roughly similar critical thinking skills and knowledge of the content they are teaching — at the very least, they demonstrate that they know the content they should be teaching in both types of schools.

When it comes to teachers' nuanced use of language, we see large and significant treatment effects between 0.34 and 0.36 standard deviations relative to control school teachers. Treatment school teachers are able to more precisely and consistently use the word "understand" or "know" when presented with sentences in which the difference between the use of one or the other word is a bit nuanced. A similar effect is observed in teachers' use of the words "science" and "technology." While these are concepts that are directly explored in the teacher training material, the word "tender" is a more neutral term. Here, teachers are presented with the statement "the principal treated children with tenderness." and were asked to determine whether an additional set of sentences

[38]These results extend to foundational learning outcomes in lower primary as well. We administered oral assessments in literacy and numeracy to ten lower primary students per class-school pair using a common assessment used in Uganda by UWEZO TWAWEZA (UWEZO, 2016). Here, again we find positive effects that are large and significant at P1. Results available upon request.

used the term "tender" in similar fashion (e.g., "the meat was very tender" is not a similar use of this term). Again, we see large treatment effects with regard to this measure on the order of 0.25 standard deviations.

6 Cost Effectiveness Analysis

To gauge the cost-effectiveness of our learning outcomes, we use Learning-Adjusted Years of Schooling (LAYS) by following the procedure outlined by Angrist et al. (2020). LAYS effectively controls for differences in education quality across countries when measuring cost-effectiveness — it seeks to render an absolute measure of cost-effectiveness for education interventions. For example, 10 LAYS per $100 reflects an increase of 10 years of high quality education per $100 spent. Angrist et al. (2020) use 3 LAYS per $100 as a threshold for "the most cost-effective programmes"; the deworming intervention in Miguel & Kremer (2004), for example, is estimated to have 5.68 LAYS per $100.

To calculate LAYS, we divide the treatment effect on aggregate PLE exam scores for 2019 by the benchmark high-performance learning rate of 0.8σ per year.[39] This gives the improvement in learning outcomes expressed in terms of equivalent high-quality schooling years, which Angrist et al. (2020) define as learning-adjusted years of schooling (LAYS). We multiply this number by the number of students exposed to trained teachers. As we discuss in section 2 above, KN trains teachers across all levels of primary schooling, though we have only measured treatment effects on learning outcomes for P6 and P7 students. The treatment effect on higher order learning outcomes using questions similar to the PLE questions (SIPRO, 2019), and without any preparation or prior warning, suggests larger treatment effects for P6 students relative to their P7 counterparts, perhaps because they were exposed to trained teachers at an earlier age. Thus, there is no reason to believe that students in lower levels somehow benefit less from the teacher training, so we do not differentiate between different grade levels and implicitly assume that the effect on lower-level students is the same as that on the students taking PLE exams in 2019.[40]

Dividing this number by per-student costs, we obtain 9.62 LAYS per $100.[41] This number is ranked in the top 5% of all interventions currently measured using the LAYS method in Angrist et al. (2020). Furthermore, the four interventions ranked higher are

[39] We use 0.51σ, which measures the treatment effect for aggregate PLE exam scores (summing over all subjects). Results available upon request, but also note the number is similar to the one presented in Figure H.1b.

[40] We could, alternatively, divide costs by the number of teachers in the training who teach P6 and P7 and re-calculate LAYS. Doing this does not change our LAYS figure dramatically.

[41] Overall annual cost of the intervention for a single cohort of teachers in Budondo is $27,682. This sum covers the original 18 treated schools with 8,541 students as measured in 2017: omitting closed schools, it decreases to 8,361 which is the number we use in our analysis. The treatment effect measured in this paper assumed 1 year of training for each of the 2 cohorts of teachers, so we multiply the annual cost by 2 to obtain our final per-student cost.

largely driven by their low costs — learning gains are likely to plateau in these settings without an improvement in the quality of teaching.[42]

Finally, we note that this estimate is likely a lower bound. We only use improvements along a single dimension of learning measures — specifically, aggregate standardized test score results — however, it is clear from our analysis that there are many dimensions through which learning is improved in this setting. It does not appear that LAYS, or any other cost-effectiveness measure to date, has adopted an approach that captures the possible multi-dimensionality of learning, so we refrain from introducing one ourselves at this time. This is particularly relevant when cost-effectiveness analysis is implemented on programs that only focus on a single learning outcome, such as reading, measured via literacy assessments. The treatment effect on aggregate PLE scores already reduces the dimensionality of learning across four PLE subjects while also leaving out improvements in critical and scientific thinking which are likely to contain learning dimensions that are not correlated with PLE results.

7 Conclusions

A novel intervention aimed at helping teachers learn how to learn has large treatment effects on the quality of student learning along many dimensions — standardized test scores improve as do our measures of higher order learning, scientific capabilities and student creativity. We outline a conceptual framework that proposes a way of understanding these changes in terms of changes in the teacher. Our evidence shows that teachers adjust their pedagogy and relationships with students by teaching in a more engaging manner, by creating an environment in class in which students ask more questions, by being more aware of details of their students' lives and by increasing the degree to which they learn from their colleagues.

In the paper, we contrast the teacher's new pedagogy, the one promoted by the PSA materials, to her old one, what is sometimes described as "banking pedagogy". This new pedagogy is also different from the kind of highly structured pedagogy which has shown some success in developing countries (Piper et al., 2014, Banerjee et al., 2017, Piper et al., 2018, Muralidharan et al., 2019). Structured pedagogy is based on the idea that for learning certain specific skills, it is important that the student goes through a set of fixed steps that are universally applicable. While this is likely to be an important feature of a high-quality educational experience, it is clear that it cannot be a substitute for promoting a deeper understanding and that this part of learning probably becomes more important as students move to higher grades. To get to this, teachers need to get students

[42] Angrist et al. (2020) Describe these four interventions as "providing information on the returns to schooling in Madagascar (Nguyen, 2008), creating school links to village councils in Indonesia (Pradhan et al., 2014), and grouping students by ability level in Kenya (Duflo et al., 2011)."

to reflect on questions that do not, or sometimes can not, have a textbook answer. To guide a learner through such an experience requires skill and human interaction through a mode that is unlikely to be replaced by a script. Our evidence suggests that it requires that the teacher herself possesses the capabilities required of someone willing to learn about the world around her.

Anecdotal reports from KN tutors as well as our own evidence also suggests that this transition to a new pedagogy moves hand in hand with changes in teacher-student relationships. As teachers begin to ask more open-ended questions, they start to realize that students are reluctant to speak up. After all, the students were accustomed to having a teacher feed them the information they absorbed in hierarchical fashion. Over time, the teacher engages the pupil in new ways, often by engaging them outside of class in village settings and becoming more aware of the challenges pupils face at home. This results in a softening of their relationship, which allows the pupils to be more open in class and facilitates the teacher's application of the new pedagogy.

Our conceptual framework suggests that a key difference between treated and control schools is that students in treated schools were taught in ways that increasingly exposed them to conceptual understandings of content as opposed to learning that promotes memorization and rote learning. Given the emphasis on measuring rote learning in standardized tests in Uganda (Burdett, 2017), we expected that the increase in PLE scores would be quite modest and were surprised to learn of the magnitude of the effect after two years of the training — one that increased the effective (division 3) pass rate by 25 percentage points (50% above control school equivalent).

The size of the impact hints at the potentially important relationship between conceptual mappings and absorption of information. There is growing research suggesting that people are not always attentive to information they are exposed to (e.g., Hanna et al. (2014)), or they ascribe incorrect meanings to information signals they receive (e.g., Michelson et al. (2021)). The proposed policy response is often to improve on the information signals (e.g., by "nudging" or labeling) to prevent mistakes in judgement. However, this presupposes that proper judgement cannot be taught and learned. Our results suggest that a better understanding of underlying concepts within a field of inquiry might promote higher absorption of information by an individual.[43]

We hope this paper will promote further research into "learning to learn" teacher training models beyond primary school settings in Uganda. Indeed, we intend to study a modified teacher training program offered by KN in secondary schools as soon as conditions allow.[44] Moreover we imagine taking this set of ideas beyond the school setting. For example we want to ask whether something similar could be applied to the relation-

[43]For example– as explored in one of the PSA materials– would an increase in a farmer's conceptual understanding of the ways in which genetic and environmental factors influence crop growth affect their ability to absorb information accurately?

[44]This paper was written in 2020 during the height of the Covid-19 pandemic.

ship between agricultural extension workers and smallholder farmers. These field officers apparently often operate under the assumption that smallholders lack knowledge of new technologies and approaches and it is their job to "explicate" this knowledge to them.[45] What would happen if instead the extension workers learned how to learn and extended this approach to learning to smallholder partners through processes of action-research?

More broadly, we note that this orientation to learning contributes to a growing conversation on how our conception of social reality shapes behaviors. Kranton (2019) argues that the assumptions laden in the policy-environment we act in often shapes the behaviors of human beings.[46] In this paper, we provide evidence that when teachers' assumption of knowledge changes in a way that is reflected in their attitudes and in their practice, then behaviors also change — this is absent of any change in incentives for the teacher.

We note, too, that this paper has left out an important feature of KN's teacher training: participants in the PSA program are encouraged to view themselves as "promoters of community well-being" and to seek ways of applying their approaches to learning to advance processes of community development. We plan to study these community level impacts in the future.

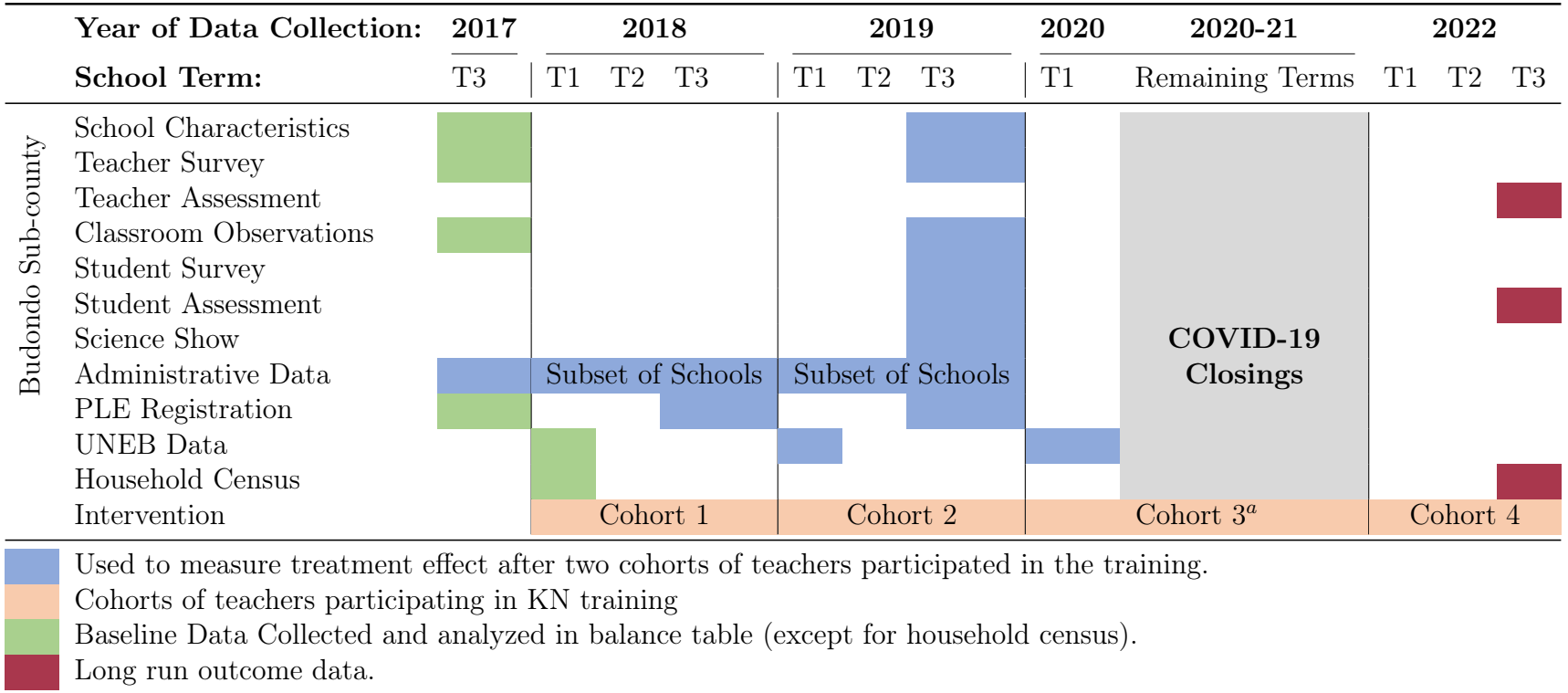
[45] We discuss the meaning of this term in section 2.2.

[46] There are many examples in Bowles (2017), the book Kranton (2019) is reviewing, that show, for example, that monetary incentives crowd out embedded norms and moral codes when they do not consider the context people may have previously used to make decisions.

Statistics	Summary Statistics								Balance Tests	
	All		Control			Treated				
	Mean	Sd	Mean	Sd	N	Mean	Sd	N	β	p value
School Characteristics										
Gov vs. Private	0.48	0.51	0.43	0.51	14	0.53	0.52	15	0.15	0.67
N Teachers	13.21	4.05	11.79	2.42	14	14.53	4.84	15	0.02	0.83
Percent P6 and P7 Teachers	0.43	0.08	0.45	0.07	14	0.41	0.08	15	-0.45	0.79
Percent Teachers Have Diploma	0.24	0.23	0.27	0.24	14	0.20	0.21	15	-0.10	0.91
Percent Any In-Service Training < 2017	0.21	0.18	0.22	0.18	14	0.20	0.19	15	-0.17	0.78
Enrolment	520.28	289.42	487.64	268.35	14	550.73	314.01	15	0.00	0.73
Pupil-Teacher Ratio	38.20	14.78	40.12	17.21	14	36.41	12.43	15	-0.01	0.71
Specification Statistics								**F Score (p value)**	1.74 (0.14)	
								Clusters	29	
Teacher Survey Outcomes										
Teacher Gender	1.35	0.48	1.35	0.48	113	1.34	0.48	132	0.00	0.99
Attended Training Last Year	0.26	0.44	0.28	0.45	113	0.23	0.43	132	-0.06	0.47
Farm Land	0.68	0.47	0.73	0.45	113	0.68	0.47	132	-0.08	0.60
Boys Better Pupils	3.19	1.52	3.32	1.51	113	3.11	1.55	132	-0.02	0.46
HT Listens	1.31	0.69	1.40	0.83	100	1.24	0.55	119	-0.07	0.30
How Satisfied	2.06	0.86	2.07	0.89	113	2.07	0.85	132	-0.01	0.83
Connect Family	1.86	0.85	1.94	0.97	113	1.84	0.74	132	-0.02	0.62
Specification Statistics								**F Score (p value)**	0.51 (0.82)	
								Clusters	29	
Teacher Dyad Outcomes										
ij Speak About Classroom Management	0.84	0.37	0.84	0.36	1,280	0.83	0.38	2,057	0.03	0.45
ij Visit Each Others Classrooms to Learn	0.79	0.41	0.81	0.39	1,280	0.77	0.42	2,057	-0.07	0.30
ij Plan Classroom Activities Together	0.65	0.48	0.68	0.47	1,280	0.63	0.48	2,057	-0.04	0.55
Specification Statistics								**F Score (p value)**	0.77 (0.52)	
								Clusters	29	
Classroom Observation Outcomes										
Share of Engaged Pupils	4.04	2.18	3.96	2.17	1,611	4.16	2.16	1,549	0.02	0.15
Activity: Q and A	0.16	0.36	0.16	0.37	1,612	0.15	0.35	1,551	-0.05	0.30
Activity: Practice and Drill	0.03	0.17	0.03	0.17	1,612	0.03	0.17	1,551	-0.03	0.78
Teacher Out of Class	0.24	0.43	0.24	0.42	1,612	0.24	0.43	1,551	0.07	0.44
Materials: None	0.50	0.50	0.50	0.50	1,598	0.49	0.50	1,536	0.05	0.48
Materials: Textbooks	0.04	0.20	0.03	0.17	1,598	0.05	0.22	1,536	0.19	0.11
Materials: Blackboard	0.36	0.48	0.35	0.48	1,598	0.36	0.48	1,536	0.05	0.48
Specification Statistics								**F Score (p value)**	1.27 (0.30)	
								Clusters	29	
Student PLE Outcomes										
PLE English	7.02	1.69	7.39	1.55	364	6.74	1.73	486	-0.08	0.05
PLE Science	6.55	1.79	6.88	1.80	364	6.31	1.75	486	-0.04	0.47
PLE Maths	7.08	1.51	7.25	1.50	364	6.96	1.50	486	0.02	0.24
PLE Social Studies	5.59	1.75	5.80	1.69	364	5.43	1.79	486	0.05	0.31
Specification Statistics								**F Score (p value)**	1.86 (0.15)	
								Clusters	27	

Notes: This table reflects balance tests of teacher, teacher dyad, and classroom outcomes and covariates. We present summary statistics of each measure, displaying means and standard deviations for the whole sample "All," the sample of teachers/teacher dyads/classroom observations in control schools "Control," and the sample of teachers/teacher dyads/classroom observations in treatment schools "Treated" (the latter two also include number of observations). We have baseline data for all schools for variables that do not use the Student Survey in their construction. For school admin data, including teacher attendance, we were in the process of collecting baseline data outside of Budondo sub-county in March 2020 when the Covid-19 related lockdowns started — we exclude these variables form balance tests here. Balance tests reflect an OLS regression with the specification $\text{Treated}_{is} = \beta X + \epsilon_{is}$ where i represents student, s represents school, X represents the vector of covariates in the rows of this table, β is the vector of coefficients associated with each covariate and ϵ_{is} is the error term clustered at the school level. Each specification is run for each data set separately, datasets are separated by the horizontal lines in the table. The F Score and number of clusters is reported for each specification. The F Score's p value (in parentheses) reports results of the null hypothesis test that coefficients are jointly orthogonal within a given specification.

Table 1: Summary Statistics and Balance Table



Notes: [a]Cohort 3 did not participate in the training as designed because groups studied in small groups to promote social distancing and tutors were not able to accompany teachers in schools to advance teacher "reflections" based on classroom observations (schools were closed and did not accept visitors).

Table 2: Timeline of Data Collection in the Study

Hypothesis:	**(H2.1): Traditional Learning Outcomes**					
	Primary Leaving Exam Results 2019					**Passthrough 2018-2019**
Outcome Variable:	English	Science	Mathematics	Social Studies	PLE Pass	P6 to P7
Treatment (*ITT*)	0.64***	0.56***	0.12	0.44***	0.24***	0.12***
	(0.13)	(0.15)	(0.15)	(0.12)	(0.07)	(0.04)
H_0 : *ITT* = 0						
p value	[0.00]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.43]	[0.00]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.01]$^{\pm\pm}$
RI p value	[0.01]$^{\pm\pm}$	[0.03]$^{\pm\pm}$	[0.55]	[0.02]$^{\pm}$	[0.04]$^{\pm}$	[0.06]$^{\pm}$
BH Critical p value (5%)	[0.01]	[0.03]	[0.05]	[0.02]	[0.03]	[0.04]
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Enum FE	No	No	No	No	No	Yes
Source of Data	UNEB	UNEB	UNEB	UNEB	UNEB	Student Survey
Unit of Observation	Student	Student	Student	Student	Student	Student
Standardized Variable	Yes	Yes	Yes	Yes	No	No
Range of Outcome Variable	[-1.07, 3.47]	[-1.29, 2.96]	[-1.24, 3.49]	[-1.54, 2.68]	{0,1}	{0,1}
Control School Mean	0.00	0.00	0.00	0.00	0.51	0.81
Clusters	29	29	29	29	29	29
Observations	899	899	899	899	964	328
Estimator	OLS	OLS	OLS	OLS	OLS	OLS

Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report p values using randomization inference ("*RI p value*") as well as the Benjamini-Hochberg (BH) "*BH Critical p value*" at the 5% level within hypothesis. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. Randomization Inference using 1,000 permutations of school-level treatment indicator within matched-pair strata.

Variable Descriptions: Units of observation in all columns are students, though the data sets and, hence, number of observations differ. The **first through fifth columns** utilize official test scores for the primary leaving exams (PLE) purchased from Uganda's National Examination Bureau in 2020, relating results from tests taken in November 2019. Students can receive scores ranging from 1 (best) to 9 (worst) for each exam. We have transformed this measure to facilitate interpretation such that 9 is the top score and 1 is the worst score (reflected in figures H.2a through H.2d). Furthermore, all measures in the first four columns have been standardized using the pooled subject-specific mean and standard deviation of control schools. Therefore, **treatment effect measures reflect standard deviation differences between treatment schools and control schools in columns one through four**. We analyze the PLE pass rate and P6 to P7 pass-through rate in **columns 5 and 6 respectively**. Both are binary measures equaling 1 if a student passed the exam/grade level and 0 otherwise. The number of observations in column five is greater than that in columns one through four because students who registered but were not present for the PLE exam are classified as students who must repeat the PLE exam in the following year. Students are divided into four "divisions" based on PLE scores, division 1 reflects high performance and division 4 is the lowest. Students receiving division 4 marks and below are considered to have the PLE (typically with an aggregate score above 28). The variable analyzed in column 6 is constructed using the student survey. For more details, see pages 20-21 of the pre-analysis plan. We standardize variables in columns one through four according to the control school mean and standard deviation.

Table 3: Preliminary Student Outcomes for Hypothesis **(H2.1)**: Traditional Learning Outcomes Increase

Hypotheses:	**(H2.2): Higher Order Learning**		**(H2.3): Science Show**	**(H2.4): Creativity**
Outcome Variables	Apply/Understand	Critical Thinking	Index (Mean)	Index
Treatment (*ITT*)	0.73***	0.45**	0.87**	0.44**
	(0.14)	(0.16)	(0.40)	(0.20)
H_0 : *ITT* = 0				
p value	[0.00]$^{\pm\pm}$	[0.01]$^{\pm\pm}$	[0.03]$^{\pm\pm}$	[0.03]$^{\pm\pm}$
RI p value	[0.00]$^{\pm\pm}$	[0.02]$^{\pm\pm}$	[0.13]	[0.35]
BH Critical p value (5%)	[0.03]	[0.05]	[0.05]	[0.05]
Pair FE	Yes	Yes	Yes	Yes
Enum FE	Yes	Yes	Yes	Yes
Source of Data	Student Assessment	Student Assessment	Science Show	Student Survey
Unit of Observation	Student	Student	Student Group	Student
Standardized Variable	Yes	Yes	No	No
Range of Outcome Variable	[-2.00, 1.68]	[-1.75, 1.95]	[1, 9.16]	[0, 20]
Control School Mean	0.00	0.00	2.63	6.21
Clusters	29	29	29	29
Observations	329	329	158	329
Estimator	OLS	OLS	Tobit	Tobit

Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report p values using randomization inference ("*RI p value*") as well as the Benjamini-Hochberg (BH) "*BH Critical p value*" at the 5% level within hypothesis. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. Randomization Inference using 1,000 permutations of school-level treatment indicator within matched-pair strata. Tobit estimator in the third column treats 1 as the lower bound and 10 as the upper bound; in the fourth column, 0 is the lower bound with no upper bound.

Variable Descriptions: The unit of observation in the first, second and fourth columns is individual students, and student groups in the third column. Each group was judged by two separate judges, thus there are 79 total groups across 29 schools in the analysis. The **first and second columns** report results from the researcher administered student assessments, measuring applied understanding and critical thinking using the procedure outlined on pages 21 and 22 of the pre-analysis plan. The science show outcome in the third column averages across the 12 outcomes measured by science show judges and described on page 25 and table 1 of the pre-analysis plan. The creativity score follows the procedure described on pages 26 and 27 of the pre-analysis plan. We standardize variables in columns one and two according to the control school mean and standard deviation.

Table 4: Preliminary Student Outcomes for Hypotheses (H2.2) through ((H2.4)): Higher Order Learning, Science Shows and Creativity

Hypothesis:	**(H1.1): Pedagogy**			**(H1.2): Effort**	**(H1.3): Learning**
Outcome Variable:	Share of Engaged Pupils	Student Inquisitiveness	Corporal Punishment	Knowledge of Student	Teacher Network
Treatment (*ITT*)	0.39***	0.06**	−0.01	0.10***	0.27**
	(0.15)	(0.03)	(0.04)	(0.03)	(0.11)
H_0 : *ITT* = 0					
p value	[0.01]±±	[0.02]±±	[0.81]	[0.00]±±	[0.02]±±
RI p value	[0.01]±±	[0.28]	[0.94]	[0.06]±	[0.14]
BH Critical p value (5%)	[0.02]	[0.03]	[0.05]	[0.05]	[0.05]
Pair FE	Yes	Yes	Yes	Yes	Yes
Enum FE	Yes	Yes	Yes	Yes	No
Grade FE	Yes	No	No	No	No
Source of Data	Classroom Observations	Student Survey	Student Survey	Stud. + Teach. Survey	Teacher Network
Unit of Observation	Classroom Snapshots	P6 Teachers	P6 Teachers	P6 Teachers	Teacher Dyads
Range of Outcome Variable	{1,2,...,6}	[0,1]	[0,1]	[0,1]	{0,1,2,3}
Control School Mean	4.41	0.22	0.53	0.62	1.80
Clusters	29	29	29	29	29
Observations	2,380	95	95	95	1,466
Estimator	Ologit	Tobit	Tobit	Tobit	OLS

Analysis Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report p values using randomization inference ("*RI p value*") as well as the Benjamini Hochberg (BH) "*BH Critical p value*" at the 5% level within hypothesis. Hypothesis categories are delineated in the top row of the table. ± suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; ±± suggests a significant discovery at the 5% level. Randomization Inference using 1,000 permutations of school-level treatment indicator within matched-pair strata. Tobit estimator in the second through fourth columns treats zero as the lower bound and one as the upper bound.

Variable Descriptions: Units of observation in the **first column** are classroom snapshots observed using the Stallings instrument — 10 "snapshots" are taking during each class; therefore, 238 classes (P4 to P6) were observed across 29 schools. "Share of Engaged Pupils" indicates the observer's assessment of the share of pupils engaged in an activity with a teacher: 1 = No pupil, 2 = One pupil, 3 = A few pupils, 4 = Half of the pupils, 5 = Most of the pupils, and 6 = All pupils (details on page 16 of the pre-analysis plan). Teachers in the second to fourth columns are restricted to those who at least 90% of P6 students listed as one of their teachers. The **second column** analyzes the percent of students who indicate that they have asked questions in the previous two school terms to the specified teacher when they do not understand a topic. The **third column** analyzes the percent of students who indicate that the specified teacher has "caned" or beaten them at some point in the previous two school terms (details on page 17 of the pre-analysis plan). The **fourth column** combines student and teacher responses regarding student attendance and the students relationship with their guardians for each student. We analyze the treatment effect on the percent of correct teacher responses (details available on pages 18-19 of the pre-analysis plan). Unit of observation in the **fifth column** is within-school teacher dyads. We sum across binary measures of teacher interactions along three dimensions where both teachers indicate the existence of a link reflecting teacher collaboration (details available on pages 19-20 of the pre-analysis plan).

Table 5: Preliminary Teacher Outcomes for Hypotheses **(H1.1)** through **(H1.3)**: Teacher Pedagogy, Teacher Effort, Teacher Learning

Source of Data	**Science Show**											
Category	**Framing**		**Experiment**			**Hypothesis**		**Measurement**			**Articulating**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment (*ITT*)	−0.79	0.39	1.34**	0.51	0.22	2.29*	2.84**	0.99	2.55**	1.83***	2.12***	2.37**
	(0.81)	(0.94)	(0.62)	(0.78)	(0.92)	(1.18)	(1.25)	(0.87)	(0.98)	(0.62)	(0.79)	(0.96)
$H_0 : ITT = 0$												
p value	[0.33]	[0.68]	[0.03]$^{\pm}$	[0.52]	[0.81]	[0.05]$^{\pm}$	[0.02]$^{\pm\pm}$	[0.26]	[0.01]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.01]$^{\pm\pm}$	[0.01]$^{\pm\pm}$
RI p value	[0.50]	[0.76]	[0.14]	[0.63]	[0.89]	[0.20]	[0.14]	[0.41]	[0.06]$^{\pm}$	[0.05]	[0.04]$^{\pm}$	[0.08]$^{\pm}$
BH Critical p value (5%)	[0.03]	[0.05]	[0.02]	[0.03]	[0.05]	[0.05]	[0.03]	[0.05]	[0.03]	[0.02]	[0.03]	[0.05]
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Standardized Variable	No	No	No	No	No	No	No	No	No	No	No	No
Range of Outcome Variable	[1,10]	[1,10]	[1,10]	[1,10]	[1,8]	[1,10]	[1,10]	[1,10]	[1,10]	[1,10]	[1,10]	[1,10]
Control School Mean	3.24	2.84	3.81	2.19	1.8	2.71	2.4	1.65	2.92	2.16	3.2	2.65
Clusters	29	29	29	29	29	29	29	29	29	29	29	29
Observations	158	158	158	158	158	158	158	158	158	158	158	158
Estimator	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit

Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report p values using randomization inference ("*RI p value*") as well as the Benjamini-Hochberg (BH) "*BH Critical p value*" at the 5% level within hypothesis. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. Hypotheses are grouped according to the categories outlined in the second row of the table. Randomization Inference using 1,000 permutations of school-level treatment indicator within matched-pair strata. Tobit estimator treats 1 as the lower bound and 10 as the upper bound across all columns.

Table 6: Granular Science Show Outcomes

		Class Attended in 2018						
Observation Set (Class)	**All**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
Panel A: Outcome Variable = Is Student Attending School in 2022?								
Treatment 2018 (*ITT*)	0.03**	0.01	0.01	0.02*	0.06***	0.07*	0.05	0.09**
	(0.01)	(0.00)	(0.01)	(0.01)	(0.02)	(0.03)	(0.04)	(0.04)
H_0 : *ITT* = 0								
p value	[0.02]	[0.21]	[0.31]	[0.1]	[0.00]	[0.06]	[0.15]	[0.03]
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Range of Outcome Variable	{0, 1}	{0, 1}	{0, 1}	{0, 1}	{0, 1}	{0, 1}	{0, 1}	{0, 1}
Control School Mean	0.89	0.99	0.97	0.93	0.84	0.8	0.76	0.69
Panel B: Outcome Variable = Is Student Attending Secondary School in 2022?								
Treatment 2018 (*ITT*)	0.1**					0.12**	0.07*	0.08*
	(0.04)					(0.06)	(0.03)	(0.05)
H_0 : *ITT* = 0								
p value	[0.02]					[0.04]	[0.07]	[0.08]
Pair FE	Yes					Yes	Yes	Yes
Range of Outcome Variable	{0, 1}					{0, 1}	{0, 1}	{0, 1}
Control School Mean	0.62					0.52	0.72	0.68
Panel C: Outcome Variable = Class 2022 - Class 2018 (*Persistence*)								
Treatment 2018 (*ITT*)	0.14***	0.09*	0.11**	0.17***	0.19***	0.17*	0.21***	0.25**
	(0.04)	(0.05)	(0.05)	(0.06)	(0.05)	(0.09)	(0.08)	(0.1)
H_0 : *ITT* = 0								
p value	[0.00]	[0.06]	[0.04]	[0.01]	[0.00]	[0.07]	[0.01]	[0.01]
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Range of Outcome Variable	{0, 1, .., 4}	{0, 1, .., 4}	{0, 1, .., 4}	{0, 1, .., 4}	{0, 1, .., 4}	{0, 1, .., 4}	{0, 1, .., 4}	{0, 1, .., 4}
Control School Mean	2.55	2.66	2.71	2.65	2.51	2.39	2.24	2.15
Unit of Observation	School-going child between grades 1 and 7 in 2018							
Source of Data	2022 Household Census (All)							
Clusters	29	29	29	29	29	29	29	29
Observations	6,056	1,264	1,077	1,081	949	712	565	408
Estimator	OLS (All)							

Analysis Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect in which a student is assigned to treatment based on their 2018 school assignment only. All observations in column 1 reflect number of observations in Panel A and C. Total observations in Panel C sums observations across columns representing P5 to P7.

Table 7: Treatment Effects on Long-Run School Enrollment and Persistence

		Subject					
Dependent Variable:	**All**	**Bio**	**Chem**	**Physics**	**Math**	**Reading**	**Language**
Panel A: All Secondary Students Matched with Census							
Treatment 2018 (*ITT*)	0.33***	0.14	0.18	0.08	0.14**	0.24*	0.20
	(0.13)	(0.14)	(0.15)	(0.14)	(0.06)	(0.13)	(0.14)
$H_0 : ITT = 0$							
p value	[0.01]	[0.32]	[0.22]	[0.56]	[0.04]	[0.07]	[0.16]
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Source of Data	2022 Secondary Student Assessment						
Unit of Observation	Secondary School Students Matched with HH Census in 2022						
Standardized Variable	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Range of Outcome Variable	[-2.69, 3.63]	[-2.12, 1.59]	[-1.98, 1.57]	[-1.66, 1.93]	[-.42, 5.60]	[-2.11, 2.08]	[-2.67, 1.06]
Control School Mean	−0.32	−0.18	−0.22	−0.07	−0.15	−0.25	−0.12
Observations	306	306	306	306	306	306	308
Estimator	OLS	OLS	OLS	OLS	OLS	OLS	OLS

Table 8: Long-Term Learning Outcomes for Students Enrolled in Secondary School in 2022

Table 9: Student Learning Effects in 2022

Grade	P4	P5	P6	P7
Panel A: Student Math Scores				
Treatment (*ITT*)	0.59***	0.43***	0.43***	0.04
	(0.15)	(0.08)	(0.15)	(0.08)
Panel B: Student English Scores				
Treatment (*ITT*)	0.45***	0.41***	0.36***	0.25***
	(0.09)	(0.13)	(0.09)	(0.06)
Panel C: Student Science Scores				
Treatment (*ITT*)	0.60***	0.33**	0.51***	0.15
	(0.13)	(0.13)	(0.12)	(0.13)
Panel D: Student Critical Thinking Scores				
Treatment (*ITT*)	0.29***	0.12	0.21***	0.26***
	(0.06)	(0.14)	(0.05)	(0.07)
Pair FE	Yes	Yes	Yes	Yes
Standardized	Yes	Yes	Yes	Yes
Observations	1,105	1,030	973	842
Clusters	27	27	27	27

Analysis Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect based the school's assignment to treatment. All outcomes are standardized within class and effects reflect % change in standard deviation units in treated schools.

Table 10: Teacher Learning Effects in 2022

	Subject	Critical-Thinking	Understand/Know	Science/Technology	Tender
Treatment (*ITT*)	0.07	0.10	0.36***	0.34**	0.25**
	(0.14)	(0.11)	(0.12)	(0.17)	(0.11)
English Teacher	−0.06				
	(0.17)				
Math Teacher	−0.06				
	(0.17)				
Pair FE	Yes	Yes	Yes	Yes	Yes
Standardized	Yes	Yes	Yes	Yes	Yes
Observations	147	251	248	248	248
Clusters	23	24	24	24	24

Analysis Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect based the school's assignment to treatment. All outcomes are standardized and effects reflect % change in standard deviation units in treated schools. Column 1 results only consist of responses from science, math or english teachers' subject assessments. Science teachers is the omitted category.

References

Acemoglu, D., et al. (2012). What Does Human Capital Do? A Review of Goldin and Katz's *The Race Between Education and Technology*. *Journal of Economic Literature*, *50*(2), 426–63.

Angrist, N., Evans, D., Filmer, D. P., Glennerster, R., Rogers, F. H., & Sabarwal, S. (2020, October). *How to Improve Education Outcomes Most Efficiently ? A Comparison of 150 Interventions Using the New Learning-Adjusted Years of Schooling Metric* (Policy Research Working Paper Series No. 9450). The World Bank. Retrieved from https://ideas.repec.org/p/wbk/wbrwps/9450.html

Arbab, F., & Lample, P. (2005). *Educational Concepts.*

ASER Centre. (2019). Annual Status of Education Report (Rural) [Computer software manual]. New Delhi — last accessed November 3, 2020. Retrieved from http://img.asercentre.org/docs/ASER%202019/ASER2019%20report%20/aserreport2019earlyyearsfinal.pdf

ASER Pakistan. (2019). Annual status of education report (provisional) [Computer software manual]. Lahore — last accessed November 3, 2020. Retrieved from http://img.asercentre.org/docs/ASER%202019/ASER2019%20report%20/aserreport2019earlyyearsfinal.pdf

Bando, R., Näslund-Hadley, E., & Gertler, P. (2019, September). *Effect of Inquiry and Problem Based Pedagogy on Learning: Evidence from 10 Field Experiments in Four Countries* (Working Paper No. 26280). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w26280 doi: 10.3386/w26280

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

Baten, J., de Haas, M., Kempter, E., & Selhausen, F. M. (2020). Educational Gender Inequality in Sub-Saharan Africa: A Long-term Perspective. *African Economic History Working Paper Series*, *54*.

Becker, G. S. (2009). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education.* University of Chicago Press, 3rd Edition.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. Retrieved from http://www.jstor.org/stable/2346101

Bold, T., Filmer, D. P., Molina, E., & Svensson, J. (2019). The Lost Human Capital: Teacher Knowledge and Student Achievement in Africa. *Policy Research Working Papers*, *8849*.

Bowles, S. (2017). *The Moral Economy.* Yale University Press.

Bradler, C., Neckermann, S., & Warnke, A. J. (2019). Incentivizing Creativity: A Large-Scale Experiment with Performance Bonuses and Gifts. *Journal of Labor Economics*, *37*(3), 793–851.

Brock-Utne, B. (2002). *Whose Education for All?: The Recolonization of the African Mind* (Vol. 1445). Routledge.

Bruhn, M., & McKenzie, D. (2009, October). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, *1*(4), 200-232. Retrieved from http://www.aeaweb.org/articles?id=10.1257/app.1.4.200 doi: 10.1257/app.1.4.200

Bude, U. (1983). The Adaptation Concept in British Colonial Education. *Comparative Education*, *19*(3), 341–355.

Burdett, N. (2017). Review of High Stakes Examination Instruments in Primary and Secondary School in Developing Countries. *Research on Improving Systems of Education (RISE) Working Paper 17/018*. Retrieved from https://riseprogramme.org/sites/default/files/publications/RISE_WP-018_Burdett_0.pdf

De La Warr, L. (1937). *Higher Education in East Africa: Notes and Correspondence of the De La Warr Commission.* Manuscripts in the Africana Collection, Makerere University Library.

Duckworth, E. (2006). *The Having of Wonderful Ideas and Other Essays on Teaching and Learning.* Teachers College Press.

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

Feierman, S. (1985). Struggles for Control: The Social Roots of Health and Healing in Modern Africa. *African studies review*, *28*(2/3), 73–147.

Flabbi, L., & Gatti, R. (2018). A Primer on Human Capital. *Policy Research Working Paper*(8309). Retrieved from http://documents1.worldbank.org/curated/en/514331516372468005/pdf/WPS8309.pdf

Freire, P. (1970). Pedagogy of the Oppressed. *New York: Continuum*.

Gilligan, D. O., Karachiwalla, N., Kasirye, I., Lucas, A., & Neal, D. (2018). Educator Incentives and Educational Triage in Rural Primary Schools. *IZA Discussion Paper Series*, *11516*.

Glewwe, P., & Muralidharan, K. (2016). Chapter 10 - Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), (Vol. 5, p. 653 - 743). Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780444634597000105 doi: https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Goldin, C., & Katz, L. F. (2009). *The Race Between Technology and Education.* Harvard University Press.

Guilford, J. P. (1967). *The Nature of Human Intelligence.* McGraw-Hill.

Hanna, R., Mullainathan, S., & Schwartzstein, J. (2014). Learning Through Noticing: Theory and Evidence From a Field Experiment. *The Quarterly Journal of Economics*, *129*(3), 1311–1353.

Hanson, H. E. (2010). Indigenous Adaptation: Uganda's Village Schools, ca. 1880–1937. *Comparative Education Review*, *54*(2), 155–174.

Hanushek, E. A., & Woessmann, L. (2012). Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation. *Journal of Economic Growth*, *17*(4), 267–321.

Jansen, J. (1989). Curriculum Reconstruction in Post-colonial Africa: A Review of the Literature. *International Journal of Educational Development*, *9*(3), 219–231.

Kolencik, P. L., & Hillwig, S. A. (2011).
In *Encouraging Metacognition: Supporting Learners through Metacognitive Teaching Strategies.* (p. 183). Peter Lang Inc., International Academic Publishers.

Kranton, R. (2019, March). The Devil Is in the Details: Implications of Samuel Bowles's The Moral Economy for Economics and Policy Research. *Journal of Economic Literature*, *57*(1), 147-60. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jel.20171463 doi: 10.1257/jel.20171463

Kremer, M. (1993). The O-ring Theory of Economic Development. *The Quarterly Journal of Economics*, *108*(3), 551–575.

Lang, J. M. (2012). Metacognition and Student Learning. *The Chronicle of Higher Education*.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing Critical Thinking in Higher Education: Current State and Directions for Next-generation Assessment. *ETS Research Report Series*, *2014*(1), 1–23.

Michelson, H., Fairbairn, A., Ellison, B., Maertens, A., & Manyong, V. (2021). Misperceived Quality: Fertilizer in Tanzania. *Journal of Development Economics*, *148*, 102579. Retrieved from http://www.sciencedirect.com/science/article/pii/S0304387820301541 doi: https://doi.org/10.1016/j.jdeveco.2020.102579

Miguel, E., & Kremer, M. (2004, January). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, *72*(1), 159-217. Retrieved from https://ideas.repec.org/a/ecm/emetrp/v72y2004i1p159-217.html

Ministry of Human Resource Development. (2020). *National Education Policy 2020.* Government of India.

Moulton, J. (2002). Uganda: External and domestic efforts to revive a derelict primary school system. *Contributions to the Study of Education*, *82*, 53–86.

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

National Curriculum Development Centre (NCDC). (2016). *Primary 6 Science Curriculum.* National Curriculum Development Centre.

Nguyen, T. (2008). Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar. *Unpublished manuscript*.

Perkins, D. N. (1992). *Smart Schools: Better Thinking and Learning for Every Child.* The Free Press.

Piper, B., Zuilkowski, S. S., Dubeck, M., Jepkemei, E., & King, S. J. (2018). Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers' Guides. *World Development*, *106*, 324–336.

Piper, B., Zuilkowski, S. S., & Mugenda, A. (2014). Improving Reading Outcomes in Kenya: First-year Effects of the PRIMR Initiative. *International Journal of Educational Development*, *37*, 11–21.

Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., & Artha, R. P. (2014, April). Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia. *American Economic Journal: Applied Economics*, *6*(2), 105-26. Retrieved from https://www.aeaweb.org/articles?id=10.1257/app.6.2.105 doi: 10.1257/app.6.2.105

Ranciére, J. (1991). *The Ignorant Schoolmaster.* Stanford University Press.

SIPRO. (2019). *Sipro Mid-term II Examination 2019: Primary Six.* The SIPRO Educational Services Ltd.

Ssekamwa, J. C. (1997). *History and Development of Education in Uganda.* Fountain Pub Limited.

Swartz, R. J., & Perkins, D. (2017). Psychology Library Editions: Cognitive Science. In R. J. Swartz & D. Perkins (Eds.), *Teaching thinking: Issues and approaches* (Vol. 24). Routledge.

Tanner, K. D. (2012). Promoting Student Metacognition. *CBE—Life Sciences Education*, *11*(2), 113–120.

The World Bank. (2020, October). *Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell us Are "Smart Buys" for Improving Learning in Low- and Middle-Income Countries?* Retrieved from http://documents1.worldbank.org/curated/en/719211603835247448/pdf/Cost-Effective-Approaches-to-Improve-Global-Learning-What-Does-Recent-Evidence-Tell-Us-Are-Smart-Buys-for-Improving-Learning-in-Low-and-Middle-Income-Countries.pdf

Uganda Ministry of Education and Sports (MoES). (2014, Jun). *Teachers Initiative in Sub-Saharan Africa (TISSA) Uganda.* Ministry of Education and Sports, Uganda. Retrieved from https://dakar.iiep.unesco.org/sites/default/files/ckeditor_files/tissa_uganda_-_2014.pdf

Uganda Ministry of Education and Sports (MoES). (2017, May). *A Teacher Incentive Framework (TIF) For Uganda.* Ministry of Education and Sports, Uganda.

United Nations Educational Scientific and Cultural Organization (UNESCO). (2020). Uganda: Education and literacy (statistics) [Computer software manual]. — last accessed November 3, 2020. Retrieved from http://uis.unesco.org/en/country/ug

UWEZO. (2016). Are Our Children Learning? Uwezo Uganda Sixth Learning Assessment Report [Computer software manual]. Kampala: Twaweza East Africa — last accessed November 3, 2020. Retrieved from http://www.uwezo.net/wp-content/uploads/2016/12/UwezoUganda2015ALAReport-FINAL-EN-web.pdf

World Bank Group. (2015). *User Guide: Conducting Classroom Observations.* World Bank Group.

World Bank Group. (2018). World Development Report 2018: Learning to Realize Education's Promise. *Washington, DC: The World Bank*. Retrieved from https://www.worldbank.org/en/publication/wdr2018

Young, A. (2018). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *The Quarterly Journal of Economics*, *134*(2), 557-598. Retrieved from https://doi.org/10.1093/qje/qjy029 doi: 10.1093/qje/qjy029

Online Appendix for Learning to Teach by Learning to Learn

Vesall Nourani Nava Ashraf Abhijit Banerjee

November 11, 2023

Appendix

Table of Contents

A Summaries of Each PSA Course in the Teacher Training

The entire PSA program is organized around the concept of a capability, conceived as developed capacity to think and to act purposefully in a particular sphere of activity. Each unit of instruction has been designed to contribute to the strengthening of one or more such capabilities. The capabilities can be broadly categorized into four areas: mathematics, science, language, and processes of community life. The spiritual dimension of human reality is explicitly discussed given its importance to the lives of most populations served by the FUNDAEC materials, and its role in endowing participant reflections with deeper purpose and meaning. The approach taken by FUNDAEC materials and tutors can be coarsely described as inquisitive and exploratory — allowing participants to be the protagonists of their own learning and not imposing one or another world-view. The materials help participants consider the validity of all perspectives by assisting them to express and verify their own thoughts on a particular matter without blindly adopting any point of view.

Table A.1: PSA Courses Studied During Intense Training Sessions

Term#	Book	Unit	Capacities Developed
1	Primary Elements of Description	Properties	To think about the nature of language and the way it is used to describe reality.
1	Matter	The Heating and Cooling of Matter	Observing and describing the transformation of material substances.
1	DSA	Basic Concepts	—
2	Dawn of Civilization	Transition to Agriculture	Reading Comprehension. Understanding human experience in the context of the evolution of civilization
2	Nurturing Young Minds	To Describe the World	Teach young children (ages 4-5) to describe the world with increasing levels of clarity.
2	DSA	Education	—
3	Promoting a Healthy Environment	Environmental Issues	Capacity to create awareness and identify solutions of community environmental issues.
3	Food Production on Small Farms	Planting Crops	Effective participation in the development, selection, and use of technology.
3	DSA	Science	—

Each of the courses in the training, along with the term during which they are studied and a coarse summary of their associated capacities, are listed and summarized in Table

A.1. There are four aspects of the courses that are worth pointing out. First, the courses chosen are designed to develop participants' facility with the scientific method by discussing and implementing applications relevant to the local context. For example, in "Planting Crops" (T3) teachers learn how to apply experimental methods to planting, use local flora to learn how to describe plant systems, etc. This allows teachers an opportunity to reflect on the ways in which exploratory methods can be introduced into the classroom using processes students themselves are familiar with.

Second, it is worth noting that the FUNDAEC courses have an explicit aim of leading the participants to think of themselves as "Promoters of Community Well-being." Thus, the focus on local application of the scientific method is aligned with this aim. The courses are also explicit about exploring both the "spiritual qualities" a promoter of community well-being might possess in addition to the relationships and qualities a promoter of community well-being seeks to engender in others. Thus, the participants in the course are attentive to how their new-found scientific capabilities can be applied to learning about the material, social and moral dimensions of community well-being. This theme is explored in some depth in the courses covered in the third term of the training. In the unit titled "Planting Crops" (T3) teachers develop an understanding of how experimental methods can be used to learn about the productivity and social impacts of diverse types of agricultural technologies. In the unit titled "Environmental Issues" (T3) teachers explore the relationship between human activity and pollutants stemming from systems of transportation, small- and large-scale manufacturing, and waste-management processes. In both courses, care is taken to allow participants to reflect on the moral implications of various approaches to agricultural production and use of resources that have potential negative spillovers.

Third, and relatedly, many of the concepts teachers learn — and capacities they acquire — are reinforced in subsequent lessons through actions teachers undertake in the communities they live in. For example, in the unit titled "Properties" (T1), teachers are exposed to the primary elements of descriptions. Specifically, they reflect on the words one might use to describe substances that exist in the world, their positions in space and their relationships with one another. Then, in the unit titled "To Describe the World" (T2), teachers are asked to organize groups of pre-school children through hands-on activities that enable them to describe the world around them with increasing levels of clarity. Using resources from their village environment — such as leaves, flowers, sticks, and pebbles — they help children understand the concept of sets, size of sets, characteristics of elements of sets, numbers, shapes, etc. many of which were previously explored by teachers in "Properties." This act of organizing village children to carry out these activities allows teachers a concrete setting in the community to apply the linguistic capacity they gained in earlier courses. Similarly, while studying "Transition to Agriculture" (T2), teachers develop the ability to describe the emergence of culture and civilization in a manner that contextualizes modern experiences by an analysis of the past. This concept is reinforced by asking teachers to interview elders in the community to learn how processes around child-rearing and agricultural production have evolved over generations in the local community.

Fourth, there are natural opportunities embedded in each course for teachers to contemplate the ways in which these capabilities can be used to improve their understanding of and responses to classroom processes. We share one such example when describing the change of phase of a substance through the texts "Properties" and the "Heating and Cooling of Matter" (T1).

Across all of the terms teachers study units in a course titled Discourse for Social Action (DSA). Many of the courses in DSA cover concepts that allow participants spaces for meta-cognitive reflection of the reasons for covering various topics the PSA curriculum. Specifically, participants consider aspects of epistemology associated with scientific methods of knowledge generation. In this vein, participants reflect on the use of assumptions in developing scientific theory, methods used by various scientific disciplines, the connection between knowledge and reality, and purposes tied to the progress of scientifically generated knowledge. Teachers engage themes that range from avoiding a tendency to seek simplistic answers, on the one hand, to the value of scientific investigation for processes of social and economic development on the other. The spaces for reflection offered during the study of DSA allow them to correlate the capacities gained in studying PSA materials to social processes relevant to the education of children.

The below sections describe the content of courses in the PSA sequence utilized by Kimanya-Ngeyo in their primary-school teacher training as of November 2018. Each course is often accompanied by a section titled "To the Tutor," that FUNDAEC addresses to the tutor who is guiding a group of participants through the course materials. The below summaries often copy text from these "To the tutor" sections as well as some select sections of the course that are more relevant to the teacher training.

These courses are continuously refined by FUNDAEC, so the summaries described below are subject to change. These summaries do not capture the fullness of the content and especially to not capture the value of the mode in which these materials are studied — by mutually exploring their implications with other peers with the aid of a tutor who facilitates the learning process. They also leave out many of the exercises and workbook activities that aim to enhance the participant's understanding of course content.

A.1 *Primary Elements of Description:* Properties

This unit was created to address a shortcoming perceived by FUNDAEC when it started to offer its courses. This was that young people often lacked facility with words and concepts needed to describe the world around them. While they could recognize various shapes and speak vaguely about the size and position of objects, they had not reflected on the meanings the words associated with these ideas convey.

The first few lessons of the unit acquaint participants with a number of concepts and words used in making descriptions at the most basic level: shape, size, and position. The difference between objective and subjective descriptions is then briefly examined, followed by a review of the general properties of matter in its solid, liquid, and gaseous phases. Some of the properties of specific substances, such as melting and boiling point, color, density, and hardness, are examined next. The concept of property is used to enter into a discussion of the qualities of human beings; of these, the qualities of truthfulness, justice, and love are treated in detail. In a similar vein, participants are asked in the final lesson to think about the characteristics of the kind of communities that, as future agents of change, they will be called upon to build.

Many ideas employed in describing the world necessarily come from various branches of knowledge, and so the lessons touch on a number of subject matters. It should be remembered, however, that the unit is primarily concerned with developing capabilities in the area of language. The words and concepts emphasized in the lessons are intended to help the student think about the nature of language. The lesson dedicated to shape, to take one example, should not be treated like an exercise in geometry; its purpose is to

encourage participants to reflect on the concept of shape and its role in expression.

The capabilities used in describing reality are not developed through knowledge of physical existence alone. FUNDAEC conceives of reality as having a social and spiritual dimension as well; the unit therefore is not content to undertake an exploration of physical properties only. By introducing higher entities — the human being and human communities — it is able to extend the discussion into the social and spiritual realms, and students are compelled from the outset of their exploration of language to think about different facets of existence and the means for describing it. It is hoped that, in this way, they will begin to form habits of the mind that avoid the tendency towards fragmentation which invariably gives rise to unnecessary contradictions and imaginary dichotomies in life.

Sections: 1) Shape 2) Size 3) Space 4) Position 5) Objective and Subjective Descriptions 6) Matter and Its General Properties 7) Phases of Matter 8) Specific Properties of Matter 9) Melting Point and Boiling Point 10) Color 11) Density and Hardness 12) Qualities of Human Beings 13) Truthfulness 14) Justice 15) Love 16) Characteristics of a Community.

A.2 *Matter:* The Heating and Cooling of Matter

This book aims at fostering in participants some of the capabilities that scientists draw upon to study and understand the world around them. It enables participants to execute a chief task of science: that of building models of the way the real world behaves. It does so by focusing analysis around a phenomenon that is easy to understand and is present in every day life: basic thermodynamics. This approach takes the participant through basic concepts required to understand simple models of thermodynamics while placing them on the threshold of understanding atomic theory. The final lessons opens the way for a general consideration of models and theories. These sections allow participants to reflect on several important, yet abstract, ideas: how are scientists conceived as members of "scientific communities" who consume and generate bodies of knowledge, not all of which are valid, that accumulate over long periods of time. They further reflect on the range of validity of scientific models and the relationship between beliefs of scientists and the models they create. Participants reflect on the ways in which scientific models evolve when presented with observations outside of the range of validity of previously accepted models.

Sections in the course: 1) What is Science? 2) Observation 3) Temperature 4) Heat 5) Force 6) Pressure 7) Variables That Determine the State of a Gas. 8) Change of Phase 9) Particles 10) Models and Theories

A.3 *Dawn of Civilization:* Transition to Agriculture

The primary purpose of the unit is to contribute to the capabilities of reading with good comprehension and expressing ideas with clarity. The story of humanity's transition from hunting and gathering to agriculture serves as the means through which this purpose is achieved. As such, the unit also seeks to enhance a range of capabilities associated with the process of civilization building. Not least among these is the capability of placing events and developments in historical context. Each of the twelve lessons in this unit (apart from the first) follow the same format. Each begins with a reading intended to help participants think about the historical developments that eventually led to the emergence

of civilization. In their entirety, the readings present a view of human existence that has two dimensions, material and spiritual. Human progress, it is suggested, is driven by knowledge of both these aspects of reality.

To convey concepts that might otherwise be too challenging, the unit introduces an imaginary study group learning about early human history and uses the dialogue between its members as a means of addressing theoretical and abstract subjects. The sections of each lesson that deal with capacities of language are "Comprehension," "Building Vocabulary," "About Language." The final four sections are "Extension" (an idea is amplified and considered in the light of new, additional information), "Reflections" (think about how a concept relates in concrete terms to the world around them), "Writing" (participants invest thought into a relevant question and formulate a cogent, well-written response). The final section is called "Investigation" and encourages them to carry out modest research based on what they have learned and to articulate their findings in writing, helping them to see early in their education that knowledge has highly practical dimensions.

The sections of the unit (in order) are: 1) Introduction 2) Human Evolution 3) Hunter-Gatherers 4) The Process of Domestication 5) What Led to Domestication 6) The Man with a Red Feather on His Head 7) Genetics 8) Material Development 9) Archaeology 10) Spiritual Development 11) Science 12) Religion.

A.4 *Nurturing Young Minds:* To Describe the World

This book focuses on developing capacity in contributing to processes of community life while simultaneously developing mathematical and language capabilities by enlisting participants in processes around children's education. In the first lesson, emphasis is placed on nurturing an attitude in which one does not view oneself as superior to another. This implies that our mode of engaging with others is one of a mode of service — readiness to learn with others, acquiring and utilizing knowledge for the upliftment and progress of the community. This course will focus on service a "promoter of community well-being" can render with children and asks participants to identify a group of 5-6 children of pre-school age that a participant can meet with frequently and take through educational activities. In order to maintain a proper posture of service, the course focuses its first lesson on the non-cognitive qualities, conceived of as "spiritual qualities," required by someone adopting a posture of service. Participants have the opportunity to reflect on the relationship between the following qualities and a mode of service: love, faith, kindness, sensitivity to others' needs, respect, patience, honesty and trustworthiness, humility and diligence.

The educational activities the participant undertakes with children is that of describing sets. They are encouraged to learn the concept of a set, the concepts of belonging to a set, elements of a set, and of the number of elements in a set (more vs. fewer). Along the way, participants think of ways they can introduce this activity with objects from their immediate environments. For example, a participant will collect stones, seeds or leaves. Then, the participant will group stones together with stones, seeds together with seeds and so on. He or she will show these groupings to the children and say "this is a set of stones," "this is a set of seeds," etc. until they have done this with a sufficient number of sets. Then, the participant will help the children do the exact same thing and to verify the presence of the various sets. The next activity invites children to form their own sets. The participant will also introduce games to the children that enable them to engage in physical activity. The teacher can refer to children as a "set of children squatting," which

will prompt the children to squat and so on.

Following these activities, it should be possible to reinforce the concept of a set by helping children understand what it means to belong to a set — i.e., to possess a certain characteristic. The teachers are asked not to use the word characteristic, but to use examples from the children's immediate environment to differentiate sets using their characteristics. E.g., a set of flowers can be divided into a set of all flowers or a set of red flowers (flowers possessing a red characteristic). Next, with the examples in front of the participant, they are encouraged to point to two sets with the same characteristic (e.g. flowers, sticks, etc.) and ask the children "which of these sets has more elements?" Repeat this step and verify the children's understanding. Many other activities are suggested including coloring sheets, games, songs and other activities that will assist the children learn the concepts associated with sets. Some activities encourage the children to think of the characteristics that describe themselves to see if they would put themselves in a set of, say, kind children. The unit emphasizes that one does not need a classroom, only a corner of a living room or place in a yard.

The next educational activity is centered around exploring "God's creation." Teachers first reflect on what it means to possess consciousness. They are asked to imagine a human being singing a beautiful tune and a bird singing a tune. Is the bird conscious of what it is doing? Is the human being? This is not meant to be a subject of discussion with the children. It is intended for the teachers so that they reflect meaningfully on what it means to claim that one "knows" something. Then, moving forward, activities are described that one could conduct with the children so they can explore the world around them using their senses. First their sight, then hearing and touch. The participant is asked to think of activities for the sense of smell and taste. As an example of one of the activities under the sense of touch the teacher is encouraged to give children an object that they will be tasked with describing according to its roughness, smoothness, softness, etc. Following this activity children are asked whether there are things that one can know without one's senses. This will allow children to reflect on the reality of non-material things.

Next, given the list of objects identified by the children, teachers will introduce them to the classification. Again, teachers need not use the term "classification," and should instead focus on helping children group objects into useful categories. A natural first step is to identify whether one of the listed items falls under the animal kingdom, mineral kingdom, plant, etc. The teachers might then encourage the children to think about differences between human beings and animals. Specifically, teachers are encouraged to talk to children about the human soul and the non-material realms of reality it inhabits. This is important because it develops children's abstract thinking ability while building this ability atop of activities that use physical and motor skills to understand the world.

Counting is also introduced in a similar vein. Reflections allow teachers to think about the value of the concrete measurement of reality that counting allows while being aware of uncountable aspects of reality that the rigidity of counting prohibits. Next, teachers are introduced to approaches of introducing addition and subtraction, space, and time and change.

The text is explicit in its exploration of spiritual reality by framing discussions around "spiritual qualities." In the context of a lesson on addition and subtraction, for example, teachers are able to introduce the concept of "more and less." The text suggests that teachers can ask children to connect the concepts of more and less to the attitudes of giving and sharing. If you share your food with your friend, do you have more or less food? Does this make you richer or poorer? Does the act of giving one child love

subtract a mothers amount of love she can give to another child? In discussing concepts associated with space, teachers reflect with children about what it means to have an "orderly" space. They are asked to explore this concept by discussing the relationship between children's belongings and their immediate environment. Do your belongings have a particular place in your room? What about the way you appear to others? Care is taken that teachers explore the benefits of orderliness that emerges from inner discipline and not from externally imposed discipline. In discussing time and change, teachers explore whether all indicators of change can be measured. Is it desirable to reduce the intellectual and spiritual growth of a person to a group of indicators that can be cleverly measured? The text states that clearly measurement might help, but it is important not to take this to an extreme. How might a teacher assist students to develop their capacity to evaluate their own process of learning?

A.5 *Promoting a Healthy Environment:* Environmental Issues

The participants of PSA courses see themselves as "promoters of community well-being." As such, they need to develop an understanding of the kinds of challenges faced by communities. These can be categorized into three types: those a community can meet using its own resources, those that a community requires help from the outside to solve and those challenges that are well beyond the possibilities of an average village to solve. Despite the scope of challenges in the third category, promoters of community well-being acknowledge that the problem cannot be ignored at the level of a microregion, the geographic unit participants are most concerned about.

This course develops a fundamental capacity required by promoters of community well-being is one in which they discuss such challenges with local populations and create awareness of the issues involved. After all, can people do anything to address their problems if they are not conscious of them and remain unaware of possible solutions? This course focuses on the question of the physical environment and the challenge of raising awareness regarding the environmental issues in a microregion. It begins by stating a simple truth: that every human activity affects the physical environment. Participants identify activities in their microregion and begin to develop an awareness of how these activities influence the physical environment. Two sections are dedicated to an analysis of pollutants that emerge out of systems of transportation. Participants research changes in transportation systems in their microregions in the last twenty years and engage in conversations with other members of the community with the purpose of raising awareness about the relationship between transportation systems and pollution. They are asked specifically to take care "to ensure that the [conversations] you create take the form of a conversation, not a sermon."

Participants also gain an understanding of possible solutions to the problem of air pollutants generated by transportation systems. These fall into three broad categories: 1) technological remedies, 2) those that relate to individual behavior, 3) actions and decisions of our governments. Promoters of community well-being can likely contribute to solutions in categories 2 and 3 and participants reflect on ways in which they would assist their peers to reflect on these types of solutions.

After exploring this theme, the course switches its attention to small-scale manufacturing and large-scale manufacturing and again encourages participants to reflect on how these activities negatively effect the physical environment and the solutions to these negative effects. These sections model a conversation among a few local manufacturers to

aid participants to think of their own roles as promoters of community well-being. In this way, participants are exposed to more complex forms of waste management and think about how wasteful substances are discharged into the air and water while others are thrown away as refuse.

A final human activity that participants analyze is that of daily individual and family activities: namely the waste generated from the consumption behavior we engage in. Participants are asked to keep a log of the types of waste they produce in their daily activities in order to recognize the relationship between this activity and the production of waste. There are three questions that are the subject of study and conversation among the participants: Can we reduce the amount of waste material generated? Can we find a beneficial use for the waste material? Can we find a safe method of disposal? These questions are discussed while also encouraging participants to be moderate and to consider the desirable goal of prosperity in their communities. Similarly to the challenges associated with transportation and small-and-large-scale manufacturing, participants consider solutions associated with reducing consumption patterns, finding beneficial uses for waste materials, and how to safely dispose of parts that cannot be re-used. In the final section, participants learn the distinction between biodegradeable and nonbiodegradeable waste. This forms the basis to understanding the production of compost in the unit titled "Planting Crops."

Sequence of sections in this unit: 1) Awareness 2) Human Activity 3) Transportation 4) Factors that Affect Air Polution 5) Solutions to Air Pollution 6) Small-Scale Manufacturing 7) Waste Management in Small-Scale Manufacturing 8) Large-Scale Manufacturing I 9) Large-Scale Manufacturing II 10) Consumption and Solid Waste 11) Solutions to Solid Waste

A.6 *Food Production on Small Farms:* Planting Crops

Technology extends the fruits of science to humanity. It is a fundamental theme throughout all courses, but is treated extensively in this unit. The capabilities explored in this unit broadly fall in the category of "science" and "processes of community life." Given the central role of agriculture, its processes and activities represent an excellent means for helping people advance in the capabilities they will need to participate effectively in the development, selection, and use of technology. Participants will do this by exploring how farming a particular rotation of crops on a small plot of land can generate high yields year after year.

The first lesson of "Planting Crops" lays a foundation for the unit as a whole. Here the concept of primary production is introduced, and through a series of short exercises participants are asked to reflect on the importance of agriculture and animal husbandry for human societies. This paves the way for a brief discussion on the historical evolution of agriculture production and on the two sources of knowledge available about this essential area of activity: modern science and technology, on the one hand, and the experience of the farmers of the world, on the other. The concept of appropriate technology follows naturally from this idea, and participants are helped to think about the implications of technological choice for the way of life of a region. It is not intended for them to reach any conclusions in this respect. They should simply gain a sense of the scope of the issues involved.

The sections in this book follow a conversation among four individuals who represent different ways of thinking about social change. A young man who wishes to help his

grandmother improve production on her small farm is seeking advice from others on how to proceed. His grandmother provides a perspective informed by traditional knowledge. A friend studying at a nearby university brings the view of specialized science of agriculture. His cousin shares ideas as a participant in the FUNDAEC courses. They seek a balanced approach to change — characterized by a healthy relationship between cultural heritage of a local population and modern science.

The sequence of ideas in the course are as follows: 1) Primary Production 2) Factors That Determine Crop Production 3) Social, Economic, and Cultural Factors 4) Introduction to Soils 5) Physical Properties of Soil 6) The Chemical Composition of Soil 7) Fertilizers 8) Soil Biology 9) Preparing the Land 10) City Gardens 11) Seeds. Unique to this unit are exercises dubbed "Practice and Experimentation," which provides participants with the opportunity to apply what they are learning and develop their skills and abilities, and "Investigation", which encourages them to broaden the scope of their practical knowledge by, for example, interviewing farmers.

The central concern of this unit is on technology. The lessons seek to raise discussion above the formula of "either-or" or even the harmonious co-existence of both. The unit attempts to show students that in every region, including their own, a learning process can be set in motion by which new knowledge is generated from the interaction between the traditional knowledge system and modern science and technology, which can then be applied to the problems of everyday life and used to promote the sound progress of the region. The hope is that students will come away from the unit with the understanding that such a learning process is one in which they, along with many others, can participate.

A.7 Example of Facilitation Style in Teacher Training

In "Properties," participants gain insights into the importance of precision of language in scientific thinking. After studying different parameters used to describe reality (e.g., temperature, color, size, etc.), teachers are asked to reflect on the concept of shape, one such parameter. In an exercise, the tutor guides a conversation aimed at responding to the veracity of the following statement: "Shape is a specific (not general) property of substances." Here is a snippet of the ensuing conversation recorded by one of the PIs:

Participant A It's false.

Tutor Can you go ahead and explain a little more why you think it's false?

Participant A As we were listing the specific properties of matter, we were listing shape and size. The general properties are common. Shape is one that is common to many substances. It seems every type of matter has shape. However, those specific properties that allow us to distinguish one substance from another, shape is not a type of property.

Tutor Hmm. Ok. How about others, what do you say?

Participant B For me, I say it's specific. If you want to distinguish a table and something else then you use its shape.

Participant C Let's say there's a table that has triangular shape. Does it mean that tables only have triangular shapes? What about oranges and apples? They have the same shape. But they're different substances.

Tutor It's quite challenging, isn't it? If you think about it, it's challenging to state the reason why you can't use shape or size to distinguish one thing or another. However, the other participant is saying I can use shape to distinguish one object from another. I can say that an orange is spherical and it helps me distinguish it from something that is not spherical.

Participant C That's true, but we're examining general vs. specific properties. Her statement is right, but we're trying to talk about those properties that are specific. When you talk about specific properties, they need to be such that they make the substance different from others.

Tutor [*sensing that participant C has grasped the concept, but others have not*] Is shape a general property of matter?

Participant D How do you distinguish a bus from the lorry? Don't you use shape?

Participant E We're not saying we can't describe objects using general properties. But there are those properties that further describe matter but they create a sense of difference.

At this point, the tutor points the participants back to the use of the term specific and general properties in the text. It seems many in the group have reached the understanding of participants E and C, but a few are still holding out.

Tutor Shape helps us describe substances, but is it a specific property of substances? We're not here for debating with anyone. We're here to explore concepts for the purposes of understanding. If you have reservations about what you are discussing and it appears that the majority is on the other side which you are not yet it's best to put a star so that you come back and try to understand. You shouldn't just accept that the majority is right. You have to understand on your own. Let's move on.

B Ugandan Educational Context

B.1 History of Schooling and Education in Uganda

One way of characterizing the history of schooling in Uganda is through the cyclical manner in which education access and education quality have been prioritized across different points in time. Consistent with global trends, the call of the moment is to enhance education quality through the primary and secondary schooling pipeline. The manner in which this increase in quality is to take shape is subject to intense public discussion in Uganda.

To situate the context of the current moment in a broader perspective, it is helpful to understand the landscape of schooling in Uganda across time — we provide a more detailed summary in appendix section B.1. Missionary groups in the late 19th century spurred a nation-wide literacy movement in Uganda spanning many tribes (Hanson, 2010). During this period, Ugandans used informal networks to self-organize into community schools, allowing children, youth, adults and elders to study literacy side-by-side, often before, during and after completing farming-related chores for the day. According to British accounts, by the early 1930s, Ugandans had organized 5,673 community schools in this manner, promoting the participation of scores of villages along the way.

Seeing a strength in Ugandans' initiative to implement formal education, multiple parties sought to move beyond basic literacy by increasing education quality: namely, British colonialists sought to bring a standard of education that prioritized observational science that was increasingly desired by the younger generations of Ugandans (Feierman, 1985, Ssekamwa, 1997). Education began prioritizing "practical" forms of knowledge such as agricultural and other technical disciplines (Bude, 1983). To implement a policy of curricular change, the Governor of the Colony began taxing the teacher "profession" and used the revenues to transition 218 of the "best" community schools to this new form. This tradeoff intended to enhance education quality at the expense of education access.[47]

In the lead-up to independence, it was clear that Ugandans again began to emphasize increased schooling access to the broader population. This required increasing the number of well-trained public servants and increasing government revenue to support trainings and schools (Ssekamwa, 1997). Numerous conferences were held across Africa during the same time to identify pathways forward that would combine indigenous schooling approaches with the best of imported versions of schooling (Jansen, 1989). Ultimately, however, the rapid pace of modernization led to the Ugandan government choosing to expand the existing (colonial) educational systems rather than experimenting with new forms of schooling (Brock-Utne, 2002).

The 1960s through the late 1980s was characterized by political instability in Uganda, which sidetracked many of the planned educational initiatives. Towards the beginning of the 1990s, Uganda began to re-prioritize educational access. They were among the

[47] This change in the curriculum also had an impact on girl participation in schooling. For example, in conversations recorded by the De La Warr commission 1937, Archdeacon Bowers, expressing the Church Mission Society's position stated "We feel very strongly that the education of girls should have a strong bias towards the building of homes. The time has not come to give them an academic education," to which Martin Kayamba, a highly ranked Tanzanian in the colonial service, responded "In our tribal life women are very important; we have women chiefs ruling people; women doctors; women agriculturalists who support their families, and their influence is very strong not only in the homes but outside. I think myself the education of women is just as important if not more than that of men." The rise in gender inequality in education during this period in Africa is also documented rigorously using econometric and statistical methods in (Baten et al., 2020).

first countries to implement a policy of universal primary education (UPE) in 1997, rapidly increasing the number of learners attending primary school — a 300% increase in enrollment was observed between 1996 and 2004. However, the familiar tension between access and quality re-emerged which was perhaps hindered by the separation of efforts to increase access and quality of education in Uganda (Moulton, 2002).

Today, close to 100% of the population enrolls in primary school. The Ministry of Education and Sports (MoES) is now intensifying its focus on increasing education quality, which is also an important factor in increasing access to secondary education.[48] The main avenue through which this focus is exercised is by considering the training and incentive framework for teachers. This is reflected in the MoES's publication on the Teacher Initiative in sub-Saharan Africa (TISSA) (MoES, 2014) and the teacher incentive framework (MoES, 2017), both of which emphasize an important role for recruiting teachers with higher credentials and also systematic incorporation of teacher trainings that can make teachers more effective, better compensated, connected to community processes, and driven by a greater sense of purpose. Importantly, these initiatives are taking place alongside a curricular reform of lower secondary school which emphasizes "competency-based" education and teacher policies that are calling for programs that promote continuous professional development. Thus, teacher trainings similar to the one offered by KN are very timely.

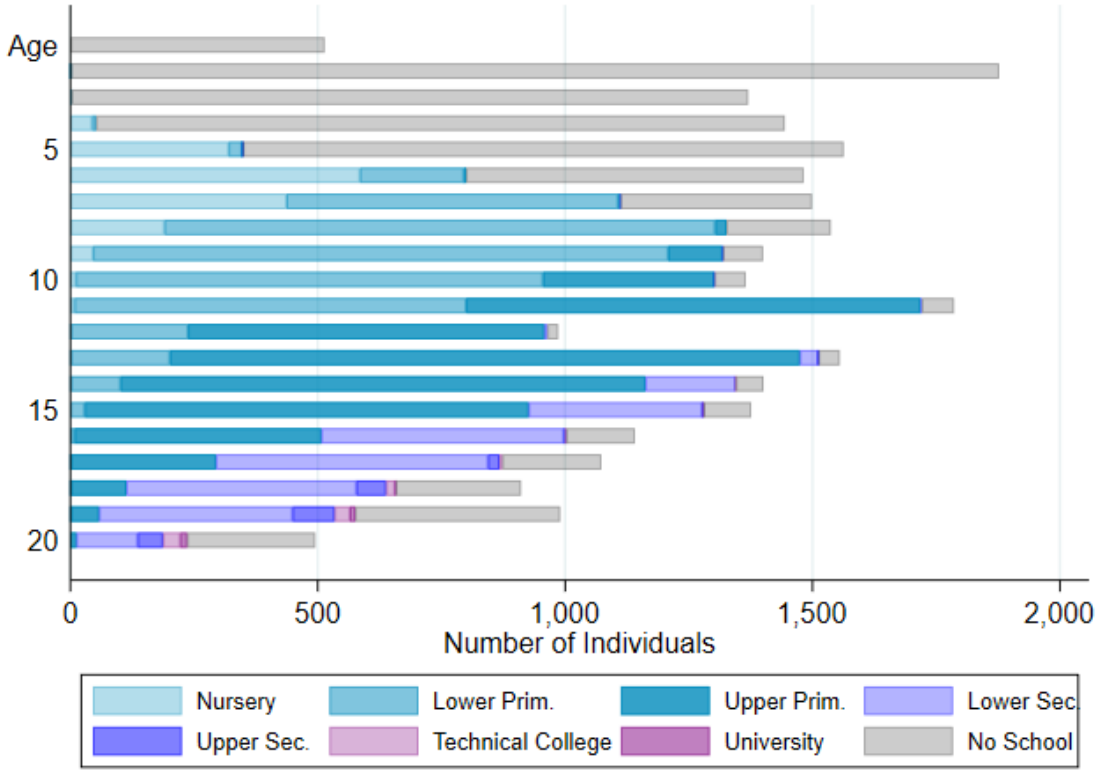
B.2 Study Context — Jinja District, Uganda

We zoom in on Budondo sub-county in Jinja district to provide a more detailed picture of the schooling landscape. In March-April 2018, we administered a community and school census across 38 villages in Budondo, aimed at providing a rich perspective on the patterns of schooling and economic activities of the households in this region. According to this census, there are 8,982 households, split across 38 villages. The mean number of members per household is 5.1 (median 5) and the mean number of children aged 16 or under, per household is 2.7 (median 2). The head of household is male in 75% of cases and has an average age of 44.1 years (median 42). Additionally, 10% of households have only one member, and both the head and spouse of the household residing in the sub-county in 68% of household cases. Each household has lived in the same village for 19.0 years on average (median 14). A vast majority of the households have homes with an iron roof (99.5%). The region is ethnically composed of Basoga, with a share of 73.5% of households claiming Basoga heritage with clusters of Banyole (6.3%), Bagisu (5.1%), Baganda (4.3%), Iteso (2.3%), and others. Most of the diversity is concentrated in the trading center of Buyala, which contains three villages 3,230 households.

We turn our attention to the nature of schooling in Budondo. First, it is worth noting that district education offices (DEO) are largely autonomous across Uganda in administering primary education. They are tasked with delivering education using the National Curriculum, and the district-level administrators have final say over the staffing of schools in their district.[49]

Of the 44,445 individuals for whom we have age-data in Budondo sub-county, 75% are below the age of 30 and 84% of children and youth between the ages of 5 and 20 attended some form of school in 2017 — be it nursery, primary, secondary or a form of higher education. Figure B.1a shows the distribution of schooling by age and category

[48] It should be noted that Uganda also implemented a universal secondary education in the early 2000s.
[49] The district administrators themselves are accountable to the Primary Education Office in MoES.



(a)

Figure B.1: Schooling in Budondo

within Budondo. A majority of individuals attend primary school by the age of 7 and the percent of individuals attending "no school" is decreasing until roughly the age of 13 when it begins to increase again. Of the individuals between the ages of 13 and 20 who are not in primary school, 33% no longer attend any form of school.

Figure B.2 shows the distribution of ages enrolled in different levels of schooling in primary school.[50] We can clearly see that there is a very large age variance in level schooling and this variation widens as students advance in school levels. For example, in Primary 1, there is significant density throughout ages 6 and 10, suggesting that either students start Primary 1 quite late, or those who start around age 5 or 6 have a tendency to repeat classes — most likely, both factors play a role. The distributions of ages in Primary 4 (the transition year between primary and middle school) and 7 Primary flattens relative to Primary 1. In Primary 7, there is high high density throughout ages 13-18. Below, we use features of the household census to show that a factor driving this process is the very large repeat rate — households indicate that students in P1 through P6 repeat the previous year's class between 14 and 18 percent of the time, a figure consistent with UWEZO (2016) and our own student survey data we describe in further detail later in the paper.[51] There is no meaningful difference in the dropout, repeat, and transition

[50]Similar distributions for secondary school are available upon request.

[51]The repeat rate in P7 is slightly lower, however, it is worth noting that Gilligan et al. (2018) argue that "Educational Triage" plays a role in the repeat rate at all levels of the primary schooling pipeline. They argue that educators encourage their weaker students to repeat or drop out classes so that the "performance" of the school in high-stakes primary leaving examinations improves. The indicator of

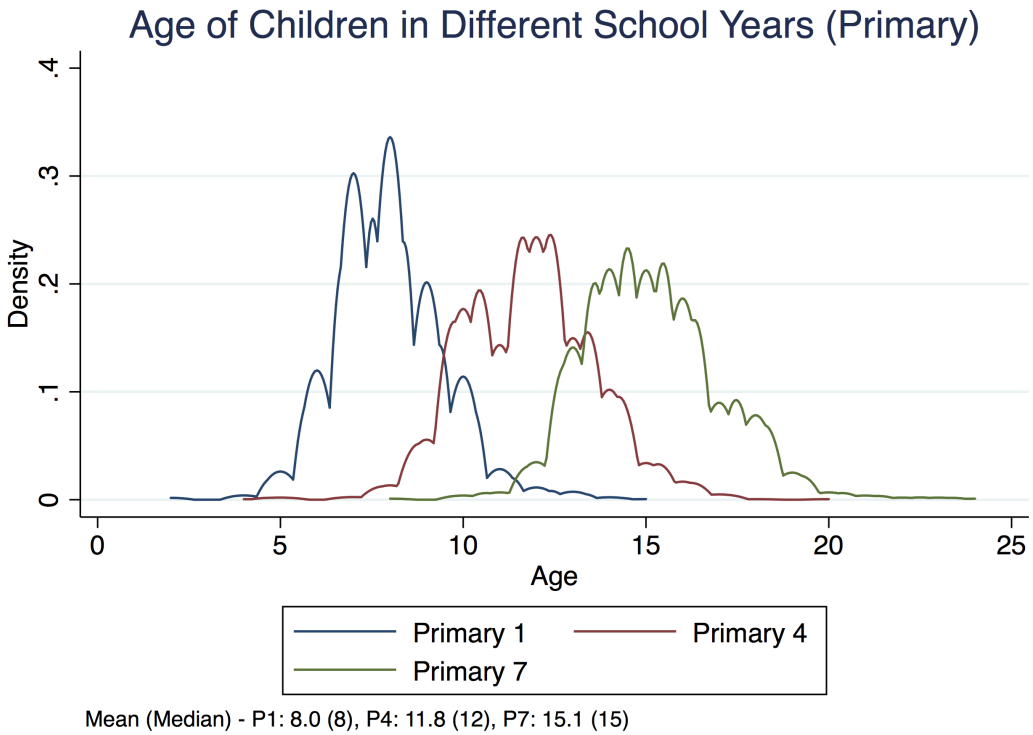


Figure B.2: Distribution of Ages Enrolled in Different Primary School Levels.

rates across girls and boys in primary school.

For each individual not in school in 2018, we also request information on the highest level of schooling that individual reached. Thus, we can analyze the distribution of the highest year of schooling achieved across four different cohorts (age 30-39; 40-49; 50-59; 60+) in Figure B.3 to understand how progression through school has changed across generations. The distribution is very bi-modal for men, showing spikes at S4 and S6 as we expect from the dropout rates discussed above. The distribution is becoming more bi-modal over time for younger cohorts of women (in older cohorts, few women attended secondary school). In summary, that data suggest an important role for enhancing education quality in the early years that could address both the ends of participation and access to schooling (especially at the higher levels of schooling) as well as quality of learning.

B.2.1 Dropout, Repetition, and Transition Rates

Table B.1 computes the dropout rate, repetition rate, and transition rate at each year of school. To do this, we take advantage of the fact that the household roster in the community census requests each child's school history in the previous year as well as the attendance at the beginning of 2018. We can define "dropout" as someone who was in school in 2017 but said they were not attending school in 2018. The "dropout rate" for a given year is then the numbers of dropouts from that year divided by the total number of students in that year. The repetition rate is defined as the number of people who report expecting to be in the same school level in 2018 as they were in 2017, divided by the total number of people in that school level in 2017. Finally, the transition rate is defined as

achievement in PLE examinations is valued by district education offices and It is believed that these measures of performance are key drivers of school choice.

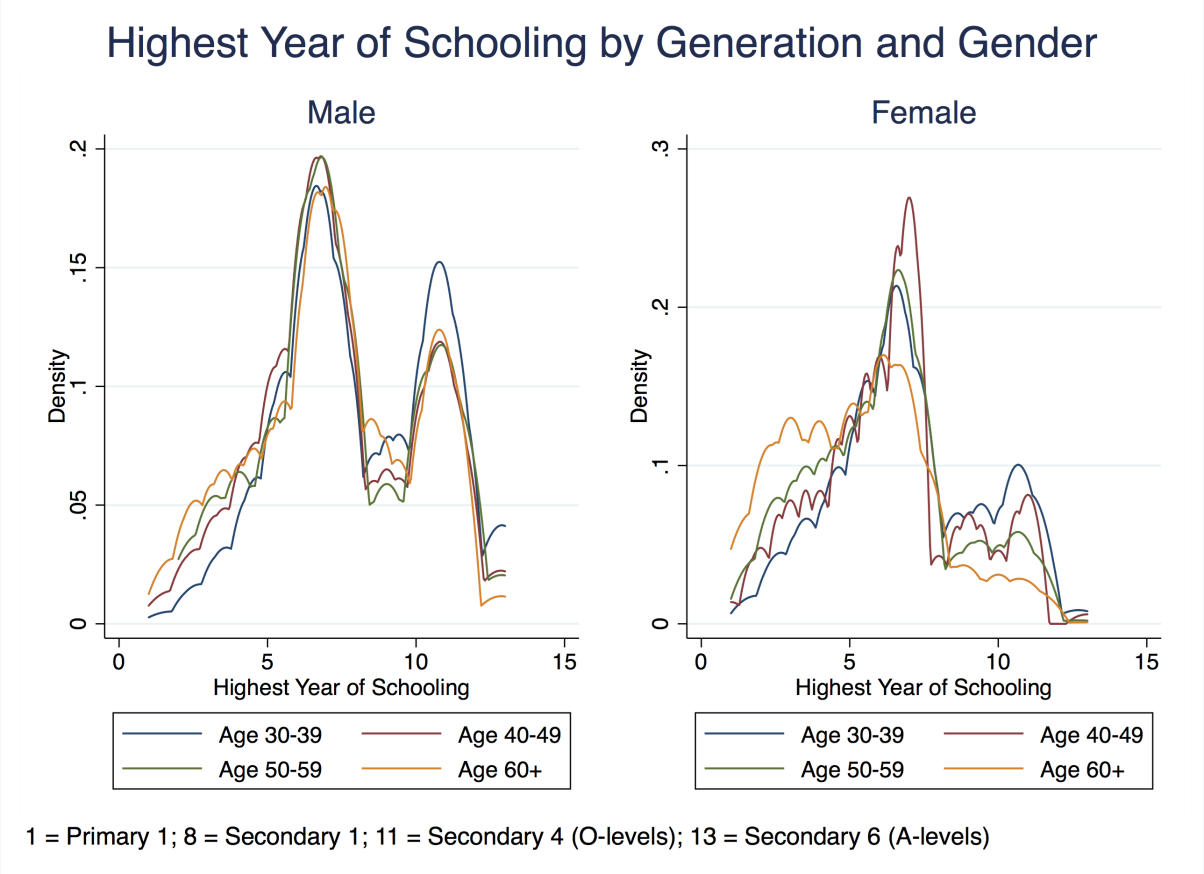


Figure B.3: Distribution of highest year of schooling achieved by generation/cohort. Generations are split into age groups of 10 years. Nearly everyone is finished with school by age 30.

the number of people who expect to be in a higher year than they were last year divided by the total number of people in that year last year.[52]

Dropout rates are relatively low — ranging between 3 and 8 percent in primary school. Dropout rates spike at S4 and S6 levels of schooling, reflective of the O-level and A-level requirements of these school levels. They are also surprisingly balanced by gender. Primary 7 has a higher dropout rate than most, but not overwhelmingly so.

Table B.2 gives a sense of the cumulative impact of these dropout rates: each row produces the estimated share of an initial Primary 1 cohort that reach the row-specified level of education. These numbers are sensitive to assumptions about how repeaters behave: if we assume they never transition then only 20% reach to S1; if we assume they always transition then 73% reach S1 (eventually). This table presents a somewhat middle-road but we hope to eventually improve these estimates by using multi-year data where we can actually track a cohort across time. With that said, the table shows that just over half of the original cohort reach secondary school (the passthrough rate is actually higher for women than men).

Finally, given our interest in parent-teacher interactions, it is also worth noting that we asked household heads to answer the following question regarding each of their children who attend school: "Have you interacted with [child's] teacher?" Just over half (53%) of the (non-missing) responses "Yes." These "Yes" respondents are equally split between

[52]It is worth nothing that the number of people who expect to be in school this year but were not in school last year is very small, thus we ignore them in this analysis.

Table B.1: School Continuation Table

	Dropout		Repeat		Transition		Number	
	Male	Female	Male	Female	Male	Female	Male	Female
P1	.03	.03	.19	.16	.77	.80	1,071	965
P2	.02	.02	.14	.08	.84	.89	1,017	888
P3	.03	.03	.18	.15	.78	.81	966	911
P4	.03	.03	.18	.17	.78	.80	986	936
P5	.05	.03	.18	.19	.75	.78	849	865
P6	.08	.05	.18	.15	.73	.79	638	705
P7	.09	.09	.08	.11	.82	.78	480	565
S1	.04	.03	.02	.03	.93	.94	354	384
S2	.03	.04	.03	.03	.92	.92	289	350
S3	.04	.05	.05	.03	.90	.92	319	317
S4	.32	.37	.07	.09	.48	.42	283	314
S5	.04	.06	.00	.02	.96	.91	68	47
S6	.38	.38	.07	.03	.37	.42	86	64

Notes: The numbers reflect estimated probabilities of dropping out, repeating, or transitioning for each school level according to survey responses from the household census. Details for how the calculations are made are described in the text. We distinguish each category of variables by male and female students. N refers to the number of individuals in that school-level in the previous year. The rows do not always sum to 1, especially in Secondary 6, because individuals sometimes report that they had yet to decide their schooling status in 2018 and also because of a small number of people who state that they transition to a *lower* class.

male and female students.

B.3 Additional Socio-Economic Context in Budondo

Table B.3 shows the distribution of occupation for each member of the household, as specified by the household head, at a series of different age categories. Budondo is commonly referred to as the "bread-basket" of Jinja. This is evident from the distribution of individuals who state that their profession is "Farmer." However, these categories are not mutually exclusive — an individual can select more than one category. The large share of people reporting "Other" in the 20-29 category are largely women.

Figure B.4 shows the main composition of ethnicities for each of the 38 villages. Basoga clearly dominates in general, though not in every single village. Figure **??** shows a similar picture, but for religions rather than for ethnicities. The distribution varies a little bit across villages, but all villages have some representation of the major religions. Just over one-third of people are Muslim, and most of the rest are Christians split across a few denominations. Figure B.6 shows the same picture for a household's main source of power. There is a fair split across most villages, and it is clear that several villages are not electrified whatsoever.

Table B.2: Cumulative School Continuation Table

	Leftover Share (M)	Leftover Share (F)
Primary 1	1	1
Primary 2	.95	.96
Primary 3	.93	.93
Primary 4	.87	.89
Primary 5	.81	.84
Primary 6	.73	.78
Primary 7	.62	.71
Secondary 1	.55	.62
Secondary 2	.52	.59
Secondary 3	.5	.56
Secondary 4	.47	.53
Secondary 5	.27	.27
Secondary 6	.26	.25
Post Sec. 6	.12	.13

Notes: Rates are sensitive to assumptions about repeaters as described in the text.

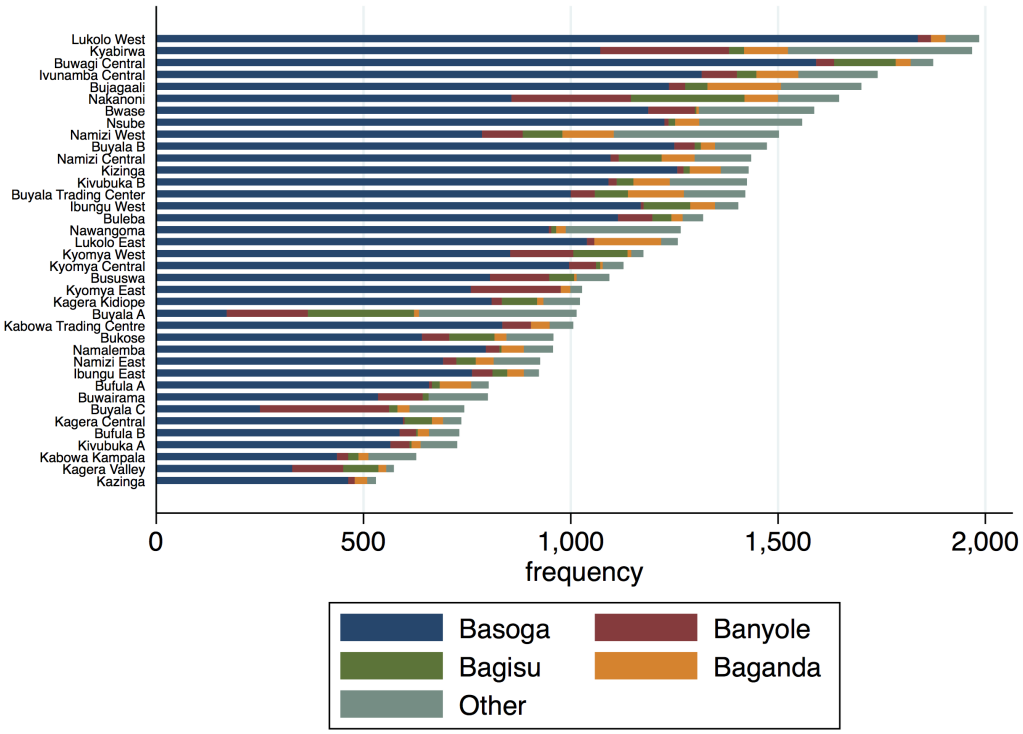


Figure B.4: Distribution of ethnicity by village. The ethnicities are grouped into the four biggest ones and 'Other' and the villages are sorted by the size of the village. Note that this is done at the individual, rather than household, level.

Table B.3: Occupations by Age

Age Category:	15-19	20-29	30-39	40-49	50-59	60+
Farmer	.12	.39	.58	.66	.7	.61
Teacher	0	.02	.03	.03	.02	.01
Taxi/Boda-boda driver	.01	.05	.07	.06	.03	.01
Truck driver	0	.01	.01	.02	.01	0
Artisan	0	0	0	0	0	0
Construction worker/Mason	.01	.05	.05	.03	.02	.01
Factory worker	0	.01	.01	.01	.01	0
Carpenter/Carver	0	.01	.01	.01	.01	0
Herbalist	0	0	0	.01	0	0
Hairdresser	.01	.03	.01	0	0	0
Health worker	0	.01	.01	.01	.01	.01
Office worker	0	.01	.01	.01	.01	0
Shop attendant/Trader	.01	.08	.11	.11	.09	.04
Baker/Cook	0	.02	.03	.02	.01	0
Priest	0	0	0	0	0	0
Elder	0	0	0	0	.01	.06
Student	.66	.11	0	0	0	0
Works in household	.02	.02	.02	.01	.01	0
Broker/Middleman/Property Master	0	0	0	0	0	0
Mechanic	.01	.03	.02	.02	.03	.01
Fishing	0	.02	.02	.01	0	0
Tailoring	0	.02	.02	.02	.02	.01
Sugar Cane Loading	.01	.02	0	0	0	0
None	.14	.17	.08	.07	.08	.23
Other(specify)	.01	.03	.03	.03	.03	.03
N	5,646	6,744	3,914	3,177	2,085	2,309

Notes: Rows need not sum to 1 as multiple occupation selections are possible.

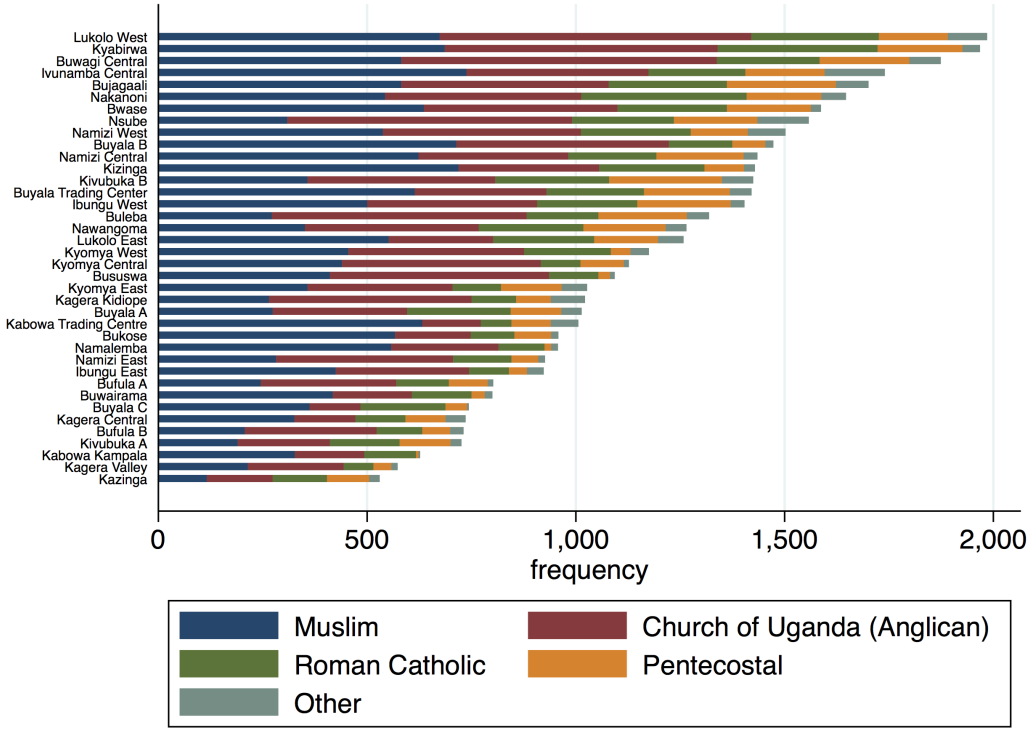


Figure B.5: Distribution of religion by village. The religions are grouped into the four biggest ones and 'Other' and the villages are sorted by the size of the village. Note that this is done at the individual, rather than household, level.

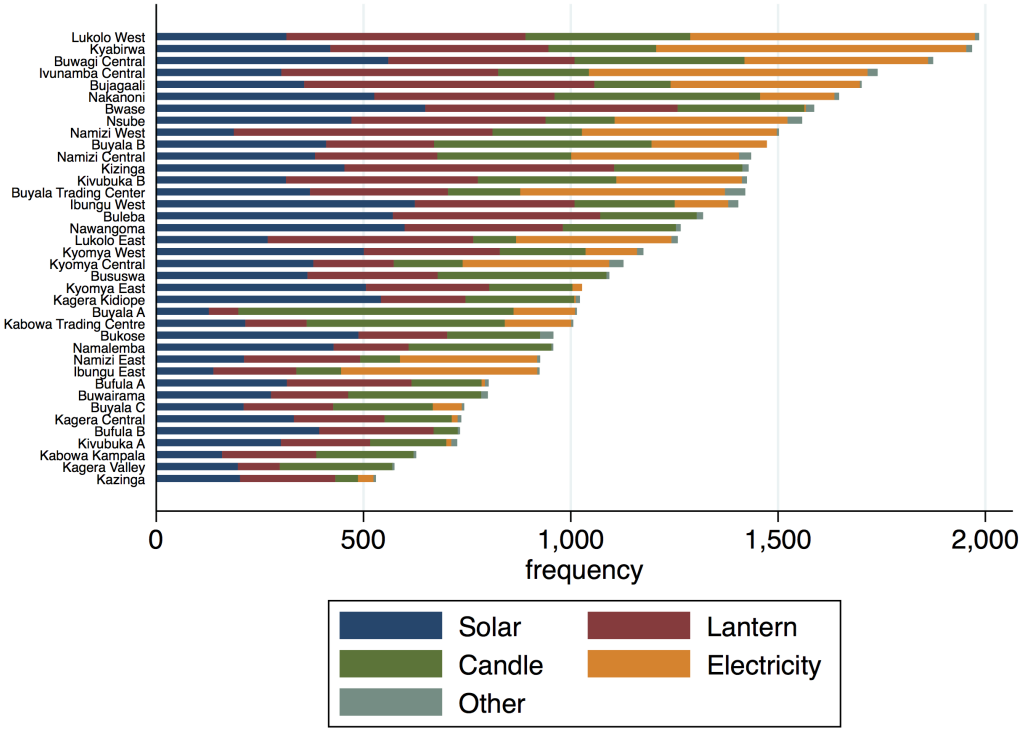


Figure B.6: Distribution of main source of power by village. The power sources are grouped into the four biggest ones and 'Other' and the villages are sorted by the size of the village. Note that this is done at the individual, rather than household, level.

C Measurement of Outcome Variables

We will organize the description of measurements according to the research questions articulated in section 2.2. Each of these measures will be constructed using data from the above-mentioned sources. All variables described below are also described in similar fashion in our pre-analysis plan. We refrain from providing summary statistics of each variable in a separate table since we present the number of observations and control school means for each variable in our results table.

C.1 (Q1) Does the teacher training change teacher pedagogy?

(H1.1) Classroom Pedagogy Improves.

Using the Stallings classroom observation tool and student survey responses, we construct measures of student engagement and participatory pedagogy in school.[53] We construct the variables under each outcome measure for hypothesis **(H1.1)** in the following manner.

- **Student Engagement *(Stallings)*:** After identifying the activity the teacher is engaging students in, the enumerator needs to specify a response to the question "How many pupils are engaged in the activity with the teachers?" The enumerator can respond with 1) "No pupil," 2) "One pupil," 3) "A few pupils," 4) "Half of the pupils," 5) "Most of the pupils," or 6) "All pupils."[54]

- **Students ask questions in class *(Teacher Outcome from Student Survey)*:** We ask students to indicate whether they asked teachers any questions during class (in terms 2 and 3) to help them with a problem they didn't understand. Respondents selecting "yes" continue to indicate for which teachers this statement is true. We sum student responses by teacher and divide by the number of students surveyed. Formally, and for each school k, let y_{ijk} be equal to one if teacher i was mentioned by student j and zero otherwise. The variable we construct is equal to $y_{ik} = \sum_j y_{ijk}/N_k$ where N_k reflects the number of students surveyed in school k. We restrict analysis of this variable to those teachers who teach P6 students as indicated by at least 90% of surveyed students mentioning a teacher in this way.

- **Corporal Punishment *(Teacher Outcome from Student Survey)*:** We gently ask students to indicate whether a teacher has "caned" or beaten him or her. We stress to the students that they do not have to respond to the question if they do not wish to, but that their responses will remain confidential no matter how they respond.[55] If the students say "yes," we ask them to indicate which teachers caned or beat him or her. We construct teacher-level outcomes in the same manner as the "Students ask questions in class" variable above.

(H1.2) Teacher Motivation and Effort:

[53] We will include TIPPS measures related to critical thinking pedagogy and pedagogy facilitating deeper learning and understanding of concepts when the data become available, likely in mid-2021.

[54] We follow the instructions in World Bank Group (2015) and emphasize that even if only one student is involved in an activity, students are coded as "engaged" if they are paying attention to the activity. For example, if a single student is asked by the teacher to read a passage out loud, all students are considered "engage" if they are at least listening attentively to the passage being read.

[55] "Caning" is a common term used for using something like a thin bamboo stem to hit a child on his or her head as a form of corporal discipline.

- **Teacher Attendance *(School Admin Data)***: Schools are required to collect daily teacher sign-in sheets, however not all schools do so. For the schools that collect this information, we sum and then digitize the days of teacher attendance per term at the teacher level.

- **Teacher Knowledge of Student *(Teacher Outcome from Student Survey)***: In the student survey we ask students in P6 to indicate their attendance at the school during the prior two school days. We additionally ask these students to indicate their relationship with the people they live with. Then, in the teacher survey, we randomly selected six of the students we surveyed and added two additional names of students who do not attend the teacher's school (hereafter dubbed "fake students"). We first ask the teacher to identify whether the named student attends the school (question 1). If the teacher responds "yes" to this question, we proceed to ask, questions 2, 3 — whether the named student was present in school during the previous 1-2 school days (one question for each school day) — and question 4 — what is the relationship that the student has with the people he or she stays with. We match teacher and student responses and create an index reflecting the accuracy of the teacher's response

  Formally, and for each school, let q_{ijk} reflect teacher i's response regarding student j for question $k = \{1, 2, 3, 4\}$ — student j's response for question k is reflected by s_{jk}. The index we construct amounts to the percent of teacher i's correct responses across all students and questions:

$$\text{Teacher Knowledge of Student}_i = \frac{\sum_j \sum_k \mathbb{1}(q_{ijk} = s_{jk})}{\sum_j \sum_k 1}, \tag{3}$$

  where $\mathbb{1}(q_{ijk} = s_{jk})$ is equal to one when the teacher's response matches the student's response and zero otherwise.

(H1.3) Teacher Collaboration:

- **Teachers learn from and collaborate with colleagues *(*Teacher Survey):** Recall that to compile our sample for teacher interviews, we acquire a complete list of teachers in each school in the second term of a given school year and interview upper primary teachers in the third term of that school year. We pre-load teacher names in each school into our teacher survey and ask each interviewed teacher to respond to the following three questions regarding classroom and school learning/collaboration with other teachers:

  1. Have you and [*colleague*] spoken about how to better manage your classrooms?
  2. Have you or [*colleague*] visited any of each others classroom to help improve teaching practices?
  3. Have you planned classroom activities together with [*colleague*]?

  Thus, for a given teacher i and colleague j we observe i's description of the symmetric relationship she has with j — $y_{ij}^{net,k}$ for questions $k \in \{1, 2, 3\}$. Importantly, if school s has a sample of N_s teachers, then we observe both i's responses ($y_{ijs}^{net,k}$) and j's response ($y_{jis}^{net,k}$) for each ij or ji pair in $\{1, 2, \cdots N_s\}$. To add precision to

our measures, we only consider affirmative responses in which ijs and jis agree.[56] In other words,

$$y_{ijs}^{net*,k} = y_{ijs}^{net*,k} = (y_{jis}^{net,k} = \text{Y} \ \cap \ y_{ijs}^{net,k} = \text{Y}) \quad \text{For all } k \text{ in } \{1,2,3\} \text{ and } i,\ j \text{ in } N_s \tag{4}$$

where $(y_{jis}^{net,k} = \text{Y} \ \cap \ y_{ijs}^{net,k} = \text{Y})$ indicates that both i and j answered "Yes" to social network question number k.

The final variable we construct for the pre-analysis plan is the dyad-specific average response for each question $y_{ijs}^{net,k}$. In other words, we analyze

$$y_{ijs}^{net} = \frac{1}{3} \sum_{k \in \{1,2,3\}} y_{ijs}^{net*,k} \tag{5}$$

C.2 (Q2) Does the teacher training change student outcomes?

(H2.1) Traditional learning outcomes increase *(PLE scores, Student Assessment and Student Survey)*:

- **Primary Leaving Examination (PLE) at the end of P7 *(UNEB and School Admin Data)*:** As described above, we collect school records on PLE registration and combine it with official data from UNEB. Lower scores indicate higher performance and range from 1 to 9 for each subject. To ease interpretation, we flip this relationship and make 9 the highest performance and 1 the lowest performance. Furthermore, we aggregate scores from each of the four subjects in a manner that reflects the official approach taken by UNEB to determine student performance. Therefore, our aggregate measure ranges from 4 (lowest performance) to 36 (highest performance) which we standardize according to the control school mean and standard deviation.

- **Student Pass-through Rates *(Student Survey)*:** In the student survey, we ask students to indicate the grade they are currently in, the grade they were in one year prior and the grade they were in two years prior. We indicate that a student has had a successful pass-through rate if the most recent year's grade is one (or more) grade ahead of the previous year's grade. Formally, and for example, a successful pass-through for student i in school s in time t (t reflects the school year) is characterized by

$$y_{ist}^{pass} = \mathbb{1}(\text{Grade}_t - 1 >= \text{Grade}_{t-1}) \tag{6}$$

where $\mathbb{1}$ reflects an indicator variable equal to one if $(\text{Grade}_t - 1 >= \text{Grade}_{t-1})$ is true and zero otherwise.[57]

[56] A subset of sampled teachers are not included in our pre-loaded survey program because they were not listed as school teachers by school administrators prior to term 3. Therefore, for these teachers we only have observations for either $y_{ijs}^{net,k}$ or $y_{jis}^{net,k}$ but not both. We exclude these teachers (whether present in i or j) from our analysis which preserves the precision of our measure and imposes symmetry on dyadic responses.

[57] We furthermore ask whether students attended the same school in the prior year. In the pre-analysis plan we will pre-commit to analyzing this variable for all students in our sample, though exploratory

- **Researcher-provided Student Assessment *(Student Assessment)*:**

(H2.2) Higher order learning outcomes increase

- **Measuring Recall, Understanding and Critical Thinking *(Student Assessment)*:** Close to 90% of assessment questions in developing countries, as analyzed by Burdett (2017), test students' ability to recall and memorize information. Thus, when analyzing student learning using official examinations such as PLE, researchers are most likely correlating student recall of facts and information with treatment status. These are important features of student learning, but not nearly enough to be prioritized to the extent that they are. We analyzed a common assessment booklet in Uganda, and tried to identify the questions that follow elements of Liu et al.'s (2014) conceptual framework for critical thinking assessments. Specifically, questions that: 1) evaluate evidence and its use, 2) analyze and evaluate arguments, 3) understand implications and consequences, 4) develop sound and valid arguments, 5) understand causation and explanation. These elements cohere with Burdett's (2017) description of reasoning/critical thinking. Burdett (2017) classifies questions into three categories of learning according to whether they measure recall, understanding (or application of knowledge), or reasoning/critical thinking. We found very few questions that address the last category of questions in SIPRO's (2019) standard assessment. This required us to construct our own questions measuring reasoning and critical thinking, which we did by following principles of critical thinking assessment design discussed in Liu et al. (2014). For further details on our approach, see appendix section D.

(H2.3) Field-based scientific competency increases *(Science Shows)*:

- **Science Shows.** The Ugandan National Curriculum Development Centre (NCDC) is pushing a curriculum transition towards a "competency-based" framework for education, starting with curriculum reform in secondary education. Recently, however, it began specifying expected competencies in students in primary training as well. Given the national emphasis on scientific competencies, and together with the Jinja District Education Office (DEO), we develop a field-based measure of testing scientific competencies in students across all of the schools in our study area. This activity is designed to measure the following competencies articulated by NCDC's (2016) curriculum guide for Primary 6 students:

  > Pupils apply correct scientific processes in investigations of various phenomena by: 1) identifying problems, 2) designing and practicing scientific investigation processes, 3) examining the evidence useful in inferences, 4) demonstrating the skills of observation, classification, accurate measurement and recording, 5) making predictions and formulating hypothesis for evidence, 6) communicating findings accurately and honestly, 7) analyzing causes and effects, 8) using a variety of sources for acquiring information, and 9) recording information with reasonable accuracy.

  We develop a 12-category rubric that measures each of these competencies on a scale of 1 to 10, with 10 reflecting better performance. Extensive details of the

analysis will analyze whether this variable responds differently when the analysis only includes students who attended the same school year after year.

approach taken by the DEO to invite the schools to participate in science shows, preparation of judges, and the events themselves are provided in appendix section E.

After the science show. The judges provide scores to the research team and the DEO. The research team aggregates the scores of each group by taking the mean across all twelve quantitative variables in the judges rubric represented in Table E.1.

(H2.4) Student creativity increases *(Student Survey)*

- **Creativity Score *Student Survey***: We utilize a well-known tool designed by psychologists to measure student creativity. We follow the procedure implemented by Bradler et al. (2019), who utilize this measure in an experimental economic setting. The method effectively involves a one minute exercise in which students generate a list of creative uses of a common object in their environment. We repeat this exercise for three objects: plastic bottles, maize plant residues, and bamboo stems. Once data are collected, researchers construct a creativity score by giving a point for each feasible use of the object and multiplying points according to the degree of uniqueness in the gathered data. Details of this procedure are outlined in appendix section F.

Additional Long-Run Student Learning Outcomes

In order to match students in secondary schools to a primary-school treatment status, we use a fuzzy merge procedure to match student identifiers in the secondary assessment with details collected in the census. We know, for example, that 3.428 students from census households attended lower secondary school in 2022. Of these, 1,206 students attend a secondary school in our new sample. Of these, 655 students attended a primary school in our sample. We are able to match

D Researcher Provided Student Assessment

The following pages display the student assessment as received by the students following enumerator administration of the student survey. Each individual students participating in the survey sits independently on a bench nearby the school. Enumerators were encouraged to tell the students to "try their best" if they had any questions regarding the survey, though such questions were minimized during the piloting phase of the student assessment.

Using SIPRO (2019), and recognizing that our sample only includes P6 students at the beginning of term 3, we borrow questions from a subset of the term 2 middle-of-term assessment questions sold to schools for the researcher-administered assessment.

These assessments rarely contain questions relevant for measuring higher order learning. To address this shortcoming, we designed several questions that intend to measure higher-order learning principles described in Liu et al. (2014). Specifically, we tried to identify one question for each of the following elements of Liu et al.'s (2014) conceptual framework for critical thinking assessments: 1) evaluate evidence and its use, 2) analyze and evaluate arguments, 3) understand implications and consequences, 4) develop sound and valid arguments, 5) understand causation and explanation. These elements cohere with Burdett's (2017) description of reasoning/critical thinking.

Appendix section D displays the format we used for the student assessments. Questions 1 through 5 were borrowed directly from SIPRO (2019) while questions 6 through 9 were generated by the research team. Students fill out the assessment on paper and enumerators enter student answers into our pre-programmed tablet-based form. Enumerators were also trained to assess the completeness of sentences and grammar structure of responses in question 4. Appendix Table D.1 demonstrates how we classified each assessment question according to Burdett (2017).

We utilize factor analysis to predict three coarse learning outcomes based on our approach to classifying assessment questions. The three outcomes reflect different dimensions of student ability: 1) recall, 2) understanding and 3) reasoning or critical thinking ability. Out of the 23 opportunities to receive points from the SIPRO (2019) exam, 52% fall under recall, 39% fall under understanding and 9% fall under reasoning.

The questions we add to the student assessment all fall under the reasoning category. However, since such questions are rare in the Ugandan context, we only use them to predict reasoning ability if they achieve a benchmark threshold of validity.[58] Specifically, since 2 of the questions from SIPRO (2019) represent the reasoning criteria, we analyze correlations between these two questions and each of the seven questions we construct ourselves. If the average correlation across the two questions that provide benchmark for our "reasoning" criteria is less than 0.05, we exclude the question from our construction of the reasoning variable.

Table D.1 reflects how each of the questions in the assessment, and the modes of student response, were classified to reflect three categories of learning: 1) student recall and recognition of information (recall), 2) student's ability to understand and apply a concept (understand), 3) students ability to analyze, and evaluate an argument (reason). Throughout, we roughly follow Burdett (2017) when analyzing each question type and classifying them for our own analysis.

[58] Indeed, Burdett (2017) shows that such questions are very rare throughout the African continent, though Uganda utilizes such questions with greater frequency than its sister countries across Africa.

					Level 1 - Recall		*Level 2 - Understand/Apply*									*Level 3 - Reason*							
Broadest Classification of Assessment Questions:					**Remember**					**Understand**					**Apply**			**Analyze**			**Evaluate**			**Create**	
Second level of classification:					Recognizing	Recall	Interpreting	Exemplifying	Classifying	Summarizing	Inferring	Comparing	Explaining	Executing	Implementing	Differentiating	Organizing	Attributing	Checking	Critiquing	Generating	Planning	Producing
Third level of classification:																							
Question	Text	SIPRO (2019)?	Subject	Classification																			
1.a)	Give any two examples of leguminous crops.	✓	Science	Recall	✓	✓																	
1.b)	Mention any two ways of caring for crops in the garden	✓	Science	Recall	✓	✓																	
2.a)	Name two examples of vitamin deficiency diseases	✓	Science	Recall	✓	✓																	
3.a)	Which part of a plant carries out reproduction?	✓	Science	Recall	✓	✓			✓														
3.b)	Which part of a plant provides attachment to branches?	✓	Science	Recall	✓	✓			✓														
4.a)	At which school was the fire outbreak?	✓	English	Understand	✓		✓				✓												
4.a)	†Proper Sentence Structure	✓	English	Recall	✓									✓									
4.b)	On which day did the fire break out?	✓	English	Understand	✓		✓				✓												
4.b)	†Proper Sentence Structure	✓	English	Recall	✓									✓									
4.c)	Whose dormitory got burnt?	✓	English	Understand	✓		✓				✓												
4.c)	†Proper Sentence Structure	✓	English	Recall	✓									✓									
4.d)	Where were the children when the fire broke out?	✓	English	Understand	✓		✓				✓												
4.d)	†Proper Sentence Structure	✓	English	Recall	✓									✓									
4.e)	Why was the police not able to save the property?	✓	English	Understand	✓		✓			✓	✓		✓										
4.e)	†Proper Sentence Structure	✓	English	Recall	✓									✓									
4.f)	What did the fire brigade officers tell the school management?	✓	English	Understand	✓		✓				✓												
4.f)	†Proper Sentence Structure	✓	English	Recall	✓									✓									
4.g)	Where was the school advised to put smoke detectors?	✓	English	Understand	✓		✓				✓												
4.g)	†Proper Sentence Structure	✓	English	Recall	✓									✓									
5.a)	How many patients were admitted on Sunday?	✓	Math	Understand	✓		✓																
5.b)	Which day had the least number of patients admitted?	✓	Math	Understand	✓		✓		✓		✓	✓	✓				✓						
5.c)	How many more patients were admitted on Wednesday than Thursday?	✓	Math	Reason	✓		✓		✓		✓	✓				✓	✓	✓					
5.d)	Find the total number of patients admitted in the whole week.	✓	Math	Reason	✓					✓							✓	✓					
6.a)	Which statement best supports Sarah's thoughts?		Critical Thinking	Reason	✓		✓				✓					✓	✓	✓					
6.b)	Which statement best supports Aminah's thoughts?		Critical Thinking	Reason	✓		✓				✓					✓	✓	✓					
6.c)	Which statement best supports both Aminah and Sarah's thoughts at the same time?		Critical Thinking	Reason	✓		✓				✓					✓	✓	✓					
6.d)	Which statement does not support either Aminah or Sarah's thoughts?		Critical Thinking	Reason	✓		✓				✓					✓	✓	✓					
7)	Sarah's Evaluation of a statement		Critical Thinking	Reason	✓		✓									✓	✓	✓	✓	✓			
8)	Aminah's Causal Reasoning		Critical Thinking	Reason	✓		✓									✓	✓	✓	✓	✓			
9)	Correlation vs. Causation		Critical Thinking	Reason	✓		✓									✓	✓	✓	✓	✓			

Table D.1: Classification of Assessment Questions

School:_____

Name:_____

Class:_____ Stream:_____

Age: _____ Sex: (Male OR Female – Circle one)

Village Where You live: _____

Name of parent/guardian: _____

Your position in class at the end of term 2: _____

---

1. A. Give any two examples of leguminous crops.

   i. _____

   ii. _____

   B. Mention any two ways of caring for crops in the gardens.

   i. _____

   ii. _____

2. A. Name two examples of vitamin deficiency diseases.

   i. _____

   ii. _____

3. A. Which part of a plant carries out the following activities?

   i) reproduction. _____

   ii) provides attachment to branches._____

   B. State any two uses of leaves to plants.

   i. _____

   ii. _____

4. **Read the passage below carefully and then answer, in FULL SENTENCES, the questions that follow.**

It was a sad Saturday morning at Mogga Primary School last term. That morning shocked everyone with the sudden burning of one of the dormitories for Primary Four children. Fortunately, by the time of the fire outbreak, all of them were in class writing weekly tests, so no children were injured but a lot of property was burnt. The Police Fire Brigade rushed to the scene to save property but failed because by the time it arrived, most properties had burnt to ashes.

The fire brigade officers advised the school management to put up enough fire fighting equipment to use in case of a fire outbreak. The burnt dormitory had only one fire extinguisher which made it hard for people to stop the fire. The school did not have smoke detectors so they were advised to put smoke detectors on every building in the school.

a) At which school was the fire outbreak?

_____

b) On which day did the fire break out?

_____

c) Whose dormitory got burnt?

_____

d) Where were the children when the fire broke out?

_____

e) Why was the police not able to save the property?

_____

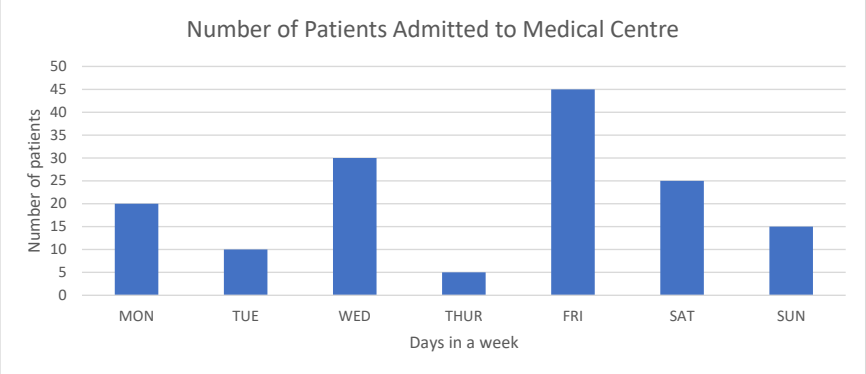
f) What did the fire brigade officers tell the school management?

_____

g) Where was the school advised to put smoke detectors?

_____

5. The bar graph below shows the number of patients admitted to the life care medical centre in a week. Study and use it to answer the questions that follow.



a) How many patients were admitted on Sunday?

_____

b) Which day had the least number of patients admitted?

_____

c) How many more patients were admitted on Wednesday than Thursday?

_____

d) Find the total number of patients admitted in the whole week.

_____

Question 6: In the survey, you shared your thoughts about Sarah and Aminah and how they think about education. Remember,

- **SARAH BELIEVES:** the purpose of education is to help people get jobs.
- **AMINAH BELIEVES** that the purpose of education is to help people make positive changes in their communities.

Think about these sentences.

1. 1 out of 5 (20%) of students repeat a class each year.
2. It is easier for someone with education to find a job.
3. Without education, it is difficult to improve your life.
4. In every village, the most helpful people are those with the most education.

According to you, which sentences belong in each of the blank spaces below?

(Choose one sentence for each blank space and write the sentence number in the blank)

a.) Which statement best supports Sarah's thoughts? _____

b.) Which statement best supports Aminah's thoughts? _____

c.) Which statement best supports both Aminah and Sarah's thoughts at the same time? _____

d.) Which statement does not support either Aminah or Sarah's thoughts? _____

Question 7: SARAH heard that the following statement is TRUE.

1. People who are educated have more skills **BECAUSE** of their education.

SARAH decides that this also means:

2. People with skills are more educated **BECAUSE** of their skills. (CIRCLE ONE)

Is **SARAH** correct in making this decision?

CORRECT NOT CORRECT

Question 8: Aminah believes the following two statements are TRUE.

1. People who are educated serve their community **BECAUSE** of their education.

2. People who serve their community create peaceful communities **BECAUSE** of their service.

Thinking only about **AMINAH**'s beliefs, would **AMINAH** say the following:

3. People create peaceful communities **BECAUSE** of their education. (CIRCLE ONE)

YES NO

Question 9:

In many parts of the world, the following statement is true:

1. Schools that have poor performance receive more teachers to help students perform.

Is statement number 2 bellow CORRECT or NOT CORRECT?

2. If the number of teachers in a school decreases, the school's performance will improve.

CORRECT NOT CORRECT

E Judging the Science Shows

Prior to the science show. In order to measure these competencies we worked with the DEO to organize science shows at each of the schools in our study and trained judges to utilize a rubric that measures student competencies along each of these criteria. We describe each of these activities in turn below.

First, NCDC (2016) articulates a set of activities that P6 students could implement in the course of a science show. In term two and three science includes lessons on natural resource conservation (with an emphasis on soil) and clean water preparation (for drinking) at home. Together with the DEO, we drafted a letter to each school that, in part, reads:

> "Each stream of primary 6 class will organize 3 or 4 groups of students who will prepare a science-oriented activity on one of two subject competencies:
>
> 1. "Experiments that compare at least two different approaches of conserving resources in the community such as soil, minerals, fuel, water, and air.
> 2. "Experiments that compare at least two different approaches of preparing clean and safe water for drinking and washing."

This letter is sent to schools towards the end of term 2 (typically July) and schools are informed that science shows will take place in the middle of term 3 (October), giving them three months to prepare. The letter states that each group of students will present the results of their experiments in front of pre-appointed science show judges and each presentation should take 10-15 minutes with an expected 5 additional minutes of questions and answers with the judges.

For schools participating in the science shows, the student surveys ask students to describe 1) the research question they are exploring in their science show project, 2) the hypothesis around the question, 3) why they think the hypothesis might be true, 4) which two approaches they will be comparing in their experiment and 5), what they should observe from these approaches according to their hypothesis. The purpose for these questions was to ensure that schools were preparing adequately for the science show and also to provide the team of researchers material that could be adapted to utilize during the training of science show judges.

The rubric provided to the judges is displayed in appendix E and is shared with each school as an attachment to the letter. We measure each of NCDC's desired competencies on a ten point scale using twelve questions. Table E.1 summarizes the rubric item used to assess each of these competencies. Our aim is to make each measure as objective a description of each competency as possible. Schools are also told that the DEO will invite a selection of community members to attend each show on the specified day and time and the highest ranking school will receive a certificate from the DEO that will also be displayed at the district office headquarters.

One week prior to the training, IPA Uganda helped identify a pool of six highly qualified enumerators to act as science show judges who participated in a training. The researchers modified student survey responses that emulated possible science shows the team would be expected to judge in the following month and asked each of the judges to provide scores for each of the items on the rubric. For three days, judges spent one hour assessing a handful of mock science shows using the form containing the rubric. Each of

Framing	
1. Identifying the problem: The pupils clearly stated the problem they are hoping to address with their project.	**2. Relevance of the problem: The pupils clearly stated why the problem is an important one to address.**
1 = Did not state the problem at all. 5 = State a problem, but not clearly. 10 = Clearly stated the problem.	1 = Did not indicate why the problem is important. 5 = Indicated why the problem is important, but not convincingly. 10 = Convincingly indicated why the problem is important.
Experiment	
3. Designing an experiment: The pupils conducted an experiment that clearly outlines how they would study the problem they identified.	**4. Designing an experiment: The experiment was very creative — it tries to learn about the problem in a way that others have tried before.**
1 = Did not conduct an experiment at all. 5 = Conducted an experiment that was not clear. 10 = Conducted a very clear experiment.	1 = Others have tried to address the problem in a similar manner. 5 = The approach taken was creative, but not original. 10 = The approach was creative and original (never seen before).
5. Designing an experiment: The pupils tested a technology not seen before. 1 = The technologies tested are known by all. 5 = The technologies are known by some but not all. 10 = The technologies were completely original and creative.	
Hypothesis	
6. Describing a hypothesis: The students had a clearly articulated hypothesis.	**7. Analyzing a hypothesis: The students clearly linked their hypothesis to the design of the experiment.**
1 = I had no idea what hypothesis the students were testing. 5 = The students mentioned a hypothesis, but it was not clear. 10 = The students mentioned a very clear hypothesis, comparing at least two groups.	1 = There was no link between the hypothesis and the experiment. 5 = The link between the hypothesis and experiment was there but not clear. 10 = It was very clear how the experiment would address the hypothesis.
Measurement	
8. Making observations: The students captured observations in a systematic manner (for example, by using a logbook).	**9. Making observations: The students made accurate measurements of their observations using relevant instruments.**
1 = There is no evidence that the students recorded observations. 5 = Students recorded observations, but not in a structured way. 10 = The observations were recorded in a very structured and systematic way.	1 = The students did not make measurements. 5 = The students made measurements that were not accurate. 10 = The students made very accurate measurements.
10. Sources of information: The students used a sufficient number of information sources to make their conclusions (for examples, they ran the experiment more than once or they measured more than one outcome.	
1 = The students did not use any sources of information to support conclusions. 5 = There were at least two sources of information in support of conclusions. 10 = The students made every effort to support conclusions with all available evidence.	
Articulating	
11. Communicating information: The students communicated their findings independently.	**12. Communicating information: The students communicated their findings accurately.**
1 = The students could not describe what they did without help. 5 = The students described their work sometimes on their own, sometimes with help. 10 = The students were the only ones describing their work.	1 = The students looked like they did not understand what they were saying. 5 = The students understood what they were saying, but it wasn't always connected to what they did and learned from their experiments. 10 = The students understood their work and connected it to their experiment.

Table E.1: Science Show: Judge's Rubric

the judge's scores was entered into an excel spreadsheet and compared with one another. A PI for the project then identified questions where there was disagreement around the responses provided by each judge. Though imperfect, this process allowed the PIs to identify judges who could articulate the reasons behind the scores they provided in a clear and sound manner. Through this process, two of the judges were selected to work as judges for the science show on behalf of the PI and the DEO.

The day of the science show. Science shows were held in two to three schools

each day from middle of October 2019 until the beginning of November 2019 (9:00 AM, 11:30 AM, and 2:30). The team of science show judges and two assistants (who also serve as backup judges) arrived at each school one hour prior to each show to assist the headteacher with logistics. In a typical science show, the group of students tasked with presenting their experiment lead the judges and community members to the site of their experiment — agricultural experiments were often implemented outdoors near the school's garden plot. Groups differed in the modality of their presentation — some selected a group leader who was tasked with presenting the work of the group, others presented their experiment as a team. Regardless of the mode of presentation, the judges were trained to ask a set of questions that they would direct to the entire team and that they would use in scoring each presentation. These question types are also articulated in the judge rubric in appendix E and are summarized by the following four items:

1. When did you start preparing for this activity? How did you come up with the question that you studied?

2. Describe how you decided to investigate the question?

3. How did you measure outcome [*specify the outcome from the experiment context*]?

4. *[Instruction: Provide an alternative explanation for the observations made by the students. Ask the students how they would respond to this explanation.]*[59]

After engaging with each of the presentations, the judges invite the teachers and head teachers to address the students and community members. They then close by expressing their appreciation for the student and school's efforts and stating that the DEO will communicate the results of the shows at a later date.

For the sake of transparency, we present qualitative comments provided by judges in Tables E.2 and E.3 for the top 10 and median 10 group performances in the science shows alongside their overall ranking. The following pages display the rubric provided to science show judges for the purpose of assessing student performance in the science shows.

[59]This exercise was incorporated into the training of the science show judges prior to the science shows. The training was thorough enough that the judges developed a standard set of alternative explanations for most student experiments.

Table E.2: Top 10 Student Performance in Science Shows

Rank	Score	Research Hypothesis	Research Experiment
Top 10 ranked Science Groups			
1	110	Filtered water is clean compared to unfiltered water	The pupils collected muddy water and a water bottle and cut off its bottom and turned it upside down. They put cotton wool at the bottom [of the bottle], added gravel, coarse sand, fine sand, and charcoal pest then poured muddy water through the bottle.
2	109	Filtered water is clean as compared to unfiltered water	They got cotton placed them in a container, added gravel stones, coast sand, fine sand,charcoal paste. Added a container of muddy water and placed on the Stand with another container below to receive the clean water.
3	100	Mulched soils are not carried away by running water compared to the soils that are left bare	The pupils obtained two soil samples and put them into two boxes. They covered one soil sample with green grass. After, they poured water onto the covered soil sample to determine how much soil will be lost and put a basin below the soil sample where water was dropping. They then poured water on the soil which was not covered with grass.
4	100	Mulched soils are not carried away by running water while compared to soils that are left bare	The pupils obtained two soil samples and placed them in two different containers. They left one of the container bare and the other covered the soil with grass. They then poured water in both containers and placed basins below to help capture the soil that was lost in both containers
5	93	Boiled water is safe for drinking compared to filtered water	Experiment 1: Decanting- They collected muddy water in a container, prepared 3 containers and poured muddy water into the first container and left it to settle for some time. They then transferred the water into the second container and left it to settle again. They then poured the water into the third container and later poured it into the kettle. They put the water on fire and left it to boil and later poured the water into a clean container. Experiment 2: Filtration collected muddy water. Students got a funnel, a clean container, and a filter funnel. Poured the water into clean container through the funnel to remove the residue. They then poured the water into the kettle to boil it. After boiling they put the water into a clean container and covered it.
6	88	Boiled water is clean and safe for drinking compared to filtered water	Collect muddy water in a container and leave it to settle as you prepare 3 clean containers. Pour the muddy water in the first container leaving the dirt below and leave it for sometime to settle. Transfer the water carefully into another basin leaving the dirt within and wait for it to settle, put the water in the third basin and then put it in the kettle then boil it. Let it to cool for sometime and then put it in Ina covered container ready for drinking Get two containers and put muddy water in one of the basins then use a sieve and pour the water into the other container and then boil the water and put it in the container covered ready for drinking.
7	86	Soils that are covered have more moisture and nutrients compared to soils that are not covered	They got maize stocks, banana leaves, and dry grass and covered the soil.
8	85	Filtered water is clean compared to unfiltered water	[The students] got muddy water and put a piece of cloth through the funnel. They then poured dirty water into the container through the funnel. They repeated the process a second time.
9	83	Wastes like banana peelings, cassava peelings and potato peelings decay to make soil fertile other than broken bottles and kaveera that instead damage the soil	The pupils collected different items like banana peelings, potato peelings, ash, cassava peels, egg shells, kaveras, broken bottles plastics , old clothes and shoes and a container having soil. They got banana peelings which they put in the soil which decayed to make manure and made the soil fertile.
10	82	Filtered water is clean compared to water that is not filtered	The pupils got a cup and placed in it mudy water , placed a clean cloth on the cup and poured water though Into another container. He removed the dirt from the cloth and then poured the water through again.

Notes: Responses were collected from science show judges as described previously in the paper. Science group participants were graded out of 120 points on their ability to conduct scientific experiments and present their findings with clarity. This table shows the research questions, processes, and hypotheses of the groups that ranked in the top 10 out of 182 according to judges. We also collected further qualitative insights on the nature of the experiments; their measurement and observation; and the students' perceptions of the importance of this research to their community, among others.

Table E.3: Median 10 Student Performance in Science Shows

Rank	Score	Research Hypothesis	Research Experiment
101-110th Ranked Science Groups			
101	23	Not defined	The pupils put soils in 4 pieces of [a] jerrican, planted beans in beans in the first jerrican, planted maize in the second jerrican, tomatoes in the third, and Sukuma in the fourth.
102	23	Not defined	The pupils collected plastic bottles and sticks. They then put holes in the bottles and arranged bottles onto the sticks. They then got nails and used them to make a dustbin out of the bottles.
103	23	Covered soil is fertile because it stops soil erosion compared to soil that is not covered	The [students] covered one plot of soil with banana leaves and dug some lines in another plot to create terraces.
104	23	Not defined	The pupils got two samples of soil, one was covered with dry grass and the other one terraces was made showing that once water comes with a high speed it is reduced to control soil erosion.
105	22	Not defined	The pupils collected soil and divided it into 3 plots, the first plot had maize and the others had no plant[s] on them. They then applied manure around the maize plant on the first plot. They then spread [manure] onto the soil of the second plot and mixed it with the soil and afterwards planted seeds and covered. In the third plot, the students dug holes and added manure into the holes, mixed with the soils and planted 4 maize grains in each hole.
106	22	Covered soil is fertile because soil erosion doesn't occur compared to the one that is not covered	Get banana leaves and grass and cover the soil. It will control soil erosion. Draw lines in The soil to make terraces
107	22	Not defined	Pupils got soil, planted egg plants and covered the soil with dry plant leaves. They also put soil in a jerrican and planted trees.
108	22	Pupils divided the land into 3 plots. Planted sweet potatoes on the first plot, banana on the second plot, maize on the third plot	
109	21	Not defined	The pupils got poles and used nails to make a tiptap stand. After they got a 3 litre jerrican and put holes into it and passed a [goose] wire through the holes on the jerrican and tied it onto the tiptap stand. They tied another goose wire on the handle of the jerrican and tied it to a pole. They then poured water into the jerrican and covered it.
110	20	Not defined	The pupils obtained dirty water into a jerrican, they put it into a glass and let it settle. They then poured the water into another glass.

Notes: Responses were collected from science show judges as described previously in the paper. Science group participants were graded out of 120 points on their ability to conduct scientific experiments and present their findings with clarity. This table shows the research questions, processes, and hypotheses of the groups that ranked in the between 70th and 79th out of 182 according to judges. We used this range of observations as those that were ranked lower had difficulty providing research questions and hypotheses and thus wouldn't have been provided greater clarity and the qualitative nature of science show scores. We also collected further qualitative insights on the nature of the experiments; their measurement and observation; and the students' perceptions of the importance of this research to their community, among others.

School:_____Group Number:_____

Date:_____ Time: _____ Name of Judge: _____

JUDGE SUMMARIZE THE GROUP'S ACTIVITY:

WHAT ARE THE NAMES OF THE PUPILS WHO PRESENTED FOR THIS GROUP:

WHICH PUPIL SEEMS TO BE THE LEADER OF THE GROUP:

WHAT WAS THEIR RESEARCH QUESTION?

WHAT PROCESS ARE THEY STUDYING?

DESCRIBE THE EXPERIMENT THAT THE PUPILS ARE UNDERTAKING:

WHAT WAS THE GROUP'S HYPOTHESIS:

WHAT DID THE PUPILS OBSERVE:

School:_____Group Number:_____

Date:_____ Time: _____ Name of Judge: _____

HOW DID THEY MEASURE THESE OBSERVATIONS:

WHAT CONCLUSIONS DID THE PUPILS REACH:

JUDGE TALLIES:

1. Identifying the problem: The pupils clearly stated the problem they are hoping to address with their project. (Choose a number between 1 and 10)
   1 = Did not State the Problem at all;
   5 = Stated a problem, but not clearly;
   10 = Clearly stated the problem

   SCORE:

2. Relevance of the problem: The pupils clearly stated why the problem is an important one to address.
   1 = Did not indicate why the problem is important to address
   5 = Indicated why the problem is important, but not convincingly
   10 = Convincingly indicated why the problem is important

   SCORE:

3. Designing an experiment: The pupils conducted an experiment that clearly outlines how they would study the problem they identified.
   1 = Did not conduct an experiment at all;
   5 = Conducted an experiment that was not clear;
   10 = Conducted a very clear experiment;

   SCORE:

4. Designing an experiment: The experiment was very creative – it tries to learn about the problem in a way that others have tried before.
   1 = Others have tried to address the problem in a similar manner.
   5 = The approach taken was creative, but not original.
   10 = The approach was creative and original (never seen before)

   SCORE:

5. Designing an experiment: The pupils tested a technology not seen before.
   1 = The technologies tested are known by all.
   5 = The technologies are known by some but not all.
   10 = The technologies were completely original and creative.

   SCORE:

6. Describing a hypothesis: The students had a clearly articulated hypothesis.
   1 = I had no idea what hypothesis the students were testing.
   5 = The students mentioned a hypothesis, but it was not clear.
   10 = The students mentioned a very clear hypothesis, comparing at least two groups.

   SCORE:

School:_____Group Number:_____

Date:_____ Time: _____ Name of Judge: _____

7. Analyzing a hypothesis: The students clearly linked their hypothesis to the design of the experiment.
   1 = There was no link between the hypothesis and the experiment.
   5 = The link between the hypothesis and experiment was there but not clear.
   10 = It was very clear how the experiment would address the hypothesis.

   SCORE:

8. Making observations: The students captured observations in a systematic manner (for example, by using a logbook).
   1 = There is no evidence that the students recorded observations.
   5 = Students recorded observations, but not in a structured way.
   10 = The observations were recorded in a very structured and systematic way.

   SCORE:

9. Making observations: The students made accurate measurements of their observations using relevant instruments.
   1 = the students did not make measurements.
   5 = the students made measurements that were not accurate.
   10 = the students made very accurate measurements.

   SCORE:

10. Sources of information: The students used a sufficient number of information sources to make their conclusions (for example, they ran the experiment more than once or they measured more than one outcome).
    1 = The students did not use any sources of information to support conclusions.
    5 = There were at least two sources of information in support of conclusions.
    10 = The students made every effort to support conclusions with all available evidence.

    SCORE:

11. Communicating information: The students communicated their findings independently.
    1 = The students could not describe what they did without help.
    5 = The students described their work sometimes on their own, sometimes with help.
    10 = The students were the only ones describing their work.

    SCORE:

12. Communicating information: The students communicated their findings accurately.
    1 = The students looked like they did not understand what they were saying.
    5 = The students understood what they were saying, but it wasn't always connected to what they did and learned from their experiments.
    10 = The students understood their work and connected it to their experiment.

    SCORE:

13. Did the pupils demonstrate their activity in a practical way? (YES ---------- NO)

IF YES, DESCRIBE _____

School:_____Group Number:_____

Date:_____ Time: _____ Name of Judge: _____

Q&A Session

INSTRUCTIONS TO JUDGES: DURING THE Q&A SESSION, ENCOURAGE RESPONSES FROM ALL OF THE STUDENTS. IN OTHER WORDS, IF ONLY ONE STUDENT RESPONDS TO THE QUESTION, PLEASE CALL ON OTHER STUDENTS AT RANDOM FROM WITHIN THE GROUP.

1. **JUDGE ASK THE STUDENTS: When did you start preparing for this activity? How did you come up with the question that you studied?**

Summarize Response:

INSTRUCTION TO JUDGE: Allow the students to freely respond, and then indicate which of the following best resembles the student response. [CHOOSE ONE]

A. The learners formulated their own questions or hypothesis to be tested.
B. Teacher suggests topic areas or provides samples to help learners formulate own questions or hypothesis.
C. Teacher offers learners lists of questions or hypotheses from which to select.
D. Teacher provides learners with specific stated questions or hypotheses to be investigated.
E. Other _____
F. Cannot respond because no clear choice observed.

2. **JUDGE ASK THE STUDENTS: Describe how you decided to investigate the question?**

Summarize Response:

INSTRUCTION TO JUDGE: The students shared in their presentation how they answered the question. What we want to know here is how they decided whether this approach to answering the question would be a good one.

A. Learners developed the procedures and protocols to independently plan and conduct a full investigation.
B. Teachers encouraged the learners to plan and conduct a full investigation, providing support along the way.
C. Teacher provided the guidelines for learners to plan and conduct part of an investigation. Some choices are made by learners.
D. Teacher provides the procedures and protocols for the students to conduct the investigation.
E. Other_____
F. Cannot respond because no clear choice observed.

School:_____Group Number:_____

Date:_____ Time: _____ Name of Judge: _____

3. **JUDGE ASK THE STUDENTS: How did you measure outcome XXX?**

Judge's Outcome identified:

Pupil's Response:

---

4. **JUDGE ASK THE STUDENTS: Provide an alternative explanation for the observations made by the students. Ask the students how they would respond to this observation.**

Judge's Alternative Explanation:

Pupils' Response:

INSTRUCTION TO JUDGE: Select one of the following.

A. The pupils were very convincing in their response and responded to this observation adequately.
B. The pupils provided a response to this observation and it was somewhat convincing.
C. The pupils provided a response to this observation but it was not convincing.
D. The pupils did not respond at all to the alternative explanation.
E. Other_____

Other Questions:

- What is the importance of this activity (you could add if they are confused about this question…. to the community)?
- _____
- _____

F Categories of Creativity score

Bradler et al. (2019) measure creative performance with the unusual uses task originally developed as Guilford's alternative uses task (Guilford, 1967). In the unusual uses task, subjects are asked to name as many unique and unusual uses as they can in a limited amount of time for an item common to their subject context. In our setting, we asked students to identify unusual uses for three different items: a plastic bottle, maize plant remains, and bamboo stems. The enumerator first demonstrated how the activity works by providing some creative uses of a plastic bag and then asking the student to share additional ideas for creative uses of plastic bags. Once the students understand the nature of the activity, they are told that they have one minute to list as many creative and uncommon uses of the next item that they possibly can. Once they are told the item, the enumerator starts a timer on a tablet that sounds an alarm when 60 seconds have passed. In the interim, the enumerator writes each creative use of an object mentioned by a student to later record into the tablet.

We evaluate students' answers using two measures of the unusual uses task: "validity" and "originality".[60] Each valid use is valued at one point. We then categorize the answers of students according to their primary functions.[61] The originality of responses is measured by the statistical infrequency of answers according to the categories they corresponded to. Specifically, we give one additional point to a valid answer if less than 10% of participants gave the same answer and allotted two additional points to a valid answer as very original if less than 1% of subjects gave that answer.[62]

Answers that were unclear were counted as invalid. Table F.1 shows the categories and percent responses for data collected in 2019for the task involving a plastic bottle. The respective tables for the tasks involving the remains of maize plants and bamboo stems can be found in the appendix (Tables F.2 and F.3 respectively).

[60] We omit "flexibility," and "elaboration," measurement items in the original creative uses task (Guilford, 1967). Flexibility reflects the variety of a student's response and is determined by counting the number of different categories into which responses fall. Elaboration reflects the level of detail in their response. Since we utilized enumerators in drawing out responses from students, we feel that these two measures are more compromised by enumerator effects than "validity" and "originality," which complicate the use of these concepts in empirical analysis even with enumerator fixed effects.

[61] Unlike Bradler et al. (2019) we do not delineate response categories *ex ante*. Rather, we anonymize responses and ask two members of our research team to categorize the universe of responses. In case of disagreements a PI intervenes and makes the final decision.

[62] Bradler et al. (2019) use thresholds of 8% and 1% respectively.

Category	Percent
Storage	40.1
Drinking	26.2
Toy	22.7
Irrigation	17.1
Planting	14.4
Income	11.4
Charcoal Lighter	11.1
Recycling	8.6
Construction	8.5
Funnel	4.7
Decoration	3.6
Plates	2.3
Musical Instrument	1.5
Stove Stand	1.4
Animal Feeders	1.3
Washing	1.1
Measurement	0.9
Scare Crow	0.8
Tool Handle	0.8
Jerrycan Repair	0.8
Rubbish Container	0.7
Filter	0.7
Curtains	0.6
Spray	0.6
⋮	⋮
Fishing	0.6
Light Protector	0.5
Science Experiments	0.5
Shoes	0.4
Life Jacket	0.3
Plastic Wire	0.3
Workout Equipment	0.3
Weapon	0.3
Control Soil Erosion	0.3
Gutter	0.2
Pipes	0.1
Make Dice	0.1
Making Bricks	0.1
Tyre	0.1
Communication Lines	0.1
Roof	0.1

Note: This table shows the categories used for the student creativity scores for the exercise involving plastic bottles. Each mention receives one point and objects mentioned less than 10% and 1% frequency receive one and two extra points respectively.

Table F.1: Categories used for creativity scores for unusual uses of plastic bottles

Category	Percent
Fire	41.8
Mulching	38.3
Fertilizer	30.3
Cooking	29.4
Animal Feed	28.9
Construction	10.2
Toy	6.1
Fencing	3.9
Roofing	3.8
Disciplinary Tool	2.3
Income	1.5
Cover	1.2
Scare Crow	1.1
Plant Support	1.0
Cleaning	0.8
Decoration	0.8
Hygiene	0.7
Making Ashes	0.7
Mats	0.7
Medicine	0.7
Straw Bed for Animals	0.6
Charcoal	0.6
Making Tools	0.5
Furniture	0.3
Weapon	0.3
⋮	⋮
Baskets	0.2
Traditional Practices	0.2
First Aid	0.2
Wind Breaker	0.2
Building Shades	0.2
Creating Salt	0.2
Creating Smoke	0.2
Sugar Cane	0.1
Goal Posts	0.1
Riding Aid	0.1
Counters	0.1
Hat	0.1
Harvest Beans	0.1
Regerminate	0.1
Maize Packaging	0.1
Saw Dust	0.1

Note: This table shows the categories used for the student creativity scores for the exercise involving maize residual. Each mention receives one point and objects mentioned less than 10% and 1% frequency receive one and two extra points respectively.

Table F.2: Categories used for creativity scores for unusual uses of maize plant residues

Category	Percent
Construction	60.4
Disciplinary Tool	35.4
Fire Wood	20.2
Fencing	11.7
Cooking	11.3
Furniture	6.4
Income	6.1
Music Instrument	4.9
Pole	4.9
Animal Feed	4.6
Decoration	4.2
Roofing	3.8
Fishing	2.8
Plant Support	2.7
Crafts	2.5
Weapons	2.3
Medicine	2.2
Dishes	1.9
Boats	1.7
Toy	1.3
Art	1.3
Praying	0.9
Mulching	0.9
Gates	0.8
Measurement	0.6
⋮	⋮

Category	Percent
⋮	⋮
Fertilizer	0.6
Charcoal	0.6
Plumbing	0.6
Counting	0.6
Rain System	0.5
Straws	0.5
Control Soil Erosion	0.4
Wiring	0.3
Ladder	0.3
Bikes	0.2
Power Generation	0.1
Charcoal Lighter	0.1
Toothbrush/Floss	0.1
Smoking Pipe	0.1
Floaters	0.1
Pounding Machines	0.1

Note: This table shows the categories used for the student creativity scores for the exercise involving bamboo stems. Each mention receives one point and objects mentioned less than 10% and 1% frequency receive one and two extra points respectively.

Table F.3: Categories used for creativity scores for unusual uses of bamboo stems

G Additional Appendix Tables

Statistics	**Summary Statistics**								**Balance Tests**	
	All		Control			Treated				
	Mean	Sd	Mean	Sd	N	Mean	Sd	N	β	p value
Teacher Survey Outcomes (Student Variables — 2019 Only)										
Student Asks Questions	0.24	0.16	0.22	0.16	124	0.25	0.15	119	0.11	0.73
Corporeal Punishment	0.43	0.29	0.42	0.27	162	0.44	0.30	169	0.12	0.54
Knowledge of Student	0.45	0.23	0.43	0.21	247	0.46	0.24	238	0.48	0.12
Specification Statistics							**F Score (p value)**		1.34 (0.27)	
							Clusters		54	
Teacher Survey Outcomes										
Teacher Gender	1.39	0.49	1.40	0.49	415	1.42	0.49	422	0.03	0.50
Attended Training Last Year	0.33	0.47	0.36	0.48	415	0.32	0.47	422	-0.05	0.32
School Has Farm Land	0.65	0.48	0.67	0.47	415	0.68	0.47	422	0.01	0.94
Boys Better Pupils Than Girls?	3.33	1.52	3.33	1.52	415	3.37	1.51	422	0.01	0.72
Head Teacher Listens to Me	1.30	0.64	1.35	0.71	382	1.25	0.56	384	-0.06	0.08
How Satisfied with Job?	2.13	0.97	2.16	1.00	415	2.13	0.95	422	-0.01	0.61
Students Connect School to Family Life	1.69	0.84	1.72	0.90	415	1.68	0.78	422	-0.01	0.72
Specification Statistics							**F Score (p value)**		0.96 (0.47)	
							Clusters		89	
Teacher Dyad Outcomes										
ij Speak About Classroom Management	0.82	0.38	0.83	0.37	3,108	0.81	0.39	3,908	-0.02	0.60
ij Visit Each Others Classrooms to Learn	0.73	0.44	0.74	0.44	3,108	0.72	0.45	3,908	0.01	0.89
ij Plan Classroom Activities Together	0.62	0.49	0.64	0.48	3,108	0.59	0.49	3,908	-0.05	0.21
Specification Statistics							**F Score (p value)**		0.87 (0.46)	
							Clusters		88	
Classroom Observation Outcomes										
Share of Engaged Pupils	4.55	1.97	4.54	1.97	4,806	4.52	2.01	4,833	0.01	0.47
Activity: Q and A	0.19	0.39	0.20	0.40	4,807	0.18	0.39	4,834	-0.04	0.15
Activity: Practice and Drill	0.05	0.22	0.06	0.23	4,807	0.05	0.22	4,834	-0.04	0.34
Activity: Assignment	0.09	0.29	0.10	0.30	4,807	0.08	0.28	4,834	-0.06	0.10
Activity: Copying	0.10	0.30	0.10	0.30	4,807	0.09	0.29	4,834	-0.04	0.24
Teacher Out of Class	0.13	0.34	0.13	0.34	4,807	0.15	0.35	4,834	0.03	0.60
Materials: None	0.42	0.49	0.42	0.49	4,792	0.43	0.50	4,813	-0.03	0.65
Materials: Textbooks	0.04	0.19	0.04	0.18	4,792	0.04	0.20	4,813	0.02	0.77
Materials: Notebooks	0.10	0.30	0.11	0.31	4,792	0.10	0.30	4,813	-0.03	0.68
Materials: Blackboard	0.40	0.49	0.40	0.49	4,792	0.38	0.49	4,813	-0.03	0.63
Specification Statistics							**F Score (p value)**		0.74 (0.69)	
							Clusters		89	

Notes: This table reflects balance tests of teacher, teacher dyad, and classroom outcomes and covariates. We present summary statistics of each measure, displaying means and standard deviations for the whole sample "All," the sample of teachers/teacher dyads/classroom observations in control schools "Control," and the sample of teachers/teacher dyads/classroom observations in treatment schools "Treated" (the latter two also include number of observations). At the time of writing, we have baseline data for all schools for variables that do not use the Student Survey in their construction. For school admin data, including teacher attendance, we were in the process of collecting baseline data outside of Budondo sub-county in March 2020 when the Covid-19 related lockdowns started — we exclude these variables form balance tests here. Balance tests reflect an OLS regression with the specification $\text{Treated}_{is} = \beta X + \epsilon_{is}$ where i represents student, s represents school, X represents the vector of covariates in the rows of this table, β is the vector of coefficients associated with each covariate and ϵ_{is} is the error term clustered at the school level. Each specification is run for each data set separately, datasets are separated by the horizontal lines in the table. The F Score and number of clusters is reported for each specification. The F Score's p value (in parentheses) reports results of the null hypothesis test that coefficients are jointly orthogonal within a given specification.

Table G.1: Balance Test for Teacher Outcomes and Covariates

Statistics	Summary Statistics								Balance Tests	
	All		Control			Treated				
	Mean	Sd	Mean	Sd	N	Mean	Sd	N	β	p value
Student Survey Outcomes										
Passthrough Rate 2018-2019	0.88	0.33	0.87	0.33	283	0.90	0.30	314	0.06	0.34
Passthrough Rate 2017-2018	0.91	0.29	0.91	0.29	282	0.92	0.28	314	0.03	0.68
Creativity Index	7.46	3.91	7.82	4.05	283	7.88	4.22	315	0.00	0.88
Specification Statistics								**F Score (p value)**	0.33 (0.80)	
								Clusters	54	
Student Assessment Outcomes										
Recall and Recognition	-0.10	2.32	0.46	2.04	280	0.68	2.19	317	0.01	0.65
Apply Understanding	-0.04	1.73	0.43	1.42	280	0.50	1.29	317	-0.01	0.83
Critical Thinking	-0.04	1.50	0.16	1.48	280	0.29	1.45	317	0.01	0.59
Specification Statistics								**F Score (p value)**	0.22 (0.88)	
								Clusters	54	
Student PLE Outcomes										
English	2.98	1.69	2.61	1.55	364	3.26	1.73	486	0.06	0.15
Science	3.45	1.79	3.12	1.80	364	3.69	1.75	486	0.05	0.40
Mathematics	2.92	1.51	2.75	1.50	364	3.04	1.50	486	-0.03	0.16
Social Studies	4.41	1.75	4.20	1.69	364	4.57	1.79	486	-0.06	0.33
Female	0.54	0.50	0.54	0.50	376	0.55	0.50	494	0.00	0.99
Specification Statistics								**F Score (p value)**	1.66 (0.18)	
								Clusters	26	

Notes: This table reflects balance tests of student outcomes and covariates. We present summary statistics of each measure, displaying means and standard deviations for the whole sample "All," the sample of students in control schools "Control," and the sample of students in treatment schools "Treated" (the latter two also include number of observations). At the time of writing, we do not have baseline data for all schools. For Student Survey Outcomes and Student Assessment Outcomes, we did not collect student data until 2019, which was two years into the intervention in Budondo sub-county. For PLE outcomes, we were in the process of collecting baseline data outside of Budondo sub-county in March 2020 when the Covid-19 related lockdowns started. Balance tests reflect a regression with the specification $\text{Treated}_{is} = \beta X + \epsilon_{is}$ where i represents student, s represents school, X represents the vector of covariates in the rows of this table, β is the vector of coefficients associated with each covariate and ϵ_{is} is the error term clustered at the school level. Each specification is run for each data set separately, datasets are separated by the horizontal lines in the table. The F Score and number of clusters is reported for each specification. The F Score's p value (in parentheses) reports results of the null hypothesis test that coefficients are jointly orthogonal within a given specification.

Table G.2: Balance Test for Student Outcomes and Covariates

Table G.3: Summary of All Pre-analysis Specifications

Dependent Variables:	Fixed Effects			Estimator	Datasets	Unit of Obs.	Notes
	Pair	Enum	Grade				
(H1.1) Classroom Pedagogy Improves							
Student Engagement	✓	✓	✓	Ologit	Stallings	Classroom	Ordered proportion of engaged students.
Critical thinking and exploration[a]	✓	✓	✓	OLS	TIPPS	Classroom	Factor analysis using classroom observations
Understanding concepts and deep learning[a]	✓	✓	✓	OLS	TIPPS	Classroom	Factor analysis using classroom observations
Students ask questions	✓	✓		Tobit	Student Survey	Teacher	Transform student response into teacher-level variable. Censored below by 0 and above by 1.
Corporal punishment	✓	✓		Tobit	Student Survey	Teacher	Transform student response into teacher-level variable. Censored below by 0 and above by 1.
(H1.2) Teacher motivation and effort increases							
Teacher attendance	✓	✓		OLS	School Admin Data	Teacher	
Teacher knowledge of student	✓	✓		Tobit	Teacher and Student Survey	Teacher	Combine teacher and student responses. Censored below by 0 and above by 1.
(H1.3) Teacher collaboration increases							

[a] indicates that we were not able to measure this outcome.

Continued on next page...

. . . continued from previous page

Dependent Variables:	Fixed Effects			Estimator	Datasets	Unit of Obs.	Notes
	Pair	Enum	Grade				
Teacher learning and collaboration	✓	✓		OLS	Teacher Survey	Teacher Dyad	Mean of responses for ij pair.
(H2.1) Traditional learning outcomes increase							
English	✓			OLS	UNEB	P7 Student	Standardized score
Science	✓			OLS	UNEB	P7 Student	Standardized score
Math	✓			OLS	UNEB	P7 Student	Standardized score
Social Studies	✓			OLS	UNEB	P7 Student	Standardized score
Pass PLE	✓			OLS	UNEB	P7 Student	Standardized score
P6 Pass-through Rate	✓	✓		OLS	Student Survey	P6 Student	Most recent year
(H2.2) Higher order learning outcomes increase							
Understanding	✓	✓		OLS	Student Assessment	P6 Student	Standardized score of factor analysis
Critical Thinking	✓	✓		OLS	Student Assessment	P6 Student	Standardized score of factor analysis
(H2.3) Field-based scientific competency increases							
Science Show Result	✓	✓		Tobit	Science Show	Student Group	Average across all scores in rubric; censored below by 1 and above by 10
(H2.4) Student creativity increases							

[a] indicates that we were not able to measure this outcome.

Continued on next page. . .

. . . continued from previous page

Dependent Variables:	Fixed Effects			Estimator	Datasets	Unit of Obs.	Notes
	Pair	Enum	Grade				
Creativity score	✓	✓		Tobit	Student Survey	P6 Student	Score measuring number and uniqueness of student ideas; censored below by 0

Note: All specifications will measure the Treatment Effect using the Intent to Treat (ITT) coefficient. All standard errors across all specifications are clustered at the school level. Acronyms/Abbreviations: UNEB = "Ugandan National Examinations Bureau"; TIPPS = "Teacher Instructional Practices and Processes System"; OLS = "Ordinary Least Squares"; P7 = "Primary Seven"; P6 = "Primary Six"; Enum = "Enumerator" or "Judge"; Obs. = "Observations". Dependent variables grouped by hypotheses described in section **??**.

Hypothesis:	**(H2.1): Traditional Learning Outcomes**					
	Primary Leaving Exam Results 2019					**Passthrough 2018-2019**
Outcome Variable:	English	Science	Mathematics	Social Studies	PLE Pass	P6 to P7
Treatment (*ITT*)	0.60***	0.53***	0.19*	0.49***	0.27***	0.14***
	(0.09)	(0.12)	(0.09)	(0.10)	(0.07)	(0.04)
H_0 : *ITT* = 0						
p value	[0.00]±±	[0.00]±±	[0.05]±	[0.00]±±	[0.00]±±	[0.00]±±
RI p value	[0.00]±±	[0.01]±±	[0.13]	[0.00]±±	[0.01]±±	[0.02]±±
BH Critical p value (5%)	[0.01]	[0.03]	[0.05]	[0.02]	[0.03]	[0.04]
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Enum FE	No	No	No	No	No	Yes
Source of Data	UNEB	UNEB	UNEB	UNEB	UNEB	Student Survey
Unit of Observation	Student	Student	Student	Student	Student	Student
Standardized Variable	Yes	Yes	Yes	Yes	No	No
Range of Outcome Variable	[-1.28, 2.44]	[-1.56, 2.48]	[-1.29, 3.15]	[-1.82, 2.35]	{0,1}	{0,1}
Control School Mean	−0.57	−0.51	−0.34	−0.48	0.46	0.78
Clusters	29	29	29	29	29	29
Observations	808	808	808	808	866	294
Estimator	OLS	OLS	OLS	OLS	OLS	OLS

Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report p values using randomization inference ("*RI p value*") as well as the Benjamini-Hochberg (BH) "*BH Critical p value*" at the 5% level within hypothesis. ± suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; ±± suggests a significant discovery at the 5% level. Randomization Inference using 1,000 permutations of school-level treatment indicator within matched-pair strata. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Variable Descriptions: Units of observation in all columns are students, though the data sets and, hence, number of observations differ. The **first through fifth columns** utilize official test scores for the primary leaving exams (PLE) purchased from Uganda's National Examination Bureau in 2020, relating results from tests taken in November 2019. Students can receive scores ranging from 1 (best) to 9 (worst) for each exam. We have transformed this measure to facilitate interpretation such that 9 is the top score and 1 is the worst score (reflected in figures H.2a through H.2d). Furthermore, all measures in the first four columns have been standardized using the pooled subject-specific mean and standard deviation of control schools. Therefore, **treatment effect measures reflect standard deviation differences between treatment schools and control schools in columns one through four**. We analyze the PLE pass rate and P6 to P7 pass-through rate in **columns 5 and 6 respectively**. Both are binary measures equaling 1 if a student passed the exam/grade level and 0 otherwise. The number of observations in column five is greater than that in columns one through four because students who registered but were not present for the PLE exam are classified as students who must repeat the PLE exam in the following year. Students are divided into four "divisions" based on PLE scores, division 1 reflects high performance and division 4 is the lowest. Students receiving division 4 marks and below are considered to have the PLE (typically with an aggregate score above 28). The variable analyzed in column 6 is constructed using the student survey. For more details, see pages 20-21 of the pre-analysis plan.

Table G.4: Windsorized Data: Test of Hypothesis **(H2.1)**: Traditional Learning Outcomes Increase

Hypotheses:	**(H2.2): Higher Order Learning**		**(H2.3): Science Show**	**(H2.4):Creativity**
Outcome Variables	Apply/Understand	Critical Thinking	Index (Mean)	Index
Treatment (*ITT*)	0.80***	0.44***	0.78***	0.79***
	(0.13)	(0.13)	(0.29)	(0.17)
H_0 : *ITT* = 0				
p value	[0.00]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.01]$^{\pm\pm}$	[0.00]$^{\pm\pm}$
RI p value	[0.00]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.04]$^{\pm\pm}$	[0.00]$^{\pm\pm}$
BH Critical p value (5%)	[0.03]	[0.05]	[0.05]	[0.05]
Pair FE	Yes	Yes	Yes	Yes
Enum FE	Yes	Yes	Yes	Yes
Source of Data	Student Assessment	Student Assessment	Science Show	Student Survey
Unit of Observation	Student	Student	Student Group	Student
Standardized Variable	Yes	Yes	No	No
Range of Outcome Variable	[-2.45, 1.40]	[-1.84, 1.89]	[1, 9.16]	[0, 20]
Control School Mean	−0.53	−0.27	2.16	5.47
Clusters	29	29	28	29
Observations	295	295	141	295
Estimator	OLS	OLS	Tobit	Tobit

Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report p values using randomization inference ("*RI p value*") as well as the Benjamini-Hochberg (BH) "*BH Critical p value*" at the 5% level within hypothesis. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. Randomization Inference using 1,000 permutations of school-level treatment indicator within matched-pair strata. Tobit estimator in the third column treats 1 as the lower bound and 10 as the upper bound; in the fourth column, 0 is the lower bound with no upper bound. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Variable Descriptions: The unit of observation in the first, second and fourth columns is individual students, and student groups in the third column. Each group was judged by two separate judges, thus there are 79 total groups across 29 schools in the analysis. The **first and second columns** report results from the researcher administered student assessments, measuring applied understanding and critical thinking using the procedure outlined on pages 21 and 22 of the pre-analysis plan. The science show outcome in the third column averages across the 12 outcomes measured by science show judges and described on page 25 and table 1 of the pre-analysis plan. The creativity score follows the procedure described on pages 26 and 27 of the pre-analysis plan.

Table G.5: Windsorized Data: Tests of Hypotheses **(H2.2)** through **(H2.4)**: Higher Order Learning, Science Shows and Creativity

Hypothesis:	**(H1.1): Pedagogy**		**(H1.2): Effort**		**(H1.3): Learning**
Outcome Variable:	Share of Engaged Pupils	Student Inquisitiveness	Corporal Punishment	Knowledge of Student	Teacher Network
Treatment (*ITT*)	0.38**	0.07***	−0.00	0.14***	0.32***
	(0.15)	(0.02)	(0.05)	(0.02)	(0.10)
H_0 : *ITT* = 0					
p value	[0.01]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.94]	[0.00]$^{\pm\pm}$	[0.00]$^{\pm\pm}$
RI p value	[0.00]$^{\pm\pm}$	[0.19]	[0.99]	[0.00]$^{\pm\pm}$	[0.08]$^{\pm}$
BH Critical p value (5%)	[0.02]	[0.03]	[0.05]	[0.05]	[0.05]
Pair FE	Yes	Yes	Yes	Yes	Yes
Enum FE	Yes	Yes	Yes	Yes	No
Grade FE	Yes	No	No	No	No
Source of Data	Classroom Observations	Student Survey	Student Survey	Stud. + Teach. Survey	Teacher Network
Unit of Observation	Classroom Snapshots	P6 Teachers	P6 Teachers	P6 Teachers	Teacher Dyads
Range of Outcome Variable	{1,2,...,6}	[0,1]	[0,1]	[0,1]	{0,1,2,3}
Control School Mean	4.23	0.18	0.48	0.57	1.66
Clusters	29	29	29	29	29
Observations	2,140	84	83	83	1,318
Estimator	Ologit	Tobit	Tobit	Tobit	OLS

Analysis Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report p values using randomization inference ("*RI p value*") as well as the Benjamini Hochberg (BH) "*BH Critical p value*" at the 5% level within hypothesis. Hypothesis categories are delineated in the top row of the table. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. Randomization Inference using 1,000 permutations of school-level treatment indicator within matched-pair strata. Tobit estimator in the second through fourth columns treats zero as the lower bound and one as the upper bound. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Variable Descriptions: Units of observation in the **first column** are classroom snapshots observed using the Stallings instrument — 10 "snapshots" are taking during each class; therefore, 238 classes (P4 to P6) were observed across 29 schools. "Share of Engaged Pupils" indicates the observer's assessment of the share of pupils engaged in an activity with a teacher: 1 = No pupil, 2 = One pupil, 3 = A few pupils, 4 = Half of the pupils, 5 = Most of the pupils, and 6 = All pupils (details on page 16 of the pre-analysis plan). Teachers in the second to fourth columns are restricted to those who at least 90% of P6 students listed as one of their teachers. The **second column** analyzes the percent of students who indicate that they have asked questions in the previous two school terms to the specified teacher when they do not understand a topic. The **third column** analyzes the percent of students who indicate that the specified teacher has "caned" or beaten them at some point in the previous two school terms (details on page 17 of the pre-analysis plan). The **fourth column** combines student and teacher responses regarding student attendance and the students relationship with their guardians for each student. We analyze the treatment effect on the percent of correct teacher responses (details available on pages 18-19 of the pre-analysis plan). Unit of observation in the **fifth column** is within-school teacher dyads. We sum across binary measures of teacher interactions along three dimensions where both teachers indicate the existence of a link reflecting teacher collaboration (details available on pages 19-20 of the pre-analysis plan).

Table G.6: Windsorized Data: Tests of Hypotheses **(H1.1)** through **(H1.3)**: Teacher Pedagogy, Teacher Effort, Teacher Learning

Source of Data	**Science Show**											
Category	**Framing**		**Experiment**			**Hypothesis**		**Measurement**			**Articulating**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment (*ITT*)	−0.46	0.46	1.03**	0.36	0.06	1.11**	1.41***	0.20**	1.41***	0.83***	1.19***	1.28***
	(0.34)	(0.34)	(0.38)	(0.29)	(0.13)	(0.41)	(0.42)	(0.08)	(0.44)	(0.21)	(0.31)	(0.45)
H_0 : *ITT* = 0												
p value	[0.19]	[0.18]	[0.01]$^{\pm\pm}$	[0.23]	[0.66]	[0.01]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.01]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.00]$^{\pm\pm}$	[0.01]$^{\pm\pm}$
RI p value	[0.42]	[0.36]	[0.06]	[0.29]	[0.71]	[0.05]$^{\pm\pm}$	[0.03]$^{\pm}$	[0.04]$^{\pm\pm}$	[0.02]$^{\pm}$	[0.02]$^{\pm\pm}$	[0.01]$^{\pm\pm}$	[0.03]$^{\pm\pm}$
BH Critical p value (5%)	[0.05]	[0.03]	[0.02]	[0.03]	[0.05]	[0.05]	[0.03]	[0.05]	[0.02]	[0.03]	[0.03]	[0.05]
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Judge FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Standardized Variable	No	No	No	No	No	No	No	No	No	No	No	No
Range of Outcome Variable	[1,10]	[1,10]	[1,10]	[1,10]	[1,8]	[1,10]	[1,10]	[1,10]	[1,10]	[1,10]	[1,10]	[1,10]
Control School Mean	2.58	2.13	3.31	1.55	1.24	2.15	1.78	1.19	2.31	1.6	2.66	2.15
Clusters	28	29	29	29	29	29	29	29	28	28	28	29
Observations	141	141	141	141	141	141	141	141	141	141	141	141
Estimator	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS

Notes: Standard errors are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report p values using randomization inference ("*RI p value*") as well as the Benjamini-Hochberg (BH) "*BH Critical p value*" at the 5% level within hypothesis. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. Hypotheses are grouped according to the categories outlined in the second row of the table. Randomization Inference using 100 permutations of school-level treatment indicator within matched-pair strata. Tobit estimator in the third column treats 1 as the lower bound and 10 as the upper bound; in the fourth column, 0 is the lower bound with no upper bound. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Table G.7: Windsorized Data: Granular Science Show Outcomes

Category:	**Days Teacher Spent on School Farm**				**Days Teacher Spent with Pupils**			
Outcome Variable:	Clearing	Planting	Weeding	Harvesting	Clearing	Planting	Weeding	Harvesting
Treatment (*ITT*)	1.21	0.90*	1.22	0.35	1.93*	2.80***	2.52**	2.14***
	(1.20)	(0.48)	(0.85)	(0.53)	(1.07)	(0.80)	(1.27)	(0.67)
$H_0 : ITT = 0$								
p value	[0.31]	[0.07]$^{\pm}$	[0.15]	[0.51]	[0.07]	[0.00]$^{\pm\pm}$	[0.05]$^{\pm}$	[0.00]$^{\pm\pm}$
BH Critical p value (5%)	[0.04]	[0.05]	[0.01]	[0.03]	[0.01]	[0.03]	[0.04]	[0.05]
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Enum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Source of Data	Teacher Survey	Teacher Survey	Teacher Survey	Teacher Survey	Teacher Survey	Teacher Survey	Teacher Survey	Teacher Survey
Unit of Observation	Teachers	Teachers	Teachers	Teachers	Teachers	Teachers	Teachers	Teachers
Range of Outcome Variable	[0, 30]	[0, 14]	[0, 20]	[0, 8]	[0, 30]	[0, 13]	[0, 15]	[0, 10]
Control School Mean	2.75	1.59	1.93	1.12	2.31	1.04	1.63	0.59
Clusters	24	24	24	24	24	24	24	24
Observations	168	168	168	168	168	168	168	168
Estimator	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit	Tobit

Analysis Notes: Standard errors in open parentheses are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report the Benjamini Hochberg (BH) at the 5% level within hypothesis. Hypothesis categories are delineated in the top row of the table. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Variable Descriptions: Variables constructed using teacher survey. In the case that a teacher indicated that the school has a school plot, the teacher was then asked "In the the current term (term 3) or previous term (term 2) how many days total have you worked on this land for: clearing, planting, weeding, and harvesting (or other)". The enumerator proceeded in step-wise fashion to ask about the number of days spent for each activity. The response for this exercise is analyzed in the first four columns of this table for each activity, respectively. Then, the same question was repeated but the teacher was asked "how many days have you worked on this land WITH YOUR PUPILS." The response for this latter exercise is analyzed in the last four columns of the table. We display this result graphically in Figure **??**.

Table G.8: Teacher Survey: Working on Land at School

Category:		**Conditional on Not Knowing**	
Outcome Variable:	Question	Inquisitive	Ignore
Treatment (*ITT*)	−0.06	0.10**	−0.06**
	(0.06)	(0.04)	(0.03)
H_0 : *ITT* = 0			
p value	[0.31]	[0.01]±±	[0.03]±±
BH Critical p value (5%)	[0.05]	[0.02]	[0.03]
Pair FE	Yes	Yes	Yes
Enum FE	Yes	Yes	Yes
Source of Data	Teacher Survey	Teacher Survey	Teacher Survey
Unit of Observation	Teachers	Teachers	Teachers
Range of Outcome Variable	{0,1}	{0,1}	{0,1}
Control School Mean	0.61	0.82	0.08
Clusters	29	29	29
Observations	230	144	144
Estimator	OLS	OLS	OLS

Analysis Notes: Standard errors in open parentheses are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report the Benjamini Hochberg (BH) at the 5% level within hypothesis. Hypothesis categories are delineated in the top row of the table. ± suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; ±± suggests a significant discovery at the 5% level. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Variable Descriptions: Variables constructed using teacher survey. "Question" indicates that a teacher responded yes (equal to 1) if they were asked a question they did not know the answer to (0 otherwise). Conditional on being asked such a question, the teacher is given an opportunity to state the most common way they respond. The enumerator codes responses according to whether the teacher says she 1) "Ignores the student and do not respond," 2) "Tell him that you do not know and will do research to find the right answer," 3) "Suggest to him or her how to investigate the answer on his or her own," 4) "Provide the best response to the student you can provide," or 5) "Other." We code the second and third responses as "Inquisitive" and set the dependent variable equal to one if the teacher provided either one of these as a response and zero otherwise. We code the first and fourth responses as "Ignore" suggesting that the teacher effectively ignores the students question. Again, the dependent variable in this column is equal to one if the teacher mentions response number one or four and zero otherwise."

Table G.9: Teacher Survey: Teacher Inquisitiveness

Source of Data:	**Science Show**										
Category:	**Experiment**			**Hypothesis**		**Measurement**			**Articulating**		**Average**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Treatment	1.72***	0.73	0.10	1.99***	2.09***	0.87***	1.41***	2.12***	2.16***	2.06***	1.36***
	(0.50)	(0.44)	(0.28)	(0.65)	(0.52)	(0.18)	(0.34)	(0.55)	(0.72)	(0.63)	(0.37)
Teacher Outcome											
Inquisitive	−1.59**	−1.36**	−0.78**	−2.11**	−1.92**	−1.47***	−1.46**	−1.74**	−1.69*	−1.97**	−1.68***
	(0.68)	(0.60)	(0.36)	(0.87)	(0.75)	(0.37)	(0.53)	(0.77)	(0.89)	(0.80)	(0.52)
Treatment × Inquisitive	1.59**	1.24*	0.86**	2.29**	2.20***	1.15***	1.27**	1.13	1.30	2.09**	1.64***
	(0.75)	(0.66)	(0.39)	(0.89)	(0.70)	(0.39)	(0.51)	(0.79)	(0.90)	(0.85)	(0.54)
Enum FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pair FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Source of Data	Science Show	Science Show	Science Show	Science Show	Science Show	Science Show	Science Show	Science Show	Science Show	Science Show	Science Show
Unit of Observation	Student Group	Student Group	Student Group	Student Group	Student Group	Student Group	Student Group	Student Group	Student Group	Student Group	Student Group
Standardized Variable	No	No	No	No	No	No	No	No	No	No	No
Range of Outcome Variable	{1,2,...,10}	{1,2,...,10}	{1,2,...,8}	{1,2,...,10}	{1,2,...,10}	{1,2,...,10}	{1,2,...,10}	{1,2,...,10}	{1,2,...,10}	{1,2,...,10}	{1,2,...,9}
Control School Mean	3.81	2.18	1.8	2.70	2.4	1.65	2.16	2.92	3.2	2.65	2.63
Clusters	29	29	29	29	29	29	29	29	29	29	29
Observations	158	158	158	158	158	158	158	158	158	158	158
Estimator	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS

Analysis Notes: Standard errors in open parentheses are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. Hypothesis categories are delineated in the top row of the table. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Variable Descriptions: Variables in this table reflect scores provided by judges to student groups during the course of the science show. Column (1) reflects the clarity of an experiment; (2) the creativity of an experiment; (3) the uniqueness of an experiment. Column (4) measures the degree to which students described their hypotheses clearly and (5) measures the degree to which the analysis of their hypothesis is congruent with their experiment. Column (6) measures the degree to which students kept records of their observations (7) the degree of accurate measurement and (8) sufficient use of resources. Column (9) measures whether students articulated their thoughts independently (without a teacher) and (10) measures whether they articulated their thoughts accurately (they understood what they were saying). Column (11) averages across all other columns.

Teacher Outcome: We present teacher outcome interacted with treatment. "Further Research" is our measure of teacher inquisitiveness. We ask teachers whether students have ever asked them a question they don't know the answer to. If they respond "Yes," we further ask them how they respond when a student asks such a question. We code teacher responses according to whether they 1) Ignore the student and do not respond; 2) Tell him that you do not know and will do research to find the right answer; 3) Suggest to him or her how to investigate the answer on his or her own; or 4) Provide the best response to the student you can provide. "Further Research" averages across all teachers in a school who responded using category 2) or 3) in the follow up to the initial question. We interact each student's response by the within-school teacher average for this variable.

Table G.10: Science Show: Heterogeneity by Teacher Inquisitiveness

Category:	**Horizontal Treatment of & Attitudes Towards Students**			**Average**
Outcome Variable:	Free to Disagree	Decide	Learn from Pupils	Horizontal
Treatment (*ITT*)	0.50**	0.20	0.76***	0.12***
	(0.23)	(0.27)	(0.27)	(0.04)
H_0 : *ITT* = 0				
p value	[0.03]±	[0.45]	[0.01]±±	[0.01]±±
BH Critical p value (5%)	[0.02]	[0.05]	[0.03]	[0.05]
Pair FE	Yes	Yes	Yes	Yes
Enum FE	No	No	Yes	Yes
Source of Data	Teacher Survey	Teacher Survey	Teacher Survey	Teacher Survey
Unit of Observation	Teachers	Teachers	Teachers	Teachers
Range of Outcome Variable	{0,1}	{0,1}	{0,1}	[0,1]
Control School Mean	0.68	0.24	0.59	0.50
Clusters	29	29	29	29
Observations	230	230	230	230
Estimator	Logit	Logit	Logit	OLS

Analysis Notes: Standard errors in open parentheses are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report the Benjamini Hochberg (BH) at the 5% level within hypothesis. Hypothesis categories are delineated in the top row of the table. ± suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; ±± suggests a significant discovery at the 5% level. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Variable Descriptions: Variables constructed using teacher survey. All are binary variables equal to one if the teacher indicates "yes" and zero otherwise. For each column variable, the questions we ask teachers are as follows: Free to Disagree → Are pupils free to disagree with the concepts you are teaching them? Decide → Are pupils free to decide how to spend time during a lesson? "Learn from pupils" indicates whether a teacher is more likely to strongly agree with the statement "I learn as much from my pupils as they learn from me." Equal to 1 if the teacher chooses "Strongly Agree" and zero otherwise (choices include: Strongly Disagree, Somewhat Disagree, Neither Agree nor Disagree, and Somewhat Agree). We omit enumerator fixed effects in the first two columns due to insufficient variation.

Table G.11: Teacher Survey: Classroom Practices Demonstrate More Horizontal Relationships with Students

Category:	**Gentle Response**		**Harsh Response**		**Ordinal Difference**
Outcome Variable:	Ordinal	Binary	Ordinal	Binary	Gentle-Harsh
Treatment (*ITT*)	0.51^{**}	0.56^{***}	-0.37^{**}	-0.43^{*}	0.53^{***}
	(0.24)	(0.20)	(0.16)	(0.22)	(0.18)
$\mathbf{H_0: ITT = 0}$					
p value	[0.04]	$[0.01]^{\pm\pm}$	$[0.02]^{\pm\pm}$	$[0.06]^{\pm}$	$[0.00]^{\pm\pm}$
BH Critical p value (5%)	[0.01]	[0.02]	[0.03]	[0.04]	[0.05]
Pair FE	Yes	Yes	Yes	Yes	Yes
Enum FE	No	No	No	No	No
Source of Data	Teacher Survey	Teacher Survey	Teacher Survey	Teacher Survey	Teacher Survey
Unit of Observation	Teachers	Teachers	Teachers	Teachers	Teachers
Range of Outcome Variable	{0,1,2}	{0,1}	{0,1,2,3}	{0,1}	{-3,-2,. . . ,2}
Control School Mean	0.60	0.43	0.46	0.28	0.13
Clusters	29	29	29	29	29
Observations	230	230	230	230	230
Estimator	Ologit	Logit	Ologit	Logit	Ologit

Analysis Notes: Standard errors in open parentheses are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report the Benjamini Hochberg (BH) at the 5% level within hypothesis. Hypothesis categories are delineated in the top row of the table. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

Variable Descriptions: Variables constructed using teacher survey. We ask the teacher "When pupils misbehave in the classroom, what are the different techniques you use to help make sure pupils listen to you and improve their behavior?" The enumerator allows the teacher to respond unmprompted and then classifies responses according to whether the teacher 1) makes efforts to understand why the poor behavior is taking place, 2) distract the pupils in a manner that will bring attention back, 3) remind them to pay attention or become serious, 4) separate them from the rest of the class, 5) hit or beat them, 6) Shout at them, 7) punishment, but not beating, or 8) other. We code responses of 1) or 2) as "gentle" responses and 5, 6, and 7 as "harsh" responses. Ordinal measures sum sums over the total number of "gentle" or "harsh" responses provided by the teacher. Binary measures equal 1 if **only** gentle or harsh responses were provided by a teacher and zero otherwise. Differences between gentle and harsh responses reflect differences in the ordinal measures. We omit enumerator fixed effects from this table due to insuficient variation within enumerator for some outcomes.

Table G.12: Teacher Survey: Teacher Response to Poor Pupil Behavior

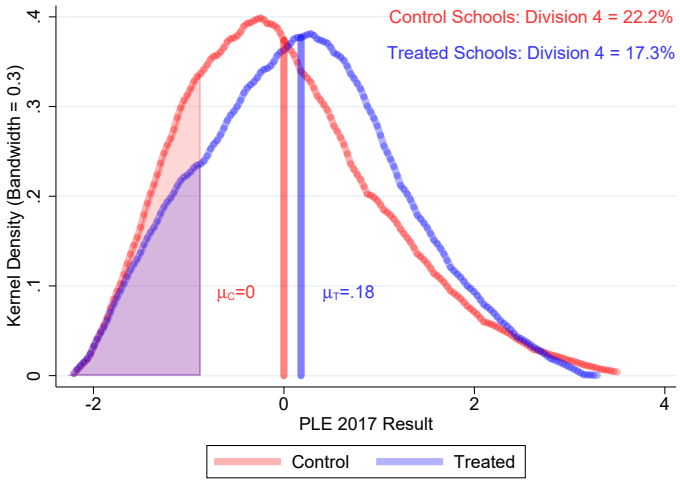
Category:	**Hunger at School**			
Outcome Variable:	Ate yesterday?	I am usually hungry during school	I am usually NOT hungry during school	Difference
Treatment (*ITT*)	0.17^{***}	-0.05	0.05	0.10^{**}
	(0.05)	(0.03)	(0.04)	(0.05)
$\mathbf{H_0}$: $\boldsymbol{ITT} = \mathbf{0}$				
p value	$[0.00]^{\pm\pm}$	[0.11]	[0.16]	$[0.04]^{\pm\pm}$
BH Critical p value (5%)	[0.01]	[0.03]	[0.04]	[0.05]
Pair FE	Yes	Yes	Yes	Yes
Enum FE	Yes	Yes	Yes	Yes
Source of Data	Student Survey	Student Survey	Student Survey	Student Survey
Unit of Observation	Students	Students	Students	Students
Range of Outcome Variable	{0, 1}	{0, 1}	{0, 1}	{-1, 0, 1}
Control School Mean	0.69	0.16	0.07	-0.08
Clusters	29	29	29	29
Observations	329	329	329	329
Estimator	OLS	OLS	OLS	OLS

Analysis Notes: Standard errors in open parentheses are clustered at the school level. *, reflects a coefficient p value from the original specification, "*p value*," less than 0.1, ** less than 0.05 and *** less than 0.01. Coefficients represent the Intent to Treat effect. We report the Benjamini Hochberg (BH) at the 5% level within hypothesis. Hypothesis categories are delineated in the top row of the table. $^{\pm}$ suggests a significant discovery, accounting for multiple hypothesis tests, at the 10% level; $^{\pm\pm}$ suggests a significant discovery at the 5% level. These results should be treated as preliminary since estimation took place in October 2020 and only includes schools from Budondo sub-county.

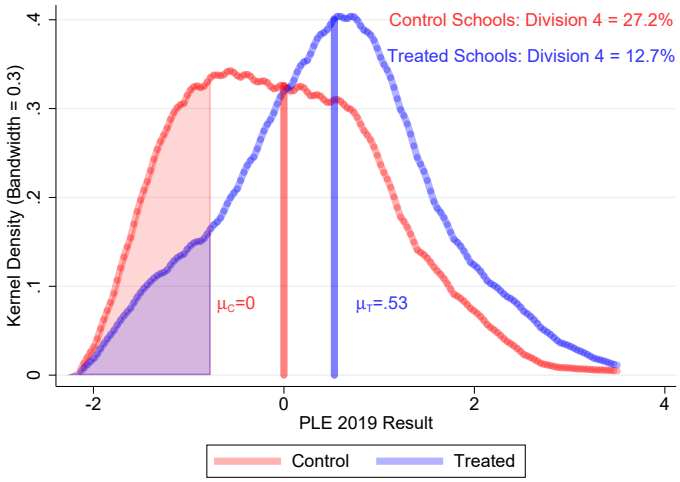
Variable Descriptions: Variables constructed using student survey. For the first variable, the student was asked "Did you eat anything at school yesterday (or Friday if today is Monday)?" A response of yes was coded as 1 and a response of no was coded as 0. For the second and third variables, the student was asked "How strongly do you agree with the following: I am usually hungry during the school day." Students who said "Strongly Agree" are coded as 1 in column two and zero otherwise. Students who said "Strongly Disagree" are coded as 1 in column three and zero otherwise. We take the difference of these two variables in column four: column three minus column two.

Table G.13: Student Survey: Eating at School

H Additional Appendix Figures



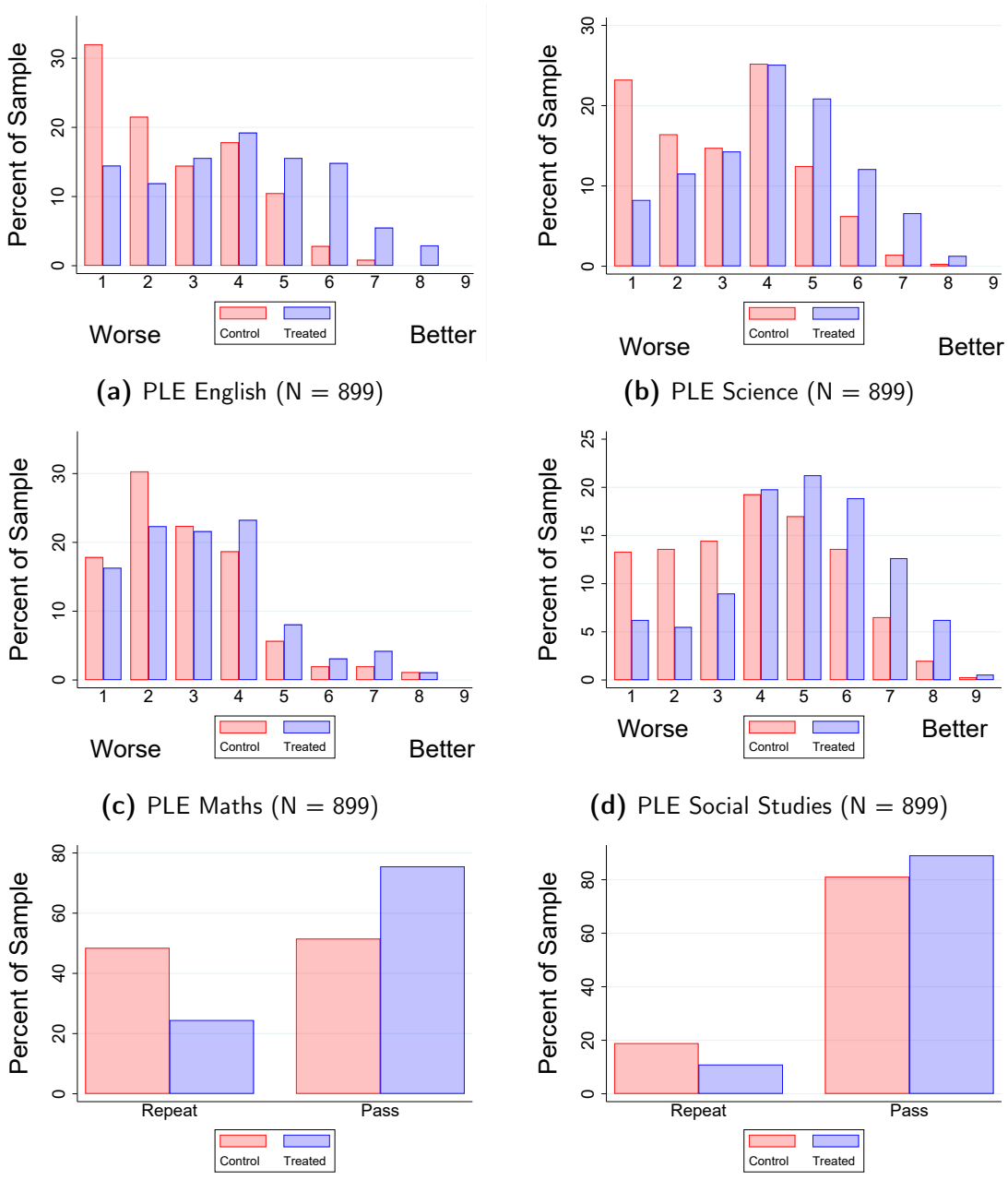
(a) 2017



(b) 2019

Notes: These graphs present kernel densities of the aggregate PLE distribution across 24 schools only. We were unable to procure student registration numbers for the PLE exam in 2017 from all of the schools in our study. Therefore, this graph shows PLE score distributions for the 24 schools in which every pair of schools from our pair-wise matching procedure provided PLE registration numbers for all P7 candidates in both 2017 and 2019. Distributions are standardized according to the control school mean and standard deviation within year. Larger values reflect higher performance. The shaded regions reflect the portion of the distribution that did not pass the division 4 threshold, which is equivalent to failing the PLE. The percentages associated with these shaded areas are printed on the body of the graph. Symbols μ_C and μ_T correspond to the year-specific mean of the control and treatment school PLE outcomes, respectively. Notice that $\mu_C = 0$ due to the standardization of PLE scores. Therefore, μ_T reflects the difference between control school and treatment school averages according to the standard deviation of control school PLE score distributions in any given year.

Figure H.1: Distribution of PLE Outcomes Across Treatment and Control Schools



(a) PLE English (N = 899)

(b) PLE Science (N = 899)

(c) PLE Maths (N = 899)

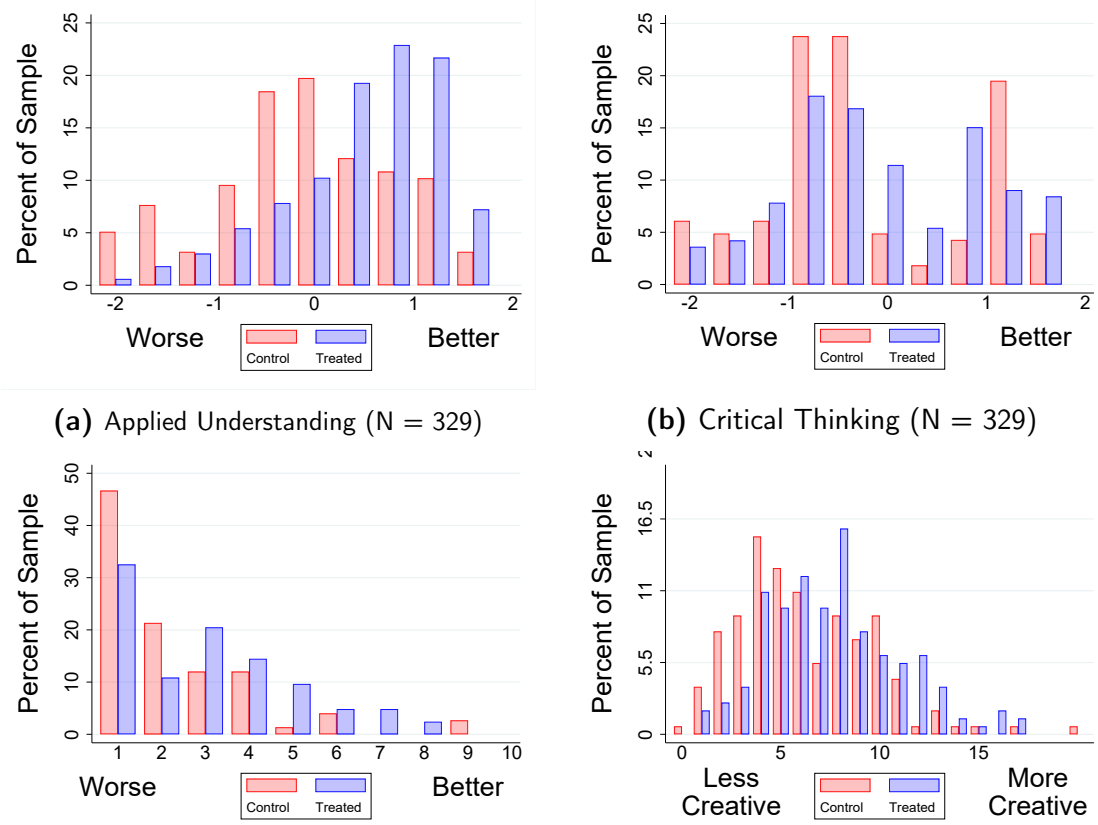
(d) PLE Social Studies (N = 899)

(e) PLE Pass (N = 964)

(f) P6 Passthrough 2018-19 (N = 329)

Notes: All variables in this figure have discrete ranges. Variables in sub-figures H.2a through H.2e are sourced from official results of the 2019 P7 Primary Leaving Examination for schools in the study sample, procurred from Uganda's National Bureau of Education (UNEB). We transform variables to ease interpetation such that larger scores reflect better performance. The variable represented by sub-figure H.2f reflects P6 students' responses when asked to confirm the class they were in during the 2018 school year. If students state that they were in P6 in 2018 then we code them as "repeat" students in 2019.

Figure H.2: Histograms of Outcome Variables in Table 3 Prior to Standardizing



(a) Applied Understanding (N = 329)

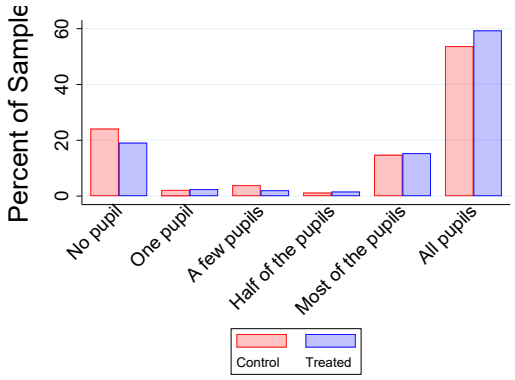
(b) Critical Thinking (N = 329)

(c) Science Show (Mean) (N = 158)

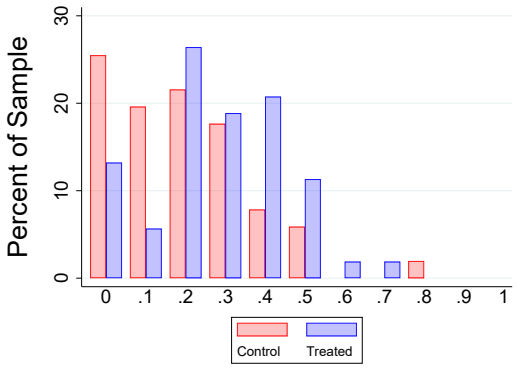
(d) Creativity Index (N = 329)

Notes: Variables represented by sub-figures H.3a through H.3c have, in principle, continuous ranges. Variables in sub-figures H.3a and H.3b were constructed using principal components analysis as detailed on pages 21 and 22 of the pre-analysis plan. Thereafter, the variables are standardized according to the control school mean and standard deviation. Sub-figure H.3c averages across all measures captured by science show judges and documented in table 1 of the pre-analysis plan. The variable in sub-figure H.3d has a discrete range and the procedure we use for constructing it is described on pages 26-27 of the pre-analysis plan.

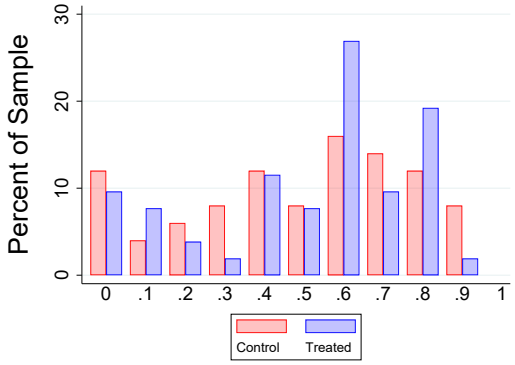
Figure H.3: Histograms of Outcome Variables in Table 4



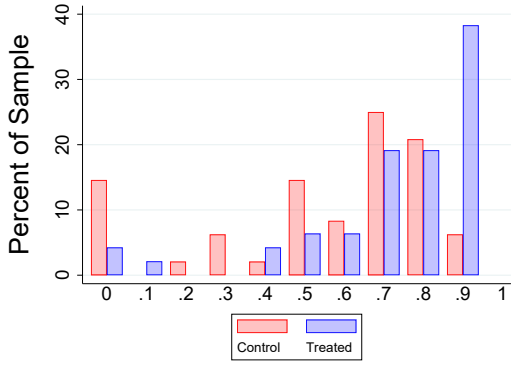
(a) Share of Engaged Pupils (N = 2,380)



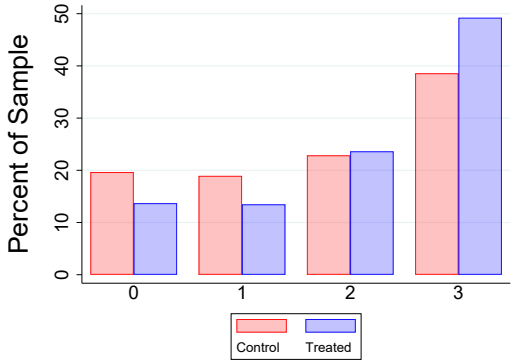
(b) Student Asks Questions (N = 104)



(c) Corporal Punishment (N = 104)



(d) Knowledge of Student (N = 104)



(e) Teacher Network (N = 1,466)

Notes: Variables in figures H.4a and H.4e have discrete ranges. Variables in remaining sub-figures are, in principle, continuous and are constructed using the student survey (and the teacher survey in the case of figure H.4d). Students respond to questions regarding specific teachers for variables reflected by figures H.4b and H.4c — we then construct a teacher-level measure for each question by summing over the students who responded in the affirmative for each teacher and dividing by the total number of students surveyed. The variable reflected in figure H.4d measures the percent of correct responses provided by P6 teachers' regarding different aspects of specific students' lives. The variable represented by H.4e sums over three dimensions of teacher interactions reflecting the degree of within-school teacher collaboration and communication.

Figure H.4: Histograms of Outcome Variables in Table 5

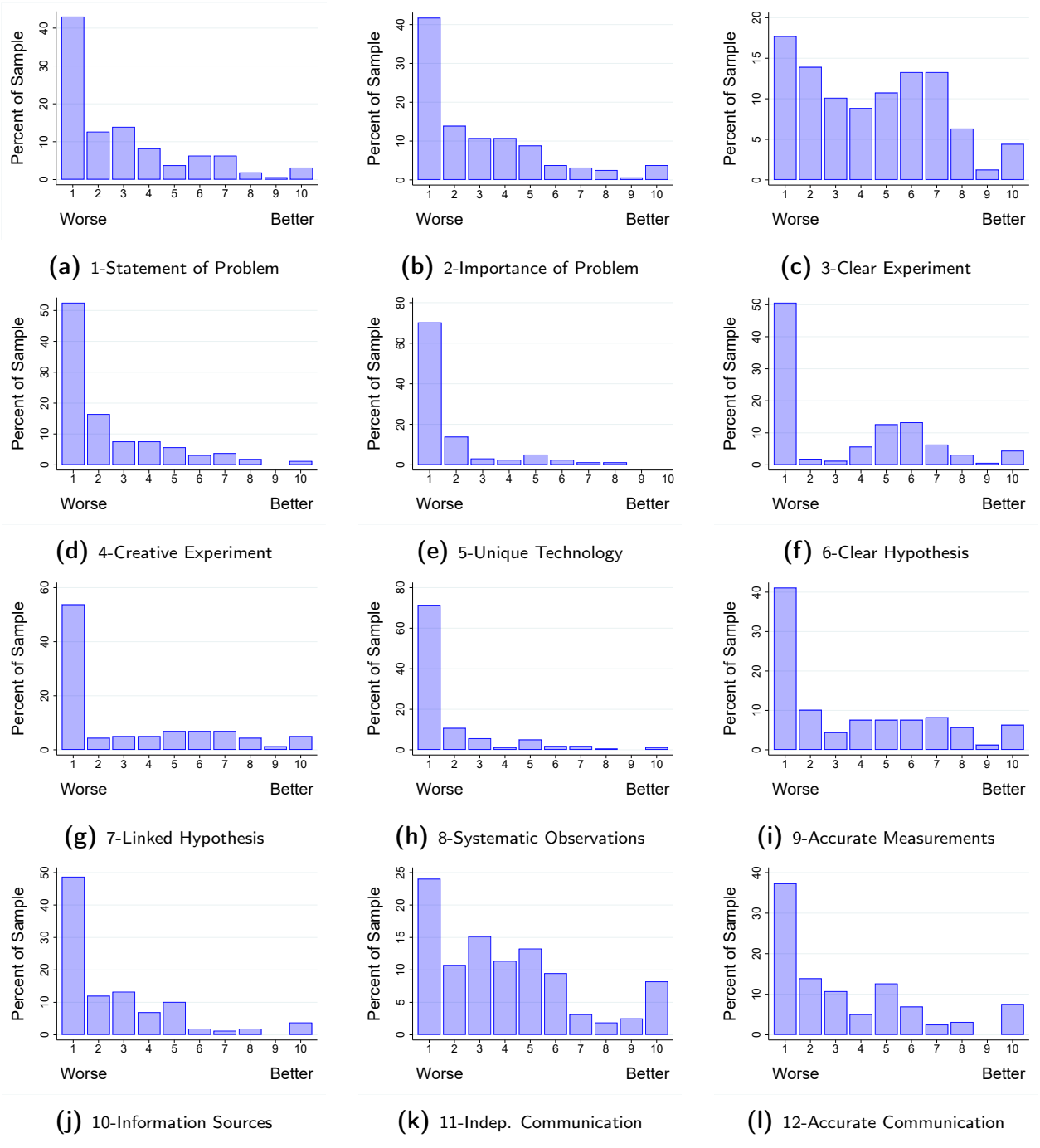
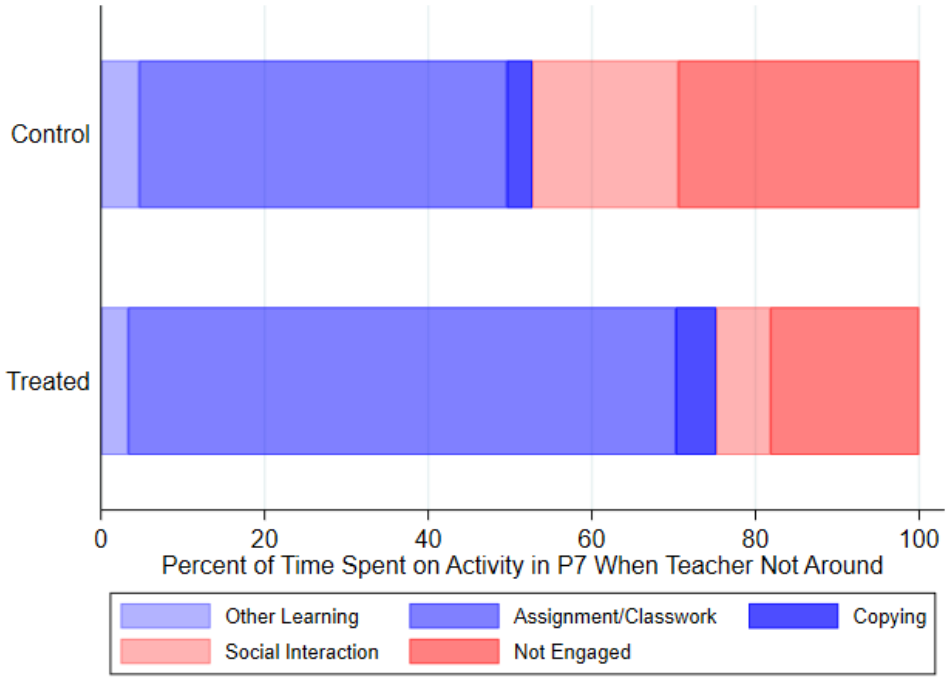
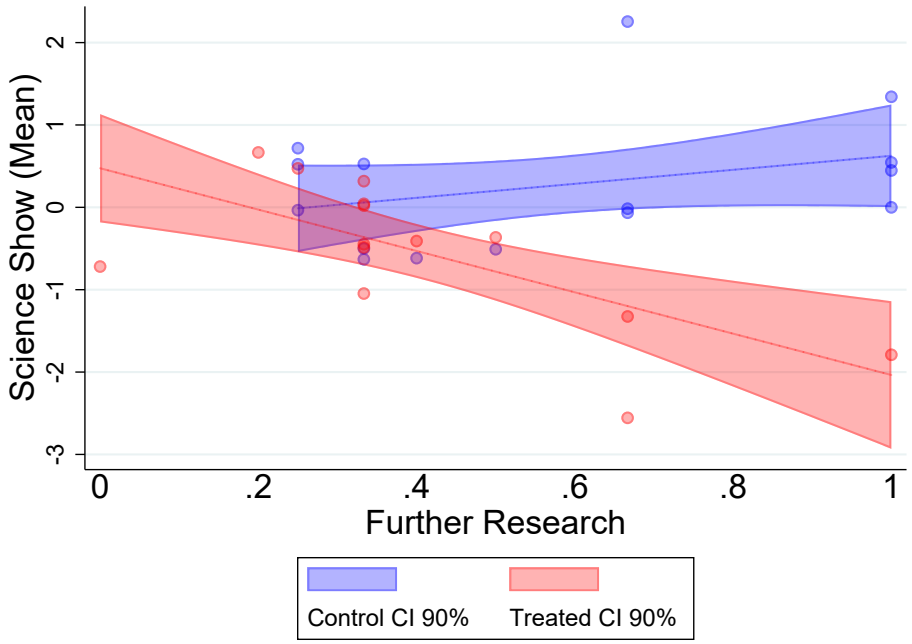


Figure H.5: Histograms of Granular Science Show Outcome Variables in Table 6



Notes: We observe classrooms in P7 when no students are engaged in an activity with the teacher. The observer categorizes the activity the students are engaged in according to whether they are engaged in learning activities (blue) or they distracted (red).

Figure H.6: Activities Observed in P7 Classrooms When Teacher is Not Around



(a) Science Show (Mean)

Y Axis Notes: Variable on y axis is a "residual" school-level average science show outcome. It is "residual" in the sense that the y axis averages the residual term after regressing the student outcome by school-pair and enumerator fixed effects.

X Axis Notes: The variable on the x axis averages teacher outcomes within school. "Further Research" is our measure of teacher inquisitiveness within school. We ask teachers whether students have ever asked them a question they don't know the answer to. If they respond "Yes," we further ask them how they respond when a student asks such a question. We code teacher responses according to whether they 1) Ignore the student and do not respond; 2) Tell him that you do not know and will do research to find the right answer; 3) Suggest to him or her how to investigate the answer on his or her own; or 4) Provide the best response to the student you can provide. "Further Research" averages across all teachers in a school who responded using category 2) or 3) in the follow up to the initial question. We interact each student's response by the within-school teacher average for this variable, using only the school-level average as an observation.

Graph Notes: We graph the scatter plot of science show group outcomes against teacher outcomes, analyzing only one observation per school by taking the average across group outcomes. We further graph the line of linear best fit within treatment status with a confidence interval of 90%.

Figure H.7: Science Show Heterogeneity by Teacher Inquisitiveness in Treatment vs. Control Schools