MATHEMATICA
CENTER FOR
INTERNATIONAL POLICY
RESEARCH AND EVALUATION

REPORT

REVISED REPORT

Evaluation Interim Report for the Georgia II Improving General Education Quality Project's School Rehabilitation and Training Activities

October 7, 2019

Ira Nichols-Barrer
Nicholas Ingwersen
Camila Fernandez
Elena Moroz
Matt Sloan

Submitted to:
Millennium Challenge Corporation
1099 Fourteenth St., NW, Suite 700
Washington, DC, 20005
Contract Number: MCC-13-BPA-0040 (CL-002)

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Matt Sloan

Reference Number: 40306.500

This page has been left blank for double-sided copying.

ACKNOWLEDGEMENTS

This report reflects the contributions of many people. From the Millennium Challenge Corporation, Jenny Heintz and Ryan Moore, the current and former technical officers for this evaluation, as well as Jenner Edelman, Sonia Shahrigan, and Marina Kutateladze, helped guide and support us throughout the project. This study would not have been possible without the contributions of many partners in the Government of Georgia and Millennium Challenge Account-Georgia. We would first like to acknowledge the wide range of implementers and coordinators from the Millennium Challenge Account-Georgia who generously shared their time and attention to help improve the quality, comprehensiveness, and depth of the study. We are grateful to Government of Georgia staff at the Ministry of Education, the Teacher Professional Development Center, and the Educational and Scientific Infrastructure Development Agency for generously sharing their time and expertise and providing important information about the project's expected outcomes, implementation, and activities. We also received indispensable support and advice from the staff of the Millennium Challenge Account-Georgia office, especially Zura Simonia, who managed the evaluation's data collector and has provided both substantive and technical expertise at every stage of the study, Nino Udzilauri, who guided evaluation planning discussions and facilitated a wide range of data collection efforts for the project's teacher and school director training activities, and Kartlos Kipiani, who generously shared planning and implementation details about the school rehabilitation activity.

This report depended on contributions from many data collection, supervisory, and support staff. We are grateful to the staff of the Institute for Polling and Marketing and the National Assessment and Examinations Centre for the successful implementation of the nationwide survey and qualitative data collection effort (in the case of the Institute) and the development and administration of student learning assessments (in the case of Centre). We also want to thank the many people who responded to our surveys and participated in in-depth interviews and focus groups. Many components of the study's data collection also would not have been possible without the deep contributions of Natia Gorgadze, who supervised data collection activities as a locally based member of Mathematica's evaluation team. Leigh Linden provided extensive analytical support throughout the design stage of this study, and, at Mathematica, Steve Glazerman and Larissa Campuzano provided technical input and useful comments on the analysis plan and draft report. We would also like to thank the editorial and administrative support staff at Mathematica.

Mathematica strives to improve public well-being by bringing the highest standards of quality, objectivity, and excellence to bear when collecting information and performing analysis for our clients. The findings in this report solely reflect Mathematica's interpretation of available information. Mathematica staff involved in analyzing the information and authoring this report did not report any conflicts of interest. The evaluation was funded exclusively by the Millennium Challenge Corporation.

This page has been left blank for double-sided copying.

CONTENTS

EXECUTIVE SUMMARY xiii

I. INTRODUCTION 1

A. Overview of evaluated activities 1

B. Literature review 4

1. Prior evidence on school rehabilitation 4

2. Prior evidence on training teachers and school directors 6

C. Impact evaluation design for the ILEI activity 8

1. Process evaluation examining program implementation and costs 9

2. Impact evaluation applying a randomized controlled trial (RCT) design 9

3. In-depth qualitative research on the effects of school rehabilitation 11

4. ILEI study sample and power calculations 11

5. ILEI evaluation timeframe 13

D. Evaluation design for the TEE activity 15

1. Performance evaluation describing program implementation and outcomes 16

2. Descriptive evaluation of teacher training, applying a matched comparison group design 18

3. TEE study population and evaluation sample 19

4. TEE evaluation time frame 20

E. Objectives of the interim report 20

II. DATA COLLECTION AND ANALYSIS APPROACH 23

A. ILEI interim study data and methods 23

1. Quantitative surveys and administrative data 23

2. Qualitative data collection 24

3. Analysis approach 25

B. TEE interim study data and methods 27

1. Quantitative surveys 27

2. Qualitative data collection 28

3. Stallings classroom observation 30

4. Analysis approach 31

III. INTERIM FINDINGS FOR THE ILEI EVALUATION 35
A. School rehabilitation program context 35
B. Physical infrastructure changes at rehabilitated schools and their perceived benefits 36
C. Changes in instructional time, facility use, and perceptions of school safety 48
1. Potential effects of infrastructure improvements on instructional time 48
2. Use of recreational facilities 53
3. Use of science laboratories 53
4. Perceptions of school safety 55
D. Changes in enrollment and school administration 56
1. Changes in enrollment 57
2. School operations and maintenance 59
IV. INTERIM FINDINGS FOR THE TEE EVALUATION 63
A. Implementation of the TEE training initiative 63
B. Training attendance and completion 64
1. Survey data 64
2. Findings from qualitative focus groups about training attendance and completion 67
C. Teacher knowledge and practices after training 67
1. First year follow-up survey data 68
2. Validating teacher survey responses 73
3. Potential impacts of the TEE training sequence on teachers 75
4. School director perceptions about the effects of training on teachers 79
5. Qualitative findings about the effects of training on teachers 80
D. School director knowledge and practices after training 84
1. School director survey data 84
2. Qualitative data from school directors and professional development facilitators 89
V. CONCLUSION 94
REFERENCES 96
APPENDIX A CONSTRUCTION OF OUTCOME INDICES 1
APPENDIX B SUMMARY OF QUALITATIVE FINDINGS FOR THE ILEI STUDY 1
APPENDIX C SUMMARY OF QUALITATIVE FINDINGS FOR THE TEE STUDY 1
APPENDIX D TEE SURVEY TRENDS FOR COHORT 1 TEACHERS 1

APPENDIX E TEE MATCHED COMPARISON GROUP ANALYSIS FOR TEACHERS COMPLETING ALL TRAINING MODULES 1

APPENDIX F TEE SUBGROUP ANALYSES 1

APPENDIX G TEE MATCHED COMPARISON GROUP ANALYSIS FOR INFREQUENT TEACHING PRACTICES 1

APPENDIX H STAKEHOLDER COMMENTS AND MATHEMATICA RESPONSES 1

This page has been left blank for double-sided copying.

TABLES

I.1 Evaluation questions for the ILEI activity and approaches to answering them ... 8
I.2 Regional rollout of the ILEI activity ... 10
I.3 ILEI MDEs for different sample sizes and compliance rates ... 13
I.4 ILEI evaluation data collection schedule ... 14
I.5 Evaluation questions for the TEE activity and approaches to answering them ... 15
I.6 TEE data collection schedule ... 20
II.1 Baseline and follow-up data collection samples in Phase I regions ... 24
II.2 ILEI qualitative data collection sample in 10 schools ... 24
II.3 TEE survey data collection samples ... 28
II.4 Description of the qualitative data collection protocols, by respondent (spring–fall 2018) ... 29
II.5 Equivalence of characteristics between teachers in first and second cohorts for full and matched analytic samples ... 32
III.1 Summary baseline characteristics of treatment schools ... 36
III.2 Comparison of infrastructure and teaching facilities in rehabilitated schools between baseline and one-year follow-up ... 37
III.3 Comparison of presence and perceptions of central heating between baseline and one-year follow-up ... 40
III.4 Comparison of quality of lighting and its effect on the learning environment at baseline and one-year follow-up ... 44
III.5 Student use of recreational school facilities ... 53
III.6 Comparison of perceptions of school safety between baseline and one-year follow-up ... 56
III.7 Average annual changes in student enrollment in schools rehabilitated in 2016 ... 58
III.8 Change in costs incurred by rehabilitated schools between baseline and one-year follow-up ... 59
IV.1 TEE activity participants and training schedule ... 63
IV.2 Teacher attendance rates in TEE training modules ... 65
IV.3 Reasons for attending or missing TEE training modules ... 66
IV.4 Correlations between the Stallings observations and related TEE teacher survey responses ... 74
IV.5 Classroom practices reported by students ... 75
IV.6 Matched comparison group analysis for practitioner teachers ... 77

IV.7 School director practices related to instructional leadership, as reported in the second follow-up survey 84

IV.8 Prevalence of classroom observations by school directors 85

IV.9 School director monitoring of teaching practices 85

IV.10 School director support for teacher professional development 86

IV.11 School director practices related to instructional inclusion 86

IV.12 Teacher reports of the instructional leadership provided by school directors in year after school director training completed 89

FIGURES

I.1 The IGEQ program logic 3

III.1 Illustration of classroom rehabilitation 38

III.2 Percentage of rehabilitated schools at baseline and one-year follow-up with ceiling or floor problems in at least one classroom 38

III.3 Percentage of rehabilitated schools at baseline and one-year follow-up experiencing problems with classroom walls 39

III.4 Student perception of classroom air quality in winter at baseline and one-year follow-up 42

III.5 Perceived effect of classroom air quality in winter on the learning environment at baseline and one-year follow-up 43

III.6 Presence of flush toilets in primary sanitary facility at baseline and one-year follow-up 45

III.7 Sanitary conditions in primary sanitary facility at baseline and one-year follow-up 46

III.8 Student comfort using sanitary facilities in school at baseline and one-year follow-up 47

III.9 Student comfort using rehabilitated sanitary facilities, by gender 48

III.10 Baseline and follow-up student absence patterns 50

III.11 Class time spent on instruction per day in the month before the baseline and one-year follow-up surveys 52

III.12 Comparison of exposure to science laboratories and demonstrations and conducting science experiments at baseline and one-year follow-up 54

III.13 Average annual student enrollment in schools rehabilitated in 2016 58

III.14 Percentage of school directors able to fully pay for school utilities after rehabilitation 60

III.15 Highest spending priority for school directors facing budget shortfalls 61

IV.1 Teaching practices related to students' critical thinking, motivation, and collaboration, as reported in the first follow-up survey 69

IV.2 Teaching practices related to tailored learning and assessing student learning, as reported in the first follow-up survey 70

IV.3 Teaching practices related to inclusion of female and minority students and ICT use in instruction, as reported in the first follow-up survey 71

IV.4 Practices related to teachers' professional development, as reported in the first follow-up survey 71

IV.5 Teaching practices related to science lessons and mathematics lessons, as reported in the first follow-up survey 72

IV.6 Teaching practices related to English lessons and geography lessons, as reported in the first follow-up survey 73

IV.7 School director perceptions of changes in teacher practices after first round of teacher training 80

IV.8 Change in time school directors reported spending on practices related to instruction or professional development after first year of training 87

IV.9 Change in time school directors reported spending on practices related to school management after first round of training 88

EXECUTIVE SUMMARY

Many of the buildings in the Georgia's public school system are inadequately maintained, dilapidated, and uncomfortable for students and teachers, particularly during the winter months. In addition, teachers and school directors in Georgia often lack access to professional development opportunities that encourage high quality instructional practices, instructional leadership, and school management. To address these issues, the Millennium Challenge Corporation (MCC) is supporting Georgia's efforts to improve educational outcomes for its students by sponsoring the Improving General Education Quality Project. The Project includes an Improved Learning Environment Infrastructure (ILEI) activity investing in school rehabilitation and a Training Educators for Excellence (TEE) activity supporting professional development by training and mentoring teachers and school directors.

This interim report describes a preliminary set of evaluation findings on the school rehabilitation activity and the teacher and school director training activity. The results presented here are intended to provide early evidence about the potential effects of project activities before the Compact comes to a close in July 2019. An endline report in 2021 will present a longer-term follow-up and examine the effects of the school rehabilitation activity through a randomized controlled trial.

Research design for the school rehabilitation activity

The ILEI activity focused on improving the learning environment in a targeted group of schools experiencing major building infrastructure problems. The program sought to improve heating, lighting, water and sanitation, recreational facilities, science laboratory facilities, and classroom conditions, with the goal of increasing students' time on task in school and ultimately improving student learning and educational attainment outcomes. In total, the program is seeking to rehabilitate up to 96 schools.

This interim report provides early evidence about the preliminary outcomes observed in the first 29 rehabilitated schools, which were completed in 2016 or 2017. The evaluation of the ILEI activity uses a mixed-methods study design with three components: (1) a process evaluation examining the program's implementation and costs, (2) a randomized controlled trial impact evaluation using a school-level stratified random assignment design, and (3) an in-depth analysis of the relationship between changes in school infrastructure and changes in the learning environment using qualitative methods in a subset of study schools.

Our process evaluation of the ILEI activity aims to answer, among others, the following questions related to program design and implementation:

1. Was the ILEI activity budgeted and planned appropriately, forecasting key risks?
2. Did the ILEI activity deliver improved facilities?
3. How was the program rolled out?
4. How much did rehabilitation differ by school?

5. What is the current and future status of facility-maintenance funding for treatment and control schools?

To provide preliminary evidence about these questions for the interim report, Mathematica conducted site visits at 29 rehabilitated schools to assess whether infrastructure improvements were delivered as designed. Ultimately, the study's endline report will supplement these observational findings with a comprehensive review of program implementation records and cost data. The review, together with semi-structured interviews with program implementers and Government of Georgia staff, will examine implementation successes and challenges and identify lessons about the implementation process that could inform similar interventions in Georgia and in other contexts.

The study's impact evaluation and in-depth qualitative analyses aim to answer the following questions related to the program's effects on school infrastructure, teachers, and students:

1. What are the impacts of the ILEI activity on the school infrastructure environment, such as regulation of classroom temperature, maintenance policy, and maintenance practice?
2. Did the Activity affect perceptions of students' and teachers' health and safety?
3. What are the impacts of the ILEI activity on teacher behavior, such as attendance and time spent teaching?
4. What were the impacts of the ILEI activity on student outcomes, such as attendance, enrollment, dropout and retention rates, time spent studying, and learning outcomes?

To estimate the impacts of the school rehabilitation activity, our study uses a rigorous school-level random assignment design. However, at the time of this interim report an insufficient number of schools had been rehabilitated to conduct an impact analysis with adequate statistical power. To provide early evidence on these research questions, the interim analysis instead focused on examining pre-post changes between the baseline learning environment (conditions observed one or two years prior to rehabilitation) and the learning environment in the first year after rehabilitation was completed. The analysis is limited to the 29 schools that had been rehabilitated by December 2017. While this sample size and early timeframe is too limited to examine impacts on learning outcomes, the interim study provided an opportunity to investigate the initial perceptions of students, teachers, and school directors about the changes occurring in rehabilitated schools.

Interim findings for the school rehabilitation activity

There was a strong pattern of improvements in the conditions of rehabilitated schools. During site visits, the research team observed significant improvements in the condition of the school's exterior and the interior's hallways, flooring, and stairs. Schools also demonstrated large improvements in aspects of physical infrastructure related to heating, lighting, air quality, and teaching facilities. The infrastructure improvements in rehabilitated schools are readily visible, particularly in classrooms. At baseline, most of the schools had at least one classroom with two or more problematic conditions present (such as cracks, water damage, mold, chipped or peeling paint, or holes in ceilings and floors). But after rehabilitation, the percentage of

schools with two or more problems in at least one classroom dropped from 72 to 7 percent for ceilings and from 72 to 0 percent for floors.

Rehabilitation systematically improved the quality of heating systems, which appeared to improve students' concentration during winter months. At baseline, about half of all observed classrooms did not have functional central heating, but after rehabilitation, all of these schools had an operational central heating system. Installing central heating coincided with substantial improvements in the number of students, teachers, and parents who reported that classrooms felt too cold on average in February (decreasing from 41 to 6 percent for students, 28 to 1 percent for teachers, and 26 percent to 1 percent for parents). Similarly, in the baseline survey, 41 percent of students reported that classroom temperatures made it more difficult to concentrate during the winter, but only 19 percent of students in rehabilitated schools (all of which had central heating) reported that this was a concern. In qualitative interviews, students consistently reported that they had felt uncomfortable because of smoke from wood stoves inside the classrooms, and that it was a relief not to have to collect wood to keep the wood stove running. Teachers also shared that before rehabilitation was completed classrooms could get so cold that students sometimes felt unwell and wanted to go home, or that parents were reluctant to send their children to school because of their discomfort in cold classrooms.

However, new central heating systems significantly increased utility costs and strained schools' operating budgets. Compared with baseline heating expenses, school directors reported that in renovated schools, the cost of heating the building in winter roughly tripled. While increases in utility costs were an expected consequence of installing or improving new building systems (especially heating and electricity systems), thus far meeting these increased costs has represented a significant challenge for school directors. After rehabilitation, 55 percent of school directors reported that they are not able to fully pay for school utilities with the available school budget, and an additional 31 percent reported that they are only able to meet these expenses on a periodic basis.

Upgraded heating systems noticeably improved air quality in classrooms, although some air quality issues remained. Before rehabilitation, the use of wood-burning stoves during the winter often harmed air quality because of poorly sealed and ventilated chimneys. Student surveys suggest that winter air quality in many classrooms did improve following school rehabilitation, with the percentage of students reporting that air quality was poor declining from 26 to 9 percent. After rehabilitation, the most common rating from students (about half of respondents) was that air quality was "fair." This rating is consistent with direct air quality measurements the survey teams made during site visits that showed other sources of air pollutants unrelated to wood-stove heating systems (potentially including such items as dust, chipped paint, or outdoor pollution sources) were still present in rehabilitated schools. Nonetheless, evidence from the one-year follow-up surveys clearly suggests that poor air quality in winter months did not affect learning as severely after rehabilitation. At baseline, nearly a third of students reported that classroom air quality affected their ability to concentrate on school work in the past month (32 percent) or disrupted classroom instruction in February (28 percent), but by the one-year follow-up, the percentages had decreased by 15 and 19 percentage points, respectively. Teachers also reported large decreases in concerns about the effects of air quality on the learning environment.

Rehabilitation also improved lighting in rehabilitated schools. Improvements to electrical systems and lighting were intended to improve the quality of teaching and the ability of students to read and learn, particularly during the winter. At baseline, at least one classroom in 79 percent of the schools had no working electric lighting, and 63 percent of students reported having difficulty reading the blackboard because of poor lighting. By the one-year follow-up, the percentage of schools without any working lighting decreased by 59 percentage points, and the percentage of students who reported having difficulty reading the blackboard dropped by half after lighting was installed (from 63 to 31 percent). Similarly, the percentage of teachers reporting that lighting was inadequate for students fell from 29 to 4 percent.

The rehabilitation program also delivered significant improvements to the sanitary facilities at rehabilitated schools. At baseline, most schools (83 percent) did not have flush toilets in their primary sanitary facility. The rehabilitation package installed flush toilets at all of these schools, although some of them encountered maintenance issues in the first year after rehabilitation. In total, 72 percent of the rehabilitated schools had fully functional flush toilet facilities throughout the building (21 percent of schools had at least one flush toilet that was not functional, and the remaining 7 percent of schools still had at least one pit latrine). Teachers and students both reported large improvements in their degree of comfort using sanitary facilities in rehabilitated schools. At baseline, most students (61 percent) said that they were never comfortable using the sanitary facilities in their school, and only 11 percent reported that they were always comfortable. Following rehabilitation, the proportions of never comfortable and always comfortable responses essentially reversed (to 11 percent saying they were never comfortable and 63 percent saying they were always comfortable). We observed very similar survey response patterns for male and female students, but qualitative focus group data suggested that these improvements were particularly beneficial for girls. Students reported that the location of renovated sanitary facilities (inside the building versus outside previously), the privacy of the stalls (with doors versus without doors previously), the presence of flush toilets using running water, and the availability of sinks with running water for handwashing were critical improvements. Students also reported that renovations had eliminated prior situations in which female students would remain in discomfort during the school day or wait to leave school to find usable toilet facilities.

The study did not find evidence of large changes in absenteeism or school enrollment following rehabilitation. We did not find a strong pattern of changes in student absenteeism at rehabilitated schools, as measured by direct attendance counts by the research team, survey data from teachers, and survey data from school directors. Attendance counts conducted by the research team on the day of each site visit did not reveal any significant changes between the baseline and follow-up attendance figures at rehabilitated schools, but teachers and school directors both reported modest improvements. Teachers reported that the average percentage of students with perfect attendance records in the past month increased from 18 percent to 23 percent, and school directors reported a modest improvement in the average absence rate during the month of February (a decrease of 4.3 percentage points). The study used administrative data to measure whether rehabilitation appeared to have increased total enrollment at these schools or changed patterns of student dropout and graduation rates. After rehabilitation, we did observe a modest increase in early-grade enrollments at rehabilitated schools (increasing total enrollment by less than 5 percent, on average), but we did not observe major shifts in these schools' dropout rates, grade-promotion rates in upper secondary school, or graduation rates.

Science laboratories provided through the rehabilitation activity are viewed very positively by teachers and students, but initial usage rates are uneven. At the time of our site visits, all but one of the rehabilitated schools had a functional science lab (less than a third of these schools had a lab at baseline). In qualitative interviews, students and teachers responded very positively to these new facilities: several interviewed teachers reported that they were changing teaching practices as a result of the improved lab facilities and resources. They described that, after the renovations, they could create more opportunities for students to become actively involved in class assignments, experiments, and discussion. On the other hand, students reported that usage rates for these labs were uneven. While survey data showed that students' exposure to lab-based learning opportunities had improved, about half of students still reported that they had little exposure to lab-based instruction, suggesting that at least some schools might lack enough trained teaching staff (or material for lab experiments) to use the new facilities consistently. After the study's interim survey data was collected, the ILEI activity rolled out additional science laboratory training activities for teachers in rehabilitated schools—we will assess whether these trainings improved laboratory usage rates in the evaluation's final report.

One of the key questions for the final study is whether the strong pattern of improvements observed in the first phase of rehabilitated schools will be sustained through the Projects' subsequent phases. At the time of this report, several months remain in the Project's implementation period; it will be critical to assess whether the ultimate number of completed schools (and the costs of rehabilitating those schools) align with the activity's original plans and cost benefit analyses. As part of the endline analysis for this evaluation, we will apply the study's randomized controlled trial design to compare the learning outcomes of students in rehabilitated schools with those of students in schools that were not rehabilitated. We will also estimate whether the investments in school rehabilitation ultimately improved students' academic achievement outcomes in the manner envisioned by the activity's logic model.

Research design for the teacher and school director training activity

The TEE activity was nationwide in scope, aiming to train all directors of schools offering secondary grades and all of Georgia's grade 7-12 teachers in the subjects of science, mathematics, English, and geography. Teachers received a series of training modules over the course of one year (with each module lasting between two and five days) and directors received the training sequence over the course of two years. Among other topics, the trainings focused on improving teachers' use of student-centered instruction practices and improving the capacity of school directors to provide instructional leadership and effective school management. The TEE evaluation relies on descriptive surveys and qualitative data collection methods to examine the potential effects of the activity. This mixed-methods study design includes two components: (1) a performance evaluation to assess the possible effects of the TEE activity on school management and classroom instructional practices using descriptive surveys and qualitative data and (2) a matched comparison group design to assess the initial impacts of the Activity's teacher training modules, also using survey data. For analyses in this interim report, the performance evaluation and the matched comparison group analysis answer research questions about the activity's implementation and initial outcomes.

The performance evaluation component of the study is designed to answer the following key research questions:

1. To what extent do school directors perceive that their instructional leadership and school management skills have changed as a result of the new training interventions?
2. To what extent do teachers perceive that their pedagogical and classroom management practices have changed as a result of the new training interventions?
3. To what extent have school directors' instructional leadership and school management practices improved?
4. To what extent have teachers' pedagogical practices (for example, conducting student-centered instruction, matching practice to subject matter, using formative assessment) and classroom management (for example, employing affirmative teaching, eliminating gender bias, increasing time on task) improved?
5. To what extent do students experience student-centered instruction, formative assessments, and classroom management practices that align with the goals of the teacher training activities?

The study also used propensity score matching to identify a comparison group for the first cohort of trained teachers. The interim analysis took place just after the first cohort completed its final training module; the comparison teachers were in the second training cohort, which had not begun the training sequence. The matching algorithm identified a comparison group with equivalent pre-intervention levels of education, teaching experience, and seniority to the trained teachers in this analysis. This ensures that a comparison of the survey outcomes of the treatment group (shortly after the training sequence was completed) to the comparison group provides a useful way to examine whether training appeared to have effects on teachers' self-reported knowledge and classroom practices. Specifically, the analysis examined the following research questions: (1) did teacher training modules improve teachers' knowledge about student-centered instruction, formative assessments, and classroom management? And (2) did teacher training modules improve teachers' willingness to use student-centered instruction, formative assessments, and classroom management?

For the interim report, the study combines quantitative survey data from teachers and school directors (collected from a geographically representative sample of 120 schools) with a wide range of qualitative data sources, including in-depth interviews with school directors, teacher focus groups, observations of pedagogical practices in the classrooms of trained teachers, and surveys from a sample of students, to measure exposure to the types of teaching practices encouraged by the training initiative. We used evidence from these data sources to assess whether the activity had plausible near-term effects on teachers' and school directors' practices that could in turn produce gains in students' learning and longer-term labor market outcomes.

Interim findings for the teacher and school director training activity

The program logic for TEE activities did not assume that teaching practices would change in the immediate aftermath of the training sequence. Rather, the activity was designed to produce initial improvements in educators' knowledge and professional development resources (through the use of teacher study groups and professional networks among directors), which would in turn change teaching practices and school management and could ultimately improve students' learning outcomes over longer periods of time. To examine whether this longer-term pattern is occurring, the endline evaluation report in 2021 will include a follow-up analysis of teacher and

school director practices up to three years after the training sequence finished. The following are key early findings from the interim analysis:

The TEE activity succeeded in implementing trainings on a nationwide scale. In total, the Activity succeeded in holding a sufficient number of training events to offer it to Georgia's whole population of school directors (about 2,000) and all of Georgia's upper-grade teachers in the subjects of science, mathematics, English, and geography (about 18,000 teachers in total). Because of its ambitious scope, the TEE activity used a phased implementation schedule, rolling out training to multiple cohorts of teachers over three years. The training sequence consisted of multiple modules (five modules for directors, and four modules for teachers), with each module lasting between two and five days. For teachers, the training sequence was held over the course of about one year for each cohort. Directors received the training sequence over the course of two years (in a single cohort). Attendance rates at the trainings were generally high. Although school directors completed the full training sequence at a higher rate (93 percent) than teachers in the first cohort (82 percent) or second cohort (55 percent at the time of this report, when makeup trainings were still being held), a large majority of both groups attended at least one training session, and nearly all of the trainees felt positively about the training experience.

After training, teachers showed a pattern of improvements in their knowledge of student-centered instruction strategies. The interim analysis showed that trained teachers became more confident in their ability to teach higher-order thinking skills and promote cooperation through group work. Trained teachers were also more confident in their ability to use lesson plans that enable differentiated instruction for students with different abilities, use formative assessments in the classroom, and create an equitable environment for girls. Each of these findings represents a statistically significant difference between the trained teachers and the matched comparison group of teachers who had not begun the training sequence; the differences represented increases of 6 to 8 percentage points on knowledge and confidence indices collected in teacher surveys.

Immediately after training, we did not find consistent evidence of changes in teachers' classroom practices. This was expected by program implementers, who designed the TEE activity to encourage changes in teaching practices over longer periods of time. The interim analysis suggests that the training did not change the classroom practices used by trained teachers in the initial period after the training sequence was completed. This finding from the study's matched comparison group analysis is also corroborated by results from surveys of trained teachers, classroom observations (with a small sample of trained teachers), and student surveys (with a convenience sample of students attending classes with trained teachers), all of which show substantial room for improvement in teachers' use of the practices encouraged in the training sequence. For example, only 6 percent of students report that they engage in collaborative group work on a daily basis, 16 percent of students report that they consistently receive the kind of short and informal assessments encouraged by the training (formative assessments), and 10 percent of teachers reported using lessons with differentiated instruction on a daily basis. On the other hand, other teaching practices were relatively strong: classroom observations, for example, revealed that teachers were effective in keeping students engaged on instructional tasks and that teachers only rarely use passive instruction techniques (such as asking students to copy written materials verbatim). In addition, a large majority of school directors (about 90 percent) reported that they do believe the training is improving classroom

instruction. We will assess if directors' more optimistic assessment of the training's effects is borne out in the study's endline data collection, which will measure teachers' practices up to three years after the training sequence ended.

In focus groups, trained teachers identified barriers to applying the training material to their classroom practice. Teachers who participated in focus groups shared two key challenges related to integrating and applying the knowledge gained during training. First, many teachers felt the amount of information in the training sequence was greater than what they could fully learn and master in a short timeframe. Second, some teachers were concerned about the difficulty of applying what they learned to their own classrooms. For example, teachers noted specific challenges related to organizing and managing collaborative group work for students (particularly in larger classrooms), and in some cases, teachers questioned whether the additional work needed to prepare lessons using differentiated instruction or to use informal assessments would be worthwhile.

There was a stronger pattern of positive changes in teachers' professional development activities. The matched comparison group analysis did show that trained teachers began to more regularly update their professional portfolios (personal records of lesson plans, approaches to curriculum, and professional development achievements). This is consistent with the evaluation's finding that a very large percentage of trained teachers (89 percent) participated in teacher study groups set up by the program to encourage greater use of the training materials and develop professional networks for teachers. According to the original program logic, it is possible that these increases in professional activities could lay the groundwork for longer-term changes in teaching practices. This is also consistent with findings from qualitative interviews with teachers designated as TEE School Professional Development Facilitators (SPDFs). In addition to the training sequence for teachers, SPDFs attended some of the instructional leadership training modules provided to school directors. SPDFs reported that the additional training in instructional leadership had helped them learn how to guide teachers in developing higher-quality lesson plans and assessment strategies.

School directors also reported that they believe the training improved their capacity for instructional leadership and school management. Most surveyed school directors believed that the training sequence improved their capacity to guide curriculum decisions, monitor teaching quality, support teachers' professional development, and manage their school operations and budget effectively. Directors' self-reported use of the practices encouraged by the training was also very strong. For example, after training, 80 percent of directors said they provide curriculum guidance on at least a monthly basis, 89 percent advise teachers on their teaching practices on at least a monthly basis, and 68 percent collect and review data on student learning on a monthly basis. In qualitative interviews, school directors reported that trainings helped them to develop school management skills related to managing schedules, human resources, and school finances. Directors particularly emphasized the value of financial training, which they said offered useful guidance on how to allocate available funds and how best to prioritize expenditures according to specific needs.

Next steps for the evaluation

This report's interim findings represent an important set of initial results and largely suggest that the pattern of longer-term effects assumed in the program logic for the ILEI and TEE activities remain plausible. But all the preliminary findings in this report are primarily descriptive in nature and limited to near-term outcomes; findings focused on medium-term outcomes, including the results from a randomized controlled trial examining the school rehabilitation activity, will be part of the study's endline report in 2021. By comparing the preliminary results summarized here with the study's ultimate findings, the overall evaluation will provide insight regarding how the key outcomes observed in rehabilitated schools, and among trained teachers and school directors, have evolved over time and across schools in different regions. Using these findings, the final report will provide important insight into whether the pattern of observed outcomes for the ILEI and TEE activities represent a cost effective set of investments, providing lessons for MCC and implementers of similar programs in Georgia and beyond.

This page has been left blank for double-sided copying.

I. INTRODUCTION

The Millennium Challenge Corporation (MCC) is supporting Georgia's efforts to improve students' educational outcomes by sponsoring the Improving General Education Quality (IGEQ) Project. The Project comprises three components. The first, the Improved Learning Environment Infrastructure (ILEI) component, invests in school rehabilitation to provide safe learning environments that include adequate facilities and heating. Second, the Training Educators for Excellence (TEE) component supports professional development by training and mentoring teachers in subjects related to science and math and by training principals to strengthen school management. Finally, the Education Assessment Support component supports Georgia's ongoing efforts to improve educational outcomes through rigorous assessments and fostering of a result-oriented education system. MCC chose Mathematica to rigorously evaluate these components to examine their impacts on both intermediate and long-term outcomes.

This interim report describes a preliminary set of evaluation findings for the ILEI (school rehabilitation) and the TEE (teacher and school director training) components. The results presented here are intended to provide evidence about the potential effects of project activities before the Compact comes to a close; the evaluation's endline report (planned for 2021) will examine the longer-term impacts and sustainability of the project's activities in the period after MCC's direct support ends.

We begin by summarizing each activity and presenting the evaluation's research questions. Later chapters discuss the scope of the data collection and analysis methods used to assess each activity and then present the study's interim findings.

A. Overview of evaluated activities

The evaluation assesses two different activities: a school rehabilitation activity focused on intensive infrastructure investments in a subset of schools and a nationwide teacher and school director training initiative designed to improve the capacities of secondary school educators in fields related to science, technology, engineering, and mathematics throughout Georgia (teacher training focused on educators in grades 7 to 12 in the subjects of science, mathematics, English, and geography).

The school rehabilitation activity aims to upgrade the quality of physical infrastructure and create an improved learning environment in program schools. Examples of potential rehabilitation areas include the following:

- Systems for heating (replacing wood stoves with central heating)
- Lighting
- Water and plumbing
- Sanitary facilities
- Recreational facilities
- Science laboratories

- Building interiors (flooring, stairs, and classroom walls)
- Building exteriors (roofing and masonry)

Through a random assignment process, the Activity selected 104 schools throughout Georgia to receive detailed rehabilitation designs. When rehabilitation was feasible, work in these schools was scheduled to take place over the course of several construction seasons (the 2015–2016 school year, the 2016–2017 school year, the 2017–2018 school year, and the 2018–2019 school year).

The one-year teacher training sequence delivered under the TEE activity was broken up into three core modules, plus an additional subject module specific to the teacher's primary teaching subject (mathematics, science, English, and geography). Each module involved an in-person training session lasting between two and five days. The three core modules for teachers covered the characteristics of a student-centered learning environment (encouraging differentiated instruction and opportunities for critical thinking and creativity); instructional and assessment strategies (lesson planning with learning objectives, and using ongoing formative assessments alongside summative assessments); and classroom management and teacher professional practice (encouraging use of collaborative group work, and encouraging teachers to engage with professional networks and teacher study groups). The TEE training sequence for school directors (the "Leadership Academy") was delivered over the course of two years, in a series of five modules addressing instructional leadership practices, staff management skills, and training in financial management related to directors' oversight of school budgets.

According to the program's logic model (Figure I.1), the inputs from rehabilitating the schools and training teachers and school directors are intended to improve students' learning outcomes, but the mechanisms for improving learning differ. In the case of school rehabilitation, the intervention aims to decrease students' and teachers' absenteeism and improve time on task during the school day, leading to improved student learning and higher educational attainment outcomes. Although it is not reflected in the program's current logic model, we believe that rehabilitating schools could plausibly improve the health and well-being of students, which might provide another pathway for the intervention to affect learning and other long-term outcomes.[1] In the case of the training activities for teachers and school directors, the program is intended to improve students' achievement outcomes by directly improving the quality of classroom instruction (through improved teaching practices) and school management (including instructional leadership from school directors).

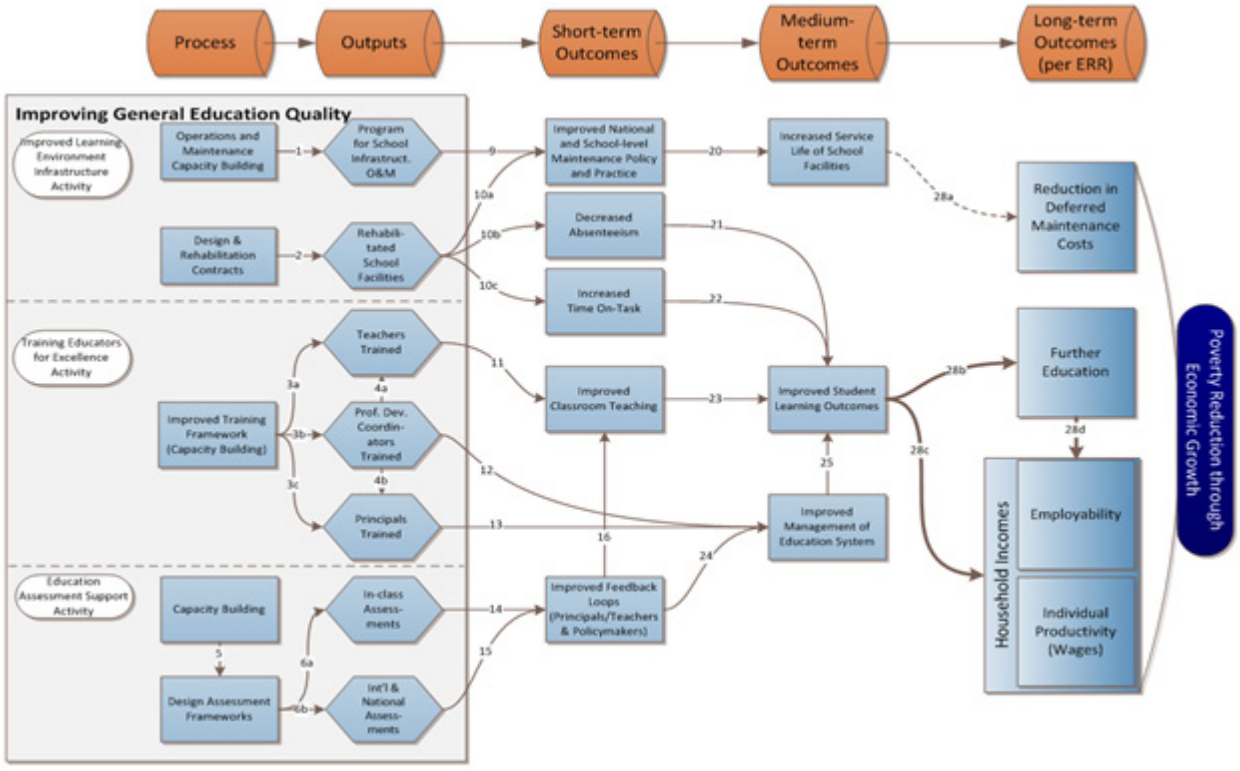
The program logic developed by MCC and Millennium Challenge Account-Georgia (MCA-G) staff presents a series of (hypothesized) causal links among program inputs and outputs and short-, medium-, and long-term outcomes that potentially support the Project's overarching goal of reducing poverty through economic growth. Each of the links represents an assumption by IGEQ program designers about how the activities will affect the Compact's beneficiaries and

[1] Children might also be exposed to poor air quality and sanitation at home, meaning that rehabilitating schools is unlikely to remove all the health risks that students face. Because treatment was assigned randomly in this evaluation, we can expect home air quality and sanitation to remain equivalent in treatment- and control-group homes at baseline and in the follow-up periods of the study. Thus, this study can attribute any health improvements observed in the treatment group to the school rehabilitation intervention.

stakeholders, which include students, teachers, school administrators, and policymakers in relevant Government of Georgia (GoG) ministries and centers. Assumptions in the program logic also provide the basis for MCC's economic rate of return (ERR) calculations for each activity.

Before the evaluation began, we assessed the plausibility of the IGEQ program logic and associated ERR calculations. To do so, the evaluation team reviewed the available evidence on the impacts of similar program designs in other contexts and discussed it extensively with local education experts and IGEQ stakeholders. These discussions included MCA-G staff, stakeholders in relevant GoG centers and ministries, and school staff interviewed during the team's site visits to schools selected for the ILEI rehabilitation program. We examined the program logic for each of the IGEQ components separately, reviewing the relevant literature on the effects of similar interventions in other contexts. We explained how an evaluation would contribute to addressing gaps in the literature and noted potential concerns about areas in which assumptions in the logic model might not hold (Nichols-Barrer et al. 2013). The various components of the interim data collection and analysis in this report are all designed to provide information about the inputs in the logic model and the potential relationships between these inputs and evaluation outcomes. Taken together, these interim findings provide preliminary evidence on whether the ultimate goals envisioned in the program logic are likely to be realized. Next, we summarize our review of the relevant literature.

Figure I.1. The IGEQ program logic



Source: MCC Georgia II Compact, Annex II.

Note: Arrows with dotted lines refer to links that MCC expects cannot be evaluated or measured. Links are uniquely numbered (e.g., "1," "2," "3a," "3b," "3c"). ERR = economic rate of return; IGEQ = Improving General Education Quality; MCC = Millennium Challenge Corporation; O&M = operations and maintenance expense.

B. Literature review

An extensive area of academic literature investigates the relationship between educational inputs and measures of student learning, educational attainment, and employment outcomes. But much less is known about the effects of these interventions in developing countries, and little empirical work exists on the education system in Georgia. In our view, the existing evidence base does not support strong predictions about the size of the program's expected impacts for either the school rehabilitation or the training activities for teachers and school directors. We summarize the relevant literature here.

1. Prior evidence on school rehabilitation

According to the ERR calculations used for the school rehabilitation activity, MCC aims for this intervention to produce the following improvements in students' long-term outcomes: a 10 percent improvement in the percentage of students transitioning into upper secondary school and a 10 percent improvement in the percentage of students transitioning into postsecondary programs. The evidence from prior studies shows great uncertainty regarding the relationship between school infrastructure inputs and all of the aforementioned outcomes. Some evaluations of school construction and rehabilitation activities found positive impacts on students' enrollment and attainment in some contexts (Burde and Linden 2013; Levy et al. 2009; Durán-Narucki 2008; Woolner et al. 2007; Bagby et al. 2014; Bagby et al. 2017) and limited to no short-term impact in other contexts (Dumitrescu et al. 2011). Very little rigorous research assesses whether a causal link exists between school rehabilitation inputs and long-run improvements in employment rates or income levels; in fact, we are not aware of any studies that tested this question using reliable empirical methods in developing countries. Measuring these long-term outcomes as part of an extended evaluation study would be a substantial contribution to the research literature and fill a significant gap in knowledge.

Past studies on school infrastructure have largely focused on, among other things, the relationship between school-building interventions or infrastructure improvements and student attendance. Specifically, researchers have tested whether attendance rates improve following upgrades to school infrastructure. Several studies in both domestic and developing country contexts have shown that improving schools' physical infrastructure can lead to an increase in school enrollment and attendance. But the impacts of infrastructure improvements likely depend on existing conditions in the affected facilities or communities. For example, if a program improves a school that is already functioning well, one would expect the benefits of the program to be relatively modest. Conversely, in a community with very limited school facilities, construction or rehabilitation programs can produce large benefits.

For example, impact evaluations of the BRIGHT program in Burkina Faso, an initiative that constructed and later expanded primary schools in 132 rural villages throughout the 10 provinces with the lowest rates of school enrollment for girls, specifically targeted communities that did not previously have ready access to a school. The evaluations found that BRIGHT schools had a positive impact on school enrollment and a large impact on test scores, primarily driven by large improvements in grade attainment (Levy et al. 2009; Kazianga et al. 2013; Davis et al. 2016). Several descriptive studies of school conditions in the United States found analogous results. A study in New York City examining the relationship between poor school facilities and various student outcomes found that students in the most deteriorated buildings attended fewer days of

school and had lower test scores in English language arts and mathematics (Durán-Narucki 2008). A pre-post case study on the effects of the renovation of a run-down elementary school in Washington, DC, found evidence of improved student attendance and test scores (Berry 2002). However, other studies show that investment in schools' physical infrastructure might improve student attendance but not necessarily in the short-term. The IMAGINE program in Niger constructed schools in 10 communities with low enrollment and primary school completion rates for girls, but—unlike the BRIGHT program implemented in Burkina Faso—many of these areas already had an existing school. Although the study did find that the newly constructed schools raised enrollment by 4.3 percentage points, it found no short-term impact on attendance rates, math test scores, or French test scores (Dumitrescu et al. 2011). But an evaluation conducted seven years after the program was implemented found that the program raised enrollment by 10.3 percentage points and attendance by 13.6 percentage points (Bagby et al. 2017).[2]

Few studies have examined the impacts of infrastructure on the amount of time spent learning tasks during the school day, and it is unclear whether school building improvements consistently lead to increases in the hours of functional instruction students receive. That said, if we assume (as shown in the rehabilitation activity's logic model) that the intervention could increase learning time, evidence suggests that, in turn, this could produce important learning gains.

Substantial evidence from the United States and developing countries suggests that increasing the time students spend on learning tasks in school can improve their test scores. For example, a randomized evaluation on the effects of short-term tutoring on cognitive and non-cognitive skills in Chile found that students from low-performing and poor schools improved their reading test scores after participating in the three-month program (Cabezas et al. 2011). Similarly, a participatory program in India trained local village volunteers on pedagogical techniques for teaching basic reading skills and subsequently tasked them to hold daily reading classes outside of school in an effort to improve the learning of village children. A randomized evaluation of the program found that the additional instruction had a positive effect on the reading skills of children who attended the camp (Banerjee et al. 2010). A great deal of research in the United States has also examined the relationship between the amount of instructional time and student learning. Studies of New York City charter schools have found that high-achieving charter schools tend to have a longer instructional year and longer school days than other charter schools (Hoxby et al. 2009; Dobbie and Fryer 2013). One of these studies found that these characteristics, coupled with frequent teacher feedback, data-driven instruction, and a focus on academic achievement, explained almost half of the variation in school effectiveness (Dobbie and Fryer 2013). A national study of the relationships between the practices of individual charter-school management organizations (CMOs) and their effects on student achievement found that CMOs with lengthened instructional hours (alongside school-wide behavior policies and more intensive teacher coaching) had larger impacts on student achievement in math and reading than other categories of CMOs (Furgeson et al. 2012).

[2] The IMAGINE program was later combined with a package of complementary interventions under the Niger Education and Community Strengthening (NECS) program, which were designed to increase access to high quality education and improve reading achievement. As a result, the impacts estimated under the 10-year evaluation reflect the combined impacts of both the IMAGINE and NECS programs.

We did not find any rigorous studies of the impact of school infrastructure in Georgia. Without evidence and knowledge on the determinants of enrollment, attendance, achievement, and attainment in the Georgian context, it is difficult to predict whether infrastructure improvements in Georgian schools will have a positive effect on student outcomes. Likewise, although studies in other countries suggest that increasing the amount of time spent on learning activities can positively affect student learning, it is unclear whether in the Georgian context teachers will be able to use additional instruction time effectively to raise students' test scores. This evaluation represents an important opportunity to fill these gaps in the research literature.

2. Prior evidence on training teachers and school directors

For the teacher and school director training activity, MCC's cost benefit analysis projected that this intervention would produce a 0.18 standard deviation improvement in student learning (in the medium-term and particularly in mathematics), ultimately resulting in a 2 percent improvement in students' future annual earnings from employment (in the long term). Prior literature has shown that training interventions can have a wide range of potential effects. In addition, many of the strongest existing studies were carried out in contexts that are not directly relevant to the TEE activity, and it is not clear whether the effects seen elsewhere will be realized by the Compact. An overview of the relevant literature follows.

Prior studies have shown an uncertain relationship between training inputs for teachers and school directors and the outcomes targeted by the intervention. Some studies show strong effects, but others do not. In the United States, an extensive literature provides rigorous evidence demonstrating that variation in teacher quality is causally linked to improvements in students' learning outcomes (for example, Chetty et al. 2011; Hanushek 2010). Rigorous studies of teacher training interventions in the United States also demonstrate that these interventions can have large effects on students' learning in some circumstances (although evidence of impacts varies across programs). The evidence for these successful programs is concentrated in earlier grade levels, and the largest learning gains (in some cases larger than 0.50 standard deviations) tend to be in studies of elementary school students in which the measured learning outcome aligned specifically with training materials (Yoon et al. 2007).

Evidence also exists from studies in developing countries that teacher training interventions can improve students' learning. Evans and Popova (2015) analyzed findings from six evidence reviews focused on education programs in developing countries. (These evidence reviews summarized results from a total of 226 separate studies.) The authors found suggestive evidence that extended teacher training programs that focus on pedagogical methods or academic subjects can have positive impacts on students' learning. In particular, the authors reported that longer-term trainings with ongoing follow-up support for teachers (the type of approach used in the one-year TEE training sequence), tended to outperform shorter-term (or one-time) training interventions with no follow-up mentoring or support. One example is the Read, Educate and Develop program in rural South Africa (Sailors 2010), which provided intensive professional development training for teachers, complete with demonstration lessons by mentors, monthly coaching visits by program staff, reflection sessions after monitoring visits, and after-school workshops for teachers. The study reported that the activity produced an improvement of 0.16 standard deviations in reading test scores.

In addition, Evans and Popova (2015) found that teacher training interventions tailored to specific academic subjects tended to be associated with larger gains in student learning. For example, when teachers in high-poverty communities in India received training on specific activities designed to improve use of literacy materials, literacy performance of early primary-grade students improved by 0.12 to 0.70 standard deviations (He et al. 2009). In contrast, a separate training program in India that provided more general guidance to teachers on how to improve students' learning in rural primary schools did not have a significant effect on learning outcomes (Muralidharan and Sundararaman 2010). This program gave teachers feedback on their students' performance at the beginning of the school year and provided a single training session focused on how to use this information to improve students' learning.

More broadly, the Evans and Popova (2015) review found that successful training and professional development interventions for teachers have had impacts on students' learning that range from 0.12 to 0.25 standard deviations. Although we do not currently plan to observe student-level outcomes in the present study design for the TEE activity (for reasons elaborated later), we believe the literature provides a useful guide regarding the range of plausible effects that the program could initially produce on teachers' and school directors' practices. In our view, it is reasonable to assume that a change of a given size in students' learning would require at least a similar (if not substantially larger) change in measures of proximate teacher-level practices related to classroom instruction and pedagogy.

The existing literature examining the effects of teacher training programs, however, might not apply directly to the TEE activity on several counts. First, there have been no large-scale, rigorous evaluations of teacher training programs in Georgia or other countries in the Caucasus region. Most prior literature focuses on studies implemented in other regions, such as sub-Saharan Africa, Asia, and Latin America, where the teacher workforce likely differs substantially from that of Georgia with respect to formal education levels and pedagogical methods. Second, the focus of TEE is on education outcomes for students in grades 7 to 12, whereas most prior studies, including all 226 studies reviewed by Evans and Popova (2015), examine training of primary-level teachers. Substantial evidence suggests that learning outcomes are more difficult to affect in later grades relative to early grades (for example, see Hill et al. 2008), so the impacts found in early-grade interventions might not apply to TEE. Finally, the large-scale national rollout of the TEE activity makes it quite different from the smaller teacher training interventions that tended to be the focus of prior impact studies. Nearly all rigorous studies on this subject focus on small, targeted programs; for example, the average number of teachers trained in the evaluations reviewed by Popova et al. (2016) was 609. The current evaluation assesses a nationwide program that aimed to train up to 18,000 Georgian-language teachers and 2,085 school directors. Carrying out the TEE activity at such a scale could pose implementation challenges that were not present in the small interventions that have been the subject of evaluation studies in the past.

Another issue that may differentiate the TEE activity is the potential timeline for observing changes in teaching practice. During the planning process for this Activity, stakeholders and implementers designed the intervention with an understanding that the potential timeline for observing changes in teaching practice could be relatively slow (2-3 years, or longer). Implementers hypothesized that teachers would be unlikely to incorporate new lesson plans and teaching approaches immediately, and would instead begin by piloting some new lesson

planning approaches during the first year and then enact changes more consistently over time once they found out which approaches appeared to be effective. We designed the evaluation and its data collection schedule to test the hypothesis that changes in teaching practices could occur over multiple years.

C. Impact evaluation design for the ILEI activity

Our evaluation for the ILEI activity uses a mixed-methods study design with three components: (1) a process evaluation examining the program's implementation and costs; (2) an impact evaluation using a random assignment design to estimate the causal impacts of rehabilitation compared with a control group, and (3) a qualitative analysis of the relationship between changes in school infrastructure and changes in the learning environment in a subset of study schools. This interim report provides preliminary results that are relevant to each of these three components; we will present the final analyses for these research questions in the study's final report in 2021 after the Compact ends and the full set of rehabilitated schools have been operating for up to two years.

Table I.1 presents the key research questions we're investigating. Our process evaluation examined outcomes related to program design and implementation; the impact evaluation examined the program's effects on school infrastructure, teachers, and students; and the study's in-depth qualitative analyses examined the relationships between rehabilitation inputs and the pattern of impacts observed in the quantitative study. The table also summarizes the data sources we will use for each component of the research.

Table I.1. Evaluation questions for the ILEI activity and approaches to answering them

Key evaluation questions	Evaluation components
Program design and implementation	**Process evaluation**
Was the ILEI activity budgeted and planned appropriately, forecasting key risks?	• Compare implementer's projected and actual cost data and examine risk assessment documents
Did the ILEI activity deliver improved facilities? How was the program rolled out? How much did rehabilitation differ by school?	• Use implementer data to compare time lines, budgets, work plans, and material use
What is the current and future status of facility-maintenance funding for schools? Do treatment schools have ongoing operations and maintenance funding to use in improved facilities? What maintenance/rehabilitation funding did control schools receive?	• Interview school directors to gather data on operations and maintenance funding and maintenance practices • Review GoG budget allocation methods to schools as they pertain to operations costs
Impacts on infrastructure, teachers, and students	**Impact evaluation (RCT) and qualitative analysis**
What are the impacts of the ILEI activity on the school infrastructure environment, such as temperature, maintenance policy, and maintenance practice? Did the Activity affect students' and teachers' perceptions of health and safety?	• Assess quality of school facilities, including observational data from enumerators on temperatures during the school day; conduct surveys and in-depth interviews with school directors regarding operations practices and equipment usage
What are the impacts of the ILEI activity on teachers' behavior, such as attendance and time spent teaching?	• Analyze teacher and student survey data; conduct in-depth interviews with teachers and student focus groups

Key evaluation questions	Evaluation components
What were the impacts of the ILEI activity on students' outcomes? What are the impacts on attendance, enrollment, dropout and retention rates, time spent studying in and out of school, and learning outcomes?	• Analyze teacher and student attendance through school visits (preferred) or administrative data; analyze time on task and teaching practices through classroom observation (video) data • Analyze student test scores
Impacts on attainment and employment	**Impact evaluation (RCT)**
What are the long-term impacts of the ILEI activity? What are the impacts on school-level student attainment (transition to secondary school and secondary school graduation) and on teacher qualifications at rehabilitated schools?	• Analyze administrative data on student attainment rates and teacher qualifications • Examine postsecondary attainment and employment outcomes using a long-term follow-up survey of students (if the study is extended beyond 2019)

Note: ILEI = Improved Learning Environment Infrastructure; RCT = randomized controlled trial.

1. Process evaluation examining program implementation and costs

For the process evaluation, Mathematica began by reviewing ILEI activity documents, including program cost data, program implementation records, and school rehabilitation design assessment reports when available. These reports document site assessments, rehabilitation recommendations, and implementation records for the program's treatment schools, and we used them to develop a basic understanding of program implementation and inputs.

As part of the final report, we will supplement the document review by conducting a series of in-depth, semi-structured interviews targeting three groups of respondents: key GoG staff; implementers, including the Activity's design contractor(s); and rehabilitation supervisors. We will develop the interview guides around numerous themes that will include respondent knowledge, attitudes, perceptions, and commitment to the ILEI activity; documentation and impressions of implementation activities; specific barriers to and challenges with rehabilitating schools; and suggestions on alternative strategies for supporting school rehabilitation efforts. We will use the major topics and themes that emerge from the review of program documents to help develop these semi-structured interview protocols. We will use these data to examine implementation successes and challenges and to document key lessons learned about implementation of school rehabilitation programs as well as implications that could help inform implementation of similar programs in other contexts.

2. Impact evaluation applying a randomized controlled trial (RCT) design

To estimate the impacts of the school rehabilitation activity, our study uses a school-level, stratified random assignment design. Schools assigned to the treatment group at minimum received detailed rehabilitation design assessments, and—when rehabilitation is feasible—treatment schools will receive the program's full set of infrastructure rehabilitation services. As part of the Compact, GoG stakeholders agreed that schools assigned to the control group will only receive business-as-usual maintenance and operations support during the life of the five-year compact (until July 2019).

To develop the random assignment procedure, we first stratified the sample of schools by region. Within regions that had a sufficient number of schools, we further stratified the sample on the following school-level characteristics:

- Minority language status (indicator for instruction primarily in Azeri or Armenian)
- Rural status (indicator for school located in a village or mountainous area)
- Average baseline test scores in math, history, and literacy

In addition, the stratification approach took into account the design status of schools in the sample in September 2014, when the first phase of random assignment took place. During the 2013–2014 school year, MCA-G hired a design contractor (Louis Berger) and partially or fully completed rehabilitation designs for several schools in the Phase I regions. No rehabilitation work took place in these schools during the 2014 summer construction season, meaning the predesigned cases could be included in the random assignment pool for this evaluation. In total, 29 program-eligible schools had existing rehabilitation designs in September 2014. To realize cost savings from this prior design work, at the request of MCA-G and MCC, the evaluation gave the predesigned schools a higher probability of being assigned to treatment (66 percent) than the schools currently lacking designs. To do so, our approach placed the pool of predesigned schools in its own separate set of region-level random assignment blocks. The study's impact analyses will adjust statistically for differences in the probability of selection into treatment associated with these predesigned strata.

This random assignment process took place in three phases that correspond to the program's staggered implementation schedule. Each of Georgia's regions was assigned to a different implementation phase (Table I.2)—this enabled the rehabilitation work in each phase to take place in a set of proximate regions, facilitating program logistics. At the beginning of a given phase, Mathematica randomly selected which schools would be eligible to receive the program from a list of schools in each region that was vetted by MCC, MCA-G, and GoG stakeholders. Mathematica completed the random assignment process for schools in the Phase I regions in September 2014, for schools in the Phase II regions in July 2015, and for schools in the Phase III regions in July 2016. We collected baseline data in the first school year following randomization for schools in each phase: 2014–2015 for Phase I schools, 2015–2016 for Phase II schools, and 2016–2017 for Phase III schools. If construction occurs as planned, the study will complete its first full analyses of the program's Year 1 and Year 2 follow-up impacts after we collect data during the 2018–2019 school year and the 2019–2020 school year, respectively.

Table I.2. Regional rollout of the ILEI activity

Phase	Regions	Number of treatment group schools	Schedule for completing rehabilitation
I	Mtskheta-Mtianeti, Racha-Lechkhumi and Kvemo Svaneti, Samtskhe-Javakheti, Shida Kartli	37	Summer 2017
II	Kakheti, Kvemo Kartli	35	December 2018
III	Guria, Imereti, Samegrelo-Zemo Svaneti	32	July 2019 (estimated)

Note: ILEI = Improved Learning Environment Infrastructure.

3. In-depth qualitative research on the effects of school rehabilitation

In addition to the process and quantitative impact evaluation, our approach also includes research designed to enrich the study's quantitative impact analyses by generating hypotheses about how school rehabilitation changes the learning environment and student outcomes. Qualitative methods provide a means of investigating potential mechanisms responsible for driving the program's impacts by collecting the type of extensive, open-ended interview and focus group data that would not be feasible to collect and analyze in all study schools. The qualitative analysis will collect data in the second follow-up year after rehabilitation in each treatment school. In total, Mathematica will select a subset of about 10 percent of the schools in the impact evaluation sample (20 schools—10 treatment and 10 control), and the local data collection firm will collect in-depth, qualitative data about program implementation and results at these schools. The data collection will pay particular attention to maintenance and operations practices, perceptions of school quality and safety, time on task, and the use of various school facilities. We will acquire this information by conducting in-depth interviews with school directors and teachers and by discussing it with secondary school students in focus groups. The in-depth interviews with school directors will assess infrastructure usage patterns, school operations, and maintenance practices; the in-depth interviews with teachers will assess how school facilities are used, time on task, perceptions of school building quality and safety, and teacher attendance. The focus group discussions with students will likewise assess how school facilities are used, time on task, perceptions of school quality and safety, and determinants of student attendance.

We expect insight from these qualitative research activities to be important and valuable, but it is important to note that qualitative methods have certain limitations. As with most qualitative research, findings from stakeholder interviews and focus groups will be illustrative and do not have the sample size to support rigorous hypothesis tests to directly estimate the program's impacts on the population being studied. We will focus on capturing how the Activity was implemented, gaining an understanding of a broad set of implementation issues from a diverse set of stakeholders and investigating the ways that school rehabilitation might affect teachers and students to improve attendance and learning outcomes. From these data, it will be possible to draw some conclusions about the potential reasons for the pattern of impacts uncovered by the impact evaluation, lessons learned in relation to implementation strategies and their potential to support school rehabilitation projects, and the potential relationships between various school infrastructure inputs and key program outcomes.

4. ILEI study sample and power calculations

To align data collection with the key outcomes envisaged in the ILEI activity's program logic, we targeted data collection efforts to students who will be in grades 9 to 12 during the study's follow-up period. Specifically, in each school, we defined the baseline study sample to be all students enrolled in grades 8 and 10 in the baseline school year. We originally planned to reinterview the students in the baseline sample in later follow-up rounds. But because of implementation delays and uncertainty regarding the final school rehabilitation schedule, many of the grade 10 students interviewed at baseline would likely have aged out of secondary school by the time rehabilitation was completed. As a result, we abandoned the original longitudinal design and instead will interview a new panel of students in the study's follow-up survey rounds (we will use the baseline data to calculate cross-sectional school-level covariates for the impact

analysis; the study will also use administrative data to track longitudinal patterns of enrollment and grade promotion across all grades). The first follow-up data collection round will survey all students enrolled in grades 9 and 11 in the year rehabilitation work is completed, and the second follow-up round will track this follow-up sample longitudinally for a second year.

We present power calculations for the study in Table I.3, showing the statistical precision provided by four illustrative sample configurations. In the benchmark scenario, we calculate the power of the study assuming all of the treatment schools that are currently scheduled to be rehabilitated will receive the program. To date, 16 of the 104 treatment schools have been excluded from the program because of implementation constraints (for example, because of structural problems with a school building that would make rehabilitation work cost prohibitive). Therefore, the benchmark scenario assumes that there will be a treatment-group compliance rate of 85 percent. However, the final number of treatment schools that will be rehabilitated between this interim report and the end of the Compact (July 2019) has yet to be determined. To reflect these possibilities, the power calculations show a variety of other scenarios regarding the rate at which Phase III schools initially assigned to the treatment group could be classified as ineligible for the program.

Depending on the final number of schools that are rehabilitated as part of the ILEI activity, we estimate that the evaluation will be able to detect statistically significant student-level impacts as small as 0.14 standard deviations in the best case and 0.19 standard deviations in the least favorable case.

Based on our review of other school construction evaluations in developing countries, we believe that the range of detectable effects shown in these scenarios represents a level of statistical precision that is adequate to detect impacts comparable with those reported for school construction in certain other contexts (Levy et al. 2009). But it is important to note that school construction interventions have not always produced sizeable short-term impacts (for example, Dumitrescu et al. 2011) and that prior studies have tended to examine wholesale construction of new school buildings rather than rehabilitation of existing facilities. Even with a minimum detectable effect equal to 0.13 standard deviations (the best-case scenario shown in Table I.3), we cannot say with confidence whether the evaluation will find significant impacts.

For the process evaluation of school rehabilitation activities, following the conclusion of the Compact, we will conduct a series of in-depth interviews targeting four groups of respondents: key GoG staff, the Activity's design contractors, rehabilitation supervisors, and the MCC and MCA-G staff involved in implementation and oversight of the rehabilitation program. Collecting information from the respondents involved in each area of implementation will enable us to develop a full picture of the planned implementation, the actual implementation, and the reasons for any divergences between the planned and actual implementations.

For the evaluation's qualitative components, the study will collect additional descriptive and qualitative data to investigate how rehabilitation affected the learning environment at study schools. Ultimately, we will draw a sample designed to obtain representative information from each of the program's 10 geographic regions in the second follow-up year after rehabilitation work has been completed. For the analyses in this interim report, we collected data from five regions (the first regions where schools were rehabilitated in 2016 and 2017), with a sample of

two schools in each region—one treatment school and one control school. In each region, we purposively selected schools to include a representative range of characteristics, such as school size and urbanicity. Within each of these schools, the local data collection firm conducted one in-depth interview with the school director, in-depth interviews with four teachers (including at least one science teacher), and two student focus groups. Each focus group included about eight randomly selected students in secondary-level grades. In total, at the conclusion of the evaluation, the qualitative sample will consist of 20 schools providing a total of 20 school director interviews, 80 teacher interviews, and 40 student focus groups. Although we believe these samples will produce meaningful descriptive data for qualitative analysis, this subsample of schools is too small to support quantitative hypothesis testing, and, as a result, we do not show power calculations for this portion of the study.

Table I.3. ILEI MDEs for different sample sizes and compliance rates

	All Phase III schools completed (no additional exclusions from treatment group)	75 percent of Phase III schools completed	50 percent of Phase III schools completed	25 percent of Phase III schools completed
Evaluation sample of schools	104 treatment 90 control	104 treatment 90 control	104 treatment 90 control	104 treatment 90 control
MDE for all schools assigned to treatment (ITT impacts)	0.12	0.12	0.12	0.12
Compliance with treatment group assignment (percentage)	85	77	69	62
Number of rehabilitated schools	88	80	72	64
MDE for rehabilitated schools (TOT impacts)	0.14	0.15	0.17	0.19

Notes: MDE calculations assume a two-tailed test with a 5 percent significance level and 80 percent power. We assume an ICC of 0.1, a school-level R-squared of 0.3, a student-level R-squared of 0.1, and an aggregate student sample comprising 30 students in grade 9 and 30 students in grade 11 enrolled at follow-up in each study school. The ICC and R-squared assumptions are based on U.S. data from school-level cluster randomized trials in education, as reported in Hedges and Hedberg (2007) and Deke et al. (2010). The student-level R-squared was assumed to be a more conservative 0.1 (versus 0.2 with a longitudinal design) to account for our cross-sectional design. However, the impact of this assumption on the estimated MDEs is minimal. TOT MDEs were calculated by dividing the ITT MDEs by the compliance rate among treatment schools (this assumes no control schools receive treatment). ILEI = Improved Learning Environment Infrastructure; MDE = minimum detectable effects; ICC = intraclass correlation; ITT = intent to treat; TOT = treatment on the treated.

5. ILEI evaluation timeframe

Table I.4 summarizes the data collection schedule. Because ILEI rehabilitation activities are occurring in multiple phases, the data collection rounds are occurring in sequence, by region (data collection for a given phase encompasses all treatment and comparison schools in the regions assigned to that phase). Note that because of program implementation delays, rehabilitation work in the Phase I regions originally scheduled to occur in summer 2015 was delayed until either 2016 or 2017. Similarly, rehabilitation work in Phase II schools originally scheduled for summer 2016 was delayed until 2018, and in Phase III schools, rehabilitation work originally scheduled for summer 2017 was delayed until the first half of 2019.

Note also that in 2019, following the end of construction in Phase III schools, Mathematica will collect additional process evaluation data beyond the surveys, student learning assessments, and qualitative data collected across the other data collection rounds. For example, for the process evaluation, the study will collect all available ILEI implementation reports and cost records after completion of rehabilitation work in Phase III.

Table I.4. ILEI evaluation data collection schedule

Collection round (Second semester of each school year).	Phase I regions (rehabilitation completed in 2016) (Mtskheta-Mtianeti, Racha-Lechkhumi and Kvemo Svaneti, Samtskhe-Javakheti, Shida Kartli)	Phase I regions (rehabilitation completed in 2017) (Mtskheta-Mtianeti, Racha-Lechkhumi and Kvemo Svaneti, Samtskhe-Javakheti, Shida Kartli)	Phase II regions (Kakheti, Kvemo Kartli)	Phase III regions (Guria, Imereti, Samegrelo-Zemo Svaneti)
2015	Baseline data collection with grades 8 and 10 students	Baseline data collection with grades 8 and 10 students	None	None
2016	None	None	Baseline data collection with grades 8 and 10 students	None
2017	One-year follow-up with grades 9 and 11 students	None	None	Baseline data collection with grades 8 and 10 students
2018	Two-year follow-up with grades 10 and 12 students Qualitative data collection	One-year follow-up with grades 9 and 11 students	None	None
2019	None	Two-year follow-up with grades 10 and 12 students Qualitative data collection	One-year follow-up with grades 9 and 11 students	One-year follow-up with grades 9 and 11 students (schools completed by February 2019)
2020	None	None	Two-year follow-up with grades 10 and 12 students Qualitative data collection	One-year follow-up with grades 9 and 11 students (schools completed after February 2019) Two-year follow-up with grades 10 and 12 students (schools completed before February 2019) Qualitative data collection

Note: ILEI = Improved Learning Environment Infrastructure.

D. Evaluation design for the TEE activity

For the TEE evaluation, our descriptive evaluation design relies on quantitative surveys and qualitative data collection methods to examine the potential effects of the training initiative.

The mixed-methods study design includes two components: (1) a performance evaluation to assess the possible effects of the TEE activity on school management and classroom instructional practices using descriptive surveys and qualitative data and (2) a matched comparison group design to assess the initial impacts of the Activity's teacher training modules, also using survey data. The performance evaluation and the matched comparison group analysis are designed to answer research questions about the program's implementation and initial outcomes; we use evidence from these analyses to assess whether the program had plausible effects on teachers' and school directors' practices that could in turn produce gains in students' learning and longer-term labor market outcomes. Although in some circumstances, matched comparison group analyses are characterized as a rigorous quasi-experimental design, in the case of this study, there were only a limited number of baseline characteristics available to match trained teachers to a comparison group (for reasons we discuss below). As a result, we believe the matched comparison group analysis should be seen as only one component of a broader descriptive evaluation design.

Table I.5 presents the research questions that each component of the TEE evaluation investigates.

Table I.5. Evaluation questions for the TEE activity and approaches to answering them

Evaluation questions	Approaches for answering them
Describe program design and implementation	**Performance evaluation**
Did the training activities embody a clearly developed theory of change? Did the TEE activity align with improvement goals and target pedagogical weaknesses identified by earlier research?	• Review program design documents, training materials, and implementation records
Was the Activity implemented as designed? What were the main challenges to implementation? Was the amount of training uniform across cohorts and subject areas? What activities did school-based professional development facilitators undertake? Did teacher study group activities occur as designed?	• Use implementers' data to compare planned time lines, budgets, and work plans to actual activities • Conduct in-depth interviews with implementers and school-based professional development facilitators
Describe teacher and school director outcomes	**Performance evaluation**
To what extent do school directors perceive that their instructional leadership and school management skills have changed as a result of the new training interventions, including project-supported collaboration with other directors in their region? Do directors report changes in attitudes toward parental engagement and community engagement? To what extent do teachers perceive that their pedagogical and classroom management practices have changed as a result of the new training interventions, project-supported collaboration with other teachers, and professional support from SPDFs?	• Analyze survey data collected from trained teachers and school directors • Analyze survey data collected from students of trained teachers • Conduct focus groups with teachers and in-depth interviews with school directors to understand perceptions of changes in performance and behavior • Analyze qualitative classroom observation data in subsample of trained teachers to describe pedagogical practices

Table I.5 (*continued*)

Evaluation questions	Approaches for answering them
To what extent have school directors' instructional leadership and school management practices improved? To what extent have teachers' pedagogical practices (for example, using student-centered instruction, matching practice to subject matter, using formative assessment) and classroom management (for example, using affirmative teaching, eliminating gender bias, increasing time on task) improved? To what extent do students experience student-centered instruction, formative assessment use, and classroom management practices that align with the goals of the teacher training activities (such as affirmative teaching, reducing gender bias, and engaging effectively with science facilities)?	• Triangulate observational data on teachers' practices with self-reported teacher survey data and student survey data
Potential effects of training on teachers	**Descriptive evaluation (matched comparison group)**
Did teacher training modules improve teachers' knowledge about student-centered instruction, formative assessments, and classroom management? Did teacher training modules improve teachers' willingness to use student-centered instruction, formative assessments, and classroom management?	• Compare the survey outcomes of teachers trained in 2016–2017 school year (Cohort 1) to a matched comparison group of teachers who will not be trained until the 2017–2018 school year (Cohort 2)

Note: SPDF = school professional development facilitator; TEE = Training Educators for Excellence.

1. Performance evaluation describing program implementation and outcomes

The performance evaluation collects information about how the TEE activity was implemented, tests whether program activities were implemented as designed, and assesses whether the practices of trained teachers and school directors align with the activities' targeted set of practices related to classroom instruction and school management. The performance evaluation analyzes several different types of data using multiple data sources, including program documentation, survey data, and qualitative research.

The study uses project reports and training databases to document the set of activities delivered (for example, the number of teachers and school directors trained and the number of schools receiving ongoing support from members of the project's training teams). To understand how the program might affect training participants and how they apply new information and skills to their work, we also collect survey data from a representative sample of teachers and school directors trained by the program. Ultimately, we collected survey data at two points in time, September 2017 (one to four weeks after the first cohort of teachers completed its sequence of four training modules) and September 2018 (one to four weeks after the second cohort completed its full sequence of TEE trainings), and will collect survey data in September 2019 (two years after the first cohort of teachers completed its training sequence, to measure longer-term outcomes).

The study also uses qualitative data to understand how the program was implemented and how the program might have changed participants' practices. As part of analyses for the final evaluation report, in-depth interviews with implementing staff and stakeholders involved with the project will help us understand project design and implementation. We also completed observation and monitoring of the teacher study groups during the program's first implementation year (the 2016–2017 school year) to measure the extent of teacher participation in study group meetings. In addition, the evaluation uses exploratory, in-depth interviews with school directors and focus groups with teachers during the program's second year (after the first cohort of teachers and school directors completed the Activity's full course of training modules) to gather more information about how the training was implemented and identify possible relationships between training activities and potential outcomes.

Finally, the study directly observed the classrooms of a sample of trained teachers delivering lessons during a regular school day. These observations occurred during the program's second implementation year (the 2017–2018 school year), and focused on a small sample of 22 teachers who completed the training sequence in September 2017 and also participated in the evaluation's teacher survey. Trained observers visited a sample of schools and conducted structured observations to measure teachers' use of instructional time, their use of materials, and their core pedagogical practices. Given the small sample size, it is important to remember that findings from these classroom observations are not representative of the broader population of trainees. However, the observation data does provide an opportunity to cross-check these teachers' survey responses against their actual practices in the classroom, and assess whether the survey data appears to be broadly consistent with survey results.

The performance evaluation does not include student learning assessments or student exams, and as a result, the evaluation does not directly measure student learning outcomes. But because of concurrent data collection activities related to the evaluation of school rehabilitation activities, it was possible to collect descriptive data from students about their perceptions of teaching practices (using a convenience sample of students surveyed in spring 2018 as part of the school rehabilitation study). As part of the interim report, we use this survey data to measure students' perceptions related to teachers' use of student-centered instruction, formative assessments, and positive classroom management practices.

In the evaluation's final report, the performance evaluation will identify implementation successes and challenges and document key lessons learned about the implementation of national-scale training programs in Georgia and implications that could help inform implementation of similar programs in similar contexts. This study component will provide in-depth information about the knowledge, attitudes, and practices of program participants. Through triangulation analyses, the performance evaluation will also assess whether the survey-reported knowledge and practices of teachers and school directors correspond with the information provided though qualitative interviews, focus groups, and classroom observations. By comparing these outcomes with the intended set of practices the program expects to encourage, the study is designed to assess whether it is plausible that the TEE training model could ultimately affect students' learning outcomes.

2. Descriptive evaluation of teacher training, applying a matched comparison group design

As part of the TEE evaluation, we also use a descriptive evaluation design to more directly examine the teacher training program's potential effects on participants. Any effort to directly estimate program impacts involves comparing outcomes for a group of participants with outcomes for a comparison or control group that does not receive the same activity in a given time period. To the extent possible, we endeavored to apply this type of design to evaluate the training program.

To examine the potential effects of the training program on teachers' knowledge and attitudes, the evaluation applies a matched comparison group design. This design compares a group of teachers who were trained during the 2016–2017 school year (Cohort 1) with a group of teachers who were trained later, in the 2017–2018 school year (Cohort 2). Specifically, in September 2017, the first cohort of trained teachers had completed the four TEE training modules, and, at that time, the second cohort of teachers had not received any training. This provided an opportunity to compare the knowledge and instructional behaviors of the trained Cohort 1 teachers against the knowledge and behaviors of Cohort 2 teachers who had not begun their training sequence. By comparing the two groups, the analysis is able to examine the potential impacts of the program on teachers' knowledge about the types of practices covered in the training intervention and teachers' attitudes toward those practices and reported willingness to use them in the future.

The purpose of a matching design is to compare a treatment group of teachers with a comparison group of teachers that credibly represents what would have occurred in the treatment group in the absence of the program. Because we did not randomly assign teachers to cohorts (for example, Cohort 1 prioritizes teachers with higher certification levels), we used a matching design out of necessity to identify a comparison group that is as similar as possible to the treatment group with respect to characteristics that are correlated both with assignment to treatment and the study's key outcomes. More specifically, our study will use propensity-score matching. In this context, a propensity score represents the probability that each teacher in the sample would have been selected to participate in the program during its first year, as estimated using data on teachers' baseline levels of teaching experience, certification, and education levels; teaching locations; and demographic characteristics. We endeavored to match Cohort 1 teachers to Cohort 2 teachers with equivalent propensity scores, thereby balancing the two groups in terms of their observed baseline characteristics.

The key methodological assumption in this design is that the propensity-score matching model accounts for all of the determinants of teacher selection in the program's first year. If the selection mechanism for the program is fully modeled by the propensity-score estimation model (that is, if the propensity score model accounts for every one of the teacher characteristics that could otherwise generate selection bias in the impact analysis), and the treatment and comparison groups are balanced in their propensity scores, such a design can produce an unbiased estimate of the program's impact.

In the case of this study, however, we have reasons to suspect that the matching analysis did not account for all of the teacher characteristics that could be sources of selection bias. In particular, the study did not collect baseline (pre-training) survey data from Cohort 1 teachers, and as a result it was not possible to directly match teachers who had equivalent pre-training knowledge levels or pedagogical practices in the classroom. Nonetheless, it is likely that accounting for the baseline attributes that the study did observe helped reduce important sources of bias (particularly in the cases of teacher age, levels of teacher education, and years of teaching experience, all of which were included in the matching model), and we believe the matching analysis does provide useful evidence about the program's potential effects. However, we do not believe that the design supports rigorous causal claims about the impacts of training. Instead, we recommend interpreting the results of the matching analysis as only one component of a broader descriptive evaluation plan.

3. TEE study population and evaluation sample

The TEE evaluation focuses on describing the outcomes of training activities delivered to school directors and teachers. The performance evaluation focuses on the first two cohorts of teachers and school directors to receive training activities in Georgian-language schools during the 2016–2017 school year and the 2017–2018 school year. Although the TEE activity is nationwide in scope and will ultimately include minority-language schools in later years, the initial cohorts of trainees prioritized staff at Georgian-language schools. Thus, the study population is limited to all Georgian-language school directors and teachers in Georgian-language schools. The study's descriptive evaluation design estimates the impacts of teacher training for the subset of Cohort 1 teachers who can be adequately matched to Cohort 2 teachers. Because the TEE activity prioritized training more senior and experienced teachers in the first cohort, the matched comparison group analysis is limited to a population of more junior and less-experienced teachers that can be matched successfully to teachers in Cohort 2.

To identify teachers and school directors for the sample, the evaluation randomly selected a geographically representative sample of 120 schools and in each school surveyed the school director and teachers in upper grades (8 to 12) in the targeted subjects of English, geography, mathematics, and science. To conduct the matched comparison impact analysis of the teacher training modules, we compared a geographically representative sample of Cohort 1 teachers who have recently completed the training sequence with a matched sample of teachers who were not yet eligible to begin the training (but received it later). We conducted the matching analysis as part of the analyses presented in this report. We summarize the matching model, as well as the baseline equivalence of the treatment group and matched comparison group, in Chapter II.

In addition to the sample of teachers and school directors, the study also conducted a small student survey module with an existing sample of students surveyed as part of the school rehabilitation evaluation in spring 2018. This survey included students in rehabilitated schools (the treatment group for the ILEI study) and students in non-rehabilitated schools (the control group for the ILEI study). Surveying these students enabled us to gather data about students' perceptions regarding the presence of targeted teaching practices in their classrooms at minimal additional cost.

We obtained qualitative data for the performance evaluation from focus groups with teachers and in-depth interviews with school directors. In addition, the performance evaluation incorporates quantitative data from classroom observations, which enable us to investigate how training has affected classroom teaching practices. High quality classroom observations are resource and time intensive; therefore, we drew only a subsample of teachers from each of the program's 11 geographic regions: we included two schools in each region for a total of 22 schools. We collected these data in the Activity's second year of implementation (the 2017–2018 school year) after the first cohort of school directors and teachers completed the full set of TEE training activities. The purposive sampling plan (described in more detail in Chapter II), was designed to produce meaningful descriptive data for qualitative analyses of teachers' and school directors' practices and to enable the qualitative study to document the presence or absence of targeted school director and teacher practices at a geographically varied subsample of schools.

For the final performance evaluation of TEE, we will combine information collected from in-depth interviews with implementers, survey data, and qualitative data to explore the relationship between training activities and the study's targeted outcomes. By collecting information from the respondents across various levels of planning, management, and implementation, we believe we can provide a full picture of the Activity's planned implementation, actual implementation, and the reasons for any differences between the planned and actual implementations.

4. TEE evaluation time frame

We collected survey data from the study's sample of school directors, Cohort 1 teachers, and Cohort 2 teachers at two points in time, September 2017 (following completion of the first teacher cohort's training modules) and September 2018 (following completion of the full training sequence for Cohort 2), and we will collect it once more on September 2019 (to measure longer-term post-training outcomes). The evaluation also conducted qualitative data collection activities in a subsample of schools during the 2017–2018 school year to further investigate possible effects of the full training sequence on the first cohort of teachers and school directors (Table I.6).

Table I.6. TEE data collection schedule

Data collection round	Cohort 1 teachers	Cohort 2 teachers	School directors and SPDFs
Surveys			
September 2017	Initial outcome survey	Baseline survey	Initial outcome survey
September 2018	Year 2 outcome survey	Initial outcome survey	Year 2 outcome survey
September 2019	Year 3 outcome survey	Year 2 outcome survey	Year 3 outcome survey
Qualitative data			
2017–2018 school year	Teacher focus groups Classroom observations		In-depth interviews

Note: The data collection also includes a convenience sample of students surveyed in March 2018. SPDF = school professional development facilitator; TEE = Training Educators for Excellence.

E. Objectives of the interim report

This report is intended to present only interim analyses regarding the evaluation's research questions. Because multiple compact activities are still being implemented at the time of this report, it would be premature to reach a final set of conclusions about the ultimate effectiveness

of these activities, and the longer-term effects of implemented activities will not be known until the evaluation period has concluded. In particular, at the time of this report, it is not possible to directly estimate the impacts of the school rehabilitation activity using the evaluation's RCT design because only a relatively small subset of schools had been rehabilitated as of the data collection round in spring 2018.

Nonetheless, the preliminary and descriptive findings presented in this report represent an important set of initial results, suggesting whether the chain of effects assumed in the program logic are plausible. Next, we explain the scope of the specific data collection efforts that we used to prepare this report together with the descriptive methods we used to conduct the study's interim analyses.

This page has been left blank for double-sided copying.

II. DATA COLLECTION AND ANALYSIS APPROACH

A. ILEI interim study data and methods

1. Quantitative surveys and administrative data

As part of the analysis for this interim report, the ILEI evaluation collected baseline and follow-up survey data on the ILEI activity's key outcomes from students, parents, teachers, and school directors. The survey data are complemented by administrative data, study-administered learning assessments, and direct observations of student attendance and school infrastructure. An MCA-procured local data collector (the Institute for Polling and Marketing, IPM) collected survey data, direct observations of attendance, and ratings of school infrastructure. The National Assessment and Examination Center (NAEC) developed and collected learning assessments for the study. Mathematica also obtained administrative data from Georgia's education management information system (EMIS) and implementation records.

Mathematica developed five data collection instruments in English: survey questionnaires of students, their parents, teachers, and school directors, as well as school building infrastructure assessments. The infrastructure assessment instrument provided the enumerators with consistent metrics for measuring school structures and systems. The infrastructure assessment teams were comprised of enumerators with engineering backgrounds who received training on how to consistently measure air quality, building systems, light levels, and temperature. Mathematica provided the technical measurement devices for this work and oversaw the training of the data collection team to ensure the protocols were carried out consistently. For example, Mathematica ensured that air quality inside classrooms was consistently measured in the same part of the classroom across all sites. Mathematica also oversaw that all air quality measurement devices, such as those for measuring levels of particulate matter and carbon monoxide, were used according to consistent protocols.

Analyses for the interim report focus on the ILEI's first phase of school rehabilitation work, which took place in the regions of Mtskheta-Mtianeti, Racha-Lechkhumi and Kvemo Svaneti, Samtskhe-Javakheti, and Shida Kartli. The data analyzed in this report are drawn from 29 schools that were rehabilitated in 2016 or 2017. Table II.1 summarizes the sample sizes for these Phase I schools.

To analyze year-to-year changes in student enrollment levels before and after rehabilitation, we also obtained administrative data from EMIS for all enrolled students at these schools. The data include anonymized information for each student who enrolled in one of the 29 rehabilitated schools rehabilitated in the first phase of the Project. The data include lists of enrolled students for six school years, from 2013–2014 through 2018–2019 (the current school year). Therefore, the data cover two school years after rehabilitation for the schools rehabilitated in 2016 (that is, during the 2016–2017 school year) and one post-rehabilitation year for the schools rehabilitated in 2017. The EMIS data also include records of each student's enrollment status in the subsequent school year (whether they remained enrolled in a rehabilitated school, dropped out of school, transferred to a school other than the rehabilitated school, or graduated from secondary school in grade 12). These measures of enrollment in the subsequent school year are available in all school years except 2018–2019.

Table II.1. Baseline and follow-up data collection samples in Phase I regions

Rehabilitation year	Survey round	Data collection dates	Number of schools	Number of students	Number of teachers	Number of school directors	Number of parents
2016	Baseline	April 30–June 7, 2015	12	638	95	12	598
	Year 1 follow-up	February 6–28, 2017	12	592	93	12	559
	Year 2 follow-up	February 5–27, 2018	12	557	89	12	516
2017	Baseline	April 30–June 7, 2015	17	1,072	134	17	994
	Year 1 follow-up	February 5–27, 2018	17	964	142	17	892

2. Qualitative data collection

The qualitative data used for the interim report came from 5 treatment and 5 control schools (10 schools in total). In each school, IPM's data collection team conducted one school director interview, four teacher interviews (10th and 12th grades; two science, one math, one foreign language), and two student focus groups (10th and 12th grades) in each school. Student focus groups included between 8 and 10 randomly selected students (with a random selected procedure designed to invite an equal number of boys and girls to participate). Data were collected in all 10 schools. Only nine school director interviews were completed, however, because one director refused to participate in the study (Table II.2).

Table II.2. ILEI qualitative data collection sample in 10 schools

Qualitative data collection method	Respondent	Number of cases per school[a]	Total
In-depth interviews	School directors	1	9
In-depth interviews	Teachers (10th and 12th grades; 2 science, 1 math, 1 foreign language)	4	40
Focus groups	Students (10th and 12th grades)	2	20

[a]The sample included 10 schools in total: 5 Phase I schools that completed rehabilitation work in 2016 and 5 Phase I control schools that were not rehabilitated.

To collect these data, Mathematica's research team (1) trained interviewers and focus group moderators on best practices in qualitative data collection, (2) provided relevant background on the study goals, (3) explained in detail the respondent-specific qualitative instruments, and (4) oversaw practice sessions (role play and in the field). The four-day training for the qualitative data collection for field staff included a review of the background and purpose of the ILEI study, a detailed presentation of each qualitative protocol, role play and peer practice, and on-site practice sessions. Interview and focus group field practice took place in schools not in the study sample.

Before the school visit, IPM contacted schools to introduce the data collection activities and schedule interviews and focus groups. Interviews and focus groups took place at the schools and were digitally audio recorded, transcribed verbatim, and translated into English. A small sample of the transcripts was randomly selected for quality assurance. Mathematica's consultant verified

those translated transcripts against the audio recordings to check accuracy of the transcription and translation process. After completing these quality assurance reviews, Mathematica staff reviewed translated transcripts and imported approved transcripts into NVIVO (a qualitative analysis software package) for analysis.

3. Analysis approach

a. Quantitative analysis

At the time of this interim report, the number of rehabilitated schools was too small to conduct a well-powered impact analysis comparing treatment schools to control schools. To provide preliminary evidence about the potential effects of rehabilitation, we instead descriptively estimate the changes in physical infrastructure in schools rehabilitated under the ILEI activity occurring between the baseline year (before rehabilitation) and the study's first follow-up year (after rehabilitation was completed). To do so, we use the following ordinary least squares (OLS) regression:

$$(1) \qquad Y_{st} = \alpha + \beta * Post_t + \gamma_s + \varepsilon_{st}$$

where Y_{st} is the outcome of interest in school s measured at time t, which is measured at either baseline or the one-year follow-up. $Post_t$ is an indicator for whether time t is the one-year follow-up (as opposed to baseline); γ_s is a set of school indicators; and ε_{st} is the random error. The estimated value of the coefficient β represents the average difference in Y_{st} within each school between baseline and follow-up survey rounds. Standard errors in the model will be clustered at the school-level using the standard Huber-White estimator to account for the possibility of correlations among schools over time. We run the same regression for teacher-, student-, and parent-level outcomes Y_{ist} that vary for individual i in school s at time t.

As we did for the analyses conducted for the baseline report, we constructed indices for most aspects of school infrastructure measured in the interim surveys. Data reduction was necessary, for several reasons. The research team collected hundreds of data items through a baseline school infrastructure assessment, student surveys, and teacher surveys. Reporting separately on each item would be impractical and could mislead readers because of the "multiple comparisons problem." This arises when researchers report the results of many hypothesis tests, where some of them are bound to be falsely rejected due to pure chance. This is the same logic whereby flipping a coin many times will eventually yield "streaks" of all heads or all tails, even if the coin is fair.

To define the key outcome indices for the evaluation, we used principal components analysis (PCA) to combine multiple measures related to aspects of school infrastructure into single indices.[3] Each index is a weighted average of related infrastructure measures, in which the

[3] A PCA is a statistical procedure that determines how a number of "factors" (in our case, related measures of infrastructure) are correlated with one another and condenses this information into linear combinations of the factors, called "principal components." Each principal component consists of a number of weights or "factor loadings" that define how much of the variation in the principal component is driven by each factor. We adopted the weights estimated for the "first principal component" to calculate our indices because, by design, the first principal component contains the set of factor weights that captures as much of the correlation between the factors as possible.

weights are aligned with measures with the highest component scores (that is, an infrastructure measure that explains a greater amount of variation across schools will receive a larger weight than measures explaining less of the variation in the sample). We further standardized the indices within the sample of schools to *z*-scores, so each index has a mean of 0 and a standard deviation of 1. Although the specific values of the indices cannot be directly interpreted, each index was coded to represent the presence of infrastructure gaps or problems and can be used to compare the infrastructure in treatment and control schools. For example, a school with a higher score on the index of physical classroom conditions would have worse conditions than a school with a lower score.

To maintain comparability to the baseline results, we used the PCA weights estimated at baseline to construct the interim indices and used the maximum values of each variable at baseline to standardize the interim indices. Tables A.1 through A.4 in Appendix A show the weights for each index included in the baseline and interim index construction. We created indices for the following aspects of school infrastructure at both baseline and interim:

- **Better condition of school building exterior.** Includes measures of the condition of the school building roof, the condition of the rain water drainage system, the condition of main entrance doors, and whether the exterior of the building is painted.
- **Better condition of interior structures.** Includes summary measures of the condition of the walls, ceilings, and floors in all classrooms and the indoor gym (if present).
- **Better condition of stairs in main school building.** Includes measures of the condition of the stairwells in the main school building, whether the stairs are level, and whether the stairs are evenly spaced (if two or more floors are present in the main school building).
- **Better air quality in classrooms.** Includes measures of the presence of particulate matter (PM) equal to or smaller than 2.5 microns in width (PM 2.5) and between 2.5 and 10 microns in width (PM 10) in parts per million (ppm) (there is extensive evidence that exposure to PM can have negative health consequences; World Health Organization 2013), the presence of carbon monoxide (CO) in ppm, and whether smoke was visible in the classroom.
- **Better condition of classroom teaching facilities.** Includes measures of whether all classrooms in a school have working lights, a lockable door, and a blackboard visible from the back of the classroom, as well as the condition of teaching equipment in classrooms.

b. Qualitative analysis

The research team developed a coding scheme to identify and parse meaningful segments of transcripts linked to the key qualitative research questions and inventory-related themes and findings across respondents. After all qualitative data were coded, the research team exported data by code and systematically reviewed the qualitative evidence pertaining to the study's research questions. Analysis focused on identifying consistent patterns and trends across transcripts (by respondent and across respondents). We also identified outliers or respondent disagreements in relation to a key theme or pattern. We documented analyses and triangulation

of findings across respondents in memos and summary tables with illustrative quotes. Appendix B contains a master table that summarizes qualitative findings for the school rehabilitation study, with illustrative quotes.

B. TEE interim study data and methods

1. Quantitative surveys

For the analyses in the interim report, the TEE evaluation collected two rounds of survey data on the TEE activity's key outcomes from teachers and school directors, as well as administrative data on training attendance from the National Center for Teacher Professional Development (TPDC). The rollout plan for the TEE teacher training sequence played an important role in determining the timing and sampling plan for the study's surveys. The TEE activity initially prioritized teachers who had passed a certification exam for their teaching subject (these teachers are classified as "senior," "lead," or "mentor" teachers), and a large majority of these more senior teachers completed the training as part of the first cohort. The remaining openings in the first training cohort were offered to teachers who had not passed the certification exam for their subject (classified as "practitioner" teachers), but a large majority of practitioner teachers completed the training as part of the second cohort. In Georgia practitioner teachers (with an average age of 52) are older than teachers who have passed their certification exam (with an average age of 46). In other words, the first cohort of trainees was both younger and more likely to have a strong grasp of their teaching subject than teachers in the second cohort.

We scheduled the data collection rounds to coincide with the two rounds of TEE teacher training. The first round of data collection (September–October 2017) took place within four weeks after the first cohort of teachers completed the TEE trainings (before the second cohort of teachers were eligible to begin their training sequence). The second survey round (September–November 2018) took place one year later, after the second round of TEE trainings was complete.

An MCA-procured local data collection firm collected the survey data for respondents in a nationwide sample of 120 schools. Schools were identified in each region of Georgia, with the number of study schools proportionate to the total number of Georgian-language schools in each region. Within regions, the study selected schools where the largest possible number of more junior practitioner-level teachers were included in the first cohort of TEE trainees. This facilitated the study's matched comparison group study design, which compared practitioner teachers in Cohort 1 to a matched sample of practitioner teachers in Cohort 2, during a period (September 2017) when the first cohort had completed its training sequence but the second cohort's trainings had not begun. In each survey round, we collected data from the director and all teachers who teach at least one of the subjects in the TEE training sequence (biology, chemistry, physics, mathematics, English, and geography) for grades 7–12 in all the sample schools. Table II.3 summarizes the sample sizes for each survey round. As with the ILEI data collection, Mathematica oversaw all the TEE data collection activities, and the local data collection firm in Georgia procured by MCA-G (IPM) implemented enumerator training with support from the research team, coordinated field work, and conducted data entry for all the evaluation's surveys. Mathematica developed two data collection instruments in English: survey questionnaires of teachers and school directors.

Table II.3. TEE survey data collection samples

Survey year	Data collection dates	Type of respondent	Survey round relative to training	Number of respondents
2017	September–October 2017	School directors	Interim survey, during training	119
		Cohort 1 teachers	Year 1 follow-up	877
		Cohort 2 teachers	Baseline	309
2018	September–November 2018	School directors	Year 1 follow-up	116
		Cohort 1 teachers	Year 2 follow-up	784
		Cohort 2 teachers	Year 1 follow-up	266

We also obtained administrative data from TPDC, which recorded teacher attendance in the core and subject modules for the first cohort of TEE teachers and school directors. For the second cohort of teachers, the data only included attendance information for the first two of the three core modules (and did not include data on the subject modules). For each training round, the data list teachers who were eligible for training in that round (including Cohort 1 teachers who had not completed training in the first round and therefore were eligible for training in the second round). The data also specify whether the teacher attended each of the training sequence's four modules (three core modules and one subject-specific module that varied in accordance with teachers' primary subject of instruction).

2. Qualitative data collection

The evaluation's qualitative data collection focused on understanding the perceptions of teachers, directors, and school professional development facilitators (SPDFs) regarding the training modules that TPDC offered, as well as on assessing the extent to which the training helped educators gain new skills. In particular, we explored the extent to which (1) teachers used the training to improve their pedagogical practices and classroom management, and (2) school directors and SPDF observed improvements in their instructional leadership and school management skills. To do this, we developed respondent-specific semi-structured focus group and interview protocols, and pre-tested instruments in schools that were not part of the study sample. Table II.4 shows the topics in each qualitative protocol.

The sample for qualitative data collection consisted of 22 schools, split into two rounds (spring 2018 and fall 2018). The sample consisted of two randomly selected study schools in each of the 11 regions in the quantitative study. In each school, IPM's data collection team conducted one school director interview, one SPDF interview, and one teacher focus group. Teacher focus groups contained 8 to 10 teachers from grades 7 through 12, and each focus group included teachers of science, math, geography, and English. (Because SPDFs often were also teachers, SPDFs were excluded from teacher focus groups.) As in the school rehabilitation study, Mathematica's research team conducted a multiday training for IPM's qualitative enumerators, including practice sessions (role play and in the field). Interview and focus group field practice took place in schools not in the study sample.

Before the school visit, IPM contacted schools to introduce the data collection activities and schedule interviews and focus groups. Interviews and focus groups took place at the schools and were digitally audio recorded, transcribed verbatim, and translated into English. After reviewing transcription and translation quality, Mathematica staff imported approved transcripts into NVIVO for analysis.

Table II.4. Description of the qualitative data collection protocols, by respondent (spring–fall 2018)

Domain	Qualitative protocol content
Director interviews	
Instructional leadership	• Perceptions about the extent to which participation in the Leadership Academy contributed to changes in directors' instructional leadership • Changes in instructional leadership practices: – Monitoring and managing teachers' performance – Monitoring and supporting teachers' instructional practices, lesson planning, differentiated instruction – Supporting teachers' professional development – Promoting inclusion and respect for diversity
School management	• Perceptions about the extent to which participation in the Leadership Academy influenced directors' school management • Changes in directors' school management: – Spending on instructional materials – Engagement with other school directors
Teacher focus groups	
Instructional leadership	• Perceptions about the TEE trainings and the extent to which participation influenced teachers' instructional leadership • Changes in instructional practices: – Using student-centered and hands-on learning – Promoting students' higher-order thinking, self-confidence, motivation, and engagement – Lesson planning, differentiated instruction, and use of summative and formative assessments – Use of ICT technology – Promoting inclusion and respect for diversity
Professional development, support, and teacher engagement	• Changes in instructional support and continuous improvement mechanisms – Feedback on instructional practices and in-classroom monitoring – Feedback on lesson planning – Support of setting goals for professional development and advancement – Implications for motivation and engagement
SPDF interviews	
Role of SPDFs	• Roles and responsibilities • Contribution of Leadership Academy trainings to role of SPDFs
Instructional leadership	• Changes in instructional leadership practices: – Monitoring and supporting teachers' instruction, lesson planning, and use of assessments – Supporting teachers' use of ICT technology – Supporting teachers' professional development – Supporting inclusion and respect for diversity

Note: TEE = Training Educators for Excellence; ICT = information and communications technology; SPDF = school professional development facilitator.

3. Stallings classroom observation

As part of data collection, we also conducted structured classroom observations using the Stallings Classroom Observation protocol to assess teachers' use of instructional time. The Stallings protocol is language- and curriculum-neutral (the protocol measures types of activities and time on task, rather than pedagogical content or subject expertise). Although the results are only descriptive, a key benefit of using the Stallings protocol is that its quantitative results are directly comparable across different types of schools and country contexts, and practices can be compared against widely recognized international benchmarks.

The observation protocol consists of gathering data in 10 brief observation periods (or "snapshots") at regular intervals during a class period. In each snapshot, the observer scans the room in a 360-degree circle, starting with the teacher, and codes in detail the following key aspects of classroom dynamics: (1) the teacher's use of class time, (2) the instructional activities taking place, (3) the materials used in the classroom, and (4) the teacher's interaction with students. The Stallings protocol provides quantitative data on the interaction of teachers and students in the classroom, as well as measures of teachers' use of class time, materials, core instructional activities, and students' engagement in academic activities. Specifically, the observation protocol yields the following measures:

- Share of class time during which teachers engaged in the core instructional activities (reading aloud, demonstration or lecture, discussion or question and answer, practice and drill, monitoring copying, and monitoring seatwork)
- Share of class time during which teachers engaged in instruction, classroom management, or other activities not related to teaching (off-task)
- Share of class time during which teachers used the following learning materials: textbook, notebook, blackboard, learning aides or manipulatives, ICT, and laboratory equipment
- Share of class time during which students engaged with the instructional activity the teacher conducted
- Share of class time during which students were off-task

Observations were conducted by enumerators who underwent a five-day training on the Stallings Classroom Observation protocol, completed practice exercises (video and on-site), and passed a certification examination with high inter-observer reliability scores. Enumerators were instructed to conduct four classroom observations in each study school (the same 11 schools visited as part of the qualitative data collection), visiting two teachers during two different class periods on nonconsecutive days. Mathematica selected the subject areas and teachers for the classroom observations, and IPM's field staff obtained consent from those teachers before observations took place.

The study conducted 44 observation sessions, observing 22 different teachers. The observations occurred in lessons taught in grades 7 through 12 (with at least one observation in each of those grades). Approximately half of the observations (22 observations) took place in a lesson teaching science subjects, and the rest took place in a lesson teaching math, English, or geography.

4. Analysis approach

a. Quantitative analysis

To explore the potential initial effects of the TEE training, we conducted an analysis comparing trained teachers to a matched comparison group of teachers who had not been trained at the time of the first survey round. We used propensity score matching to identify untrained teachers in Cohort 2 whose baseline characteristics were similar to those of practitioner teachers in the first cohort of trainees.

We estimated the probability that each teacher belongs to the treatment group (the "propensity score") using a logit regression that included as predictors each of the baseline (pre-training) variables available to the study. The matching variables were (1) status as a practitioner-level teacher (the sample was limited to practitioners, because nearly all Cohort 2 teachers were practitioners); (2) years of teaching experience; (3) gender; (4) subjects taught (math, science, geography, or English); and (5) grades taught (7 through 12). We then used these propensity scores to generate matching weights for the sample of practitioner teachers in the comparison group, using a kernel matching estimator.[4] Teachers with estimated propensity scores that fell outside the range of common support between the treatment and comparison groups (that is, treatment teachers whose propensity scores were above the highest propensity score in the comparison group, or below the lowest propensity score in the comparison group) were excluded from the matching analysis. After restricting the analysis sample to practitioner teachers, all teachers in the treatment group fell within the range of common support (that is, all Cohort 1 practitioner-level teachers in the sample were successfully matched to teachers in Cohort 2).

After matching, there were no significant differences between the treatment and matched comparison groups on any of the baseline characteristics used in the matching model, or on a separate variable indicating whether the teacher attended a non-TEE training in the last 12 months (Table II.5). The propensity score matching approach eliminated large and statistically significant baseline differences between the two groups in practitioner status, years of teaching experience, subjects taught, grades taught, and attending other training.

While the matching algorithm removed observed baseline differences between the two groups of teachers, there are still reasons to suspect that the matching analysis did not account for all teacher characteristics that could be sources of selection bias in the interim analysis. In particular, because the study did not collect baseline (pre-training) survey data from Cohort 1 teachers, it was not possible to identity teachers with equivalent pre-training knowledge levels or pedagogical practices in the classroom. Because it is possible that the study may not support rigorous causal claims about the impacts of training, we recommend interpreting the results of the matching analysis as only one component of a broader, descriptive evaluation plan.

[4] To estimate the matching weights, we used an Epanechnikov kernel matching estimator with a 0.05 bandwidth.

Table II.5. Equivalence of characteristics between teachers in first and second cohorts for full and matched analytic samples

	Full sample			Matched practitioner sample		
	Cohort 1 Mean	Cohort 2 Mean	Difference	Cohort 1 Mean	Cohort 2 Mean	Difference
Practitioner teacher	0.65	0.90	-0.25**	1.00	1.00	0.00
Years of experience as teacher	24.7	22.0	2.7**	26.8	26.1	0.7
Male	0.09	0.12	-0.03	0.11	0.12	-0.01
Subjects taught						
Science	0.40	0.28	0.12**	0.49	0.52	-0.03
Math	0.30	0.29	0.01	0.27	0.28	-0.01
English	0.19	0.39	-0.20**	0.12	0.11	0.01
Geography	0.16	0.08	0.07**	0.18	0.15	0.03
Grades taught						
7	0.52	0.48	0.04	0.51	0.50	0.02
8	0.71	0.64	0.07*	0.76	0.77	-0.01
9	0.68	0.58	0.11**	0.74	0.75	-0.01
10	0.66	0.50	0.16**	0.68	0.67	0.01
11	0.64	0.49	0.15**	0.66	0.67	0.00
12	0.36	0.32	0.04	0.34	0.32	0.02
Attended training other than TEE in last 12 months[a]	0.50	0.41	0.09**	0.42	0.43	-0.01
Sample size	**877**	**309**		**573**	**279**	

Note: Means in "Matched practitioner sample" are weighted using weights estimated using propensity score matching. Differences between Cohort 1 and Cohort 2 means and *p*-values of those differences were estimated using OLS regressions of whether in Cohort 1 on each characteristic.

[a]Attendance in other training was not used in propensity score matching.

After we identified the matched analytic sample, our analysis examined differences between the post-training knowledge and practices of Cohort 1 teachers (using survey data collected in September 2017, less than one month after Cohort 1 had an opportunity to complete the training sequence), and the baseline, pre-training knowledge and practices of the matched comparison group (who had not begun their training sequence at the time of the September 2017 survey round). To do this, our interim analysis used the following OLS regression with matching weights:

$$(2) \qquad Y_{psr} = \alpha + \beta * Cohort1_p + \delta * X_p' + \gamma_r + \varepsilon_{psr}$$

where Y_{isr} is an outcome for practitioner teacher *p* in school *s* in region *r* measured in the first follow-up survey; $Cohort1_p$ is a binary indicator that is 1 if practitioner teacher *p* is in the first cohort of trainees (the treatment group) and 0 if he or she is in the second cohort (the comparison group who were not eligible for training before the first follow-up survey); X_p' is the set of individual characteristics of practitioner teacher *p* used to estimate the propensity scores; γ_r is a set of binary indicators for each region *r*; and ε_{psr} is a random error term. The parameter of

interest in equation (2) is β, which gives the estimated difference in the outcome Y_{psr} between the treatment and comparison groups. Standard errors in the model were clustered at the school-level using the standard Huber-White estimator to account for the possibility of correlations among individuals within the same schools.

Because the survey included a range of knowledge measures, several of which are likely to be correlated with each other, we constructed summary indices of knowledge outcomes for different domains of teaching practices. As in the index construction process used for school building quality indices in the ILEI evaluation, we constructed teacher knowledge indices using PCA to estimate the weights and standardized the indices to facilitate comparisons across domains. Tables A.1 through A.4 in Appendix A show the weights for each index.

In addition to the matched comparison group analysis in this report, the interim analysis examined whether the outcomes of trained teachers and school directors changed between the September 2017 survey round and September 2018 survey round. That is, the descriptive analysis examined trends between the first and second follow-up year, for Cohort 1 teachers and school directors. We used the following OLS regression:

$$(3) \qquad Y_{it} = \alpha + \beta * Round2_t + \gamma_i + \varepsilon_{it}$$

where Y_{it} is the outcome of interest of individual i measured at time t. $Round2_t$ is an indicator for whether time t is at the second survey-round (as opposed to the first survey-round); γ_i is a set of indicators for individual respondents; and, ε_{it} is the random error. The estimated value of the coefficient β represents the average difference in Y_{it} for each individual between the first and second follow-up survey rounds. Standard errors in the model were clustered at the individual-level using the standard Huber-White estimator.

b. Qualitative analysis

To analyze data from qualitative school director interviews and teacher focus groups, the research team used a protocol very similar to the approach described above for the school rehabilitation study. The team developed a coding scheme to identify meaningful segments of transcripts linked to the TEE evaluation's core qualitative research questions. After the data were coded, we exported the transcript data by code and systematically reviewed the qualitative evidence pertaining to each of the study's research questions. Analysis focused on identifying consistent patterns and trends across transcripts (by respondent and across respondents). We also identified outliers or respondent disagreements in relation to key themes, such as the extent to which teachers used TEE training and guidance to improve their teaching practices. Appendix C contains a master table that summarizes findings from the TEE qualitative data, with illustrative quotes.

This page has been left blank for double-sided copying.

III. INTERIM FINDINGS FOR THE ILEI EVALUATION

A. School rehabilitation program context

As of December 2017, a total of 29 schools had been rehabilitated under the ILEI activity: 12 of these were completed in 2016 and 17 were completed in 2017. Most of the rehabilitated schools were in Shida Kartli (62 percent), with an additional 21 percent in Samtskhe-Javakheti, 10 percent in Racha-Lechkhumi and Kvemo Svaneti, and 7 percent in Mtskheta-Mtianeti (Table III.1). The results of our interim analysis focus on the outcomes observed in this first phase of rehabilitated schools.

In accordance with the program's targeting criteria for eligible schools, all treatment schools had substantially higher enrollment and a lower ratio of school building size to school enrollment than other schools in rural areas of Georgia. Among treatment schools, the rehabilitated schools in Phase I (the sample we examine in this interim report) had somewhat lower enrollment than the schools in later phases, resulting in lower baseline levels of utilization for the schools' existing building space. The percentage of socially vulnerable students in the Phase I rehabilitated schools (33 percent) was similar to the percentage observed in the other rural schools in Georgia (30 percent), and is also somewhat larger than the percentage in the Phase II and Phase III treatment schools (21 percent).

In the ILEI baseline report (Nichols-Barrer et al. 2017), we used the baseline survey data in treatment schools to examine whether the ILEI activity was likely to reach the expected number of students. According to baseline enrollment levels in schools scheduled for rehabilitation, the data suggested that at least 10 percent fewer students might benefit from the Activity than the number assumed in the preliminary ERR estimates (45,500 students). For this report, we used EMIS administrative enrollment data to (1) create an updated and more accurate estimate of the number of students in rehabilitated schools and (2) estimate the change in the number of expected students between the baseline data collection and the 2018–2019 school year. In the baseline data, we had estimated that there would be total enrollment of approximately 40,679 students across all treatment schools (9,834 of whom were enrolled in the 29 treatment schools that had been rehabilitated through 2017). Using the EMIS data, we observed that approximately 10,023 students were enrolled in the 29 rehabilitated schools in the year rehabilitation was completed, and that this number increased to 10,185 in the 2018–2019 school year. This suggests that the baseline survey data may have underestimated total enrollment in the rehabilitated schools at baseline by approximately 1.9 percent. In other words, it remains likely that the total enrollment in rehabilitated schools will fall short of the program's original expectations, but (if all of the planned treatment schools are rehabilitated) the shortfall may be a few percentage points smaller than what we estimated in the baseline report two years ago.

Table III.1. Summary baseline characteristics of treatment schools

	Rehabilitated treatment schools in Phase I regions (interim report sample)	Treatment schools to be rehabilitated in Phase II and Phase III	All schools in rural areas of Georgia
Number of schools	29	67	1,567
Average total enrollment	341.4	433.7	158.9
Average school building size (m^2)	2,294	2,440	1,807
Ratio of school building size (m^2) to school enrollment	8.0	6.2	18.4
Percentage of socially vulnerable students	33	21	30
Average number of socially vulnerable students	113.4	91.4	47.7
Regional distribution of schools (percentage)			
Adjara	0	0	12
Guria	0	6	5
Imereti	0	30	20
Kakheti	0	30	10
Kvemo Kartli	0	22	14
Mtskheta-Mtianeti	7	0	4
Racha-Lechkhumi and Kvemo Svaneti	10	0	3
Samegrelo-Zemo Svaneti	0	12	13
Samtskhe-Javakheti	21	0	12
Shida Kartli	62	0	7

Note: Average total enrollment, average school building size, and percentage of socially vulnerable students were estimated using 2014 administrative education management information system (EMIS) data. The sample of other schools in rural areas of Georgia summarized in this table excludes schools in the cities of Tbilisi and Batumi (because urban areas are not eligible for the program) and schools in the disputed regions of Abkhazia and Tskhinvali. Schools outside Batumi in the Adjara region are excluded from the evaluation because implementers, Millennium Challenge Corporation, and Millennium Challenge Account-Georgia decided to exclude the region from random assignment.

B. Physical infrastructure changes at rehabilitated schools and their perceived benefits

The ILEI activity is designed to upgrade the quality of the physical infrastructure of program schools (for example, building interiors, lighting, heating, water and plumbing, lavatories, science laboratories), to create a better learning environment and improve educational outcomes. In this section, we examine changes in the physical infrastructure of the Phase I program schools that were rehabilitated in 2016 and 2017 to assess the degree to which rehabilitation improved physical infrastructure and the learning environment. Through survey data (interviews at all rehabilitated schools) and qualitative data (in-depth interviews and focus groups at a subset of five treatment schools and five control schools), we also examine perceptions of how these changes affected the learning environment.

Rehabilitation investments produced a clear pattern of substantial improvements in the physical infrastructure of rehabilitated schools. We observed significant improvements in the condition of (1) the exterior of the school building, (2) interior physical infrastructure of

classrooms and indoor gyms, and (3) the stairs in the main school building (Table III.2). Classroom teaching facilities, which included measures of whether classrooms had working lighting, lockable doors, and visible blackboards, as well as of the quality of teaching equipment, also significantly improved (a change of 1.37 standard deviations). We also observed a large decrease in the presence of actively used outdoor recreation areas, which fell from 86 percent of schools at baseline to 34 percent of schools at follow-up (improvements made to indoor gyms may have also made outdoor spaces relatively less appealing, particularly in the winter month of February when data was collected). Finally, we observed a large increase in the presence of science labs: at the time of our site visits, all but one of the Phase I rehabilitated schools had a functional science lab (fewer than a third of these schools had a lab at baseline).

Table III.2. Comparison of infrastructure and teaching facilities in rehabilitated schools between baseline and one-year follow-up

	Baseline mean	Follow-up mean	Difference	*p*-value	Baseline N	Follow-up N
Better condition of school building exterior (*z*-score)	0.04	1.18	1.14**	0.00	29	29
Better condition of walls, ceilings, and floors in all classrooms and indoor gym (*z*-score)[a]	-0.08	2.33	2.41**	0.00	29	29
Better condition of stairs in main school building (*z*-score)[b]	0.02	1.20	1.18**	0.00	28	28
Better condition of classroom teaching facilities (*z*-score)	-0.23	1.37	1.60**	0.00	29	29
School has an indoor gym (p.p.)	1.00	0.93	-0.07	0.32	29	29
School has an outdoor recreation area (p.p.)	0.86	0.34	-0.52**	0.00	29	29
School has a science laboratory (p.p.)	0.31	0.97	0.66**	0.00	29	29

Notes: Differences between baseline and follow-up means and *p*-values of those differences were estimated using multivariate ordinary least squares regressions of a one-year follow-up survey indicator on each measure of infrastructure. The regressions included indicator controls for each school (not reported). Standard errors were clustered at the school level. Follow-up means were regression adjusted (estimated by adding the baseline mean to the regression-estimated difference). "*z*-scores" are constructed indices, standardized to observed values in the full baseline evaluation sample. "p.p." indicates that the reported means and differences were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

[a]Index of interior building structures included conditions in all classrooms and indoor gyms, if present.

[b]The analysis of conditions of school stairs was restricted to the 28 schools (of 29 total schools) with at least two floors in the main school building at baseline.

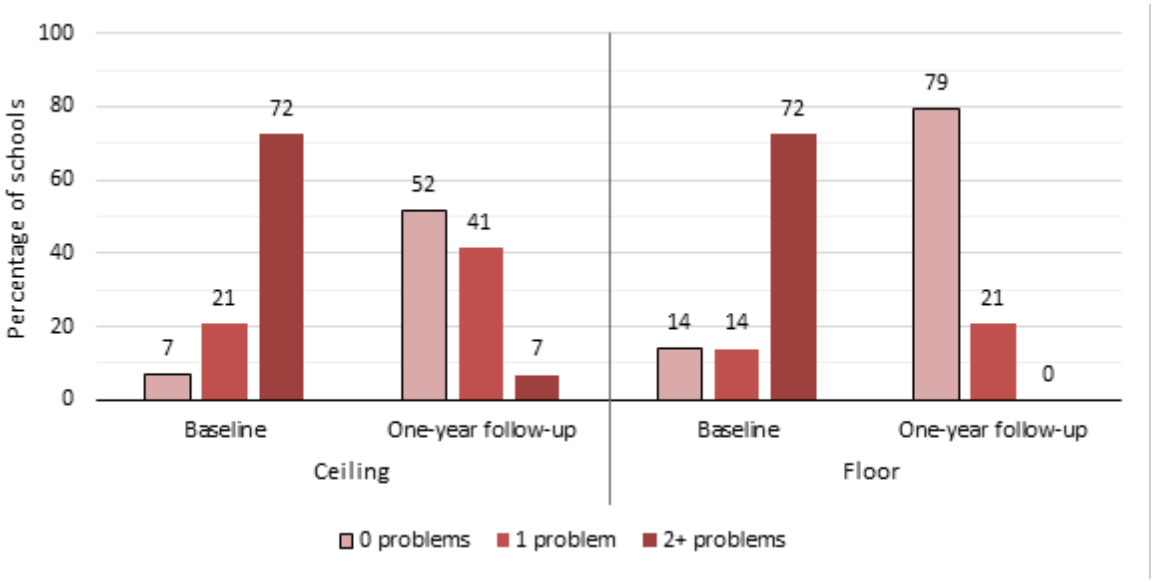
These infrastructure improvements in rehabilitated schools are readily visible in classrooms (illustrated in Figure III.1). Figure III.2 presents the distribution of problematic conditions observed in the ceiling or floor in any classroom in each rehabilitated school. In each survey round, interviewers reported whether they observed any of five problematic conditions in each structural element (cracks, water damage, mold, chipped or peeling paint, or holes in the case of classroom ceilings). At baseline, all rehabilitated schools had at least one classroom with infrastructure problems, and most of the schools had at least one classroom with two or more problematic conditions. However, after rehabilitation, the number of ceiling and floor problems dropped dramatically. For example, the percentage of schools with no problematic ceiling conditions increased from 7 to 52 percent, and the percentage with no problematic floor

conditions increased from 14 to 79 percent. Similarly, the percentage of schools with two or more problems in at least one classroom dropped from 72 to 7 percent for ceilings and 72 to 0 percent for floors.

Figure III.1. Illustration of classroom rehabilitation

Classroom before rehabilitation Rehabilitated classroom

Figure III.2. Percentage of rehabilitated schools at baseline and one-year follow-up with ceiling or floor problems in at least one classroom



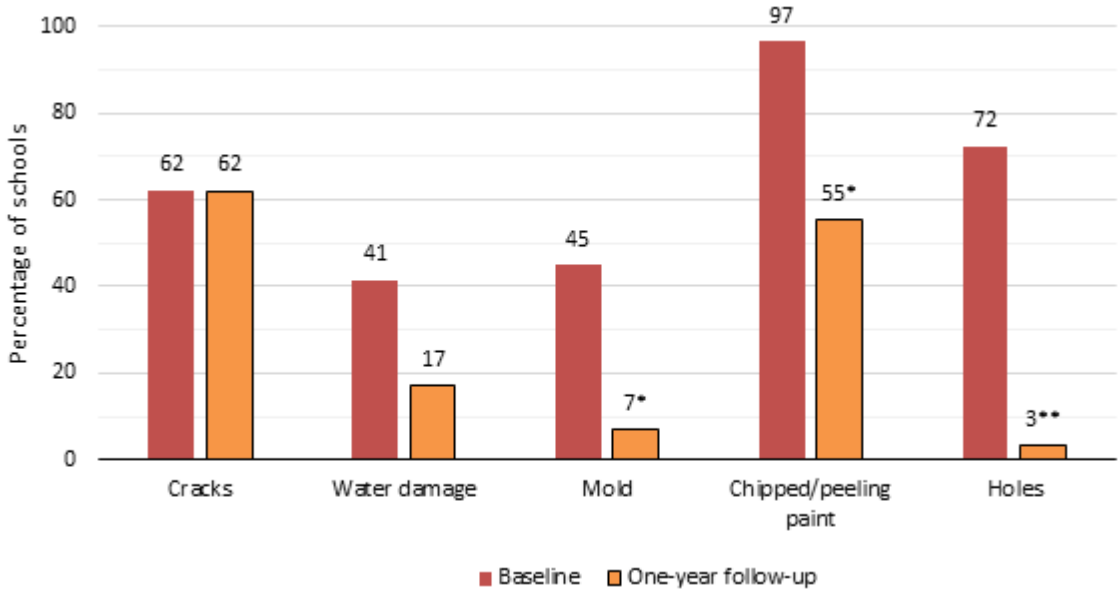
Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia School Infrastructure Surveys (2015, 2017, 2018).

Notes: Samples included 29 schools rehabilitated in 2016 and 2017.

We observed a similar pattern of improvements in classroom walls. Figure III.3 presents the prevalence of a number of problematic conditions observed in the walls in any classroom in each rehabilitated school. There were significant reductions in the percentage of schools in which classroom walls had mold (42 to 7 percent), chipped or peeling paint (97 to 55 percent), or holes in their walls (72 to 3 percent). We also observed a decline in observed water damage (from 41 to 17 percent), but this difference was not statistically significant. Despite the rehabilitation

efforts, most schools did still have at least one classroom with some form of visible crack (62 percent, unchanged from baseline). Conversations with program staff suggest that the rehabilitation effort prioritized the most serious flaws with classroom walls, leaving more superficial issues unaddressed in some cases.

Figure III.3. Percentage of rehabilitated schools at baseline and one-year follow-up experiencing problems with classroom walls



Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia School Infrastructure Surveys (2015, 2017, 2018).

Notes: Samples included 29 schools rehabilitated in 2016 and 2017.

**/* indicates that differences between baseline and one-year follow-up were significant at the 1/5 percent levels.

Rehabilitation also improved the quality of heating in classrooms during winter months. The baseline data collection confirmed that inconsistent classroom heating was a widespread problem in treatment schools before rehabilitation (Table III.3). At baseline, approximately half of all observed classrooms did not have functional central heating, and roughly half of the schools (15 of 29) had at least one classroom that was not connected to central heating. However, after rehabilitation, all 29 schools had an operational central heating system, and, in 28 of the 29 schools, every classroom was connected to the heating system on the day of the research team's site visit. The program also extended central heating to nearly all indoor gyms, increasing the percentage of indoor gyms with central heating from 38 percent at baseline to 90 percent at follow-up.

The nearly universal expansion of central heating to classrooms coincided with a large decrease in the number of students, teachers, and parents who reported that classrooms felt too cold, on average, in February (decreasing from 41 to 6 percent for students, 28 to 1 percent for teachers, and 26 to 1 percent for parents). Similarly, in the baseline survey, 41 percent of students reported that classroom temperatures made it more difficult to concentrate during the winter; after rehabilitation, however, 19 percent of students reported that this was a concern. We also observed a large improvement in the percentage of teachers who reported that classroom

temperatures in winter negatively affected their ability to teach (decreasing from 17 percent of teachers at baseline to only 3 percent in the one-year follow-up survey).

Table III.3. Comparison of presence and perceptions of central heating between baseline and one-year follow-up

	Baseline mean	Follow-up mean	Difference	*p*-value	Baseline N	Follow-up N
Classrooms						
Lacked functional central heating	0.50	0.03	-0.46**	0.00	101	114
Schools						
At least one classroom without functional central heating	0.52	0.03	-0.48**	0.00	29	29
Indoor gym lacked functional central heating[a]	0.62	0.10	-0.52**	0.00	27	27
Students						
Feels classroom is too cold, on average, in February	0.41	0.06	-0.35**	0.00	1,431	1,430
Feels temperature negatively affected ability to concentrate in February	0.41	0.19	-0.22*	0.01	1,491	1,402
Parents						
Feels classroom is too cold, on average, in February	0.26	0.01	-0.25**	0.00	1,319	1,384
Teachers						
Feels classroom is too cold, on average, in February	0.28	0.01	-0.27**	0.00	221	235
Feels temperature negatively affected ability to teach in February	0.17	0.03	-0.14**	0.00	224	234

Notes: Differences between baseline and follow-up means and *p*-values of those differences were estimated using multivariate ordinary least squares regressions of a one-year follow-up survey indicator on each measure. The regressions included indicator controls for each school (not reported). Standard errors were clustered at the school level. Follow-up means were regression adjusted (estimated by adding the baseline mean to the regression-estimated difference). The reported means and differences were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

[a] The analysis of central heating in indoor gyms was restricted to the 27 schools (of 29 total schools) with an indoor gym in the one-year follow-up.

In addition to the direct effects of low temperatures on classroom learning, the type of heating system may affect air quality in ways that have an impact on the learning environment. In particular, using wood-burning stoves during the winter may harm air quality in measurable ways, especially if classroom-specific stoves and their chimneys were poorly sealed and ventilated. As part of the building survey, enumerators collected measurements of small particulate matter (PM 2.5) and coarse particulate matter (PM 10) in February during the one-year follow-up site visits.[5] PM 2.5 and PM 10 are byproducts of wood- or coal-fire heating systems and can pose health risks at high levels (World Health Organization 2013). WHO

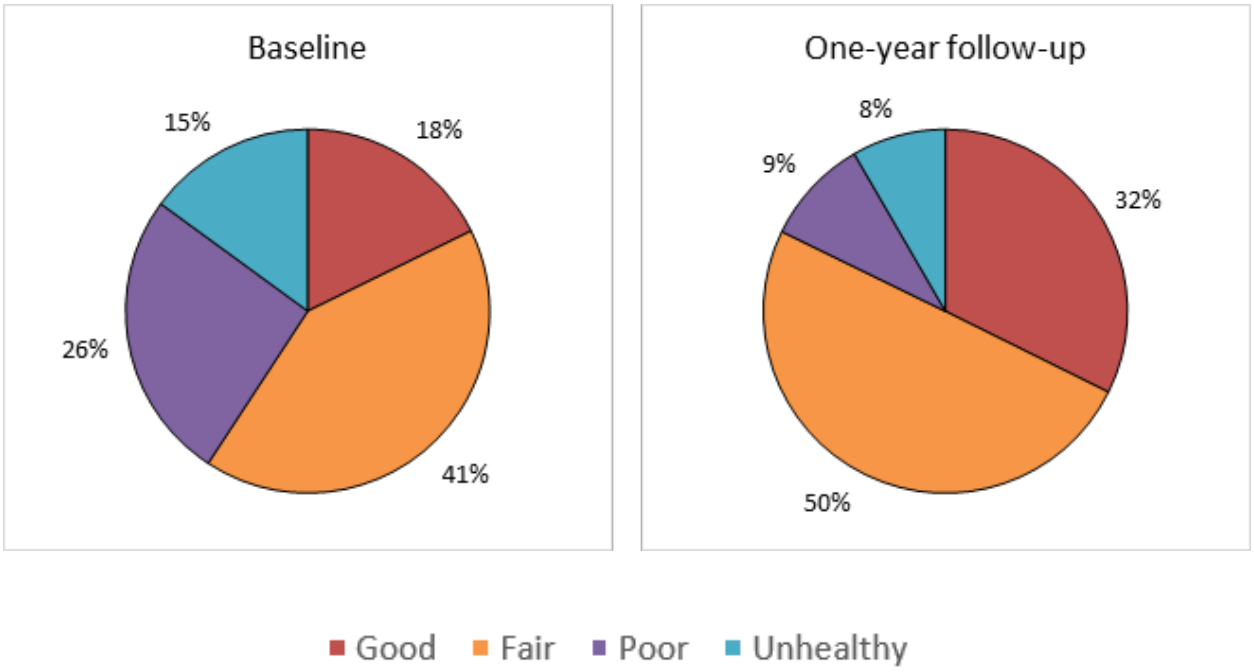
[5] Enumerators also collected measurements of PM 2.5 and PM 10 at baseline, but this was done in April and May. Therefore, the likelihood that classrooms would be heated at the time of the baseline survey was lower than at follow-up, so the air quality measurements were not directly comparable across survey rounds.

guidelines recommend keeping long-term exposure at or below 10 ppm for PM 2.5 and at or below 20 ppm for PM 10.

After rehabilitation, most classrooms still had levels of PM 2.5 and PM 10 that were somewhat higher than the WHO recommendations for long-term exposure, even though wood-burning stoves had been removed. The median classroom at rehabilitated schools had PM 2.5 levels of 12 ppm and PM 10 levels of 24 ppm; nearly all rehabilitated schools (89 percent) still had at least one classroom that exceeded the WHO's guidelines. However, more acute air quality problems (of the type associated with nearby pollution sources and poor ventilation) were less common: 20 percent of classrooms had PM levels that were more than double the WHO recommendations (PM 2.5 above 20 ppm or PM 10 above 40 ppm), and classrooms with PM levels this high were present in only a third of rehabilitated schools (34 percent for PM 2.5 and 31 percent for PM 10). This suggests that moving classrooms to cleaner heating sources was not sufficient to completely remove air quality issues at the rehabilitated schools. However, it remains possible that the impact analysis in the study's final report (comparing treatment schools to control schools) may reveal that there were significant improvements in the number of classrooms with acute air quality problems, on a relative basis.

In this sample, we cannot directly compare air quality at rehabilitated schools with baseline data because the research team did not visit Phase I schools during winter months at baseline. However, student survey results suggest that winter air quality in many classrooms did improve after school rehabilitation—but that there was still room for improvement (Figure III.4). We observed an increase in the percentage of students who believed that classroom air quality was "good" (from 18 to 32 percent), with corresponding decreases in the percentage answering that air quality was "poor" (26 to 9 percent) or "unhealthy" (15 to 8 percent). We observed similar decreases of 25 and 30 percentage points in the number of teachers and parents, respectively, who believed that classroom air quality was "poor" or "unhealthy" (not shown). However, after rehabilitation, the most common rating from students (approximately half of respondents) was that air quality was only "fair"—this is consistent with observations during site visits that sources of air pollutants unrelated to wood stove heating systems were likely still present. This finding is also consistent with the qualitative data from focus groups at a subset of rehabilitated schools: in two focus groups, students described concerns about stagnant air or unpleasant smells, especially in classrooms that do not have windows or in months when it is too cold to open windows and let in fresh air.

Figure III.4. Student perception of classroom air quality in winter at baseline and one-year follow-up

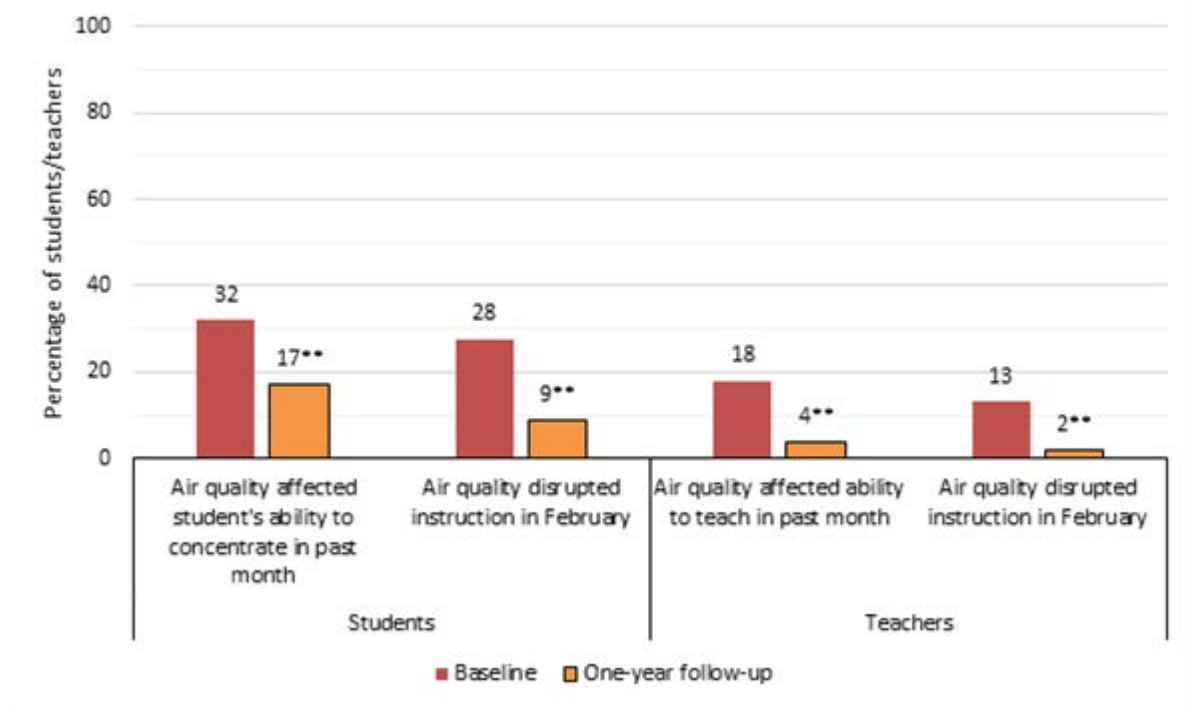


Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia Student Surveys (2015, 2017, 2018).

Notes: Samples included 1,557 students interviewed at baseline and 1,389 students interviewed at one-year follow-up in schools rehabilitated in 2016 and 2017.

Evidence from the one-year follow-up surveys also suggests that, after rehabilitation, poor air quality in winter months did not affect learning as severely. At baseline, nearly a third of students reported that classroom air quality affected their ability to concentrate on schoolwork in the past month (32 percent) or disrupted classroom instruction in February (28 percent); by the one-year follow-up, however, the percentages had decreased substantially (from 32 to 17 percent for the concentration outcome, and 28 to 9 percent for the disruption outcome) (Figure III.5). Teachers also reported large decreases in concerns about the effects of air quality on the learning environment.

Figure III.5. Perceived effect of classroom air quality in winter on the learning environment at baseline and one-year follow-up



Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia Student and Teacher Surveys (2015, 2017, 2018).

Notes: Samples included 1,514 students interviewed at baseline, 1,441 students interviewed at one-year follow-up, 238 teachers interviewed at baseline, and 234 teachers interviewed at one-year follow-up in schools rehabilitated in 2016 and 2017.

**/* indicates that differences between baseline and one-year follow-up were significant at the 1/5 percent levels.

Rehabilitation also greatly improved the quality of lighting in rehabilitated schools (Table III.4). Improvements to electrical systems and lighting were intended to improve the quality of teaching and the ability of students to read and learn, particularly during the winter. At baseline, there was no working electric lighting in at least one classroom in 79 percent of the schools, and 65 percent of students in the baseline survey reported having difficulty reading the blackboard because of poor lighting. To learn more about how these lighting issues affect the learning environment, we conducted focus groups with students in non-rehabilitated schools in the evaluation's control group (where these lighting problems still remain). Students in these control schools reported that sunlight is the primary source of light in many classrooms, and in some weather conditions the amount of natural light is not adequate. During the winter months, when there is less sunlight, students in control schools find it more difficult to see the blackboard or complete assignments. For example, students in one control school reported having to use their cell phone light to be able to read in class.

In contrast, in rehabilitated schools most of these lighting problems were addressed. By the one-year follow-up, the percentage of schools without working lighting decreased by 59 percentage points, and the percentage of students who reported having difficulty reading the blackboard because of poor lighting dropped by half (from 63 to 31 percent). In addition, student

reports that lighting made it difficult to read or negatively affected their ability to concentrate were virtually eliminated (decreasing to 5 and 6 percent, respectively). Similarly, the percentage of teachers reporting that lighting was inadequate for students fell sharply, from 29 to 4 percent.[6]

Table III.4. Comparison of quality of lighting and its effect on the learning environment at baseline and one-year follow-up

	Baseline mean	Follow-up mean	Difference	*p*-value	Baseline N	Follow-up N
Schools						
At least one classroom without working lighting in school	0.79	0.21	-0.59**	0.00	29	29
Students						
Ever have difficulty reading because of lighting	0.28	0.05	-0.23**	0.00	1,618	1,509
Ever have difficulty reading blackboard because of lighting	0.63	0.31	-0.32**	0.00	1,651	1,487
Feels lighting negatively affected ability to concentrate on schoolwork in February	0.19	0.06	-0.13**	0.00	1,522	1,459
Teachers						
Feels lighting is insufficient for students	0.29	0.04	-0.24**	0.00	227	234

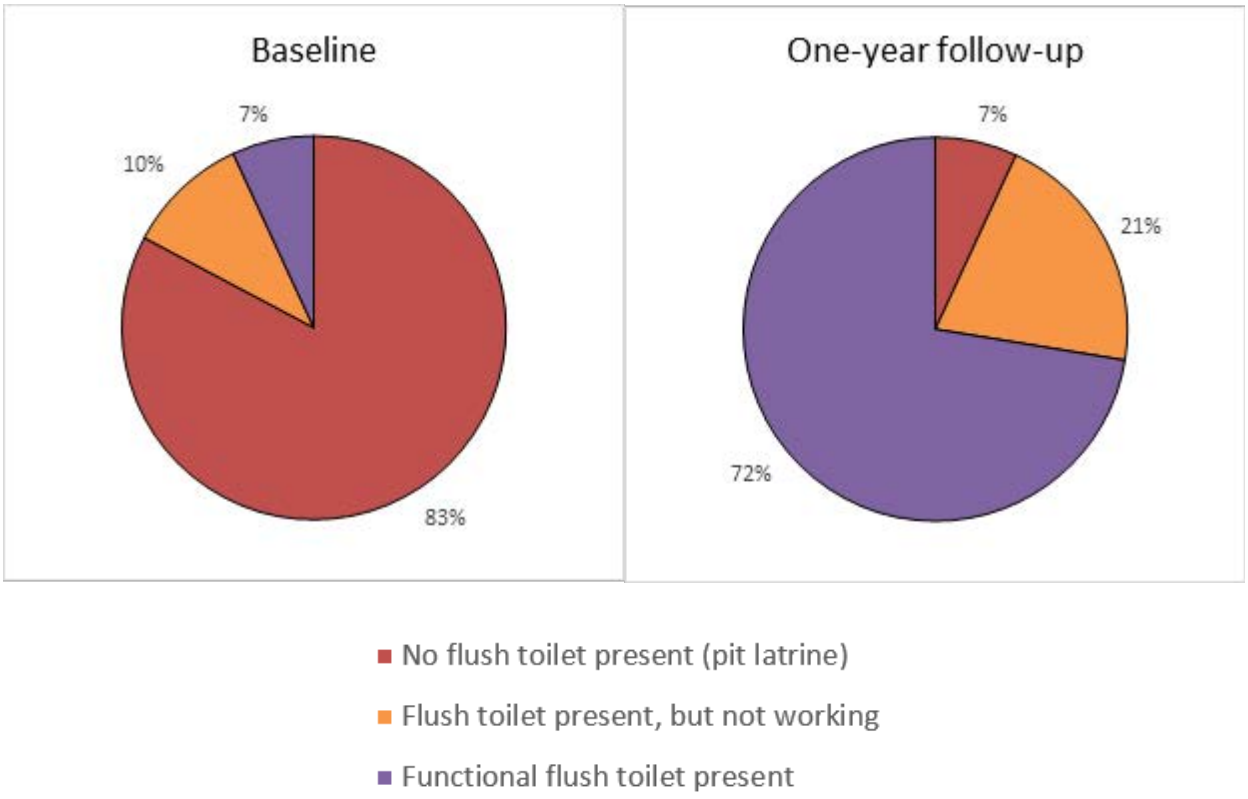
Notes: Differences between baseline and follow-up means and *p*-values of those differences were estimated using multivariate ordinary least squares regressions of a one-year follow-up survey indicator on each measure. The regressions included indicator controls for each school (not reported). Standard errors were clustered at the school level. Follow-up means were regression adjusted (estimated by adding the baseline mean to the regression-estimated difference). The reported means and differences were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

The rehabilitation program also delivered significant improvements to the sanitary facilities at rehabilitated schools. As Figure III.6 shows, most schools (83 percent) did not have flush toilets in the primary sanitary facility at baseline, and an additional 10 percent of schools had flush toilets that were not functional. By the one-year follow-up survey, however, 72 percent of schools had functional flush toilets (an increase of 65 percentage points).

[6] Survey enumerators also measured light levels at the desk furthest from windows in each classroom. We found a modest pattern of improvement in light levels, measured relative to the standard light level cutoff often recommended for classrooms (300 lux). The percentage of classrooms meeting this standard increased by 8 percentage points after rehabilitation (from 35 to 43 percent), but the difference was not statistically significant.

Figure III.6. Presence of flush toilets in primary sanitary facility at baseline and one-year follow-up

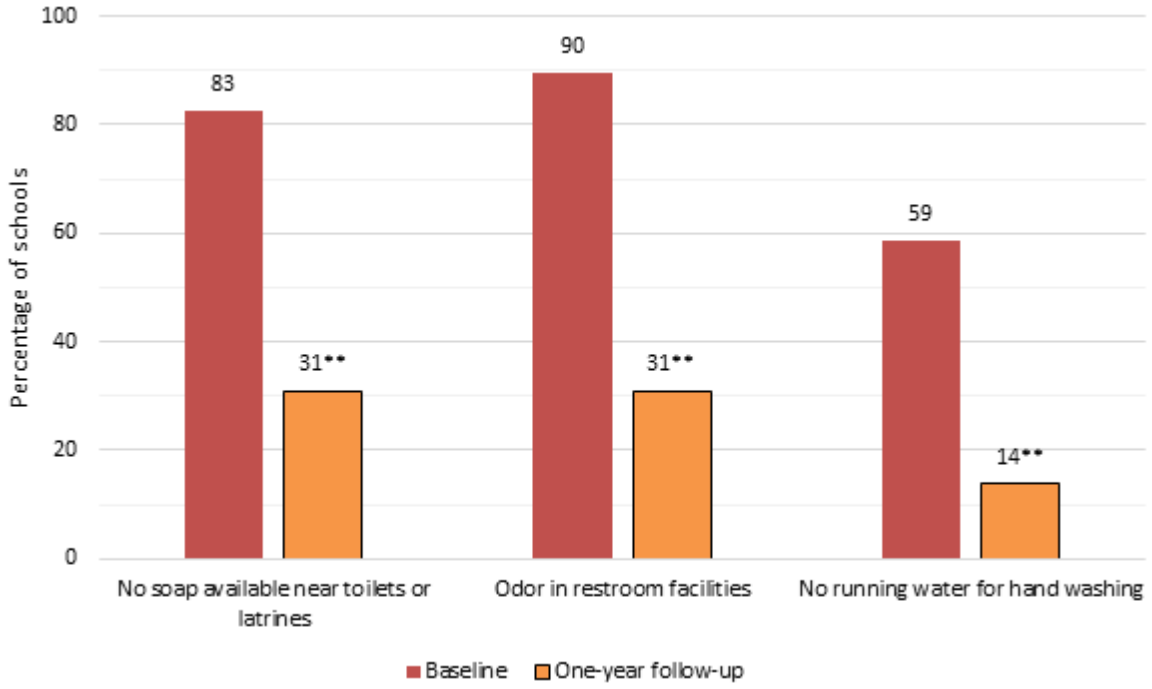


Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia School Infrastructure Surveys (2015, 2017, 2018).

Notes: Samples included 29 schools rehabilitated in 2016 and 2017.

After rehabilitation, we also observed substantial improvements in the sanitary conditions and cleanliness of toilet facilities. At baseline, most schools did not have soap available in or near school toilets or latrines (83 percent) and had an odor in the sanitary facilities (90 percent). In addition, at baseline, 59 percent of schools did not have running water for hand washing near the toilets or latrines. By the one-year follow-up survey, all three of these measures had improved by a large amounts, but there was still room for improvement (Figure III.7). At the time of the research team's site visits, nearly a third of rehabilitated schools still lacked soap near the toilets or latrine, or had an odor in their toilet facility (the presence of odors is consistent with the fact that 7 percent of rehabilitated schools still had pit latrines, and 21 percent had at least some flush toilets that were not in full working order). These survey findings were corroborated by qualitative data from student focus groups in a subset of rehabilitated schools. In two of the focus groups, students reported that they believed the renovations of toilet facilities had not been fully completed; in most focus groups, students mentioned that new sanitary facilities are generally kept clean throughout the day, but a few students also reported that soap and toilet paper were not replenished consistently in their schools.

Figure III.7. Sanitary conditions in primary sanitary facility at baseline and one-year follow-up



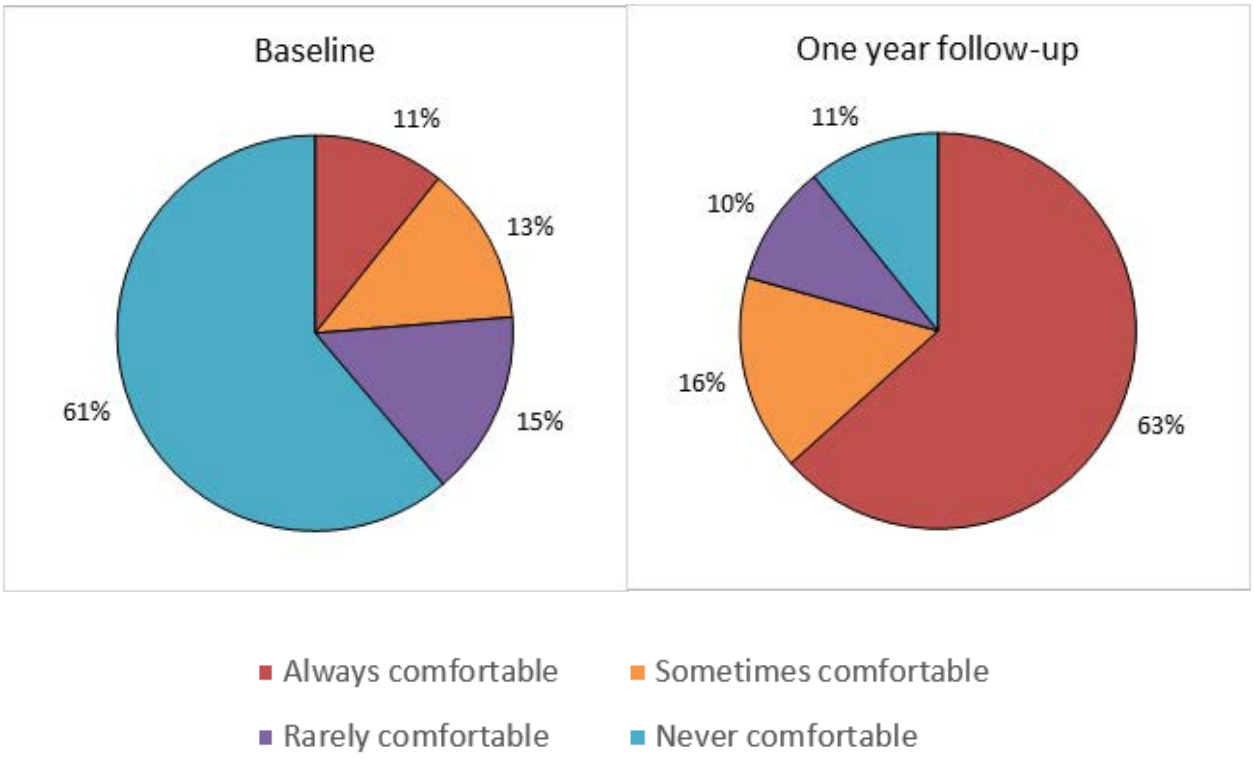
Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia School Infrastructure Surveys (2015, 2017, 2018).

Notes: Samples included 29 schools rehabilitated in 2016 and 2017.

**/* indicates that differences between baseline and one-year follow-up were significant at the 1/5 percent levels.

Teachers and students reported large improvements in their degree of comfort using sanitary facilities in rehabilitated schools. At baseline, most students (61 percent) said that they were never comfortable using the sanitary facilities in their school, and only 11 percent reported that they were "always comfortable" (Figure III.8). After rehabilitation, the proportions of "never comfortable" and "always comfortable" responses essentially reversed (to 11 percent saying they were never comfortable, and 63 percent saying they were always comfortable). We observed similar response patterns for male and female students: for both genders, two-thirds of respondents reported always being comfortable using the rehabilitated facilities (Figure III.9). Teachers also reported large improvements, with the percentage saying they were "always comfortable" increasing from 26 percent to more than 90 percent (not shown).

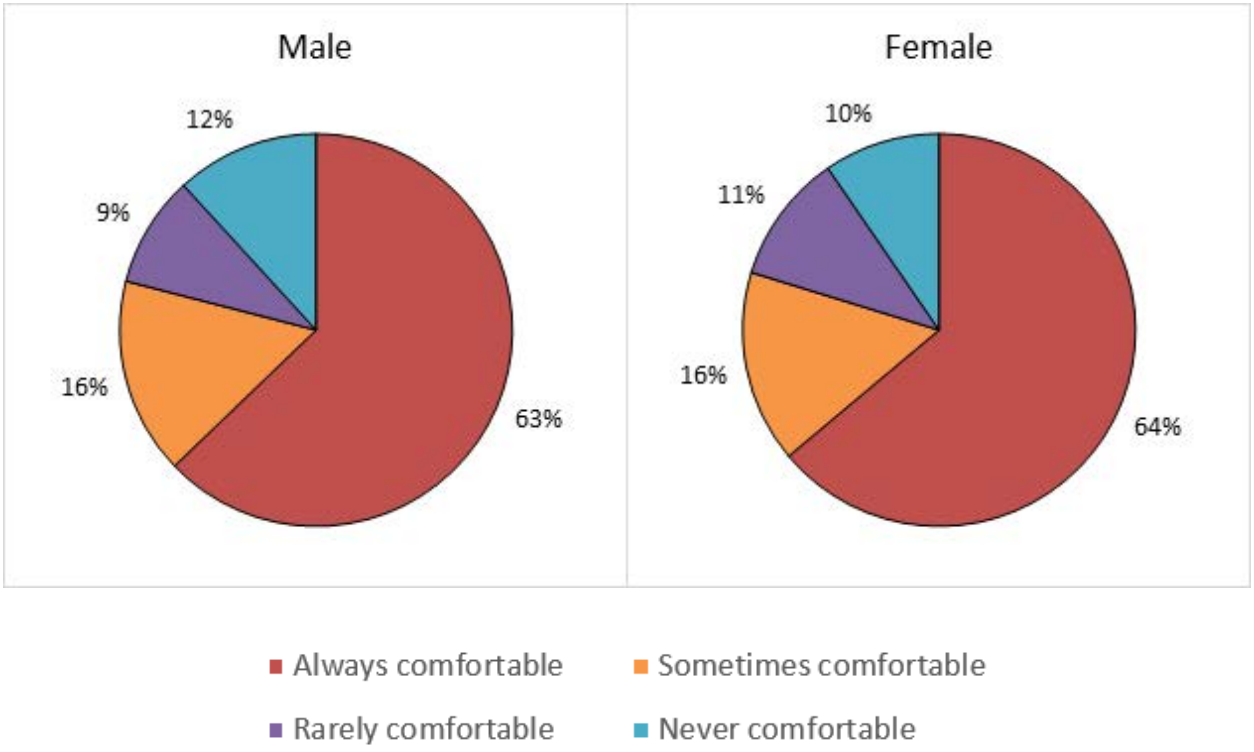
Figure III.8. Student comfort using sanitary facilities in school at baseline and one-year follow-up



Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia Student Surveys (2015, 2017, 2018).

Notes: Samples included 1,587 students interviewed at baseline and 1,366 students interviewed at one-year follow-up in schools rehabilitated in 2016 and 2017.

Figure III.9. Student comfort using rehabilitated sanitary facilities, by gender



Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia Student Surveys (2015, 2017, 2018).

Notes: Samples included 729 male and 637 female students interviewed in one-year follow-up in schools rehabilitated in 2016 and 2017.

Although survey responses about sanitary facilities were similar for boys and girls, qualitative data from students and school staff suggest that these improvements were particularly beneficial for girls. In focus groups, students reported that the location of renovated lavatories (inside the building versus outside, as previously), the privacy of the stalls (with doors versus without doors, as previously), the presence of flush toilets using running water, and the availability of sinks with running water for hand washing were critical improvements for students. Students and teachers noted that these improvements were especially helpful for girls, and that the renovations had eliminated prior situations where female students would remain in discomfort during the school day or wait to leave school to find usable toilet facilities. In control schools that were not rehabilitated, these problems remained widespread, and girls in particular expressed concerns about the lack of doors or door locks: they reported often asking classmates to accompany them to guard the doors so someone else would not open them.

C. Changes in instructional time, facility use, and perceptions of school safety

1. Potential effects of infrastructure improvements on instructional time

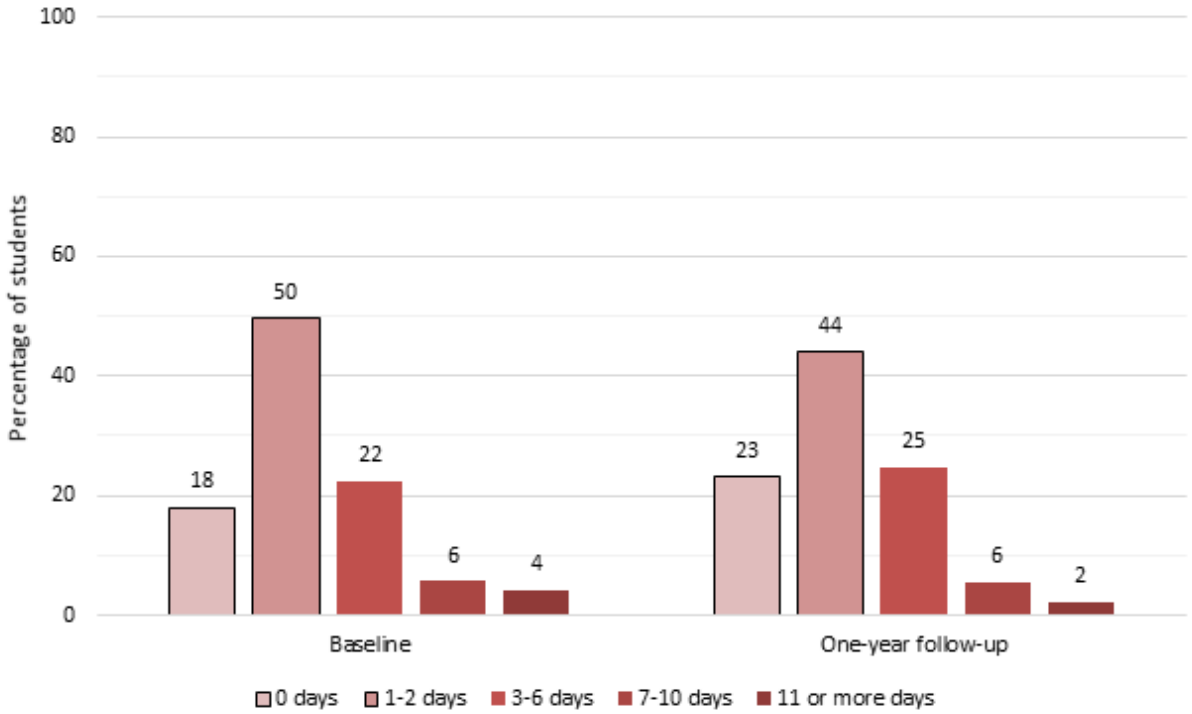
The ILEI program logic assumes that improvements in the school environment (such as the physical infrastructure, lighting, heating, and air quality in classrooms and other school facilities)

will allow students to increase the amount of instructional time they receive (by reducing absenteeism or by increasing the amount of classroom time spent on focused instruction and learning). In addition, rehabilitation was intended to enrich school experiences through the use of specialized science laboratory teaching facilities and gyms.

We did not find a strong pattern of changes in student absenteeism at rehabilitated schools, as measured by direct attendance counts by the research team, survey data from teachers, and survey data from school directors. Attendance counts that the research team conducted on the day of each site visit did not reveal any significant changes between the baseline and follow-up attendance figures at rehabilitated schools, but teachers and school directors both reported modest improvements (Figure III.10). Teachers reported the average percentage of students with perfect attendance records in the past month (increasing from 18 to 23 percent) and a modest decrease (from 50 to 44 percent) in the percentage of students with one or two absences in the previous month. School directors also reported a modest improvement in the average absence rate during February (a decrease of 4.3 percentage points) (not shown).

In addition to changes in absenteeism, rehabilitation may have increased instructional time by improving the amount of time students can spend concentrating on learning tasks. Students did report a consistent pattern of improvements in their ability to concentrate and use time well in the classroom. As discussed earlier, after rehabilitation, we observed statistically significant decreases in the percentage of students who reported that classroom heating, air quality, and lighting negatively affected their ability to concentrate and study in the winter (see Table III.3, Figure III.5, and Table III.4).

Figure III.10. Baseline and follow-up student absence patterns



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Student Surveys (2015, 2017, 2018).

Notes: Sample included 1,468 teachers interviewed at baseline and 1,433 teachers interviewed at one-year follow-up in schools rehabilitated in 2016 and 2017.

Qualitative data from student focus groups and interviews with school staff suggest that heating system improvements may have been particularly helpful in improving the quantity of instructional time that students experienced. As part of the qualitative data collection, we asked students and teachers to rank which infrastructure improvement mattered most to them, and heating system improvements were consistently ranked very highly. In renovated schools, teachers rated science labs and heating systems as the most important changes; students ranked sanitary facilities (toilets) and heating systems as the most important renovations.

Students consistently reported that heating system repairs improved the amount of instructional time at school (that is, time spent focused on instructional activities in the classroom). They also consistently reported that the prior system of using wood stove heating in classrooms often did not function properly and generated uncomfortable smoke inside the classrooms. Students were well aware of the health hazards related to continuous exposure to smoke, and some of them reported they or their peers had suffered respiratory illnesses. Students also felt that maintenance of the wood stove was a burden, because they often had to collect wood outdoors to keep the wood stove running.

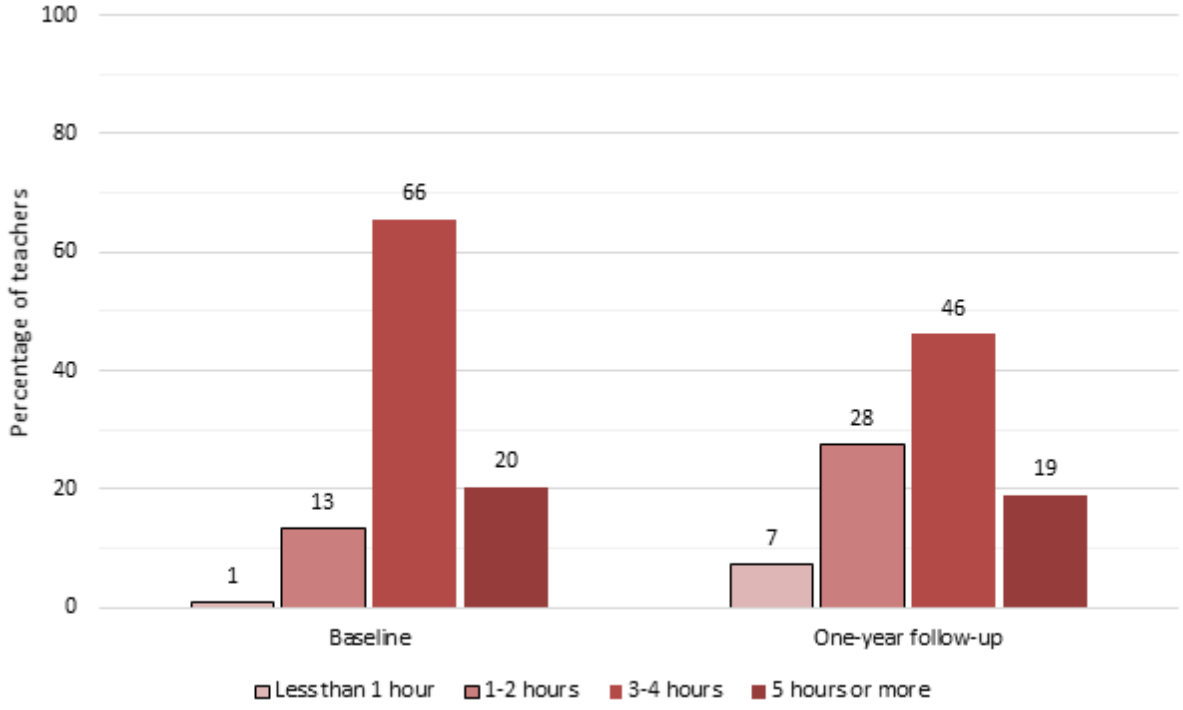
In qualitative interviews, teachers agreed that improvements in air quality and temperature had substantially benefited the learning environment. For example, one teacher said classrooms at her school used to be very smoky inside, making her students' eyes (and her eyes) irritated and teary. She said the greatest improvement at her school was that the new heating system had eliminated the smoke. Other teachers said that, before rehabilitation was completed, classrooms could get so cold that students sometimes felt unwell and wanted to go home, or that parents sometimes were reluctant to send their children to school because of the discomfort in cold classrooms. Directors of renovated schools also consistently agreed that the new central heating systems had improved the learning conditions for children. They also highlighted that these systems maintain adequate temperature not only in classrooms, but also in corridors, lavatories, buffet, sports hall, and cloakrooms, making areas of the school that did not have heating before more useful and inviting.

Although teachers agreed that heating system improvements had increased instructional time during lessons, teacher survey data also showed an unexpected decline in the average daily number of hours a given teacher spends delivering lessons to students (Figure III.11). Specifically, after rehabilitation, fewer teachers reported spending three to four hours per day on instruction (66 versus 46 percent), and more teachers reported spending one to two hours per day on instruction (13 versus 28 percent).

There are many possible reasons for this change in teachers' average hours of instruction. Currently, it is not clear whether the rehabilitation activity directly caused the pattern. If the change was related to rehabilitation, it is possible that these changes could be related to disruptions caused by teachers adjusting to new school facilities, or the process of accommodating existing or new students into the new school building after rehabilitation work was completed. However, these changes could have occurred for other reasons unrelated to rehabilitation (such as changes in school management or the number of hours teachers spent in training or professional development). Because the sample of teachers in the study is not longitudinal (at baseline, the study surveyed grade 8 and 10 teachers; at follow-up, it surveyed grade 9 and 11 teachers), it is also possible that the pattern in the survey data is reflecting simple

differences in the typical staffing arrangements at different grade levels. For example, if it is more common to staff the school with teachers on a part-time schedule in grades 9 and 11 than in grades 8 and 10, this could contribute to the pattern we observed. In the study's final report, we will investigate this pattern more directly by comparing outcomes in rehabilitated schools to a control group of non-rehabilitated schools: this will shed light on whether the rehabilitation activity caused any changes in teachers' instructional time.

Figure III.11. Class time spent on instruction per day in the month before the baseline and one-year follow-up surveys



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Teacher Surveys (2015, 2017, 2018).

Notes: Sample included 227 teachers interviewed at baseline and 232 teachers interviewed at one-year follow-up in schools rehabilitated in 2016 and 2017.

One other potential barrier to increasing instructional time pertains to the lack of food service in school cafeterias. The ILEI activity did not address provision of school meals, and evidence from student focus groups suggests that a lack of cafeteria food service could be a barrier to increasing instructional time in rehabilitated schools. In qualitative focus groups, students in rehabilitated schools (and students in control schools that were not rehabilitated) reported that it was common for the school to lack cafeteria services, meaning students must leave the school premises to buy food. In many cases, students reported that they do not have enough time during recess to walk to a store or café to buy food. Teachers in both treatment and control schools recognized this as a problem, particularly because when students leave the school to buy food, they tend to miss class time. Some students also noted that they skip classes or leave school early because they get hungry or thirsty, and do not have a place to buy food during the school day.

2. Use of recreational facilities

Rehabilitation did not change overall usage rates for recreational facilities; after rehabilitation, however, there was a large decline in the percentage of students using outdoor recreational facilities on a weekly basis. Before rehabilitation, there was a high usage rate for indoor recreational facilities (consistent with the fact that, at baseline, all rehabilitated schools had at least some type of indoor gym); after rehabilitation, there was not a statistically significant change in the usage rate for indoor gyms. However, there was a 40 percentage point decline in the percentage of students reporting that they used outdoor recreational facilities each week (Table III.5). The large drop in usage rates for outdoor space is likely related to differences in the timing of the baseline data collection (in April and May, when weather was relatively warm) and the follow-up data collection (in February, during winter). The change also corresponds to the significant decrease in the percentage of schools with outdoor recreational spaces (noted in Table III.2) and may also reflect the possibility that improvements made to indoor gyms made outdoor spaces relatively less appealing.

Student focus group data suggest that students are, in fact, using indoor recreational spaces more intensively after rehabilitation. Students highlighted the benefits of having an indoor recreational space with adequate heating and clean air that is suitable for use during winter months. Students can play indoor sports more often than before, and, in some schools, they also have had opportunities to use new types of recreational equipment and learn new sports (such as tennis). Students' perceptions of the renovated gymnasiums and sport facilities were very positive, and suggest that these facilities have contributed to increased student engagement and positive attitudes toward schooling. For example, some students now stay in school after classes end to play sports.

Table III.5. Student use of recreational school facilities

	Baseline mean	Follow-up mean	Difference	*p*-value	Baseline N	Follow-up N
Student uses an indoor gym at least once in an average week	0.90	0.86	-0.03	0.48	1,643	1,489
Student uses an outdoor recreation area at least once in an average week	0.74	0.34	-0.40**	0.00	1,600	1,373
Student uses an indoor gym or an outdoor recreation area at least once in an average week	0.94	0.87	-0.07	0.07	1,681	1,518

Notes: Differences between control and treatment means and *p*-values of those differences were estimated using multivariate ordinary least squares regressions of treatment status on each measure of baseline infrastructure. The regressions included indicator controls for the probability of selection in the intervention group that was assigned to the randomization strata of each school (not reported). Standard errors were clustered at school level. Follow-up means were regression adjusted (estimated by adding the baseline mean to the regression-estimated difference). The reported means and differences were in percentage points, with a range between 0 and 1.

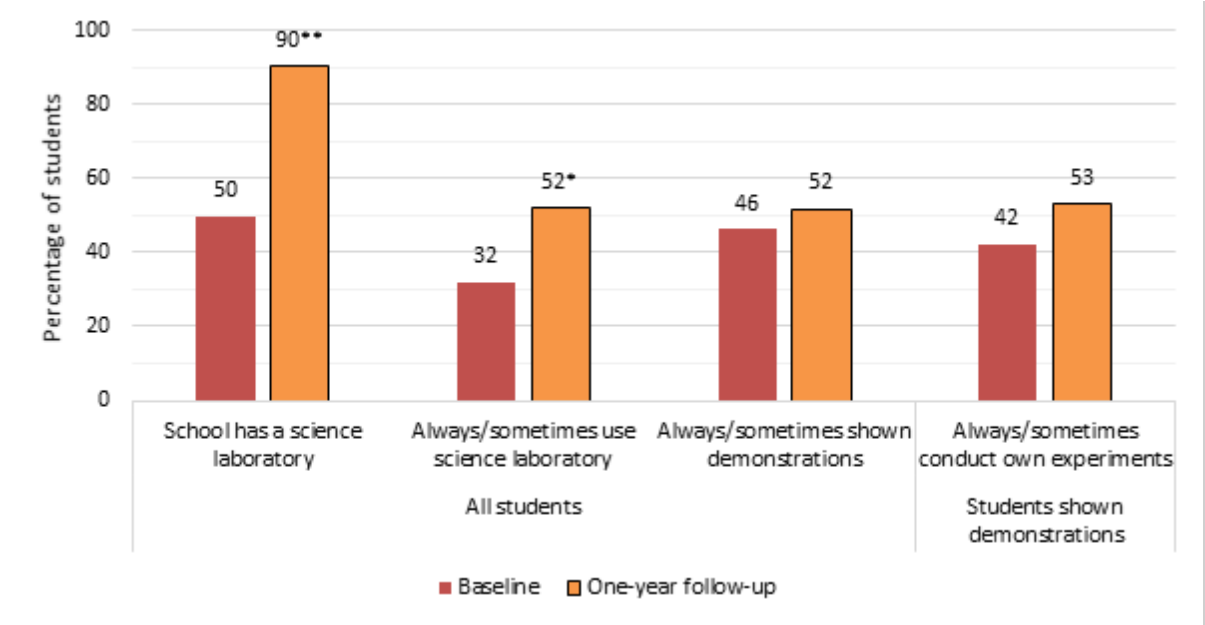
**/* indicates that differences were significant at the 1/5 percent levels.

3. Use of science laboratories

As mentioned earlier, the ILEI activity delivered new science laboratories and science equipment as part of the rehabilitation package. This resulted in significant improvements in students' exposure to science laboratories, but we did not observe a significant change in the

percentage of students receiving science demonstrations or participating in experiments (Figure III.12). After rehabilitation, the percentage of students whose school has a science laboratory increased by 40 percentage points (to 90 percent), and the percentage of students who reported "always" or "sometimes" using a science lab increased by 20 percentage points. On the other hand, we did not observe a statistically significant change in the percentage of students reporting that they "always" or "sometimes" receive science demonstrations (52 percent at follow-up) or "always" or "sometimes" participate in experiments (53 percent at follow-up). The reason for the lack of changes in these areas could be that the first follow-up surveys took place shortly after rehabilitation was completed. Beginning in summer 2017 (after the first wave of data collection analyzed in this report), the ILEI activity delivered formal training to science teachers in rehabilitated schools addressing topics such as science lab safety and lab-based instructional practices. As part of the evaluation's final impact analyses, we will assess whether science instruction outcomes improved between the first and second follow-up year after rehabilitation, after these trainings took place.

Figure III.12. Comparison of exposure to science laboratories and demonstrations and conducting science experiments at baseline and one-year follow-up



Sources: Baseline and Follow-up Millennium Challenge Corporation Georgia School Infrastructure and Student Surveys (2015, 2017, 2018).

Notes: Samples included 1,641 students interviewed at baseline and 1,461 students interviewed at one-year follow-up in 29 schools rehabilitated in 2016 and 2017. "Always/sometimes conducted own experiments" only included students who reported that teachers rarely, sometimes, or always demonstrated science experiments (1,332 students interviewed at baseline and 1,044 students interviewed at one-year follow-up).

**/* indicates that differences between baseline and one-year follow-up were significant at the 1/5 percent levels.

Qualitative data from students and teachers (collected in the second follow-up year, in a subset of rehabilitated schools) provide additional insights about how stakeholders reacted to the provision of science labs and used the new facilities.

In the second follow-up year, students in treatment schools reported that science laboratories helped them feel more motivated to attend lab sessions and more engaged with science subjects. Students stated that newly installed school labs are better resourced than what they had previously, citing access to more sophisticated equipment and new protective gear. Before the renovation, students said that there were few lab resources and that any science equipment available would have to be shared with several students. Students were also well aware of the advantages of having labs and equipment for hands-on learning. For example, they described using models to learn about human anatomy, and running chemistry and physics experiments with "real" substances. They stated that the science resources for hands-on learning have helped them learn and remember lessons more effectively.

In qualitative interviews, teachers of science subjects also consistently reported that new laboratory facilities were beneficial. Several teachers interviewed reported that they were changing teaching practices as a result of the improved lab facilities and resources. They described that, before the renovations, they relied primarily on teacher-led demonstrations (this is consistent with findings from teacher interviews in control schools, where teachers consistently identified a lack of on-site science facilities as a major concern). In rehabilitated schools, science teachers felt that they now can create more opportunities for students' active involvement in class assignments and discussion. These teachers noticed that hands-on learning opportunities have led to better student learning and have helped promote cooperation and soft skills. In addition, they have identified new opportunities to give students more difficult or complex assignments, and students are willing to pursue challenging academic tasks. Some science teachers also reported that they felt more motivated to improve instruction after the investment in new laboratories; they felt more engaged with their work, and more valued in their role as teachers.

In several focus groups, however, students reported that there was a need for more laboratory instructors or materials for experiments to support lab assignments. In one treatment school focus group, students noted that they had not used the lab as much as they had hoped, because their school did not have a lab specialist to lead instructional activities that made full use of the lab's equipment. In another school, students noted they rarely use the chemistry lab because they lack the chemicals and other raw materials needed to conduct experiments. These findings are consistent with the survey data showing that science laboratories appear to be underutilized in approximately half of schools.

4. Perceptions of school safety

Improvements in school safety could help make students and parents more willing to persist in schools to the end of secondary school, as well as increase the motivation and professional satisfaction levels of teachers and school directors. After rehabilitation, we observed a strong pattern of improvements in perceptions of school safety among students, their parents, teachers, and school directors. At baseline, only about half of students and parents believed that their overall school, school classrooms, or school stairwells were safe; after rehabilitation, the percentage of students and parents reporting that school facilities were safe increased to nearly 90 percent for all three measures (Table III.6). We also observed improvements in the view of safety among teachers (with perceptions of safety rising above 95 percent after rehabilitation).

Similarly, the percentage of school directors who believed that their school was safe rose from 59 to 100 percent.

Table III.6. Comparison of perceptions of school safety between baseline and one-year follow-up

	Baseline mean	Follow-up mean	Difference	*p*-value	Baseline N	Follow-up N
Students						
Agrees that the school is safe and healthy	0.47	0.86	0.39**	0.00	1,614	1,470
Feels very safe in the classroom	0.49	0.86	0.38**	0.00	1,643	1,509
Feels very safe using stairwells	0.49	0.89	0.41**	0.00	1,618	1,488
Parents						
Agrees that the school is safe and healthy	0.66	0.96	0.29**	0.00	1,498	1,413
Feels that students are very safe in the classroom	0.46	0.90	0.43**	0.00	1,482	1,401
Feels that stairwells are very safe	0.44	0.89	0.45**	0.00	1,384	1,326
Teachers						
Agrees that the school is safe	0.71	0.96	0.25**	0.00	228	235
Agrees that the school is healthy	0.77	0.96	0.20**	0.00	226	234
Feels very safe in the classroom	0.76	1.00	0.24**	0.00	225	235
Feels that students are very safe in the classroom	0.76	0.99	0.23**	0.00	220	235
Feels very safe using stairwells	0.55	0.98	0.44**	0.00	225	235
Feels that students are very safe using stairwells	0.66	1.00	0.34**	0.00	217	234
School directors						
Agrees that the school is safe	0.59	1.00	0.41**	0.00	29	29

Notes: Differences between baseline and follow-up means and *p*-values of those differences were estimated using multivariate ordinary least squares regressions of a one-year follow-up survey indicator on each measure. The regressions included indicator controls for each school (not reported). Standard errors were clustered at the school level. Follow-up means were regression adjusted (estimated by adding the baseline mean to the regression-estimated difference). The reported means and differences were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

D. Changes in enrollment and school administration

In addition to changing perceptions about the school and patterns of time use for students and teachers, school rehabilitation may have important implications for school operations. If rehabilitation produced changes in total enrollment levels (by attracting more students to the school or increasing the number of students persisting to higher grade levels), this may have had implications for the school's teacher-student ratios, staffing, and budgets. In addition, rehabilitated schools may have experienced substantial changes in the costs of operations and maintenance (such as expenses related to school heating, water use, and maintenance costs for school building repairs). We examined these issues using administrative data and the study's school director survey.

1. Changes in enrollment

For schools rehabilitated in 2016, we used longitudinal administrative data to investigate how student enrollment changed in the years before and after rehabilitation. School rehabilitation may have affected enrollment levels for several reasons: for example, schools with better conditions may have attracted new students from surrounding schools, or caused students to persist further in their secondary education instead of dropping out of school and entering the workforce. Here we examine these issues descriptively using longitudinal administrative data—as part of the evaluation's final report, we will use these administrative data sources to directly estimate the impacts of rehabilitation on enrollment outcomes and on persistence to graduation from secondary school in all rehabilitated schools. For this interim analysis, we focus only on the 12 schools that were rehabilitated in 2016, because the school years covered in the enrollment data only included one post-rehabilitation year for the 17 schools rehabilitated in 2017 (limiting the extent to which we could explore patterns of entering and exiting students).

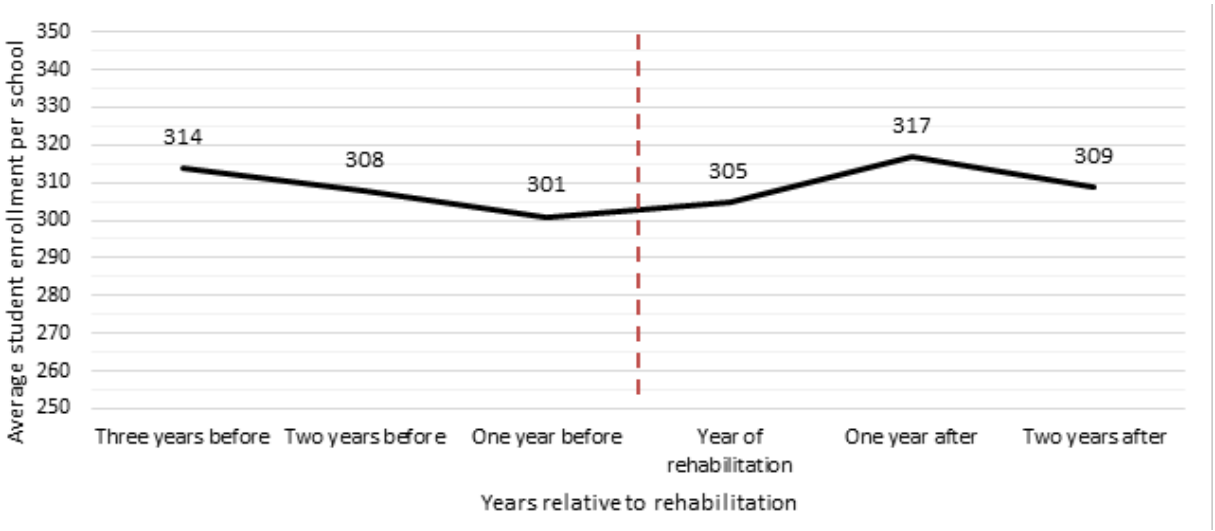
For schools rehabilitated in 2016, we did not find a strong pattern of enrollment increases but it appears that rehabilitation helped to reverse a prior trend of declining enrollments. At these schools, there was a pattern of declining enrollment in the years before rehabilitation, followed by small increases in enrollment after rehabilitation (Figure III.13). The average number of students enrolled in the schools rehabilitated in 2016 was declining by around six students per year before rehabilitation; after rehabilitation, average enrollment levels began to level off or increase.

These changes in enrollment were relatively modest. For example, the change from the lowest observed enrollment year in these schools (an average of 301 students, in the year before rehabilitation) to the highest observed year (317 students, in the year after rehabilitation) represents an increase of only 5 percent in average enrollment levels. The observed enrollment level two years after rehabilitation is similar to the level of enrollment at these schools three years before rehabilitation took place. In other words, these changes did not increase building utilization rates beyond the levels the schools had experienced in the recent past.

Changes in total enrollment could be driven by general demographic changes across student cohorts, changes in student enrollment patterns (more students entering the school in early grades or transferring in from other schools) or by changes in the patterns of exiting students (fewer students dropping out, transferring out to other schools, or graduating). To examine these patterns, in each school year, we gathered data on (1) the average number of students who entered the rehabilitated schools and (2) the average number of students who exited the rehabilitated schools before the school year started.

The change in enrollment patterns at rehabilitated schools appears to be predominantly driven by new enrollments, rather than changes in school dropout rates or graduation rates (Table III.7). After rehabilitation, the average number of new students entering the school each year increased by more than 50 percent (from 30 to 46 students per school). This increase was driven primarily by more students entering in grades 1 and 2, meaning that rehabilitation appears to have influenced more parents of young students to select the rehabilitated schools over alternatives in their region. Changes in the pattern of exiting students were minor: there were only small shifts in the number of students who dropped out, transferred out, or graduated.

Figure III.13. Average annual student enrollment in schools rehabilitated in 2016



Source: Georgia education management information system (EMIS) enrollment administrative data (2013–2014 through 2018–2019 school years).

Notes: Sample included total enrollment for 12 schools rehabbed in 2016. The red dashed line indicates the transition from the pre-rehabilitation period to the year that rehabilitation began.

Table III.7. Average annual changes in student enrollment in schools rehabilitated in 2016

	Average number of students per school: Entered rehabilitated schools[a]			Average number of students per school: Left rehabilitated schools[b]				Average number of students per school: Net change in total enrollment
	Total	Grades 1 and 2	Grades 3 through 12	Total	Dropped out	Trans-ferred out	Graduated	
Two years before rehabilitation	33	25	8	38	7	7	24	-5
One year before rehabilitation	30	24	6	37	6	8	23	-7
School year of rehabilitation	46	34	12	42	7	9	26	5
One year after rehabilitation	45	36	10	34	4	8	23	12

Note: Samples included each student enrolled in schools rehabilitated in 2016 in each school year between 2013–2014 and 2018–2019 in Georgia's education management information system (EMIS) administrative data. EMIS administrative data included measures of students' status in previous school year (excluding the 2013–2014 and 2018–2019 school years) and students' status in subsequent school year (excluding the 2018–2019 school year).

[a] Students who "entered rehabilitated schools" include (1) students who entered school for the first time, (2) students who had dropped out in the past and returned to school, and (3) students who transferred into a rehabilitated school.

[b] Students who "left rehabilitated schools" were enrolled in a rehabilitated school in the previous school year but left because they (1) dropped out of school, (2) transferred to a school outside of the 2016 and 2017 rehabilitation schools, or (3) graduated from senior secondary school.

2. School operations and maintenance

Although enrollment patterns did not appear to change dramatically at rehabilitated schools, renovated buildings did produce large changes in the costs of operating and maintaining school infrastructure. In particular, the costs of utilities increased substantially after rehabilitation, and these costs were a widespread concern among school directors. As Table III.8 shows, by far the largest utility expense in winter months is for heating (making up 89 percent of total costs at baseline), followed by electricity (8 percent) and water (3 percent). After rehabilitation, the cost of each of these utilities roughly tripled (with statistically significant increases for both heating and electricity), suggesting that these utilities were being used much more extensively after rehabilitation.

Most directors reported that it is difficult to find funds to pay for these increased utility costs. After rehabilitation, 55 percent of school directors reported that they are "never able to fully pay for school utilities with available school budget," and an additional 17 and 14 percent reported "rarely" or "sometimes" being able to pay (Figure III.14). Similarly, approximately two-thirds of school directors reported that their overall school budget was insufficient to cover both utilities and maintenance costs, in addition to costs related to teaching activities (like teachers' salaries).

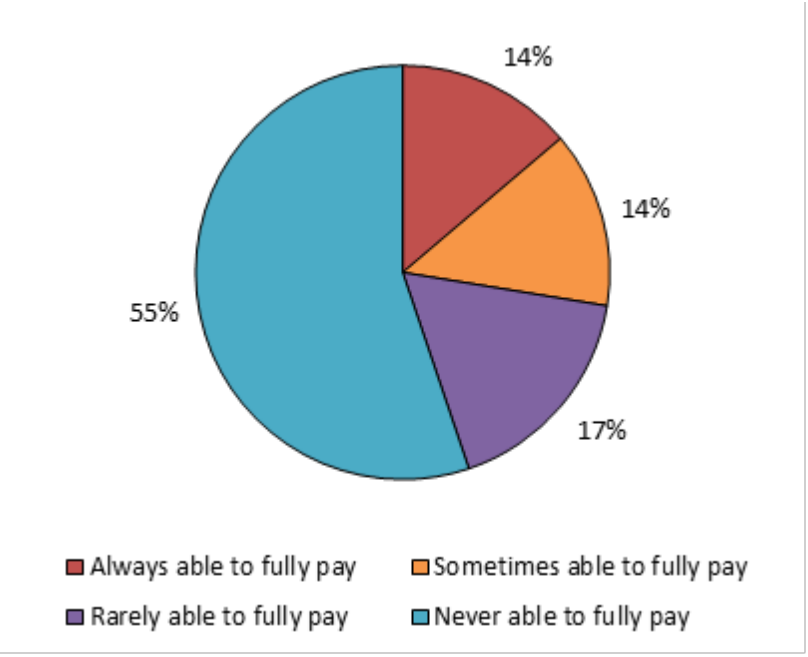
Table III.8. Change in costs incurred by rehabilitated schools between baseline and one-year follow-up

	Baseline mean	Follow-up mean	Difference	*p*-value	Baseline N	Follow-up N
Heating costs for the month of February (in Georgian lari)	1744	4761	3017**	0.00	29	29
Electricity costs for the month of February (in Georgian lari)	168	492	324**	0.00	29	29
Water costs for the month of February (in Georgian lari)	52	170	119	0.22	27	29

Notes: Differences between control and treatment means and *p*-values of those differences were estimated using multivariate ordinary least squares regressions of treatment status on each measure of baseline infrastructure. The regressions included indicator controls for the probability of selection in the intervention group that was assigned to the randomization strata of each school (not reported). Standard errors were clustered at school level. Follow-up means were regression adjusted (estimated by adding the baseline mean to the regression-estimated difference).

**/* indicates that differences were significant at the 1/5 percent levels.

Figure III.14. Percentage of school directors able to fully pay for school utilities after rehabilitation



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence School Director Surveys (2017, 2018).

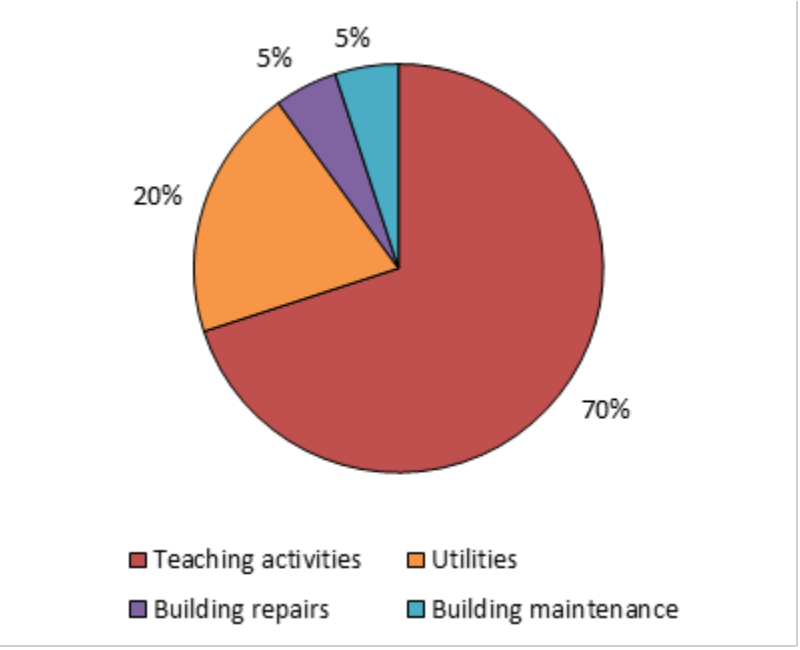
Notes: Sample included 29 school directors interviewed at one-year follow-up in schools rehabilitated in 2016 and 2017.

Although most directors reported that they face at least periodic shortfalls in paying for utilities, only 7 percent of directors reported ever turning off utilities (such as heating systems) in February to help reduce these costs (not shown). This was despite the fact that most directors said that they prioritized other aspects of the school budget above utility expenses. Among the directors who said that their overall budgets were insufficient to cover all of their costs, only 20 percent reported that utilities were their highest budget priority (Figure III.15). The school budget priority ranked most highly was teaching activities (selected by 70 percent of directors), followed by utilities (20 percent) and building repairs or maintenance (10 percent).

In qualitative interviews (conducted in a subset of rehabilitated interviews), directors reported a similar pattern of significant challenges related to the costs of operating renovated heating systems. Many school directors at rehabilitated schools highlighted concerns that renovated heating systems had generated increased operating costs for schools during winter months. These directors shared concerns about the trade-offs in deciding how to run the renovated heating system at their schools. For example, one director emphasized that, for the system to operate smoothly, it was now important for both the electrical and heating systems to run continuously on a 24-hour basis. This school director experimented with turning the heating off after the school day to save on costs, but felt that it was necessary to keep the temperature constantly warm to afford students and teachers the benefit of warm classrooms early in the morning and throughout the day. In other schools, directors delayed turning on the heating system in the fall, to reduce costs. In one treatment school focus group, for example, students

stated that they would like the heating system to be turned on earlier in the year. At this school, the heating system was turned on in mid-November, but students hoped it could be operated earlier in the year when the temperature started dropping.

Figure III.15. Highest spending priority for school directors facing budget shortfalls



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence School Director Surveys (2017, 2018).

Notes: Sample included 20 school directors who reported that school budget was not sufficient to cover utilities and maintenance at one-year follow-up in schools rehabilitated in 2016 and 2017.

The budget shortfalls related to utility costs at rehabilitated schools could have longer-term consequences. For example, if directors cannot keep up with utility payments, they may face cost pressures to avoid using rehabilitated heating, water, or electricity systems as intensively in the future. We plan to investigate the longer-term consequences of these challenges in the second year after rehabilitation (as well as the extent to which this pattern in Phase I schools is repeated at Phase II and Phase III schools) as part of the evaluation's final report.

This page has been left blank for double-sided copying.

IV. INTERIM FINDINGS FOR THE TEE EVALUATION

A. Implementation of the TEE training initiative

The TEE training initiative was nationwide in scope, aiming to train all directors of schools offering secondary grades and all of Georgia's grade 7-12 teachers in the subjects of science, mathematics, English, and geography. The Project used a phased implementation schedule that rolled out the training to multiple cohorts over three years: the initiative included two cohorts of school directors, two cohorts of SPDFs, and three cohorts of teachers. Table IV.1 presents a description of the targeted trainees and training years for each cohort under the original program design.

The first cohort of directors and SPDFs received training in the "Leadership Academy" training sequence from 2016 through 2018. In this first cohort, training participants included one school director and one SPDF from all 1,872 Georgian-language schools in the country (totaling 3,744 trainees). To accommodate the linguistic needs of the directors and SPDFs from the 213 minority-language schools in Georgia—that is, schools using Azeri, Armenian, or Russian as the primary language of instruction—the program also conducted training for a second cohort of minority-language directors and SPDFs starting in 2017.

Table IV.1. TEE activity participants and training schedule

Cohort	Types of trainees	Training period	Number of targeted trainees
Leadership Academy (school directors and SPDFs)			
Cohort 1	Georgian-language trainees	2016–2018	1,872 school directors
			1,872 SPDFs
Cohort 2	Minority-language trainees	2017–2018	213 school directors
			213 SPDFs
Teacher training			
Cohort 1	Georgian-language senior, lead, and mentor teachers	2016–2017	5,261 senior, lead, and mentor teachers
	Some Georgian-language practitioner teachers		3,768 practitioner teachers
Cohort 2	Remaining Georgian-language practitioner teachers	2017–2018	7,016 practitioner teachers
Cohort 3	Minority-language teachers	2018–2019	69 senior, lead, and mentor teachers
			2,108 practitioner teachers

Source: IREX (2016a) and IREX (2016b).

Note: The 2,177 minority-language teachers targeted for Cohort 3 of the teacher training consisted of 912 Azeri-speaking teachers, 904 Armenian-speaking teachers, and 361 Russian-speaking teachers. SPDFs = school professional development facilitators.

The implementers similarly split the trainees for the teacher training intervention into two Georgian-language cohorts and a third minority-language cohort. The program separated the Georgian-language trainees into two cohorts to accommodate the large number of Georgian-language teachers. In addition, program implementers targeted teachers with more seniority in the first cohort, in an effort to group teachers with similar skill levels together and provide an opportunity for more senior teachers to accumulate professional development credits. According

to the Georgian government's professional development scheme for teachers, more senior teachers are classified based on their ability to pass a certification exam for their teaching subject. Upon passing, these "senior," "lead," or "mentor" teachers are eligible to earn promotions and salary increases based on their number of accumulated professional development credits, including the type of credits offered to teachers who completed the TEE training sequence. In total, the program targeted all 5,261 senior, lead, and mentor teachers in Georgian-language schools for inclusion in the first cohort.

Teachers with lower levels of seniority ("practitioner" teachers) were split between the two Georgian-language cohorts—3,768 practitioner teachers were targeted in the first cohort, and the remaining 7,016 practitioner teachers were targeted in the second cohort. These teachers had not passed the government's certification examination in their subject at the time the TEE activity began, meaning they were not eligible to receive increases in salary immediately upon earning professional development credits (in other words, they had a weaker incentive to attend the TEE trainings). Since practitioner teachers (with an average age of 52) are actually older than teachers who have passed the certification exam (more senior teachers have an average age of 46), the first cohort of trainees was both younger and more likely to have a strong grasp of their teaching subject than teachers in the second cohort.

The third cohort of minority-language teachers targeted a total of 69 senior, lead, and mentor teachers and 2,108 practitioner teachers. The first cohort of Georgian teachers was trained between fall 2016 and fall 2017. The second cohort was trained between fall 2017 and fall 2018. The third cohort completed its training between fall 2018 and fall 2019.

B. Training attendance and completion

Because of the TEE activity's ambitious nationwide rollout schedule, one of the evaluation's most important implementation questions was whether it would be possible to manage the logistics of attracting a nationwide pool of thousands of educators to complete the full training sequence. This sequence included multiple training modules scheduled over the course of an entire year for teachers, and two years for school directors. We used survey data to independently measure attendance patterns at the trainings for school directors and teachers, and explored what motivated participants to attend (or not attend) the training in greater depth by using survey instruments, qualitative interviews (in the case of directors), and focus groups (in the case of teachers).

1. Survey data

Both rounds of the TEE evaluation survey collected data about school director and teacher participation in the TEE training sequence. Survey data demonstrate that attendance among the first cohort of school directors was high. During the first year of training, 98 percent of school directors attended at least one training module. Rates of attendance were also high (96 percent or greater) for each of the five specific training modules they received over the course of two years. By the time of the follow-up survey in fall 2018, after the end of the training sequence, 93 percent of school directors had attended all five of the training modules.

Although a similarly high proportion of teachers reported in surveys that they attended at least one of the training modules, the training completion rate for teachers was lower than the training completion rate for school directors. Table IV.2 shows training attendance rates for teachers in both the Year 1 and Year 2 follow-up surveys. By the end of the first round of teacher trainings (the training sequence offered to the first cohort of teachers), 89 percent of teachers reported attending at least one of the training modules; however, only 64 percent attended all four of the training modules in the sequence. The attendance rate for each of the individual modules ranged from 75 to 81 percent. After the second year of teacher training (assessed in the study's fall 2018 survey round), we observed somewhat lower attendance rates for teachers in the second cohort: 55 percent of teachers in the second cohort completed the training sequence. However, by fall 2018 the training completion rate had increased substantially for teachers in the first cohort, from 64 to 82 percent. This likely occurred because teachers in the first cohort were given an opportunity to join teachers in the second cohort and attend any training modules they may have missed during the first training round.

In addition to the self-reported training attendance rates collected in the study's TEE evaluation surveys, we also examined administrative attendance data provided by TPDC. The attendance rates for the first cohort of teachers recorded in the TPDC data were similar to those estimated using the study's survey data, suggesting that the two data sources are likely to be reasonably accurate.

Table IV.2. Teacher attendance rates in TEE training modules

			Core modules				
	All modules	Any modules	All core modules	Module 1	Module 2	Module 3	Subject modules[a]
Cohort 1							
Survey data							
After 2016-2017 round of training	64%	89%	70%	80%	83%	78%	77%
After 2017-2018 round of training	82%	95%	86%	93%	92%	87%	88%
TPDC administrative data							
After first round of training	67%	89%	72%	82%	82%	81%	76%
Cohort 2							
Survey data							
After 2017-2018 round of training	55%	78%	60%	73%	67%	62%	68%

Note: Samples included 1,186 teachers from the survey data and 16,147 teachers from the TPDC administrative data.

[a]The subject modules included separate modules for science (chemistry, biology, and physics); mathematics; English; and geography teachers.

As part of the teacher survey, we asked respondents about their motivations for attending (or not attending) the TEE training modules (Table IV.3). Nearly all of the teachers who attended training modules believed that the training would improve their teaching practice. Over half of the attendees also reported that earning professional development credits motivated their attendance as well. In contrast, fewer than half of the attendees reported attending because they

had confidence in advance that the quality of the training would be good. Further, fewer than 10 percent reported that the school director or Ministry of Education requirements were among the reasons they decided to attend. The reasons teachers gave for attending training modules were similar across both training rounds, including for teachers in the first cohort who attended modules in the second training round.

Among the relatively small number of teachers who did not attend any trainings, the most commonly reported reasons for opting out were competing obligations or personal issues, including health or family issues. In addition, other less common reasons cited by some teachers included a perception that they had not been invited, or a belief that the training would not improve their teaching practices. Few teachers reported that nonattendance was due to a lack of training requirements, not earning enough professional development credits, or believing that the quality of the training would be poor.

Table IV.3. Reasons for attending or missing TEE training modules

		After second training round	
	After first training round All teachers in Cohort 1	Cohort 1 teachers who did not complete training in first round	All teachers in Cohort 2
Reasons attended any training modules (n=1,245)			
School director required teacher to attend	7%	5%	5%
Ministry of Education required teacher to attend	8%	3%	3%
To earn professional development credits	58%	62%	59%
Believed quality of training would be good	37%	36%	35%
Believed training would improve teaching practices	89%	88%	89%
Reasons did not attend any training modules (n=147)			
Was not invited to attend training	25%	12%	31%
Was not required to attend training	1%	0%	0%
Would not earn enough professional development credits	2%	3%	0%
Believed training would be too time-consuming	8%	6%	0%
Believed quality of training would be poor	2%	0%	0%
Believed training would not improve teaching practices	8%	15%	10%
Prevented by other obligations or personal issues[a]	53%	48%	48%
Time or location was inconvenient[a]	5%	3%	10%
Believed they were too old to attend[a]	1%	9%	0%
Participated in training in the past[a]	0%	6%	3%
Other reasons[a]	2%	0%	0%

Note: Samples included 777 Cohort 1 teachers who attended training modules in the first round, 234 Cohort 1 and 234 Cohort 2 teachers who attended training modules in the second round, 85 Cohort 1 teachers who did not attend any training modules in the first round, 33 Cohort 1 teachers who did not attend any training modules in the first or second round, and 29 Cohort 2 teachers who did not attend any training modules in the second round. TEE = Training Educators for Excellence.

[a]Categorized from open "other reason" responses provided by the respondents.

2. Findings from qualitative focus groups about training attendance and completion

Teachers who participated in focus groups stated that they were motivated to participate in the TEE training because it was a valuable asset for their professional development trajectory. During the focus groups, teachers highlighted the value of the credits earned by participating in training. They were motivated not only by the knowledge gained through the training modules, but also by the credits they earned by completing the training and using the training materials. This suggests that the TEE training was well aligned with two sources of teachers' motivation: (1) improving teaching skills and (2) earning credits for career advancement. The ability to earn credits appears to have served as an incentive for teachers' completion of the training and use of materials thereafter. Although teachers did not raise the issue of partial training attendance explicitly, they did bring up some challenges that might have affected their ability to complete the training. Commuting time to training locations, as well as the amount of time required to attend the training sessions, read materials, and prepare assignments might have deterred some teachers from completing the training requirements.

While teachers' perceptions of the TEE training were generally positive, some teachers cited reasons why they were not motivated to complete the full training sequence. Most teachers who participated in focus groups viewed the training as high quality and useful in helping teachers improve their teaching practice. According to teachers who participated in the focus groups, the TEE training (or "Millennium Training," as they referred to it) offered them an opportunity to learn about new teaching methods that they had not been exposed to in the past. Most teachers felt all the training modules were informative. For some teachers, the use of technology and pedagogical resources, lesson planning, and subject-specific content were particularly useful. Overall, teachers expressed satisfaction with the quality of the trainers and the content of the training modules. Although most teachers were pleased with the pedagogical content of the training, a few teachers expressed dissatisfaction with some aspects of the training—namely, that (1) commuting to training locations placed an undue burden on their time, (2) the length of training sessions exceeded their attention span and ability to absorb information, (3) much of the training content was not differentiated by subject area and some methods were not relevant to all subject areas, and (4) some topics were redundant with information that teachers felt they already knew.

C. Teacher knowledge and practices after training

Our performance evaluation of the teacher training initiative relies primarily on descriptive analyses that use teacher survey data to measure teacher knowledge and practices in the period after the training sequence was completed. The goal of the study is to assess whether the self-reported practices among teachers are consistent with the program's theory of change—which states that the training will improve teacher knowledge of inclusive, student-centered instruction methods (both in general and specifically in English, math, geography, and science), which will then (over a period of several years) improve teachers' classroom instruction in ways that can ultimately improve students' learning outcomes.

Because there are limitations to self-reported teacher survey data—teachers may not accurately report all of the practices they actually use—the study sought to examine these outcomes in multiple ways. First, we examine data from teachers in the first year after completing the training sequence, to measure the extent to which teachers are using the types of

practices encouraged by the program. Next, we present the results of a comparison group analysis that more directly estimates the near-term effects of training by contrasting a similar group of trained teachers (immediately after finishing the training sequence) and untrained teachers who had not begun the training sequence. Finally, we explore the potential mechanisms behind the pattern of observed survey outcomes, using qualitative data from focus groups of trained teachers.

1. First year follow-up survey data

First, we investigated the extent to which trained teachers are using the types of practices encouraged by the TEE training sequence. We assessed this using survey data from 1,186 teachers in the fall of the school year after the training sequence ended for their cohort. The characteristics of teachers in the survey sample are summarized in Chapter II (see Table II.5). We report these post-training results separately for Cohort 1 and Cohort 2, for two reasons. First, in our survey sample (and the overall training program) the two cohorts differed substantially in terms of the seniority levels of trained teachers (35 percent of Cohort 1 teachers were more junior, practitioner-level teachers, while 90 percent of Cohort 2 teachers were practitioners), and responses to the training material may have been different among more senior teachers. Second, because the training for Cohort 2 teachers took place in the second year of implementation, it is possible that a more experienced cadre of implementers and trainers may have been more effective in delivering the training material for the second cohort. Examining the two cohorts separately provides some descriptive evidence about both of these potential patterns.

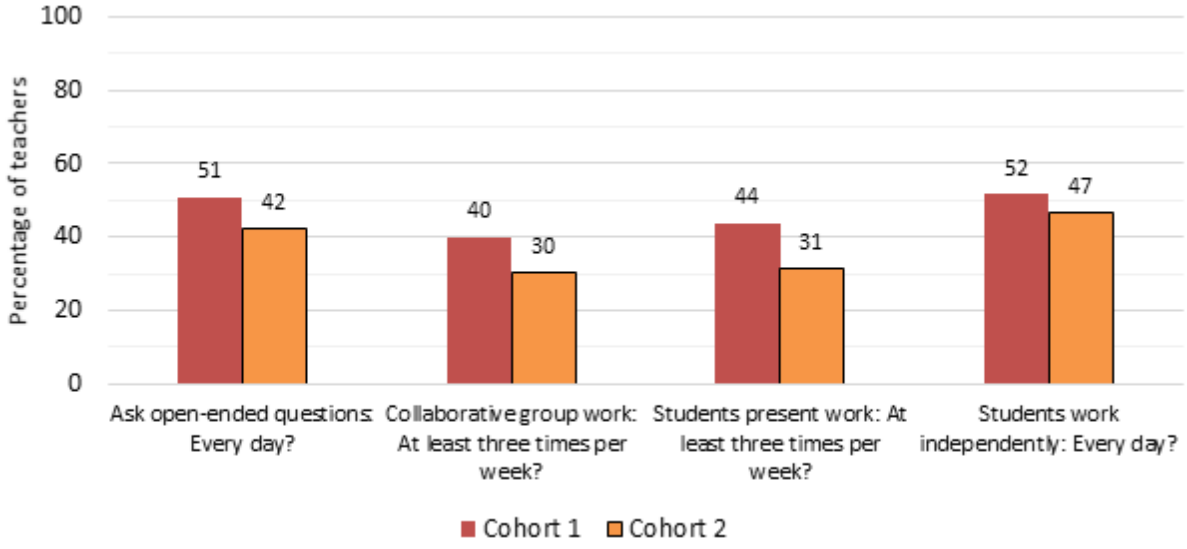
As we discuss below, the teacher survey data should be interpreted with caution because the teachers may not have accurately reported their practices in all cases. In addition, the program's theory of change did not predict that teaching practices would change in the immediate aftermath of the training sequence. Indeed, among the teachers in the survey sample and across both cohorts, we found evidence that many of the practices that were encouraged by the TEE training sequence were only being applied to a limited extent in classrooms. According to teacher survey data, most of the teaching practices that were emphasized by the TEE training modules were being used by less than half of the teachers in the study sample on a consistent basis, in the first year after the training sequence ended.

For example, roughly half of the teachers reported that they were consistently using practices related to students' critical thinking skills, motivation, and collaboration (Figure IV.1). Specifically, about half of the teachers in both cohorts reported asking open-ended questions every day, while between 30 and 40 percent of teachers reported using collaborative group work at least three days per week. Fewer than half reported that students present work in class at least three days per week, and roughly half of teachers reported that their students work independently on a daily basis.

Tailoring lesson plans and teaching practices to meet students' individual needs, and assessing student learning on an ongoing basis (through formative assessments) were both important components of the TEE training. However, both types of activities remained uncommon among teachers in their day-to-day teaching (Figure IV.2). Ten percent of teachers reported developing lesson plans that include differentiated instruction on a daily basis, only around 20 percent of teachers reported working with struggling students every day. Nearly 40 percent of teachers reported preparing lesson plans to (1) achieve specific learning goals every

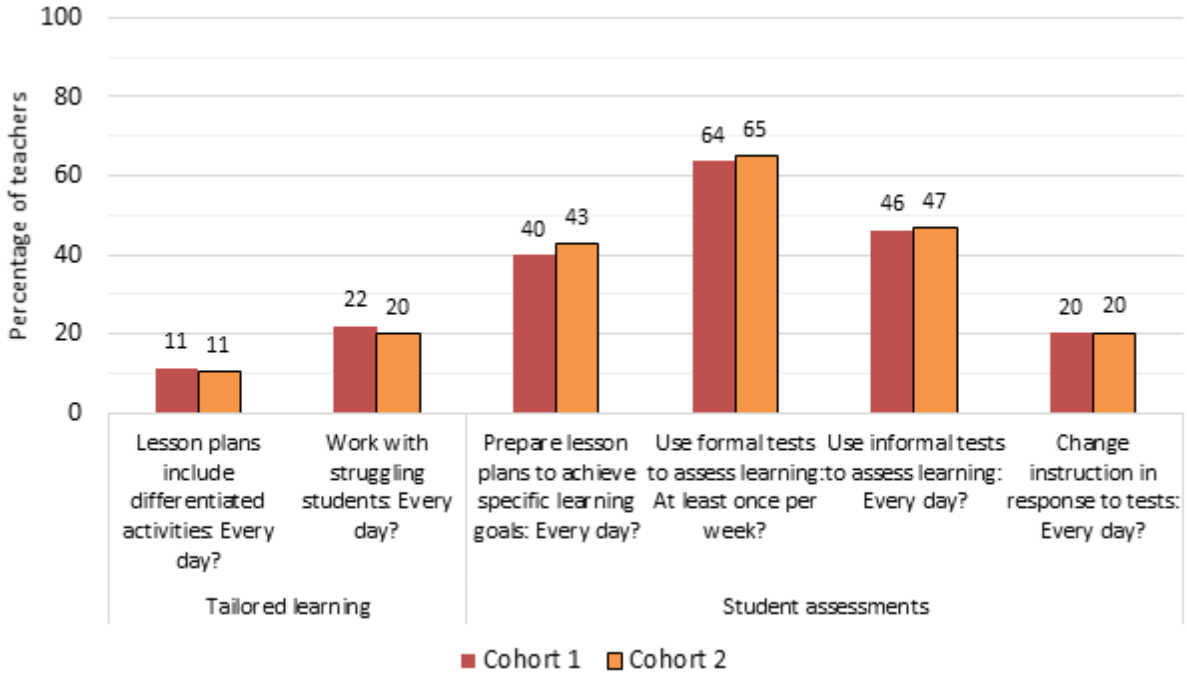
day and (2) use informal testing to assess student learning every day. And while over 40 percent of teachers reported using informal testing to assess student learning at least once per week, only 20 percent of teachers reported changing their lesson plans based on the results of formal and informal assessments.

Figure IV.1. Teaching practices related to students' critical thinking, motivation, and collaboration, as reported in the first follow-up survey



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Teacher Surveys (2017).
Note: Sample included 877 Cohort 1 teachers and 223 Cohort 2 teachers.

Figure IV.2. Teaching practices related to tailored learning and assessing student learning, as reported in the first follow-up survey

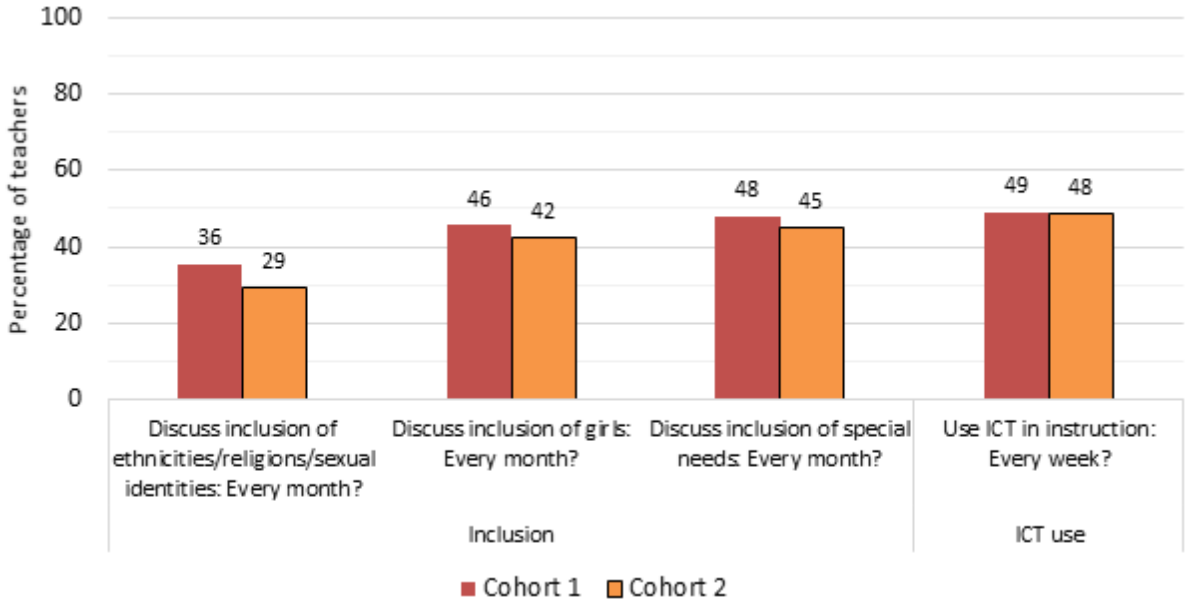


Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Teacher Surveys (2017).
Note: Sample included 877 Cohort 1 teachers and 223 Cohort 2 teachers.

Two additional areas emphasized by the TEE trainings were (1) encouraging the use of information and communications technology (ICT) and (2) instructional inclusion for disadvantaged groups. Surveyed teachers also demonstrated that there was room for improvement in these areas. Fewer than half of the teachers reported discussing inclusion with students every month or using ICT in instruction every day. Between 30 and 50 percent of teachers in both cohorts reported holding discussions with students about inclusion of students with different ethnicities, religions, or sexual identities; female students; or students with special needs (Figure IV.3). In addition, the percentage of teachers who reported using ICT during instruction every week was slightly below 50 percent for both cohorts of trainees.

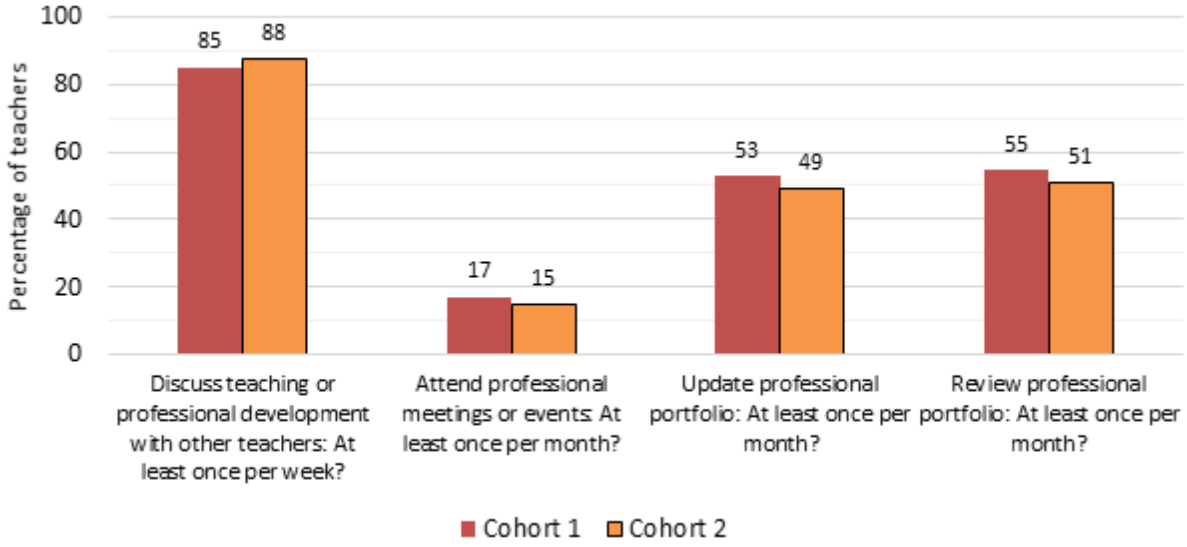
Holding substantive discussions with fellow teachers and, to a lesser extent, using professional portfolios were common ways that teachers reported managing their professional development—practices that were encouraged and supported through the TEE training initiative. Although over 80 percent of teachers reported discussing teaching practices or professional development with other teachers at least once per week, fewer than 20 percent reported attending a professional meeting or event in the last month (Figure IV.4). However, about half of the teachers reported reviewing and updating their professional portfolio on a monthly basis.

Figure IV.3. Teaching practices related to inclusion of female and minority students and ICT use in instruction, as reported in the first follow-up survey



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Teacher Surveys (2017).
Note: Sample included 877 Cohort 1 teachers and 223 Cohort 2 teachers.

Figure IV.4. Practices related to teachers' professional development, as reported in the first follow-up survey

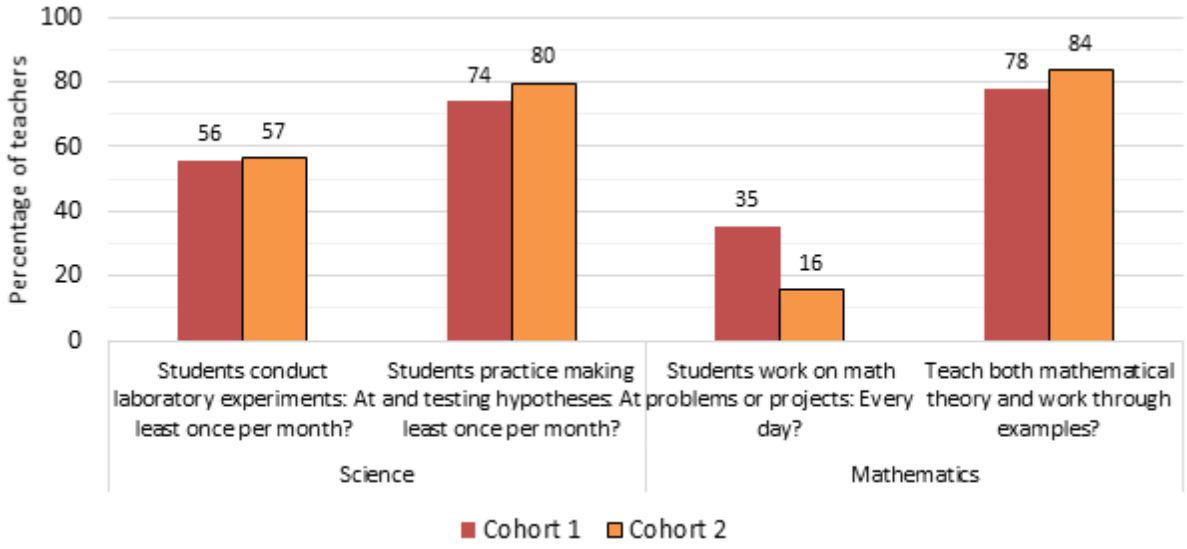


Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Teacher Surveys (2017).
Note: Sample included 877 Cohort 1 teachers and 223 Cohort 2 teachers.

Teachers reported that they commonly used many of the subject-specific teaching practices promoted by the TEE subject trainings. Over 50 percent of science teachers reported that students conducted lab experiments and over 70 percent reported that students made and tested hypotheses at least once per month (Figure IV.5). Among math teachers, about 80 percent reported teaching both theory and working through example problems. In addition, they reported that 24 percent of class time on average was spent teaching theory (not shown). However, only 35 percent of Cohort 1 and 16 percent of Cohort 2 math teachers reported that students worked on math problems or projects every day (this difference between the two cohorts is not statistically significant, due to the smaller sample size in the subject-level subgroups).

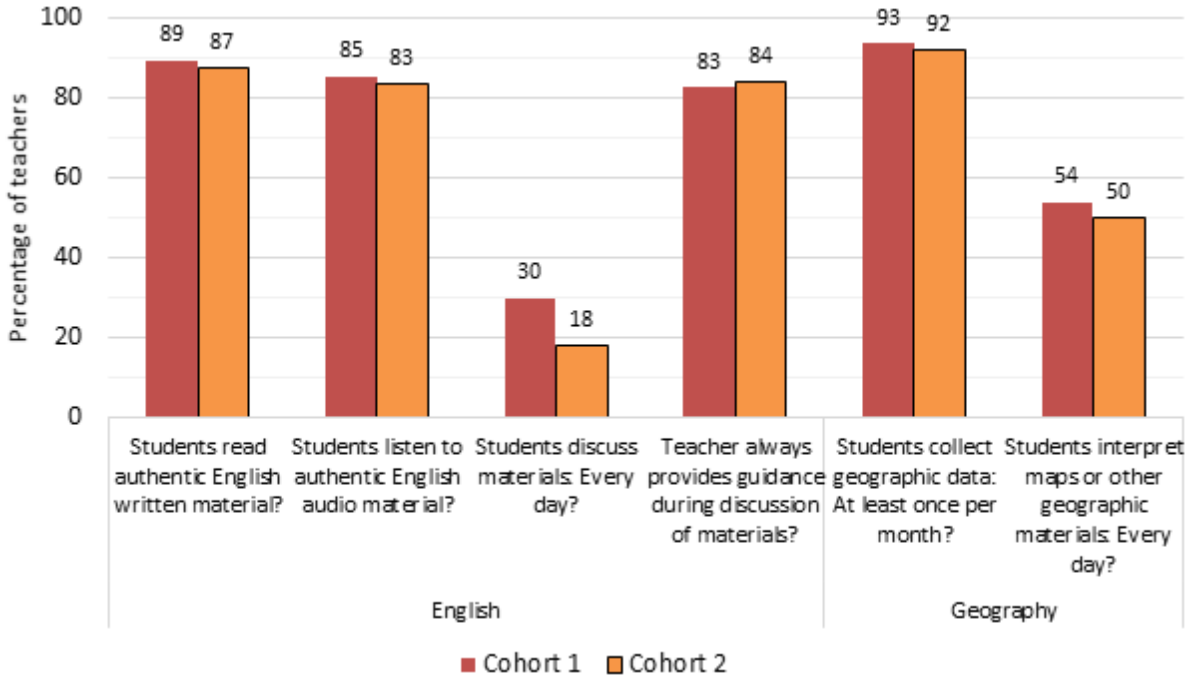
The reported use of authentic English materials in English classes was quite high in the first follow-up survey. Between 87 and 89 percent of English teachers reported that their students read authentic written materials, while between 83 and 85 percent reported that their students listened to authentic audio materials (Figure IV.6). However, the reported frequency of student discussion of these materials was fairly low: only 30 percent of Cohort 1 English teachers and 18 percent of Cohort 2 English teachers reported that students discussed materials every day. For geography, over 90 percent of students reported collecting geographic data at least once per month, while about half of the geography teachers reported that their students interpreted maps or other geographic data every day.

Figure IV.5. Teaching practices related to science lessons and mathematics lessons, as reported in the first follow-up survey



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Teacher Surveys (2017).
Note: Sample included 606 to 614 science teachers and 468 to 478 mathematics teachers.

Figure IV.6. Teaching practices related to English lessons and geography lessons, as reported in the first follow-up survey



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Teacher Surveys (2017).
Note: Sample included 299 to 302 English teachers and 243 geography teachers.

As of the time of this interim report, we have also collected a second follow-up round of data for teachers in the first cohort. The outcomes reported by those teachers in the second follow-up survey (see Appendix D) were generally consistent with the outcomes reported in the first follow-up survey. In additional subgroup analyses (see Appendix F), we also examined the pattern of survey outcomes for different types of teachers. We found that more senior teachers used TEE-related practices at higher rates than practitioner teachers, but outcomes did not vary substantially by age-level or by whether teachers completed the final subject-specific module in the training sequence.

Because of the phased rollout of the teacher training, the second follow-up round of data is not yet available for the second cohort of teachers. In addition, the study will be collecting a third round of follow-up survey data for the Cohort 1 teachers in fall 2019. At this stage it is too early to draw final conclusions about longer-term changes in post-training outcomes among trained teachers; we will explore these trends in greater depth in the evaluation's final report.

2. Validating teacher survey responses

In addition to collecting survey data from teachers, the study also conducted classroom observations with a small sample of teachers and collected survey data from students and school directors, to assess if data from these other sources was consistent with teachers' survey responses. For the classroom observations, we used the Stallings Classroom Observation protocol to gather firsthand evidence of how teachers and students spend time in the classroom,

with a sample of 22 teachers (each observed on two separate days). The key purpose of collecting Stallings observation data was to help assess whether the information provided by the teacher surveys could be independently corroborated. In a formal survey setting, teachers may have been reluctant to report their actual practices (or found it difficult to fully understand the intent and meaning of survey items about more technical pedagogical topics, such as the difference between formal and informal learning assessments). To examine these issues, we compared the Stallings observation data to the self-reported survey data provided by the teachers whose classrooms were observed.

Broadly speaking, the Stallings observation data with this subsample of teachers showed a pattern of practices that was mixed, with some areas strong performance and other areas with room for improvement. For example, in observed classrooms teachers only rarely relied on lower-quality passive instruction practices (such as reading aloud), and much higher amounts of time were spent on active instruction techniques that allow for student participation. On the other hand, nearly a quarter of the time in observed lessons was spent on classroom management tasks rather than instruction tasks, meaning teachers could be using classroom time more efficiently. More information summarizing the classroom observation data can be found in Appendix C.

To compare the classroom observation data to survey data, we estimated the correlation between several practices measured in the Stallings protocol and survey data that captured related practices (Table IV.4). On average, in this sample we did not find strong correlations between teachers' self-reported practices and the observed practices measured by the Stallings protocol. Although it's possible that the small sample of teachers observed by the study team could differ in important ways from the full survey sample, these results suggest that survey findings using teachers' self-reported practices should be interpreted with caution, as they may not correspond strongly with actual practices for all teachers. It is also possible that the lessons observed with the Stallings protocol could have differed in important ways from the general practices teachers use on a more typical basis when their instruction is not being observed.

Table IV.4. Correlations between the Stallings observations and related TEE teacher survey responses

		First follow-up survey	Second follow-up survey
Stallings classroom observation categories	TEE teacher survey measures	Correlation coefficient	Correlation coefficient
Instruction with ICT	Frequency of presenting lessons using ICT	0.12	0.11
Students working in groups	Frequency of students participating in collaborative group work	0.01	0.02
Students working in groups	Percentage of average class day students participate in collaborative group work	-0.01	-0.14

Note: Samples included 22 teachers who were interviewed in both the first and second follow-up surveys and observed in the Stallings classroom observations. TEE = Training Teachers for Excellence. ICT = information and communications technology.

**/* indicates that differences were significant at the 1/5 percent levels.

In contrast, survey data obtained from students was more strongly consistent with the self-reported practices from teacher surveys. We collected data from students about the prevalence of

various teaching practices and compared their responses to the responses of teachers. Generally speaking, the student data confirmed findings from the self-reported teacher surveys that there was substantial room for improvement in the areas of collaborative group learning and informal assessments. The evaluation team surveyed a sample of students about instructional practices in spring 2018 (the same sample of students whose survey findings were discussed in Chapter III). A majority of these students (72 percent) reported that they participated actively in the classroom in terms of speaking to teachers in response to questions. A substantial minority (37 percent) reported that they presented their work to other students at least once per week (Table IV.5). Both of those findings were fairly well aligned with the amount of time spent on class-wide discussions in the Stallings observation data and with teachers' self-reported practices. However, only 6 percent of students reported that they participated in collaborative group work on a daily basis and only 16 percent of students reported that they received informal learning assessments at least three times per week. These percentages were somewhat smaller than what would have been implied by the teacher survey data alone, although there was room for improvement in the rates of these practices reported by teachers as well.

Group work and informal assessments were both emphasized as important pedagogical strategies in the TEE trainings' original logic model. However, it is clear from both the teacher and student survey data that teachers' use of these practices was not widespread in the first year after training. It remains to be seen if the practices will be more widely adopted in later years, after both teacher cohorts complete the training sequence and have time to continue developing their skills. In addition, even if there is room for further improvement it remains possible that the training did have positive effects on teachers' knowledge and use of these recommended practices. We examine those potential effects more directly in the remainder of this chapter.

Table IV.5. Classroom practices reported by students

Classroom practice	Percentage
Groups of students work together during class: Every day	6%
Students present their work to the rest of class: At least once per week	37%
Teachers lecture without students speaking: At least three times per week	12%
Students answer teachers' questions: At least three times per week	72%
Teachers call on or encourage students to speak in class: At least three times per week	51%
Students take a short test or quiz (fewer than 20 minutes): At least three times per week	16%
Students take a full-day test or quiz: Less than once per week (but more than never)	42%

Note: Sample included 2,789 students interviewed in spring 2018 as part of the school rehabilitation survey sample.

3. Potential impacts of the TEE training sequence on teachers

To explore the potential effects of the TEE trainings, we conducted propensity score matching to identify practitioner teachers in the first cohort who had similar baseline (pre-intervention) characteristics to practitioner teachers in the second cohort. At the time of the interim analysis shown below (in fall 2017, shortly after Cohort 1 completed its training sequence), the second cohort had not begun its training round. At that point in time, comparing the self-reported knowledge and practices of Cohort 1 teachers to the matched comparison group

provides descriptive evidence about the potential effects of the training program (the matching approach is discussed in more detail in Chapter II).

However, it is important to remember that at baseline this treatment and comparison group could still have differed in fundamental ways that were not captured by the matching process. Specifically, the study did not collect baseline data on the pre-training knowledge levels and instructional practices of Cohort 1 teachers. Instead, the matching algorithm was limited to a broader set of characteristics (teachers' age, gender, education, teaching subject, and seniority level, meaning all of the teachers in the matching analysis were practitioner teachers) that may not have been sufficient to identify a fully equivalent comparison group and produce causally valid impact estimates. In addition, we conducted this matching analysis for all of the Cohort 1 practitioner teachers, regardless of whether they completed the full training sequence; in other words, the analysis was only focused on the effects of receiving an invitation to attend the TEE training sequence.[7]

Among Cohort 1 practitioner teachers, we did find evidence that the training sequence increased teachers' self-reported knowledge of targeted teaching practices and self-reported confidence in using these practices. Table IV.6 presents the regression-adjusted differences between the matched sample of Cohort 1 and Cohort 2 practitioner teachers. Across all domains of TEE-relevant teaching practices, both Cohort 1 and Cohort 2 teachers were confident in their knowledge of practices related to those domains. For example, between 87 and 92 percent of Cohort 2 teachers felt confident in their knowledge of practices related to critical thinking, increasing student motivation, and increasing student collaboration. Although the baseline levels of knowledge among Cohort 2 teachers were high, we found evidence of positive impacts of the TEE training on several measures of knowledge among Cohort 1 teachers, including building higher-order thinking, promoting cooperation through group work, creating lesson plans with different activities, including formative assessments in lesson plans, and creating an equitable learning environment for girls.

Because there is good reason to assume that confidence across knowledge measures are likely highly correlated within practice domains, we conducted additional analyses using standardized indices of knowledge that we constructed to capture confidence within each domain in a single measure.[8] We found a statistically significant impact of 0.24 standard deviations on

[7] We also conducted a second analysis that examined the effects of fully completing the set of four TEE training modules (akin to a "treatment-on-the-treated" analysis) by focusing on Cohort 1 practitioner teachers who completed the full training sequence. Most of the significant differences we observed in the primary intent-to-treat analysis presented in Table IV.6 are no longer statistically significant in this analysis—with the exception of respondents feeling confident in their knowledge of how to create an equitable learning environment for girls, Cohort 1 teachers being less likely to set time aside for students to work independently every day, and Cohort 1 math teachers being more likely to have students work on math problems or projects every day. We present the full results of the treatment-on-the-treated analysis in Appendix E.

[8] To construct the indices, we used PCA to combine multiple knowledge measures within each domain into single indices. Each index is a weighted average of related knowledge measures in which the weights are aligned with measures with the highest component scores (that is, a knowledge measure that explains a greater amount of variation across teachers will receive a larger weight than measures explaining less of the variation in the sample). We further standardized the indices within the sample of teachers to *z*-scores, so each index has a mean of 0 and a standard deviation of 1. The weights for each index are shown in Appendix A, Tables A.4 through A.7.

our index of knowledge related to critical thinking, motivation, and collaboration. Differences on the other aggregate knowledge measures were not statistically significant.

In theory, increased confidence and knowledge among teachers should translate into changes in classroom teaching practices over time. This interim analysis could only examine whether there were any immediate changes in teaching practices, and program implementers designed the TEE activity to encourage changes in teaching practices over longer periods of time (future rounds of data collection will examine trends in these practices up to two years after training completion). At least in the initial month after the end of the training sequence, we did not find evidence of training impacts for most of the self-reported teaching practices measured in the survey. With the exception of updating professional portfolios at least once per month (where we observed a 10 percentage point increase) and whether English teachers always provide guidance during discussions of class materials (16 percentage point increase), we found no significant differences between the practices conducted by practitioner teachers in Cohort 1 and their matched sample of practitioners in Cohort 2. Specifically, we did not find evidence that the training affected teachers' self-reported use of (1) practices related to critical thinking, motivation, and collaboration; (2) practices related to tailoring learning to students' individual needs; (3) practices related to assessing student learning; (4) practices related to discussing inclusion with students; (5) ICT in classroom instruction; or (6) practices related to teaching math or science.[9]

Table IV.6. Matched comparison group analysis for practitioner teachers

	Trained teachers	Untrained teachers	
	One month after training	Baseline before training	Difference
Practices related to critical thinking, motivation, and collaboration			
Knowledge of related practices			
Confident in teaching to motivate and encourage?	0.96	0.92	0.03
Confident in teaching to build self-confidence?	0.95	0.91	0.04
Confident in teaching to build higher-order thinking?	0.96	0.90	0.06*
Confident in promoting cooperation through group work?	0.96	0.87	0.08**
Standardized weighted index (z-score)	0.06	-0.18	0.24*
Ask open-ended questions: Every day?	0.45	0.47	-0.02
Ask open-ended questions: Percentage of class time (p.p.)	24.5	23.0	1.5
Collaborative group work: At least three times per week?	0.38	0.36	0.02
Collaborative group work: Percentage of class time (p.p.)	32.0	34.0	-1.9
Students present work: At least three times per week?	0.39	0.39	0.01
Students present work: Percentage of class time (p.p.)	26.1	28.5	-2.4
Students work independently: Every day?	0.48	0.53	-0.06

[9] As an additional falsification test, we also estimated the differences between the matched samples on measures of *school director* instructional leadership and classroom observation. Because school directors are shared by Cohort 1 and Cohort 2 teachers, we do not expect that the teacher training would impact the school director practices experienced by the matched Cohort 1 and Cohort 2 samples. As expected, we found no significant differences in the reported director practices experienced by teachers in the matched samples.

	Trained teachers	Untrained teachers	
	One month after training	Baseline before training	Difference
Practices related to tailoring lessons to student needs			
Confident in knowledge to create a lesson plan with different tasks?	0.92	0.86	0.06*
Lesson plans include differentiated activities: Every day?	0.10	0.07	0.03
Work with struggling students: Every day?	0.22	0.17	0.05
Practices related to assessing student learning			
Knowledge of related practices			
Confident in conceptualizing measureable learning objectives?	0.92	0.89	0.03
Confident in using formative assessments during lessons?	0.96	0.92	0.04
Confident in including formative assessments in lesson plans?	0.94	0.87	0.07*
Standardized weighted index (*z*-score)	0.03	-0.05	0.08
Prep lesson plans to achieve specific learning goals: Every day?	0.40	0.43	-0.03
Use formal tests to assess learning: At least once per week?	0.64	0.64	0.00
Use informal tests to assess learning: Every day?	0.40	0.37	0.03
Change instruction in response to tests: Every day?	0.20	0.21	-0.02
Practices related to inclusion			
Knowledge of related practices			
Confident in creating equitable learning environment for girls?	0.93	0.86	0.07*
Confident in creating equitable learning environment for special needs?	0.87	0.85	0.02
Confident in creating unbiased learning environment?	0.97	0.95	0.02
Standardized weighted index (*z*-score)	0.05	-0.08	0.12
Discuss inclusion of ethnicities/religions/sexual identities: Every month?	0.39	0.36	0.02
Discuss inclusion of girls: Every month?	0.51	0.52	0.00
Discuss inclusion of special needs: Every month?	0.50	0.51	-0.01
Practices related to ICT use			
Confident in knowledge of using ICT in instruction?	0.92	0.90	0.02
Use ICT in instruction: Every week?	0.51	0.51	0.00
Practices related to professional development			
Discuss teaching/professional development with other teachers: At least once per week?	0.84	0.82	0.02
Attend professional meetings or events: At least once per month?	0.19	0.17	0.02
Update professional portfolio: At least once per month?	0.55	0.45	0.10*
Review professional portfolio: At least once per month?	0.56	0.48	0.08
Practices related to teaching science courses			
Students conduct laboratory experiments: At least once per month?	0.53	0.60	-0.07
Students practice making or testing hypotheses: At least once per month?	0.72	0.70	0.02
Practices related to teaching mathematics courses			
Students work on math problems or projects: Every day?	0.34	0.21	0.13
Teach both mathematical theory and work through examples?	0.75	0.68	0.06
Class time spent teaching mathematical theory: Percentage of class time (p.p.)	23.5	24.2	-0.7

	Trained teachers	Untrained teachers	
	One month after training	Baseline before training	Difference
Practices related to teaching English courses			
Students read authentic English written material?	0.86	0.83	0.02
Students listen to authentic English audio material?	0.79	0.75	0.03
Students discuss materials: Every day?	0.27	0.37	-0.09
Teacher always provides guidance during discussion of materials?	0.86	0.70	0.16*

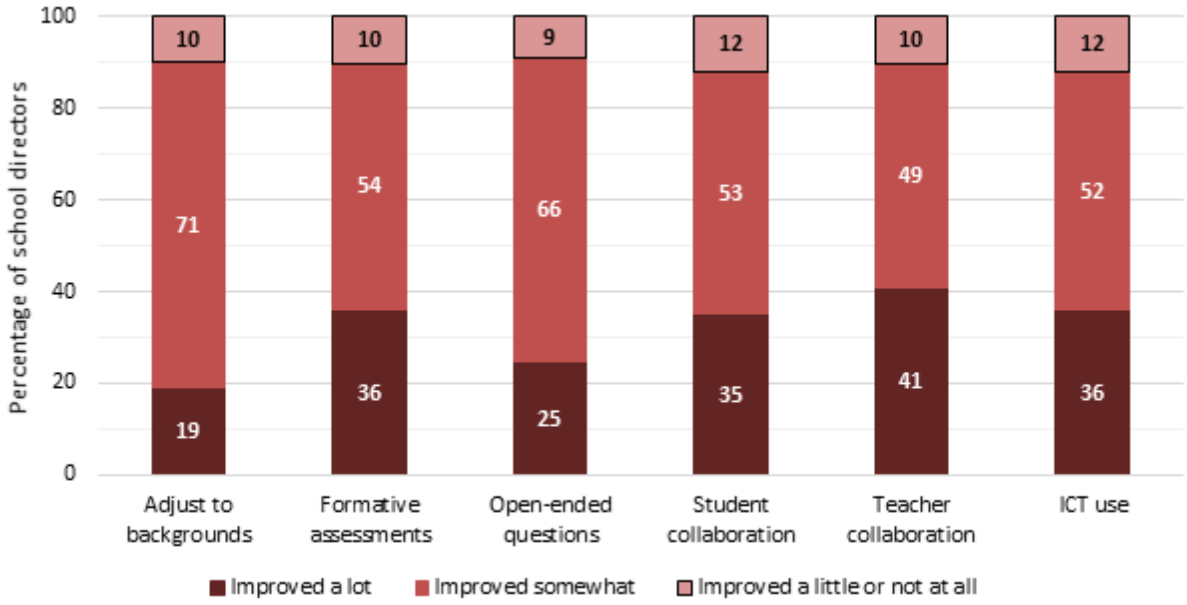
Note: Samples included 573 Cohort 1 and 279 Cohort 2 practitioner teachers. Differences between Cohort 1 and Cohort 2 means and *p*-values of those differences were estimated using multivariate ordinary least squares regressions with weights estimated by using propensity score matching. Details of the matching are presented in Chapter II. The regressions included all controls used to conduct the propensity score matching, as well as indicators for region (not reported). Standard errors were robust to heteroscedasticity. The standardized weighted knowledge indices were estimated by using principal components analysis (PCA). We present details of the PCAs in Appendix A. We restricted the matching analyses to outcomes with a comparison sample of at least 25 respondents. The geography measures did not reach this threshold and were therefore excluded from the analysis. "p.p." indicates that the reported means and differences were in percentage points, with a range between 0 and 100. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

4. School director perceptions about the effects of training on teachers

Although the teacher survey data did not provide evidence that the TEE trainings had immediate effects on teaching practices, most school directors reported that they did perceive improvements among teachers after the TEE trainings were completed. Figure IV.7 presents the proportion of school directors who reported that teachers at their school had improved a lot, improved somewhat, or improved a little or not at all for a number of different practices related to the TEE training. Nearly all school directors reported that teachers had at least improved somewhat in all six practice areas, with a substantial minority reporting that they had observed large improvements. For example, 19 percent of school directors reported large improvements in whether teachers understood and responded to individual student backgrounds and abilities. In addition, a quarter of school directors reported large improvements in whether teachers asked open-ended questions, asked students to explain their reasoning, and encouraged debate among their students. A larger percentage of school directors (between 35 and 41 percent) reported large improvements in the other four practice areas.

Figure IV.7. School director perceptions of changes in teacher practices after first round of teacher training



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence School Director Surveys (2017).

Note: Sample included 218 to 220 school directors interviewed in 2017 and 2018. Adjust to backgrounds refers to the practice of understanding individual student backgrounds and abilities and adjusting lesson plans in response. Formative assessments refers to the practice of using formative assessment strategies to assess student learning. Open-ended questions refers to the practice of asking open-ended questions, asking students to explain the reasoning behind their responses, and encouraging debate. Student collaboration refers to the practice of providing opportunities for students to work collaboratively in small groups. Teacher collaboration refers to the practice of collaborating with colleagues to improve teaching and professional development. ICT use refers to the practice of using information and communications technology in instruction.

Another potentially important finding from qualitative interviews with school directors pertains to potential "spillover effects" of the training from Cohort 1 teachers to Cohort 2 teachers who had not been trained yet. Some directors reported that the trainings benefited all teachers, including those who had not been trained, because the trained teachers were sharing their new knowledge with other teachers. If this pattern of knowledge-sharing between Cohort 1 teachers and Cohort 2 teachers was widespread, then it could help explain why the matched-comparison group analysis only found modest differences between the two groups. In our analysis, any spillover benefits from the trained teachers to the matched comparison teachers would attenuate the differences between the two groups, and potentially mask the training's full effects.

5. Qualitative findings about the effects of training on teachers

Focus groups with trained teachers provide additional insights on the potential reasons why improvements in teachers' knowledge and confidence-levels may not have translated into immediate changes in teaching practices. While in general many teachers understood the potential benefits of the practices encouraged by the TEE training, in some cases teachers had

reservations about the amount of effort required to implement these practices on a consistent basis in their own teaching.

During focus groups, teachers discussed using student-oriented teaching, demonstrations, and activities to promote students' critical thinking skills. Most teachers who participated in the focus groups reported using the student-centered teaching methods they learned in the TEE training on at least an occasional basis. Several teachers reported using group activities more frequently and discussed the benefits for students, including building autonomy, communication skills, and leadership. Teachers also stated that they now offer students more opportunities for "learning by doing"—for example, conducting science experiments or going on field trips to observe nature. They also described using demonstrations as a teaching strategy and encouraging active student participation in class through group projects and student oral presentations. Teachers noticed that student-centered teaching led to students being more analytical, creative, engaged, and motivated.

This finding, however, was not fully consistent with findings from the quantitative survey, which suggested that the training did not impact teachers' self-reported use of collaborative group activities (although the survey data did suggest that the training increased teachers' confidence in their ability to use collaborative group work effectively). One possible explanation for this is that most teachers in the survey were already doing some group work in some form before the training. The training material could nevertheless have changed the quality of group work in the classroom or the types of instructional activities where teachers considered group work to be valuable and appropriate. It is also important to remember that the focus groups were only held with a small subset of teachers; therefore, focus group responses may not be representative of the broader population of trained teachers.

Some teachers experienced difficulties in applying knowledge gained in the training to their classroom practice and in keeping up with the amount of new information they received during trainings. Teachers who participated in focus groups shared two key challenges with integrating and applying the knowledge they gained during training. First, many teachers felt the amount of information they were presented with and expected to process in a short time frame was greater than they could manage. Second, some teachers had difficulties applying what they learned or practiced during trainings in their own classrooms; teachers noted specific challenges related to organizing group work. Some teachers thought it was difficult to implement teamwork or group projects due to the large number of students in their classrooms. They understood the concept and had opportunities to practice it with peer teachers during trainings, but they had difficulty applying it in large classrooms (for example, 30 students).

Although several teachers reported using differentiated instruction effectively, other teachers raised objections to using tailored learning practices. Teachers who participated in the focus groups discussed the benefits and drawbacks of a wide range of tailored learning strategies encouraged in the TEE training sequence. In particular, teachers said they are now using a range of differentiated instruction strategies, including (1) activating prior knowledge to identify students' level and knowledge gaps; (2) breaking down tasks in smaller units of increasing complexity, which allows students to build skills incrementally; (3) adjusting the level of difficulty of tasks according to student levels and then scaffolding tasks for those who need help to complete a task; (4) involving students that have mastered a learning objective to help

those who have not yet mastered it; and (5) using collaborative learning activities to mix students with a variety of achievement levels, so less-skilled students can learn from more skilled students in small groups.

Some teachers shared that after training they were more aware of the variation in students' learning styles and made efforts to take those differences into account when planning and conducting lessons. Further, several teachers noted that these differentiation strategies helped them keep students engaged and increased their motivation. Some teachers believed these practices would be particularly beneficial for lower-achieving students, who might otherwise fall behind or become disengaged.

A subset of teachers, however, expressed objections to the differentiated instruction approach. They pointed out that differentiated instruction required extensive preparation from teachers prior to class to ensure that activities of varying levels of difficulty catered to students' needs during class. In addition to being labor-intensive, some teachers noted that this style of instruction is not always feasible, particularly in larger classes (25 to 30 students). A few teachers also mentioned that differentiated instruction could have drawbacks for lower-performing students. They noticed that some students don't like to be treated differently than their peers and can feel excluded or ashamed when they are not assigned the same tasks as higher-achieving students. Some teachers suggested that differentiated classrooms (tracking) might be an alternative approach that is more convenient for teachers and could also serve some students better.

Some teachers in the focus groups stated that the TEE training helped them improve their lesson planning; others noted that the lesson planning approach recommended in the trainings was difficult to implement and time-consuming. Although lesson planning was not unique to this training and some teachers had studied that topic in previous professional development opportunities, a common finding across teacher focus groups was that the TEE approach to lesson planning helped them learn the logic behind sequences of learning activities, assess whether or not an activity was working, and implement course corrections during lessons by modifying or replacing activities. Some teachers stated that detailed lesson plans helped them conduct more effective lessons. In addition, they said that sharing specific learning goals with students helped them stay on track in completing the lessons and helped them manage class time effectively. In spite of these virtues, some teachers also said that the level of detail expected in the TEE lesson planning approach was difficult to implement and required extensive and burdensome levels of preparation prior to class. In fact, some teachers questioned the value of the time invested in lesson planning and debated whether that time would be better spent in face-to-face instruction or one-on-one support for students.

Teachers widely reported that the TEE training encouraged the use of formative assessments, but a subset of teachers expressed resistance to using informal assessments that afford students more autonomy. Several teachers who participated in the focus groups stated that before the training they primarily used structured summative assessments, but after training they started using quicker and less-structured formative assessments more frequently. They believed the training helped them improve the use rubrics (criteria) as part of formative assessments and that the assessment process was now more transparent and reliable in instances when teachers shared these rubrics openly with students. Further, according to some teachers,

formative and other types of assessments afforded students greater autonomy and demonstrated on an ongoing basis that teachers were highly involved with their learning, which could in turn increase students' motivation and engagement. These responses were consistent with findings from the study's survey data that trainees became more confident in their ability to use formative assessments—although, the survey data did not show evidence that formative assessments were used more frequently by teachers after the training.

There was lack of consensus, however, among teachers about how, when, and if (at all) to use self-assessments. Although some teachers stated that they learned to use peer assessments and self-assessments and found those mechanisms to be valuable in gauging students' learning, other teachers disagreed strongly: their primary concern was that student self-assessments could lead to inaccurate information about students' achievement levels. One director noted that assessments have been the Achilles' heel for teachers, which suggests that teachers might need further support in this area.

Qualitative data showed that the TEE training introduced new professional development opportunities that enabled teachers to build networks with peer teachers and resulted in increased teacher motivation. According to teachers who participated in the focus groups, the training helped them build a community of practice (or professional learning community) with teachers from different schools. Teacher study group meetings were well attended. According to attendance records provided to Mathematica by TPDC, the average attendance rate at these study group meetings was approximately 89 percent. In a series of field observations of study group meetings among the first cohort of study group participants (the study observed five study group meetings during the first quarter of 2017), the research team observed active participation levels. However, the quality of the discussions, the participants' enthusiasm, and the level of teacher participation varied widely across study group meetings. For example, in one study group meeting some teachers were generally passive and seemed unwilling to contribute to the group discussion. This was in sharp contrast to another group meeting in which all teachers were highly engaged with each other's ideas and participated in a dynamic and vivacious discussion of lesson planning and other topics. The quality of study group meetings seemed to depend upon whether facilitators included discussion topics aligned with teachers' interests and needs, presented practical or actionable knowledge for teachers, organized varied activities such as small group assignments and presentations, and allowed teachers to take ownership of the discussions with an adequate level of guidance and structure. In addition, study groups organized by teaching subject seemed to be more productive and focused than mixed-subject study groups.

In focus group discussions, teachers also noted that study groups that mixed teachers with different seniority levels were highly valued: young teachers learned from older ones and vice versa. By teachers' accounts, the feeling of professional support and community engendered by these study groups also seemed to have increased teacher motivation and improved relationships among teachers. However, many teachers stated that they still needed additional opportunities to receive individual coaching or to consult with expert teachers in their specific subject areas and grade levels.

D. School director knowledge and practices after training

This section explores the knowledge and practices of school directors reported in the first and second follow-up surveys, which took place in fall 2017 during rollout period for the TEE Leadership Academy training (first follow-up) and in fall 2018, after the two-year training sequence was completed (second follow-up). We also conducted in-depth qualitative interviews with a subset of these school directors to investigate how the training activities related to changes in directors' instructional leadership and school management.

1. School director survey data

We begin by analyzing the self-reported practices provided during the first and second follow-up school director surveys. These include practices related to improving the quality of instruction, promoting inclusion, observing classroom instruction, monitoring teaching practices and student learning, and managing teachers' professional development. Although we used OLS regressions with individual fixed effects to estimate differences between the two survey rounds, we found no significant differences between the two survey rounds for any of the practices we analyzed, so we only present means for the second follow-up survey in the rest of this section.

Nearly all of the surveyed directors reported using the instructional leadership practices (presented in Table IV.7) on a monthly basis; nearly half reported using them every week. Over 80 percent of school directors also agreed that they played a large role in resolving disagreements between teachers (not shown). For each of the practices related to improving the quality of instruction in classrooms, between 78 and 89 percent of directors reported using the practices at least once a month and around half (or slightly less) reported using the practices weekly. The most common practice was providing advice on teaching, which was practiced at least once a week by 46 percent of directors in the first follow-up survey (not shown) and by 50 percent of directors in the second round.

Table IV.7. School director practices related to instructional leadership, as reported in the second follow-up survey

	At least once per week	At least once per month
Discuss policies, instruction, or learning with parents or community	49%	85%
Provide curriculum guidelines to teachers	41%	80%
Provide advice on teaching practices	50%	89%
Help teachers develop specific learning goals	40%	78%
Discuss ICT use in lessons with teachers	52%	85%

Note: Sample included 111 school directors interviewed in both the first and second follow-up surveys.

School directors reported regularly performing observations of classrooms to evaluate instruction, which usually included discussions with teachers both before and after an observation. On average, directors reported conducting 273 observations during the school year at the time of the second follow-up survey (2018–2019). In addition, the vast majority of directors reported meeting with teachers before an observation (92 to 97 percent) and providing feedback within two days of an observation (79 to 85 percent) (Table IV.8). Nearly all directors

agreed that these observations improved instruction (but only 45 percent strongly agreed with this claim).

Table IV.8. Prevalence of classroom observations by school directors

	Second follow-up survey
Number of classroom observations conducted this school year	273
Believed classroom observations helped improve instruction	
Strongly agreed	45%
Agreed	96%
Usually met with teachers before classroom observations?	97%
Usually provided feedback to teachers within two days of classroom observation?	85%

Note: Sample included 111 school directors interviewed in both the first and second follow-up survey.

Most school directors also reported that they collected and reviewed data on teacher instruction, student learning, and school accounting and budgeting practices on a monthly basis. Most of the directors reported collecting and reviewing data on teacher instruction (67 and 63 percent, respectively) and on student learning (77 and 68 percent, respectively) every month (Table IV.9). There was also a sizeable minority of directors who reported collecting student learning data every week (29 percent). However, only around 15 percent reviewed student data every week and only 12 percent collected data on teacher instruction weekly. Nearly all directors also reported reviewing their school's accounting and budgeting practices at least once a month (with a third doing so on a weekly basis). In addition, most directors reported discussing training topics with other directors at least once per month, which suggests that the program may have had some success in encouraging directors to utilize one another as ongoing sources of learning.

Table IV.9. School director monitoring of teaching practices

	At least once per week	At least once per month
Collected data on teacher instruction	16%	67%
Reviewed data on teacher instruction	12%	63%
Collected data on student learning	29%	77%
Reviewed data on student learning	15%	68%
Reviewed accounting and budgeting practices	34%	83%
Discussed training topics with other school directors	6%	51%

Note: Sample included 111 school directors interviewed in both the first and second follow-up survey.

Directors reported having high levels of support for teacher professional development. More than 80 percent of directors reported discussing career advancement paths with teachers, and around 75 percent reported discussing professional portfolios with teachers at least once a month (Table IV.10). There was similarly strong support for encouraging networking among teachers: between 78 and 84 percent of directors helped organize teacher group discussions every month and between 73 and 78 percent discussed attending working group meetings with teachers at least once a month. In addition, a sizeable proportion (between a third to a half) of directors who reported conducting these activities at least monthly reported doing so every week.

Table IV.10. School director support for teacher professional development

	At least once per week	At least once per month
Discussed career paths and professional advancement	49%	87%
Discussed professional portfolio use	27%	75%
Helped organize teacher group discussions	35%	84%
Discussed attending teacher working group meetings	24%	73%

Note: Samples included 111 school directors interviewed in both the first and second follow-up survey.

Most school directors reported supporting instruction that was more inclusive of female students and minorities, but it was rare to have a program designed to encourage female or minority participation specifically in science or math. All of the directors believed that their school was welcoming and safe for all students, with over two-thirds strongly believing this (Table IV.11). However, few of them had a program in place to support or encourage female students (13 percent) or minority students (12 percent) to study science and math. On the other hand, between 52 and 58 percent of schools had a program to support or encourage female students to play sports. In addition, close to 90 percent of directors reported holding discussions with teachers at least once a month about providing individualized lessons for students with special needs.

Table IV.11. School director practices related to instructional inclusion

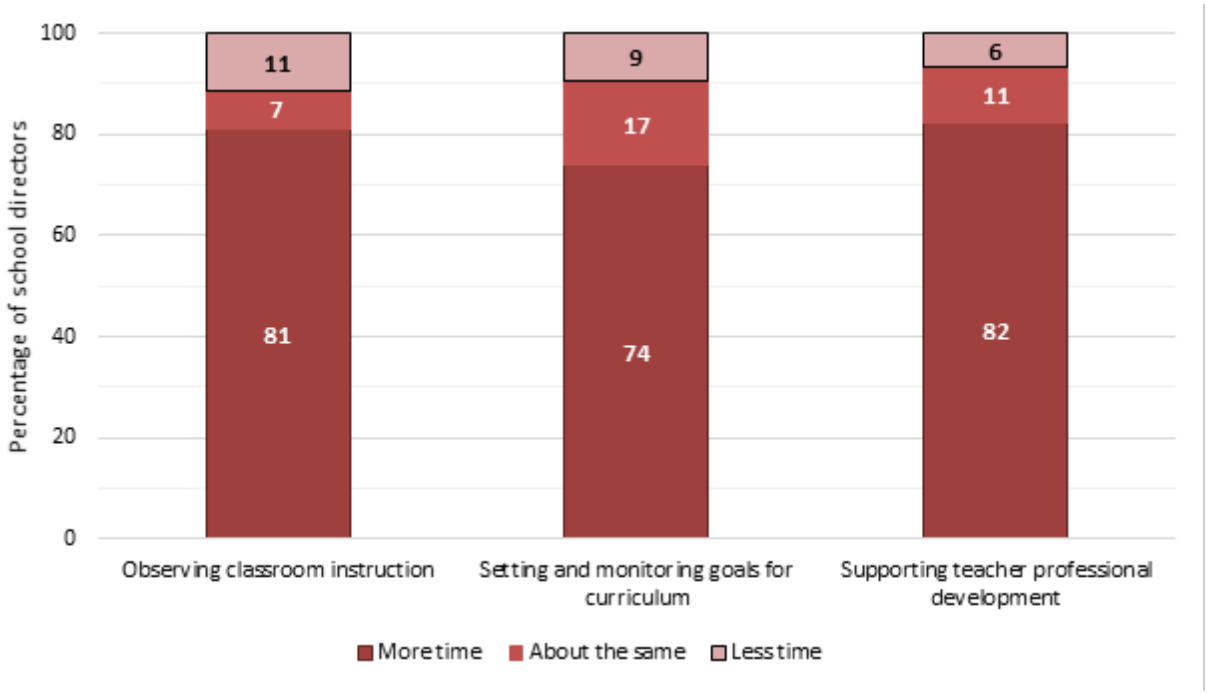
	Second follow-up survey
Had a program to support or encourage female students to study science and math?	13%
Had a program to support or encourage minority students to study science and math?	14%
Had a program to support or encourage female students to play sports?	58%
Believed school is welcoming and safe for all students	100%
Discussed individualized lessons for special needs students with teachers: At least once per month?	88%

Note: Samples included 111 school directors interviewed in both the first and second follow-up survey.

We also used survey data to explore school directors' reports of how their time use changed in the school year after the Leadership Academy was first implemented. These measures provided suggestive evidence of how the training may have impacted how school directors spent their time.[10] For all three types of activities, most directors reported that they spent more time on the activity in the school year after training than they did in the school year before training (Figure IV.8). This suggests, for example, that the frequent classroom observations reported in Table IV.8 may be due in part to the TEE training that the directors received.

[10] About a third of school directors in this sample (36 percent) had not yet completed the full training sequence by the first follow-up survey, because of the phased rollout of training sessions. Ultimately, by the second follow-up survey 93 percent of the directors in this sample had completed the full training sequence.

Figure IV.8. Change in time school directors reported spending on practices related to instruction or professional development after first year of training

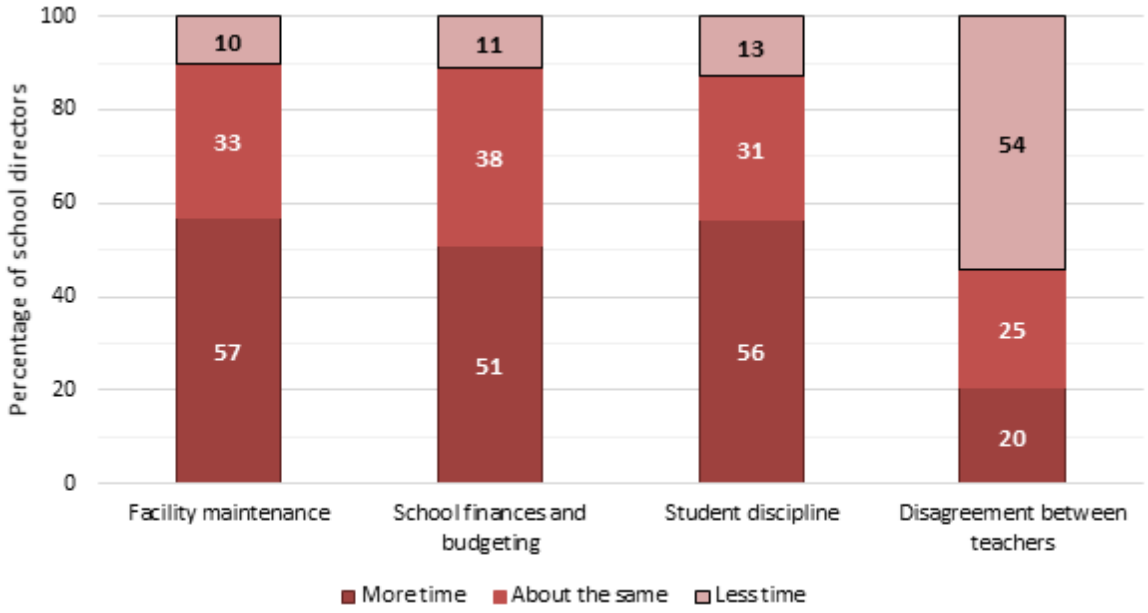


Source: Millennium Challenge Corporation Georgia Training Educators for Excellence School Director Surveys (2017).

Note: Sample included 218 to 219 school directors interviewed in 2017 and 2018.

Self-reported changes in time use for school management, discipline, and conflict resolution also improved in the year after training began, but the changes were less pronounced than for instruction and professional development. About half of the directors reported spending more time on facility maintenance, school budget, and finances, while about a third reported no changes (Figure IV.9). Generally speaking, these increases were less pronounced than the changes directors reported for instructional leadership and professional development, which was consistent with the emphases of the TEE activity and its theory of change. In terms of discipline and conflict resolution, about half of the directors reported spending more time on student discipline and less time resolving conflicts between teachers after the training. This finding may be consistent with the changes that directors reported in other aspects of their time use—for example, if directors were spending more time directly observing classroom instruction, then the amount of time spent on direct interactions with students might also increase.

Figure IV.9. Change in time school directors reported spending on practices related to school management after first round of training



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence School Director Surveys (2017).

Note: Sample included 218 to 219 school directors interviewed in 2017 and 2018.

All of the measures related to school directors were self-reported. In the absence of a baseline survey (or a phased implementation of the director training that would make it possible to compare trained directors to a group representing the counterfactual), all of the analyses in this section are purely descriptive and cannot be interpreted as impact estimates for the TEE training. However, we were able to cross-check the survey data with teacher survey data and in-depth school director interviews. The teacher data is presented below, and those qualitative findings are presented in the next section of the report.

Table IV.12 presents information from teachers about school directors' instructional leadership in the year after school director training was completed (from the second follow-up survey). The results are generally consistent with the positive story reported by school directors themselves. For example, most teachers reported receiving curriculum guidelines (71 percent), advice on teaching practices (64 percent), or help developing learning goals from school directors (54 percent) every month. However, these percentages are somewhat lower than those reported by school directors (80, 89, and 78 percent, respectively). Consistent with school director reports, most teachers suggest that their school directors are actively involved in teachers' professional development, with 77 percent reporting that their school director organizes group discussion at least once a month. Teachers report approximately three-quarters of study teachers have been observed in the current school year (during the fall). Among teachers who were observed, 80 percent met with the school director before the observations to discuss them and 98 percent met with the director after the observations to receive feedback. Finally, nearly all

of the teachers who were observed (97 percent) reported that the classroom observations had a positive impact on their teaching.

Table IV.12. Teacher reports of the instructional leadership provided by school directors in year after school director training completed

	Second follow-up survey
School director provided guidelines for curriculum: Every month?	71%
School director provided advice on teaching practices: Every month?	64%
School director helped develop learning goals: Every month?	54%
School director organized group discussions for teachers: Every month?	77%
School director observed instruction using classroom observations in current school year?	78%
Number of classroom observations conducted in current school year	2.5
Teacher believes that classroom observations had positive impact?[a]	97%
Teacher met with school director before classroom observations to discuss observations?[a]	80%
Teacher met with school director after classroom observations to receive feedback?[a]	98%

Note: Samples included 1065 teachers interviewed in 2018 survey round. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

[a] Sample restricted to 828 teachers whose school director had conducted a classroom observation of their instruction.

**/* indicates that differences were significant at the 1/5 percent levels.

2. Qualitative data from school directors and professional development facilitators

This section describes the directors' and SPDFs' perceptions about training, as well as qualitative findings about the ways in which the Leadership Academy training modules might have changed directors' and SPDFs' practices. (Appendix C shows a summary of findings and illustrative quotes.)

Qualitative data from principals and SPDFs suggest that they felt that they gained leadership skills and also improved their ability to lead instructional improvement at their school through the Leadership Academy trainings. Directors and SPDFs who participated in focus groups stated that they were very satisfied with the Leadership Academy training and felt that the training modules were relevant to their work. Both directors and SPDFs agreed that the "21st Century Schools" and "Shared Leadership" modules were the most interesting and helpful. More so than school directors, SPDFs also praised the "Formative Assessment," "Teacher Mentoring," and "Adult Learning" modules.

Directors reported that the Leadership Academy trainings enabled them to improve the quality of their instructional leadership. Directors who participated in interviews said they've become more aware of their role in supporting teachers' instruction, beyond school management. Directors reported that they were supporting teachers' instructional practice more than they did before the training and were observing their classes regularly. For example, one school director developed a monitoring system and a system of incentives as mechanisms to encourage improved teacher instruction. Newer and less experienced school directors stated that they benefited greatly from interacting with other school directors through Leadership Academy activities.

Directors discussed a number of efforts they made to help teachers improve their instructional practices, mainly through lesson monitoring and feedback sessions. Directors reported engaging with teachers in more varied ways after the Leadership Academy trainings—such as through classroom observations, consultations with individual teachers, faculty meetings, or director-led trainings. Some directors said that teachers' instruction changed in important ways at their schools. They believed that teachers were using passive, lecture-based instruction less often and were using group activities (as well as individual work and work in student pairs) more often. As one director noted, the current approach to teaching was more student-centered, whereas before it was teacher-centered and somewhat authoritarian. Directors observed that teachers were using more interactive lessons, encouraging students to raise their hands, and promoting a less restrictive classroom climate. Nonetheless, a few directors reported that implementation of the three-phase lesson plan (that is, defining a topic, identifying activities, and conducting assessments) and sequencing activities during class was not yet done optimally and that some teachers may need more support to master those practices.

During interviews, school directors highlighted notable changes in the approach to lesson planning and differentiated instruction; however, directors' views differed on the level of structure and detail that lesson plans should have. Some directors stated that lesson planning was currently done in ways that were very different from the way it was done before the trainings. Previously, lesson planning was based on a template that teachers updated with the lesson date—but neither the content nor the methodology was adjusted. In contrast, lessons plans now tend to be results-oriented and created collaboratively, by integrating input from the director and peer teachers. This has fostered a collaborative culture at schools, which has improved teamwork and staff satisfaction.

There was substantial variability in the way directors assessed teachers' lesson plans. For example, one director stated that he prefers being flexible with the lesson planning—he does not require that teachers spend too much time writing very detailed lesson plans, but rather encourages teachers to spend time choosing activities carefully and thinking through how they'll conduct each activity. Other directors attended lessons and observed whether actual classes matched what was described in the lesson plan. In addition, some directors focused on compliance with the national curriculum and conformity to the current standards.

Some directors noted that teachers were motivated to improve and show ownership of their instructional role after the training because they felt like they were part of a larger improvement process. Directors also believed that as a result of the trainings teachers were more open to receiving feedback and to pursuing opportunities for professional development. The improvement in teachers' morale seemed to be related to a closer and more collaborative relationship between the director and the teachers, which enabled both groups to give and receive constructive feedback. Some directors said that the trainings also benefited other teachers at their school who did not participate in trainings because the trained teachers were sharing their new knowledge with the other teachers. One way in which some directors gauged the school's progress and the results of their engagement with teachers was the number of teachers registered for professional development exams. Some directors believed that the number of teachers pursuing higher seniority levels increased. For example, one director reported that 3 teachers registered for the exams in 2017, while 14 teachers registered in 2018.

Directors reported that they were now more inclined to engage with other directors in collaborative initiatives and problem-solving. Most school directors reported participating in the TEE activity's quarterly professional development meetings with other local school directors. Directors stated that these meetings provided a platform for directors to build a community of practice with their peers. They said the activities and topics conducted during the meetings provided a useful way to share best practices and allowed them to discuss concerns about specific issues arising in each school. Further, some directors collaborated and shared resources with directors from different schools. For example, one director mentioned participating in a joint effort with two other directors of Kazreti schools to conduct a mixed performance of Azerbaijani and Georgian dances. Similarly, he shared the computer lab with a school in Bolnisi that did not have computers, so their students could take the school exit exams.

Directors also reported that the trainings helped them develop school management skills. Directors stated that the trainings helped them improve their time, human resource, and financial management. They highlighted the benefits of financial training, which offered useful guidance on the allocation of available funds and how to prioritize expenditures according to needs under the voucher financing system. Further, trainings offered a space for directors to develop "soft" management skills—including, assertive but nonaggressive communication, persuasion, and effective delegation—that gave them self-confidence and made them better leaders. Some directors noted that these training topics were particularly helpful in earning buy-in from older teachers and managing their resistance to pedagogical innovations.

Director interview data also indicated that some directors changed their school budget spending, as compared to previous years. Several directors discussed increasing their spending on pedagogical resources; a few also reported spending more on teacher incentives. Directors said that introducing newer and more detailed lesson plans as well as differentiated instruction techniques increased the need to provide instructional planning time and new instructional resources for teachers. After seeing changes in teachers' goals for professional advancement, the directors also invested more resources in supporting higher levels of training and certification. These supports included providing information about training opportunities; helping teachers' attend trainings (for example, arranging for substitute teachers to allow primary staff to attend trainings); and directly funding travel or tuition costs for training, in some cases.

Directors' perspectives on inclusion and diversity were mixed—these issues seem to be somewhat controversial and directors did not always agree with the training content. Although some directors stated that they were now more aware of gender issues in education, others said that their traditional perspective on gender issues were not affected by the trainings in a meaningful way because they already had the right approach to gender equity. A few directors who were interviewed stated that they purposefully sought gender balance in their schools and were aware of damaging stereotypes about young girls in STEM. They stated that gender equity should be promoted through breaking down stereotypes, which is what the training materials focused on. Some of the directors also implemented school-wide activities to promote gender equity among students. For example, one director led a project in which students researched women inventors and then posted portraits of famous women inventors around the school. With respect to diversity and inclusion, some directors emphasized their attempts to create an inclusive environment at their schools. For instance, one director organized an "International Tolerance Day" to convey the message of "no discrimination, no oppression." On the other hand, other

directors stated that the amount of time that the training spent on inclusion issues was excessive, and they wished more time had been devoted to topics (such as school management) that they found more valuable.

SPDFs perceive that their role is to provide instructional leadership and help teachers improve their instructional skills. SPDFs observed that as teachers transitioned to the TEE model they faced challenges in some areas—notably, lesson planning and differentiated instruction. Although the training showed teachers the importance of differentiated instruction in meeting students at their actual skill level, SPDFs said they continued to discuss the topic with teachers because of the need for further improvement in this area. SPDFs also stated that they were actively reviewing their lesson plans and helping teachers refine them. Some SPDFs had little awareness of the lesson planning techniques and didn't quite know how to help others in this regard before the trainings; after the trainings, they became more self-confident and able to advise them on designing lesson plans. SPDFs reported that the new approach to lesson planning required the ability to formulate a well-defined goal (or set of goals) and to creatively choose the activities that would meet those goals, noting that it is no longer acceptable for teachers to use standard templates of lesson plans across subjects or grade levels as they sometimes did in the past. These findings were consistent with teachers' and directors' perspectives on lesson planning and differentiated instruction.

SPDFs highlighted the purposeful use of different types of assessments and assessment criteria as a key benefit of the Leadership Academy. Some SPDFs noted that they've improved the way in which they conduct student assessments. Before the trainings, they did not use assessment criteria for each lesson or particular theme. Instead, they used a single assessment criterion and adapted it as needed. After the trainings, they realized that each lesson required separate assessment criteria; however, they felt this practice was labor-intensive and time-consuming. Similarly, formative assessments were not used regularly previously, but in training they learned how to implement dynamic and engaging formative assessments to gauge student learning.

Serving as liaisons between school directors and teachers and assisting teachers in achieving professional development goals were other important dimensions of the SPDFs' role that were supported by the Leadership Academy training. According to SPDFs who were interviewed, the training enabled them to better assist teachers in achieving professional development goals. They also played a role in incentivizing participation of teachers in trainings, attending trainings, organizing teacher workshops, and transferring knowledge to teachers who did not attend the TEE trainings. According to the SPDFs, teachers were now more willing to ask for support and more open to receiving constructive feedback from them. The SPDFs also reported that the TEE training and the learning partnerships teachers built with one another helped increase teachers' motivation and professional aspirations.

This page has been left blank for double-sided copying.

V. CONCLUSION

This report presented interim findings for the evaluation of the Georgia II Compact's school rehabilitation activity (ILEI) and teacher and school director training activity (TEE). The objective of the interim report and findings is to summarize preliminary evidence pertaining to each of the study's key research questions before the implementation period for these activities has officially ended: an initial draft of this report was shared with Government of Georgia stakeholders, MCC and MCA-Georgia staff, and implementing staff for their review and comment about six months before Georgia II Compact concludes in July 2019. A final report, planned for 2021, will provide additional evidence about the medium-term effects of these programs after the implementation period has ended.

For the school rehabilitation activity, the interim analysis found a strong pattern of improvements in the condition of the first 29 schools that were rehabilitated. Students, parents, teachers and school directors reported that the learning environment had improved in meaningful ways that are consistent with the program logic for the Activity. Although we did not observe major changes in student absenteeism, school enrollment, or dropout rates, findings from surveys and qualitative interviews do suggest that improvements in heating systems, air quality, lighting, and sanitary facilities may have improved the conditions in classrooms in important ways that directly facilitate instruction (particularly in winter months).

For the TEE activity, it is important to remember that the program logic did not assume that teaching practices would change in the immediate aftermath of the training sequence. Instead, the program was designed to produce rapid improvements in teachers' knowledge and their professional development resources (through the use of teacher study groups and other professional networks), which would in turn produce changes in their teaching practices and ultimately improve students' learning outcomes over longer periods of time. To examine whether this pattern is actually occurring, the final evaluation report will include a longer-term follow-up analysis of teachers' and school directors' practices up to three years after the training sequence was completed.

While it is too early to draw firm conclusions about changes in teaching practices, the interim evaluation clearly showed that the TEE activity succeeded in implementing the program on a nationwide scale. School directors had higher attendance rates at the offered trainings than teachers did, but a large majority of both groups attended one or more training sessions, and nearly all of the trainees felt positively about the training experience. In terms of the training's potential effects, we also found a fairly consistent pattern of improvements in teachers' self-reported knowledge of student-centered instruction strategies in the initial period after training, approximately one month after finishing the full sequence. However, outside of professional development activities (where we found a stronger pattern of improvements), the interim analysis did not reveal consistent evidence of short-term changes in teachers' classroom practices. Although school directors reported that they believed the training was improving classroom instruction (and they also reported that there could have been spillover benefits from trained teachers to untrained teachers at their schools), we did not observe an immediate quantitative pattern of improvements in teachers' self-reported practices. There is currently substantial room for improvement in teachers' use of the types of practices encouraged in the

training sequence, and it remains an open question whether these practices will show a pattern of improvement over a more extended period of time.

This interim report presents important initial results revealing that in general the pattern of anticipated medium- or long-term effects assumed in the program logic for the ILEI and TEE activities remains plausible. By comparing the preliminary results summarized here with the study's endline findings, the overall evaluation will yield insights on how the key outcomes observed in rehabilitated schools, and among trained teachers and school directors, have evolved over time and across schools in different regions. This will in turn enable a clearer final assessment of whether the medium-term outcomes projected for these programs were observed in practice. All of the preliminary findings in the interim report are also primarily descriptive in nature; the final report will include impact findings from the study's more rigorous random-assignment evaluation design for the school rehabilitation activity, and examine outcomes after the full implementation period is complete. Using these findings, the final report will also examine whether the pattern of observed outcomes for the ILEI and TEE activities suggests that investments in the two sets of activities were cost-effective, providing lessons for implementers of similar programs in Georgia and beyond.

REFERENCES

Bagby, Emilie, Anca Dumitrescu, Cara Orfield, and Matt Sloan. "Niger IMAGINE Long-Term Evaluation." Washington, DC: Mathematica Policy Research, October 24, 2014.

Bagby, Emilie, Kristine Bos, Anca Dumitrescu, Nicholas Ingwersen, and Matt Sloan. "Niger NECS Impact Evaluation Report." Washington, DC: Mathematica Policy Research, July 17, 2017.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India." *American Economic Journal: Economic Policy,* vol. 2, no. 1, 2010, pp. 1–30.

Berry, Michael. "Healthy School Environment and Enhanced Educational Performance. The Case of Charles Young Elementary School." Washington, DC: Carpet and Rug Institute, January 2002.

Bruns, Barbara and Javier Luque. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, DC: World Bank, 2015.

Burde, Dana, and Leigh Linden. "Bringing Education to Afghan Girls: A Randomized Control Trial of Village-Based Schools." *American Economic Journal: Applied Economics*, vol. 5, no. 3, July 2013, pp. 27–40.

Cabezas, Veronica, Jose Cuesta, and Francisco Gallego. "Effects of Short-Term Tutoring on Cognitive and Non-Cognitive Skills: Evidence from a Randomized Evaluation in Chile." Working paper. Cambridge, MA: Jameel Poverty Action Lab, May 2011.

Chetty, Raj, John N. Friedman, and Johnah E. Rockoff. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." National Bureau of Economic Research Working Paper Series, working paper no. 17699. Cambridge, MA: National Bureau of Economic Research, December 2011.

Davis, Mikal, Nick Ingwersen, Harounan Kazianga, Leigh Linden, Arif Mamun, Ali Protik, and Matt Sloan. "Ten-Year Impacts of Burkina Faso's BRIGHT Program." Washington, DC: Mathematica Policy Research, August 29, 2016.

Deke, John, Lisa Dragoset, and Ravaris Moore. "Precision Gains from Publically Available School Proficiency Measures Compared to Study-Collected Test Scores in Education Cluster-Randomized Trials." Document No. PR 10-80. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, October 2010.

Dobbie, Will, and Roland G. Fryer, Jr. "Getting Beneath the Veil of Effective Schools: Evidence from New York City." *American Economic Journal: Applied Economics,* vol. 5, no. 4, 2013, pp. 28–60.

Dumitrescu, Anca, Dan Levy, Cara Orfield, and Matt Sloan. "Impact Evaluation of Niger's IMAGINE Program." Washington, DC: Mathematica Policy Research, September 13, 2011.

Dunteman, George H. *Principal Components Analysis*. Newbury Park, California: SAGE Publications, 1989.

Durán-Narucki, Valkiria. "School Building Condition, School Attendance, and Academic Achievement in New York City Public Schools: A Mediation Model." *Journal of Environmental Psychology,* vol. 28, no. 3, 2008, pp. 278–286.

Evans, David K., and Anna Popova. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." World Bank Policy Research Paper 7203. Washington, DC: World Bank Group, February 2015.

Furgeson, Joshua, Brian Gill, Joshua Haimson, Alexandra Killewald, Moira McCullough, Ira Nichols-Barrer, Bing-ru Teh, Natalya Verbitsky Savitz, Melissa Bowen, Allison Demeritt, Paul Hill, and Robin Lake. "Charter-School Management Organizations: Diverse Strategies and Diverse Student Impacts." Cambridge, MA: Mathematica Policy Research, January 2012.

Hair, Joseph F., Ralph E. Anderson, Ronald L. Tatham, and William C. Black. *Multivariate Data Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall, 1998.

Hanushek, Eric. "The Economic Value of Higher Teacher Quality." National Center for the Analysis of Longitudinal Data in Education Research Working Paper 56. Washington, DC: Urban Institute, December 2010.

He, F., L. Linden, and M. Macleod. "A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trial." Working paper. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab, Massachusetts Institute of Technology, 2009.

Hedges, Larry V., and E. C. Hedberg. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Education Evaluation and Policy Analysis*, vol. 29, no. 1, 2007, pp. 60–87.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives*, vol. 2, no. 3, 2008, pp. 172–177.

Hoxby, Caroline M., Sonali Murarka, and Jenny Kang. "How New York City's Charter Schools Affect Student Achievement: August 2009 Report." Cambridge, MA: New York City Charter Schools Evaluation Project, September 2009.

Kazianga, Harounan, Dan Levy, Leigh L. Linden, and Matt Sloan. "The Effects of 'Girl Friendly' Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso." *American Economic Journal: Applied Economics*, vol. 5, no. 3, 2013, pp. 41–62.

Levy, Dan, Matt Sloan, Leigh Linden, and Harounan Kazianga. "Impact Evaluation of Burkina Faso's BRIGHT Program." Washington, DC: Mathematica Policy Research, June 2009.

Muralidharan, Karthik, and Venkatesh Sundararaman. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *The Economic Journal*, vol. 120, no. 546, 2010, pp. F187–F203.

Nichols-Barrer, Ira, Matt Sloan, Ken Fortson, and Leigh Linden. "Program Logic Assessment for the Georgia Improving General Education Quality Project." Final report submitted to the Millennium Challenge Corporation. Washington, DC: Mathematica Policy Research, December 2013.

Nichols-Barrer, Ira, Nicholas Ingwersen, Elena Moroz, and Matt Sloan. "Baseline Report for the Georgia Improving General Education Quality Project's School Rehabilitation Activity." Final report submitted to the Millennium Challenge Corporation. Washington, DC: Mathematica Policy Research, December 2017.

Popova, A., D.K. Evans, and V. Arancibia. "Training Teachers on the Job: What Works and How to Measure It." World Bank Group: Policy Research Working Paper. Washington, DC: World Bank, 2016.

Sailors, H. "The Effects of First- and Second-Language Instruction in Rural South African Schools." *Bilingual Research Journal*, vol. 33, no. 1*,* 2010, pp. 21–41.

Stevens, James. *Applied Multivariate Statistics for the Social Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.

WHO Regional Office for Europe "Health Effects of Particulate Matter: Policy Implications for Countries in Eastern Europe, Caucasus and Central Asia." Copenhagen, Denmark: WHO Regional Office for Europe, 2013.

Woolner, Pamela, Elaine Hall, Steve Higgins, Caroline McCaughey, and Kate Wall. "A Sound Foundation? What We Know About the Impact of Environments on Learning and the Implications for Building Schools for the Future." *Oxford Review of Education,* vol. 33, no. 1, 2007, pp. 47–70.

Yeh, S.S. "The Cost-Effectiveness of Five Approaches for Raising Student Achievement." *American Journal of Evaluation,* vol. 28, no. 4, 2007, pp. 416–436.

Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley. "Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement." *Issues & Answers Report*. Washington, DC: Regional Education Laboratory Southwest, 2007.

This page has been left blank for double-sided copying.

APPENDIX A

CONSTRUCTION OF OUTCOME INDICES

This page has been left blank for double-sided copying.

This appendix discusses an approach to reducing the amount of data to be presented in the report. This data reduction is needed for several reasons. The research team collected hundreds of data items through a school infrastructure assessment, student surveys, and teacher surveys. Reporting separately on each item would be impractical and could potentially mislead readers because of something known as the multiple comparisons problem. This problem arises when researchers report the results of a large number of hypothesis tests, and some of them are bound to be falsely rejected due to pure chance—the same logic whereby flipping a coin many times will eventually yield "streaks" of all heads or all tails, even if the coin toss is not rigged. As described in Sections III.3 and III.4, we reduce the amount of data on which to report by constructing indices for aspects of school infrastructure and knowledge of teaching practices.

Each index is a weighted average of three or more measures collected in the interim surveys related to the same topic. We identified the weights assigned to each of the related measures (or factors) using a principal components analysis (PCA) (see, for example, Dunteman 1989). This method of index construction assigns a greater weight to those measures related to the underlying topic that explain a greater amount of the variation in the topic across the sample (and less weight to those measures that explain less of the variation). PCA examines how a number of factors are correlated with one another and condenses this information into linear combinations of the factors called "principal components," equal to the number of factors. We adopted the weights estimated for the "first principal component" because, by design, PCA captures as much of the correlation between the factors as possible in the first principal component and therefore accounts for the largest amount of variability in the related measures. Finally, we standardized all components of the indices to range from 0 to 1 and also standardized the final weighted indices to a standard normal *z*-score using the mean and standard deviation of the index in the full sample.

A. ILEI evaluation

Tables A.1 through A.4 present the "factor score" and "factor loadings" of the first principal component estimated for each index presented in the ILEI evaluation section of the interim report.[11] To maintain comparability to the baseline results, we used the results of the PCA we conducted for the ILEI baseline report (weights and values used to standardize the components and indices) to construct the indices for the interim report. The factor score is equal to the proportion of variance explained by the principal component, multiplied by the number of factors in the principal component. Thus, the factor score can be interpreted as the number of variables' "worth" of variance captured by the first principal component (for example, a factor score of 2 means that the component captures two variables' worth of variance). The factor scores for the first principal components we estimated ranged from 1.47 to 2.26, so all of the first principal components captured more than one variable's worth of the variance between the factors. In other words, all of our constructed indices had more explanatory power than any single factor would have in isolation.

The factor loadings for a particular principal component are defined as the correlation between each factor and the principal component. We adopted the baseline factor loadings of the first principal component as weights to construct our interim indices. Following Stevens (1992)

[11] By construction, the first principal component has the highest factor score in the PCA.

and Hair et al. (1998), we adopted a cut-off of 0.40 to evaluate whether each factor has practical significance and excluded one factor that did not meet this cut-off.[12] (We excluded a measure of whether the main school building is painted from the "Better condition of school building exterior" index presented in Table A.1 because its factor loading was only 0.34.) As a result, all of the factor loadings used to construct the indices are larger than 0.40.

Table A.1. Results of first principal component for PCA of "Better condition of school building exterior" index

	Factor loadings
Number of problems not observed in roof of main school building (0–5)[a]	0.59
Condition of rain water drainage system on the roof of main school building (ranked 1–5)[b]	0.65
Condition of main entrance doors of main school building (ranked 1–5)[b]	0.47
Measures excluded because factor loading was below 0.40 threshold:	
Exterior of main building is painted	
Factor score	**1.47**
Proportion of variance explained by first principal component	0.49

Sources: Baseline MCC Georgia School Infrastructure Surveys (2015, 2016, 2017).

Notes: Sample included 192 schools.

[a] Problems included (1) cracks, (2) water damage, (3) rotten or deteriorated material, (4) mold, and (5) holes.

[b] Ranked categories included (1) no rain drainage system, (2) dilapidated (nonfunctional), (3) poor condition, (4) fair condition, and (5) perfect condition.

Table A.2. Results of first principal component for PCA of "Better condition of walls, ceilings, and floors in all school classrooms and indoor gym" index

	Factor loadings
Smallest number of problems not observed in walls in all classrooms and indoor gym in school (0–5 problems)[a]	0.59
Smallest number of problems not observed in ceilings in all classrooms and indoor gym in school (0–5 problems)[a]	0.59
Smallest number of problems not observed in floors in all classrooms and indoor gym in school (0–5 problems)[b]	0.54
Factor score	**1.96**
Proportion of variance explained by first principal component	0.65

Sources: Baseline MCC Georgia School Infrastructure Surveys (2015, 2016, 2017).

Notes: Sample included 194 schools.

[a] Problems included (1) cracks, (2) water damage, (3) mold, (4) chipped or peeling paint, and (5) holes.

[b] Problems included (1) unevenness, (2) cracks, (3) holes, (4) water damage, and (5) missing floor material/tiles.

[12] Hair et al. (1998) suggest different cut-offs for different sample sizes and suggest a cut-off of 0.40 for a sample size of 200—close to the size of our full sample of schools (194).

Table A.3. Results of first principal component for PCA of "Better condition of stairs in main school building" index

	Factor loadings
Number of problems not observed in stairs in main school building (0–4 problems)[a]	0.47
Stairs are level	0.63
Stairs are evenly spaced	0.62
Factor score	**2.26**
Proportion of variance explained by first principal component	0.75

Sources: Baseline MCC Georgia School Infrastructure Surveys (2015, 2016, 2017).

Notes: Samples included 188 schools with two stories.

[a] Problems included (1) unstable rails, (2) visible cracks, (3) holes in steps, and (4) missing steps.

Table A.4. Results of first principal component for PCA of "Better condition of classroom teaching facilities" index

	Factor loadings
All classrooms in school have working electric lights	0.51
All classrooms in school have lockable doors	0.61
All classrooms in school have a blackboard visible from the back of the classroom	0.40
Smallest number of types of class equipment that function properly reported by teachers in school (0–4 types of equipment) [a]	0.46
Factor score	**1.63**
Proportion of variance explained by first principal component	0.41

Sources: Baseline MCC Georgia School Infrastructure and Teacher Surveys (2015, 2016, 2017).

Notes: Sample included 194 schools.

[a] Types of equipment included (1) desks, (2) chairs, (3) blackboard/whiteboard, and (4) instructional materials.

B. TEE evaluation

Tables A.5 through A.7 present the "factor score" and "factor loadings" of the first principal component estimated for each index presented in the TEE evaluation section of the interim report. The factor scores for the first principal components we estimated ranged from 2.09 and 3.15, so all of the first principal components captured more than one variable's worth of the variance between the factors. As with the construction of the ILEI indices, we adopted a cut-off of 0.40 to evaluate whether each factor has practical significance; all factors met this cut-off and were included.

Table A.5. Results of first principal component for PCA of "Knowledge of practices related to critical thinking, motivation, and collaboration" index

	Factor loadings
Confident in teaching to motivate and encourage?	0.52
Confident in teaching to build self-confidence?	0.52
Confident in teaching to build higher-order thinking?	0.49
Confident in promoting cooperation through group work?	0.47
Factor score	**3.15**
Proportion of variance explained by first principal component	0.79

Sources: MCC Georgia TEE Teacher Surveys (2017).

Notes: Sample included 791 teachers.

Table A.6. Results of first principal component for PCA of "Knowledge of practices related to assessing student learning" index

	Factor loadings
Confident in conceptualizing measurable learning objectives?	0.54
Confident in using formative assessments during lessons?	0.59
Confident in including formative assessments in lesson plans?	0.60
Factor score	**2.38**
Proportion of variance explained by first principal component	0.79

Sources: MCC Georgia TEE Teacher Surveys (2017).

Notes: Sample included 722 teachers.

Table A.7. Results of first principal component for PCA of "Knowledge of practices related to inclusion" index

	Factor loadings
Confident in creating equitable learning environment for girls?	0.60
Confident in creating equitable learning environment for special needs?	0.58
Confident in creating unbiased learning environment?	0.56
Factor score	**2.09**
Proportion of variance explained by first principal component	0.70

Sources: MCC Georgia TEE Teacher Surveys (2017).

Notes: Sample included 722 teachers.

APPENDIX B

SUMMARY OF QUALITATIVE FINDINGS FOR THE ILEI STUDY

This page has been left blank for double-sided copying.

In this appendix, we present a summary of the findings of the qualitative analyses we conducted for the ILEI study. Table B.1 presents key qualitative findings, along with triangulation results across different stakeholders and illustrative quotes.

Table B.1. Summary of qualitative findings for Improved Learning Environment Infrastructure study

Key findings from qualitative data	Triangulation of findings by stakeholder			*Examples of qualitative evidence: Illustrative quotes*
	Students' perceptions	Teachers' perceptions	Directors' perceptions	
Improved heating systems and air quality				
Substantial improvement in temperature and air quality with new heating systems in rehabilitated schools	X	X	X	*There was no heating here before, [there] was a wood stove...now children come and study with more motivation.//Woodstoves, for example, often [were] not burning properly and we were cold.*
Students are less exposed to smoke in confined spaces for extended periods of time	X	X		*We used [o have] wood stoves for heating, and of course we had to supply the wood. But there [during the period of winds in February] the smoke was coming in the classroom. And in such cases, we had a feeling that our eyes were burning. Kids, as well as teachers, had same situation. And this kind of things was an obstacle for teaching. Now we think we are in fairy tale. We [have] normal conditions.*
Students and teachers feel more comfortable in classrooms with improved air quality (smoke free) and adequate temperature	X	X		*It is a big plus that temperature does not depend on weather. Two years ago [we did not have] such comfort. And we are very happy and thankful what we have now. The most important [thing] is that, in such a good environment, we want to conduct better lessons and the students are motivated to learn better.*
New heating systems has decreased the burden on students to collect wood, and maintain the wood stove running in the classroom	X			*When we had the stove often there was smoke and it was affecting us badly. And we had to bring wood as well, and it was hard because it was heavy.*
New heating system has decreased disruptions in class time relate to ventilation, or extremely low temperature during the winter months	X	X		*We have a central heating system everywhere and we have never had even a single issue of failing conducting lessons because of malfunction heating system or any other problem*

Key findings from qualitative data	Triangulation of findings by stakeholder			*Examples of qualitative evidence: Illustrative quotes*
	Students' perceptions	Teachers' perceptions	Directors' perceptions	
Students are less likely to be absent on cold days during the winter	X	X	X	*The furniture and the heating have had a positive impact. At that time, the teachers became ill ... they often missed the lessons, as well as the students...In February and January two or three years ago, most of the students did not come to school. There were many absences, and now we do not have a lot of absences.*
Students are more motivated and enjoy school more	X	X	X	*First of all the environment is so great that pupils are getting motivated to attend the lesson.*
Running new heating systems is more expensive than wood stoves			X	*Before installing the central heating, we use to spend 4500 GEL, now, since we have central heating, our costs per month doubled and maybe tripled, we paid 9990 GEL for heating in December.*
Water and sanitation				
Substantial improvement in sanitary facilities for students	X	X	X	*There was no sink in the old toilet. We washed your hands in the yard. Now there is [a] sink, soap, and napkins….Restrooms are tidy, clean. Better conditions compared to what it used to be...*
New sanitary facilities allow more privacy	X	X		*Toilets [are now] inside the building. Before, when toilets were outside, the whole school was informed when somebody went to toilet; everybody saw it. [Now] you can go downstairs quickly to the toilet and go back.*
New sanitary facilities are in the buildings instead of outdoors	X	X		*[The] environment is very important. When you enter the school yard you have desire to work better; [to] do everything for children. We have good conditions, walls, heating, toilets inside the building, before, when toilets were outside.*
Students now have access to sinks and running water, and most new sanitary facilities have soap and toilet paper	X	X	X	*There was no sink in the old toilet, we washed your hands in the yard. Now there is sink, soap and napkins. Better condition compared to what it used to be.*
New sanitary facilities are clean and students feel comfortable using them	X	X		*Children [can] wash hands often, [and] there are all necessary items in [the] toilet. Everything is clean. We also [have] the lavatory more comfortable than before.*

Key findings from qualitative data	Triangulation of findings by stakeholder			*Examples of qualitative evidence: Illustrative quotes*
	Students' perceptions	Teachers' perceptions	Directors' perceptions	
New sanitary facilities are well ventilated; improved uncomfortable issues with stagnant bad odor in the older facilities	X		X	No smell. It is always ventilated, there is a ventilation. They are equipped with air extractor fans.
Science and computer labs				
Lab facilities are well equipped and functional	X	X	X	*Chemistry, Physics, everything has changed, and the biology classroom is simply perfect*
Students have more and better opportunities for hands-on learning through lab experiments in science subjects	X	X	X	*The teacher would let us do almost every experiment given in the book, which made classes more interesting.*
Students are more motivated to learn science subjects	X	X	X	*Of course we use the lab. We use the reagents and instruments provided by the Millennium Foundation. The interest and motivation of our students has increased, since we regularly conduct interesting projects. The students themselves discover problems in the nature and set strategies to solve them.*
Students have more opportunities for collaborative peer work through science experiments	X	X	X	*Yes, now that we practically participate, do things with our own hands we better understand and remember what we learn.*
Renovated laboratory facilities have improved teaching practices, and teacher motivation		X	X	*They like the environment a lot. You know, the lessons are much more interesting for them in this environment. The teachers are more motivated as well and the children... We have many improvements and innovations in the school and these can be the most motivating factors.// The environment itself forms your attitude [towards] work. You feel more [valued], and you think you can do more.*
Sports facilities				
Substantial improvement in sports facilities for students	X	X	X	*Since this school has been renovated, children come to school with more joy, playground is also well done and conditions are better.*
Students value and take advantage the new sports facilities; playing sports more often and trying out new sports	X		X	*Now we are in the gym more often. In the winter it's too cold to go outside. In the summer it's very sunny and we avoid [going] outside….Now I am interested in Tennis more// Me too. I like basketball. We did not used to play at all.*

Key findings from qualitative data	Triangulation of findings by stakeholder			*Examples of qualitative evidence: Illustrative quotes*
	Students' perceptions	Teachers' perceptions	Directors' perceptions	
New sports facilities decreased risks of injuries	X		X	*The hall was damaged and the stadium was not accessible. If you fell [down there], you would get hurt.*
Perceived association between school renovations and improved teaching and learning				
Better classroom environment (temperature, air), laboratories and pedagogical materials have improved teachers' work satisfaction and pedagogical practices.	X	X	X	*When we have good conditions, support from the principal, resources, and a nice environment we [can] show more and more new things to children. ….[For example] the laboratory. I use this resource. It is warm in laboratory; children take off their coats; seat in comfortable chairs. We have a projector and computer. We watch movies, [and] conduct some discussions... Personally, I am more active than I was 5-6 years ago.*
Science teachers rely less on teacher-led demonstrations and lectures, give students more opportunities for hands-on learning, and cooperate more with other teachers	X	X	X	*Learning process is more joyful. The lessons are conducted with the new methods// We share laboratory with chemistry and biology and we cooperate with each other.*
Perceived improvement in students achievement level (or grades)	X	X	X	*We have better grades now. // [The lab] helped. For example, when preparing for the certification exams, physics and chemistry teachers, have been actively using the laboratory inventory and conducting the lessons to the fullest degree with the student. And their results showed that the knowledge they received, and the practical work they conducted, contributed to their success.*
Increased student engagement with academic activities (e.g. homework completion, participation in classwork)	X	X	X	*This renovation affects children as well. The school was ruined, desks were ruined, and children were not motivated to study. After the renovation, they started to learn better.// Well, first of all, their motivation has increased, which is the starting point for everything. Our students are happy and there is the willingness for lessons to be more active. This enabled me to discover students, which need to do more during the lesson than we used to. They are given more homework, or more complex home assignments. They expressed their desire to know the specialties better.*

Key findings from qualitative data	Triangulation of findings by stakeholder			*Examples of qualitative evidence: Illustrative quotes*
	Students' perceptions	Teachers' perceptions	Directors' perceptions	
Improved school climate and student sense of belongingness contribute to student learning	X	X		*They [students] are happy. They like new school, they [say] to children in other schools that their school is cool. Everybody in the region wants to have a school like ours. They like our school and our children like our school. // I'm very satisfied with what we have now, with what was done. I am not a new teacher and I have been working as a teacher for a long time. I'm have worked in very difficult conditions, when there was no heating, and we were not able to take off our coat [during] lessons. I had a dream to conduct lesson without coat and to have an opportunity to move normally.*

This page has been left blank for double-sided copying.

APPENDIX C

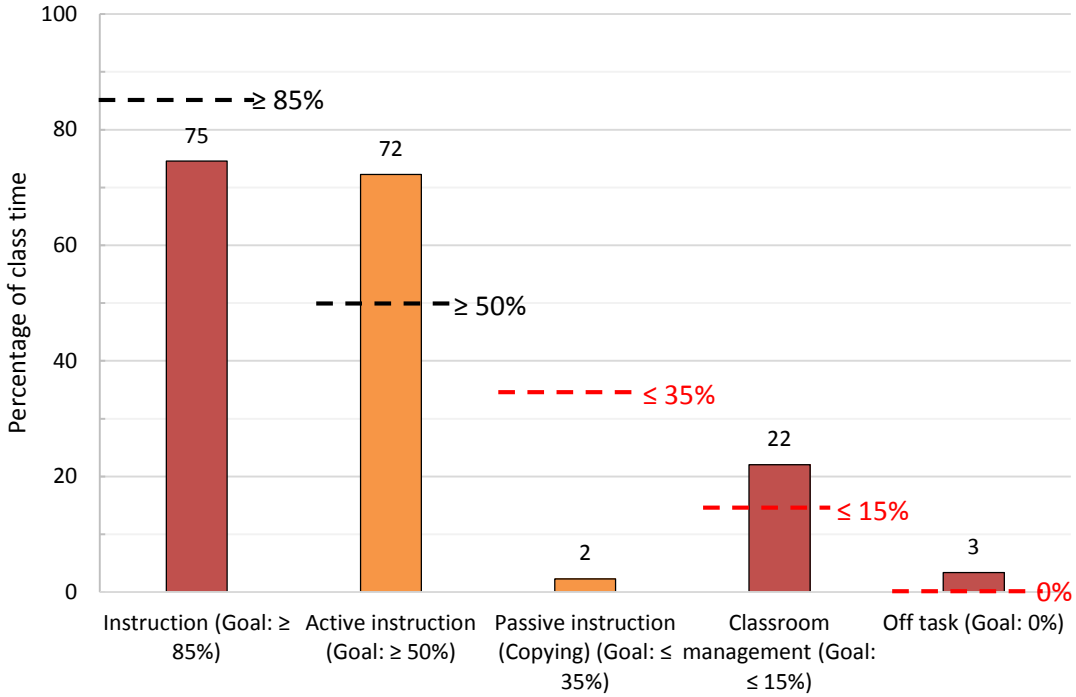
SUMMARY OF QUALITATIVE FINDINGS FOR THE TEE STUDY

This page has been left blank for double-sided copying.

In this appendix, we present a summary of the findings of the qualitative analyses we conducted for the TEE study, including classroom observation activities with a sample of 22 teachers and the study's qualitative data collection activities (teacher focus groups and in-depth interviews with school directors).

We used data collected with the Stallings Classroom Observation protocol to provide firsthand evidence of how teachers and students spend time in the classroom. Figure C.1 presents the percentage of observed class time that teachers spent on instruction; classroom management; or activities unrelated to instruction or classroom management—that is, activities that were off task. Teachers spent 75 percent of class time on instruction, almost all of which was spent on active instruction (72 percent) as opposed to passive instruction (such as copying written materials). Activities related to classroom management made up 22 percent of class time. Teachers were off task only 3 percent of the time. Broadly speaking, these results were reasonably well aligned with internationally recognized benchmarks for the Stallings rubric (Bruns and Luque 2015), although there appeared to be room for improvement in increasing the total amount of time spent on instruction (10 percent below benchmark) and decreasing the amount of time spent on classroom management (7 percent above benchmark). The proportion of class time spent on active instruction (for example, lectures and teacher-led activities) exceeded the recommended benchmark. This is a positive outcome, indicating that only a limited amount of time was being spent on passive instruction activities that are less likely to be effective (such as copying written materials without any active teacher engagement in the lesson).

Figure C.1. Percentage of class time spent on instruction and other tasks during the Stallings observations

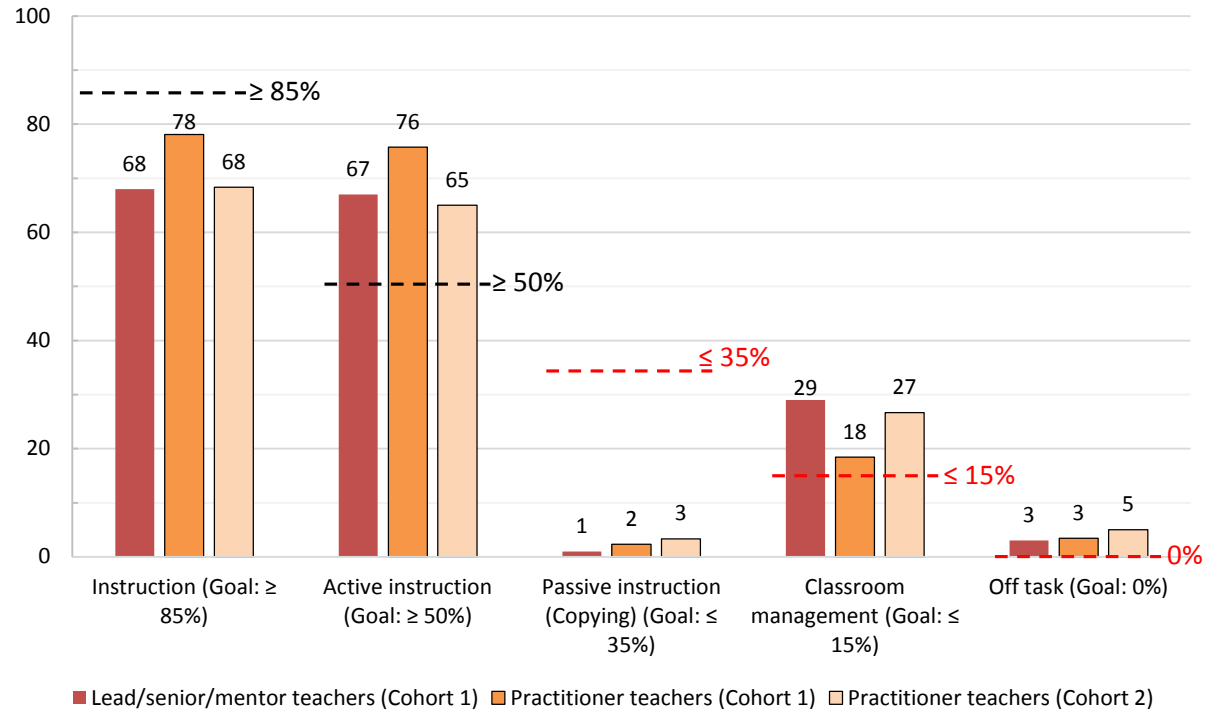


Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Stallings Observations (2018).

Note: Sample included 22 teachers observed in 2018.

These observation outcomes may have differed depending upon the teacher cohort. As mentioned previously, the two cohorts differed substantially in terms of the seniority levels of trained teachers. In addition, the training for Cohort 2 teachers took place in the second year of implementation, which could have differed from the first implementation year as implementers and trainers became more experienced. Examining the two cohorts separately provides some descriptive evidence about both of these potential patterns. In practice, however, we found fairly similar patterns of time use among lead, senior, and mentor teachers compared to practitioner teachers in both the first and second cohorts (Figure C.2). Practitioner teachers in the first cohort appeared to spend slightly more time on active instruction and less time on classroom management than the other teachers; however, the differences between them were only about 7 to 9 percentage points (and the sample sizes for these subgroups were fairly small). None of these groups met the recommended benchmark of spending at least 85 percent of class time on instruction.

Figure C.2. Percentage of class time spent on instruction and other tasks, by seniority status



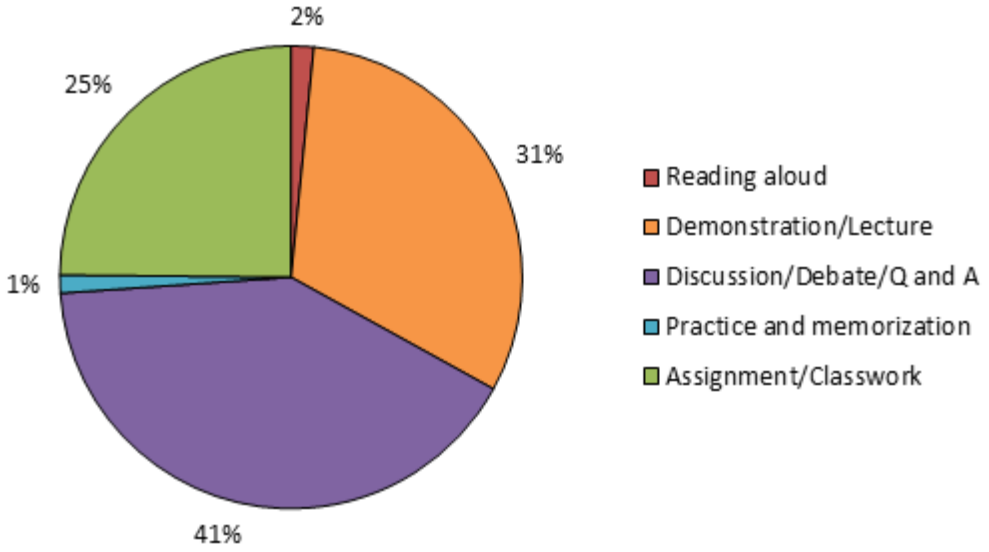
Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Stallings Observations (2018).

Note: Sample included 22 teachers observed in 2018.

Of the time spent on active instruction, teachers spent the most time (41 percent) on discussion, debate, or Q and A with the students (Figure C.3). An additional 31 percent was spent on demonstration or lecturing, while 25 percent was spent on assignments or classwork. Only 3 percent of active instruction was spent reading aloud to students or working on practice and memorization. Most of the time spent on classroom management consisted of non-disciplinary tasks with students present (67 percent). Roughly a quarter of management time was

spent on tasks without students being present (Figure C.4). Only 7 percent of time spent on classroom management involved disciplining students.

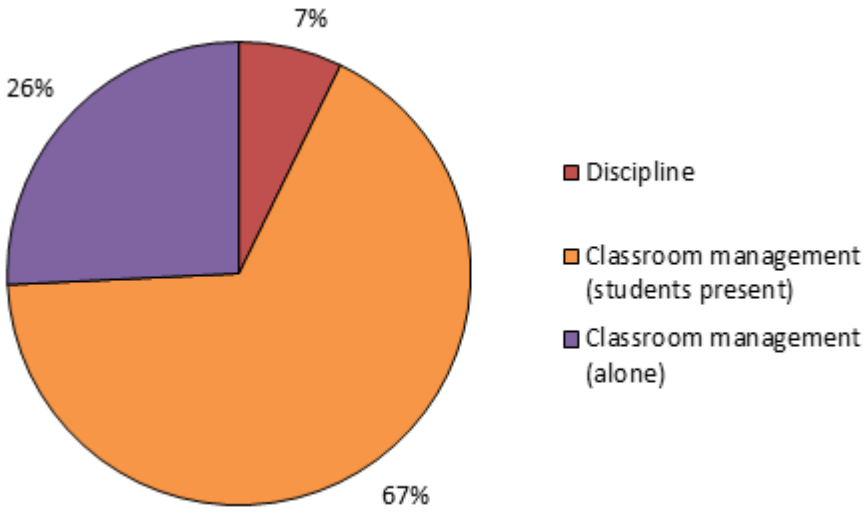
Figure C.3. Percentage of active instruction time spent on different activities



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Stallings Observations (2018).

Note: Sample included 22 teachers observed in 2018.

Figure C.4. Percentage of classroom management time spent on different activities

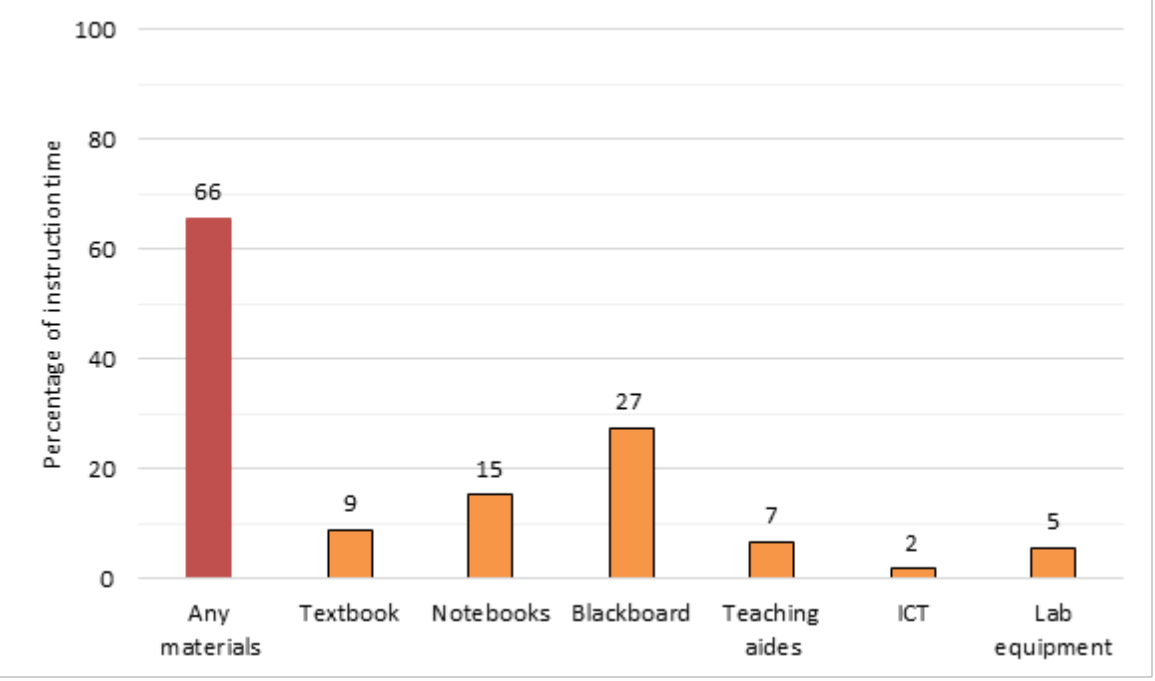


Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Stallings Observations (2018).

Note: Sample included 22 teachers observed in 2018.

The Stallings observations also included data on the teaching materials used by the teacher during each instructional activity, as well as the approximate numbers of students involved in each activity or who were off task. Two-thirds of instructional activities involved the use of some kind of teaching material (Figure C.5). Of these, the most common were blackboards (27 percent) and notebooks (15 percent). Less than 10 percent of instruction involved the use of textbooks (9 percent), teaching aides (7 percent), or lab equipment (5 percent). ICT use was uncommon, making up only 2 percent of the time spent on instruction in observed classes.

Figure C.5. Percentage of instruction time spent using different teaching materials



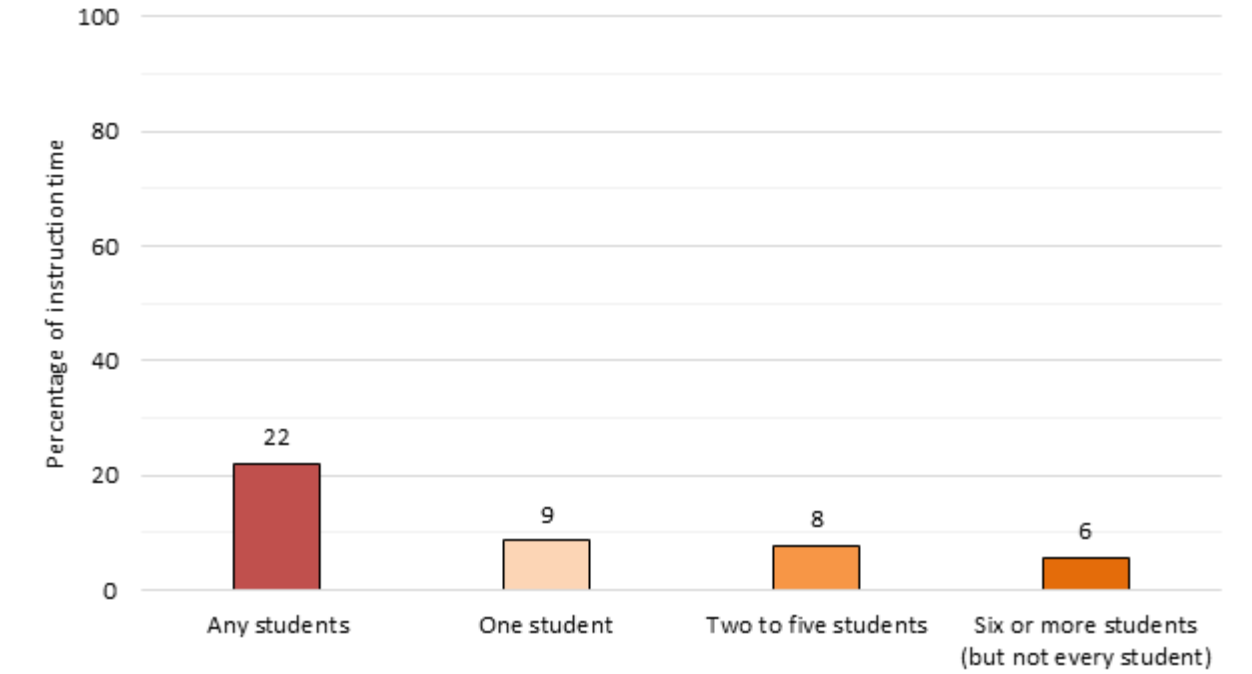
Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Stallings Observations (2018).

Note: Sample included 22 teachers observed in 2018.

All students were engaged in the lesson for most of the instructional time. Specifically, during 78 percent of instructional time all students were engaged in the task; for the remaining 22 percent of instructional time, the number of students who were off task was usually limited. One student was off task for 9 percent of instructional time, two to five students were off task for 8 percent of instructional time, and more than five students were off task for only 6 percent of instructional time (Figure C.6).[13]

[13] The average class size in the Stallings observations was approximately 14 students, so 6 students or more comprised 43 percent or more of students in the average classroom.

Figure C.6. Percentage of time with students off task during classroom instruction



Source: Millennium Challenge Corporation Georgia Training Educators for Excellence Stallings Observations (2018).

Note: Sample included 22 teachers observed in 2018.

In the remainder of this appendix, we also summarize the study's qualitative data and findings from interviews and focus groups with teachers, directors, and SPDFs. Table C.1 presents key qualitative findings, along with triangulation results across different stakeholders and illustrative quotes.

Table C.1. Summary of qualitative findings for Training Educators for Excellence study

Key findings from qualitative data	Triangulation of findings by stakeholder			Examples of qualitative evidence: *Illustrative quotes*
	Teachers' Perceptions	Directors' Perceptions	SPDFs' Perceptions	
To what extent have school directors' instructional leadership and school management skills improved?				
Leadership academy training increased directors' self-confidence, and improved their ability to manage human resources and finances at school.		X		*Up to date, every module has served as a reference manual for me. I am well-aware of the things that need to be improved in my school and I think that everything is given in those manuals*
Directors are better able to support teachers' instruction, observe lessons, provide enriching feedback, and sustain a community of practice at their schools.	X	X		*I personally conduced trainings with the teachers, so that they can work together. I conducted an open lesson myself according to the material I went through during the training, so that they know what's better.*
More active engagement with other school directors		X		*It was interesting in the sense that we shared our concerns with each other. We shared information about what didn't work in our school or, on the contrary, what did work, though we didn't expect it to work out well. Generally speaking, the most important thing about those meetings is that it allows me to share the experience of my colleagues. There atmosphere there is such that there is no chance you will leave without getting at least one advice. And you never feel ashamed of the problems your school [is] facing, and you feel free to discuss them.*
Positive changes in their schools increased teacher motivation (morale), and student engagement	X	X	X	*Higher quality of instruction and visual improvement of the school building make students eager to come to school and learn, and teachers are more enthusiastic to deliver their classes.*
Increased school-wide efforts of some directors to promote student diversity and gender equity		X		*We create an inclusive environment, and introduced it in his school. We organized events like the 'international tolerance day' to convey the message of: 'No discrimination, no oppression'*

Key findings from qualitative data	Triangulation of findings by stakeholder: Teachers' Perceptions	Directors' Perceptions	SPDFs' Perceptions	Examples of qualitative evidence: *Illustrative quotes*
Increased school spending due to greater need for pedagogical resources and teacher incentives strategies		X		*The implementation of certain activities certainly requires more resources. Teachers need more financial assistance on part of the school administration in order to implement certain projects.// Because these trainings made teachers realize the importance of diversified lessons, which naturally increases the need for diversified resources. When a teacher tells you that he/she needs this or that resource to deliver good classes, you should provide these resources. Consequently, the levels of spending have increased.*
To what extent have teachers' pedagogical practices and classroom management improved?				
Teachers learned student-oriented teaching methods, and have noticed the benefits of those methods on students' learning. Lessons based on solely on teachers lecture are no longer encouraged or accepted.	X	X		*The instructional process has improved and it becomes evident during the monitoring of the instructional process. And it is reflected in everything'.// It was not a child-oriented approach. We treated [students] in an authoritarian manner. We were leading the lesson process; the teacher was leader and children were secondary.*
Improved class time management, and more frequent use of active-learning and student-centered activities. Teachers have noticed students are more engaged in class as a result.	X	X		*It enabled me to better control my time, implement several activities during one lesson, and make the class more joyful and interesting for the students.*
More frequent use of collaborative group assignments.	X			*We use the projects very well; we make them for almost all topics. They [students] make presentations...The teams assess each other and identify their mistakes.*

Key findings from qualitative data	Triangulation of findings by stakeholder: Teachers' Perceptions	Directors' Perceptions	SPDFs' Perceptions	Examples of qualitative evidence: *Illustrative quotes*
More frequent use of differentiated instruction strategies, and improved ability to adjust activities to students' skill level.	X	X	X	*We start with easier tasks, so that they are involved too, then we get on to the more difficult ones and they get into it little by little. The apt ones get involved too. //Of course there are talented and less talented children in the class. We have to take it into the consideration, and we have to give one [type of] task to [the] talented child and another one to the less talented one. [For] evaluation it is the same, because if the less talented child [achieves] something, you have to see it and appreciate it. Everyone [is evaluated] according to [their] skills and capabilities.*
Improved lesson planning, moving away from a standard (rote) format to more detailed lesson plans. Preparing detailed lessons plans is challenging and time consuming for teachers, but many feel it's resulted in more effective lessons, and better class time management.	X	X	X	*The 'planning' topic was the most interesting, [and] that session has enabled us to conduct an effective lesson. //I thought I was doing it well, but after the last training, which was about planning, I saw how to plan a lesson. When you have planned everything, when you have choose right activities, when you have selected resources in a right way, everything is much more better; the result is better. I was very glad with lesson and with myself and students also showed me better result.*
Teachers views on the use of formative assessments were somewhat mixed. Many teachers learned new ways of implementing formative assessments, and are using them; but others don't find them valuable or believe they take away from instructional time.	X	X	X	*Assessment has been the Achilles heel for teachers. The more trainings we attend, the better we get at evaluation, because the evaluation system is now different…//There are rubrics, evaluation criteria. So, the trainings hugely benefited us to improve these.//And the rubrics, I personally have learn a lot about the holistic and analytic rubrics, how to construct them, which one is holistic and which one is analytic and then… The formative assessment turned out not to be what we thought it was. We found out that the formative assessment is a verbal, sentence…*

Key findings from qualitative data	Triangulation of findings by stakeholder: Teachers' Perceptions	Directors' Perceptions	SPDFs' Perceptions	Examples of qualitative evidence: *Illustrative quotes*
Some efforts are focused on supporting Azerbaijani students, but diversity and inclusion remains a somewhat controversial issue for some teachers.	X			*We often have 4, 5, 6 ethnical Azeri children in the class. I can't say I treat them differently or provide any kind of special encouragement. I just think that all children are equal. The greatest encouragement in my opinion is that everyone is treated equally and no one is distinguished. // It often happens that we study religion topics on a lesson. So, I always say that every religion is acceptable for us, moreover, we respect other religions and we do not assume anyone's religious believes.*
The knowledge gained, and the collaborative relationships built with other teachers during trainings has enriched their work and increased their motivation	X	X		*My motivation was raised after these trainings. Because, when you learn much, regarding the teaching process and the student-oriented methods, you try to use it in the class. After these trainings, with so much information, you want to keep up with everything, not to stay behind, and be more successful in the future… You want to do more.*
To what extent have SPDFs' ability to support teachers improved?				
SPDFs review and help refine teachers' lesson plans			X	*The present-day lesson plan has a better formulated goal. And if you, as a teacher, have this ambition to write a well-defined goal, it means that you give consideration to many things, you think how to bring it all up to standard. You should apply all your creativity in order to find a relevant aspect when dealing with each particular class. It's completely ruled out that you elaborate a lesson plan for the 6th grade and later use the same plan in some other class. It's a very specific approach.*
SPDFs support teachers in finding ways to use differentiated instruction in subject areas and are supporting teachers to use them more effectively.			X	*Therefore differentiated approach is really necessary and essential. And it largely depends on teacher's skillfulness how accurately he/she will 'diagnose' what this or that student needs to do, which particular student group he/she fits into, and how to tailor a lesson individually to each student. I consider a teacher's great achievement if a student is happy that the lesson is tailored to his/her needs.*

Key findings from qualitative data	Triangulation of findings by stakeholder: Teachers' Perceptions	Directors' Perceptions	SPDFs' Perceptions	Examples of qualitative evidence: *Illustrative quotes*
SPDFs learned new ways to implement assessments, and are helping teachers use them regularly.	X		X	*Yes, we developed rubrics (criteria sheets). For example, we jointly elaborated rubrics for various types of activities, like independent seatwork, summary writing etc. We agreed on them, and tailored them to our subjects. Then, we distributed them among the departments. At first all the teachers worked together, and then each department tailored those rubrics specifically to their subject. For example, I elaborated a 10-point test; other teachers wanted it to be a 15-point test, but the overall work, style, diagrams or analysis were done in a single form. And I think that it was helpful. We more or less manage to assist them.*
SPDFs are better able to assist teachers in achieving professional development goals			X	*So, that's the way I assist them, and it would have been hardly possible [without] those trainings. It was the training that allowed us to do that....They also know the specifics of stage-to-stage transition, and they are capable of making relevant conclusions.*
Improved supportive relationships between SPDFs and teachers			X	*Teachers used to avoid asking me for support, they were shy, but now we work together, we know that we should not criticize each other and have friendly environment, teachers do not avoid asking questions any more. We just discuss the plans they make, it's like we don't criticize their work we discuss and give the recommendations and therefore we have really good results.*

APPENDIX D

TEE SURVEY TRENDS FOR COHORT 1 TEACHERS

This page has been left blank for double-sided copying.

In this appendix, we present how the teaching practices that were a focus of the TEE trainings changed among teachers in Cohort 1 between the first- and second-year follow-up surveys. This analysis provides evidence of whether practices promoted by the TEE training improved or deteriorated between the first and second year after training.

Table D.1 presents regression-adjusted differences in practices of Cohort 1 teachers between the first and second follow-up surveys. We did not find a systematic pattern of differences between the first and second follow-up results in this sample; nearly all of the differences between the two survey round are not statistically significant. The endline evaluation will examine data from a third survey round, to test whether more pronounced trends developed after an additional year (two years after the conclusion of the TEE training sequence).

Table D.1. Changes in practices of Cohort 1 teachers between first and second TEE follow-up surveys

	First follow-up survey	Second follow-up survey	Difference
Practices related to critical thinking, motivation, and collaboration			
Ask open-ended questions: Every day?	0.51	0.49	-0.02
Ask open-ended questions: Percentage of class time (p.p.)	26.0	26.2	0.2
Collaborative group work: At least three times per week?	0.40	0.35	-0.05
Collaborative group work: Percentage of class time (p.p.)	35.1	33.9	-1.2
Students present work: At least three times per week?	0.44	0.41	-0.02
Students present work: Percentage of class time (p.p.)	27.6	28.5	0.9
Students work independently: Every day?	0.52	0.51	-0.01
Practices related to learning tailored to student needs			
Lesson plans include differentiated activities: Every day?	0.11	0.09	-0.02
Work with struggling students: Every day?	0.22	0.16	-0.06*
Practices related to assessing student learning			
Prepare lesson plans to achieve specific learning goals: Every day?	0.40	0.36	-0.04
Use formal tests to assess learning: At least once per week?	0.64	0.64	0.01
Use informal tests to assess learning: Every day?	0.46	0.45	-0.01
Change instruction in response to tests: Every day?	0.20	0.21	0.01
Practices related to inclusion			
Discuss inclusion of ethnicities/religions/sexual identities: Every month?	0.36	0.34	-0.02
Discuss inclusion of girls: Every month?	0.46	0.44	-0.01
Discuss inclusion of special needs: Every month?	0.48	0.49	0.01
Practices related to ICT use			
Use ICT in instruction: Every week?	0.49	0.48	-0.01
Practices related to professional development			
Discuss teaching/professional development with other teachers: At least once per week?	0.85	0.86	0.02
Attend professional meetings or events: At least once per month?	0.17	0.18	0.01
Update professional portfolio: At least once per month?	0.53	0.50	-0.03
Review professional portfolio: At least once per month?	0.55	0.53	-0.01
Practices related to teaching science courses			
Students conduct laboratory experiments: At least once per month?	0.56	0.50	-0.05
Students practice making or testing hypotheses: At least once per month?	0.74	0.75	0.01
Practices related to teaching mathematics courses			
Students work on math problems or projects: Every day?	0.35	0.30	-0.05
Teach both mathematical theory and work through examples?	0.78	0.76	-0.02
Class time spent teaching mathematical theory: Percentage of class time (p.p.)	23.9	28.2	4.3
Practices related to teaching English courses			
Students read authentic English written material?	0.89	0.87	-0.03
Students listen to authentic English audio material?	0.85	0.87	0.02
Students discuss materials: Every day?	0.30	0.22	-0.08
Teacher always provides guidance during discussion of materials?	0.83	0.75	-0.07
Practices related to teaching geography courses			
Students collect geographic data: At least once per month?	0.93	0.95	0.02
Students interpret maps or other geographic materials: Every day?	0.54	0.50	-0.03

Table D.1 (*continued*)

Note: Samples included 688–784 Cohort 1 teachers. We estimated differences between first- and second-year follow-up means and *p*-values of those differences using multivariate ordinary least squares regressions with indicators for each teacher. Standard errors are clustered at the teacher level. "p.p." indicates that the reported means and differences were in percentage points, with a range between 0 and 100. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

Table D.2 presents changes between the first and second follow-up surveys separately for lead/senior/mentor teachers and practitioner teachers. As with the overall results in Table D.1, we did not find a systematic pattern of differences between teachers' responses in the first and second follow-up surveys in either subgroup.

Table D.2. Changes in practices of Cohort 1 teachers between first and second TEE follow-up surveys, by seniority status

	Difference	
	Lead/senior/ mentor teachers	Practitioner teachers
Ask open-ended questions: Every day?	0.01	-0.04
Ask open-ended questions: Percentage of class time (p.p.)	-1.0	0.9
Collaborative group work: At least three times per week?	-0.08	0.0
Collaborative group work: Percentage of class time (p.p.)	-3.0	-0.3
Students present work: At least three times per week?	-0.03	-0.02
Students present work: Percentage of class time (p.p.)	-0.2	1.5
Students work independently: Every day?	-0.04	0.01
Lesson plans include differentiated activities: Every day?	-0.02	-0.02
Work with struggling students: Every day?	-0.05	-0.06
Prepare lesson plans to achieve specific learning goals: Every day?	-0.05	-0.04
Use formal tests to assess learning: At least once per week?	-0.02	0.02
Use informal tests to assess learning: Every day?	-0.03	0.01
Change instruction in response to tests: Every day?	0.01	0.01
Discuss inclusion of ethnicities/religions/sexual identities: Every month?	0.00	-0.03
Discuss inclusion of girls: Every month?	0.04	-0.04
Discuss inclusion of special needs: Every month?	0.04	-0.01
Use ICT in instruction: Every week?	0.03	-0.03
Discuss teaching/professional development with other teachers: At least once per week?	-0.03	0.04
Attend professional meetings or events: At least once per month?	0.02	0.01
Update professional portfolio: At least once per month?	-0.02	-0.04
Review professional portfolio: At least once per month?	0.00	-0.03
Students conduct laboratory experiments: At least once per month?	-0.04	-0.06
Students practice making or testing hypotheses: At least once per month?	0.04	0.01
Students work on math problems or projects: Every day?	-0.03	-0.07
Teach both mathematical theory and work through examples?	-0.07	0.02
Class time spent teaching mathematical theory: Percentage of class time (p.p.)	5.5	3.5
Students read authentic English written material?	-0.08	0.05
Students listen to authentic English audio material?	0.02	0.02
Students discuss materials: Every day?	-0.12	-0.03
Teacher always provides guidance during discussion of materials?	-0.03	-0.13
Students collect geographic data: At least once per month?	-0.03	0.03
Students interpret maps or other geographic materials: Every day?	-0.09	-0.01

Table D.2 (*continued*)

Note: Samples included 245–273 Cohort 1 lead/senior/mentor teachers and 443–511 Cohort 1 practitioner teachers. We estimated differences between first- and second-year follow-up means and *p*-values of those differences using multivariate ordinary least squares regressions with indicators for each teacher. Standard errors are clustered at the teacher level. "p.p." indicates that the reported means and differences were in percentage points, with a range between 0 and 100. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

This page has been left blank for double-sided copying.

APPENDIX E

TEE MATCHED COMPARISON GROUP ANALYSIS FOR TEACHERS COMPLETING ALL TRAINING MODULES

This page has been left blank for double-sided copying.

Over a third of the practitioner teachers in the first training cohort did not complete the full training sequence during the first round of training. Because of this, it is possible that the estimated impacts of the TEE training estimated in our "intent-to-treat" analysis (presented in Table IV.6) were attenuated due to the fact that some teachers in the treatment group did not benefit from the entire training sequence. To explore this issue, we also examined the effects of TEE training on Cohort 1 teachers who completed the full training sequence, matching them to Cohort 2 teachers who eventually would go on to complete the training sequence the following year. In Table E.1, we present the regression-adjusted differences in teaching knowledge and practices after the first training year of a matched sample of Cohort 1 practitioner teachers who completed all of the training modules in the first training round and Cohort 2 practitioner teachers who completed all of the training modules in the second training round. These results are akin to those produced by a "treatment-on-the-treated" analysis; it compares the outcomes of Cohort 1 teachers who had received the full treatment with those of a comparison group of Cohort 2 teachers. We restricted the Cohort 2 teachers to those who later went on to attend all of the training modules in the second training round to try to control for unobserved factors that led some teachers to complete the training modules and led others to fall short (for example, because of motivation or ability). As in the matching we conducted for the primary analysis, we used propensity score matching to identify a comparison group that was equivalent to the "treatment-on-the-treated" group of teachers, with respect to the following baseline characteristics: (1) years of teaching experience; (2) gender; (3) subjects taught (math, science, geography, or English); and (4) grades taught. After matching was completed, there were no statistically significant differences between the two groups on any of the variables included in the matching.

While we did not find strong evidence that the full training sequence increased teachers' self-reported knowledge of targeted teaching practices and self-reported confidence in using these practices, the results are somewhat difficult to interpret because of the smaller size (and therefore reduced statistical power) of this alternative sample. We found statistically significant evidence of positive impacts of the TEE training on only one measure of knowledge (creating an equitable learning environment for girls). However, we also found descriptive evidence that Cohort 1 teachers who took the whole training sequence were more confident in their knowledge as measured by the standardized indices of knowledge that we constructed for three of the domains, but these differences were not statistically significant: the pattern is somewhat difficult to interpret, because the sample size in the "treatment-on-the-treated" analysis is smaller than the sample used for the study's primary analyses, and the analysis did not have enough statistical power to detect effects that are of a similar magnitude to the effects we found for the full sample.

Similar to the primary analysis, we did not find evidence of training impacts for most of the self-reported teaching practices measured in the survey. With the exception of students working independently every day (where we observed a 15 percentage point decrease) and whether students in math classes work on math problems and projects every day (20 percentage point increase), we found no significant differences between the practices conducted by practitioner teachers in Cohort 1 who completed the training sequence and their matched sample of practitioners in Cohort 2. Overall, these results suggest that incomplete training participation by some Cohort 1 practitioner teachers is not driving the pattern of potential effects we presented in Chapter IV.

Table E.1. Matched comparison of practices of practitioner teachers in Cohorts 1 and 2 who completed the TEE training modules

	Cohort 1	Cohort 2	Difference
Practices related to critical thinking, motivation, and collaboration			
Knowledge of related practices			
Confident in teaching to motivate and encourage?	0.96	0.95	0.01
Confident in teaching to build self-confidence?	0.97	0.93	0.03
Confident in teaching to build higher-order thinking?	0.97	0.94	0.04
Confident in promoting cooperation through group work?	0.97	0.92	0.04
Standardized weighted index (*z*-score)	0.14	-0.04	0.18
Ask open-ended questions: Every day?	0.44	0.45	-0.01
Ask open-ended questions: Percentage of class time (p.p.)	25.2	24.71	0.5
Collaborative group work: At least three times per week?	0.37	0.31	0.06
Collaborative group work: Percentage of class time (p.p.)	31.9	36.06	-4.2
Students present work: At least three times per week?	0.40	0.47	-0.06
Students present work: Percentage of class time (p.p.)	24.5	29.15	-4.6
Students work independently: Every day?	0.50	0.65	-0.15**
Practices related to learning tailored to student needs			
Confident in knowledge to create a lesson plan with different tasks?	0.94	0.91	0.03
Lesson plans include differentiated activities: Every day?	0.09	0.10	-0.01
Work with struggling students: Every day?	0.20	0.18	0.03
Practices related to assessing student learning			
Knowledge of related practices			
Confident in conceptualizing measurable learning objectives?	0.92	0.87	0.05
Confident in using formative assessments during lessons?	0.97	0.93	0.04
Confident in including formative assessments in lesson plans?	0.95	0.90	0.05
Standardized weighted index (*z*-score)	0.05	-0.09	0.14
Prepare lesson plans to achieve specific learning goals: Every day?	0.42	0.39	0.03
Use formal tests to assess learning: At least once per week?	0.65	0.70	-0.05
Use informal tests to assess learning: Every day?	0.44	0.44	0.00
Change instruction in response to tests: Every day?	0.21	0.25	-0.04
Practices related to inclusion			
Knowledge of related practices			
Confident in creating equitable learning environment for girls?	0.96	0.88	0.08*
Confident in creating equitable learning environment for special needs?	0.89	0.87	0.02
Confident in creating unbiased learning environment?	0.97	0.98	0.00
Standardized weighted index (*z*-score)	0.13	0.00	0.13
Discuss inclusion of ethnicities/religions/sexual identities: Every month?	0.39	0.33	0.06
Discuss inclusion of girls: Every month?	0.51	0.51	-0.01
Discuss inclusion of special needs: Every month?	0.52	0.50	0.02
Practices related to ICT use			
Confident in knowledge of using ICT in instruction?	0.93	0.91	0.01
Use ICT in instruction: Every week?	0.53	0.48	0.05
Practices related to professional development			
Discuss teaching/professional development with other teachers: At least once per week?	0.84	0.84	0.00
Attend professional meetings or events: At least once per month?	0.18	0.13	0.05
Update professional portfolio: At least once per month?	0.58	0.52	0.06
Review professional portfolio: At least once per month?	0.57	0.58	-0.01

	Cohort 1	Cohort 2	Difference
Practices related to teaching science courses			
Students conduct laboratory experiments: At least once per month?	0.55	0.67	-0.12
Students practice making or testing hypotheses: At least once per month?	0.75	0.72	0.03
Practices related to teaching mathematics courses			
Students work on math problems or projects: Every day?	0.34	0.14	0.20*
Teach both mathematical theory and work through examples?	0.76	0.61	0.16
Class time spent teaching mathematical theory: Percentage of class time (p.p.)	23.7	21.1	2.6
Practices related to teaching English courses			
Students read authentic English written material?	0.97	0.85	0.12
Students listen to authentic English audio material?	0.91	0.69	0.22
Students discuss materials: Every day?	0.24	0.38	-0.15
Teacher always provides guidance during discussion of materials?	0.91	0.77	0.14

Note: Samples included 349 Cohort 1 and 140 Cohort 2 practitioner teachers. We estimated differences between Cohort 1 and Cohort 2 means; we estimated *p*-values of those differences using multivariate ordinary least squares regressions with weights estimated using propensity score matching. Details of the matching are presented in Chapter II. The regressions included all of the controls used to conduct the propensity score matching, as well as indicators for region (not reported). Standard errors were robust to heteroscedasticity. We estimated the standardized weighted knowledge indices using principal components analysis (PCA). We present details of the PCAs we conducted in Appendix A. We restricted the matching analyses to outcomes with a comparison sample of at least 25 respondents. The geography measures did not reach this threshold, so we excluded them from the analysis. "p.p." indicates that the reported means and differences were in percentage points, with a range between 0 and 100. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

This page has been left blank for double-sided copying.

APPENDIX F

TEE SUBGROUP ANALYSES

This page has been left blank for double-sided copying.

This appendix presents three subgroup comparisons of teaching practices in the first year after the training sequence ended for each cohort. The subgroup comparisons we examined are: (1) practitioner teachers as compared with senior, lead, or mentor teachers; (2) teachers with less than 20 years of teaching experience and teachers with 20 or more years of teaching experience; and (3) teachers who attended a TEE subject module and teachers who did not attend a subject module (but did attend at least one core module).

Table F.1 presents comparisons of teaching practices between teachers with different seniority levels. Although practitioner teachers were "less senior" in terms of their professional qualifications, on average they were older than teachers who had the senior, lead, or mentor qualification levels (mean age in the sample was 52.2 years for practitioner teachers and 46.7 years for senior, lead, or mentor teachers). Senior teachers were significantly more likely to conduct teaching practices related to critical thinking, motivation, and collaboration, and they were also more likely to use formal tests to assess learning at least once a week, to have students listen to authentic English audio materials (among English teachers), and to have students interpret maps and other geographic materials every day (among geography teachers). However, practitioner teachers performed better with practices related to discussing inclusion of ethnicities/religions/sexual identities (by 7 percentage points) and inclusion of girls (by 12 percentage points). They were also more likely to attend professional meetings or events at least once a month.

In Table F.2, we present the differences between teachers with more (20 or more years) or less (less than 20 years) teaching experience. Overall, the teachers in our sample were highly experienced: around 60 percent of the teachers had been teaching for 20 years or longer. Consistent with the fact that practitioner teachers were older on average (see above), practitioner teachers were more likely to have 20 year or more of experience (63 percent) than senior, lead, or mentor teachers (50 percent). Results for practices related to critical thinking, motivation, and collaboration were mixed for less experienced teachers: compared to more experienced teachers, they were less likely to ask open ended questions or have students work independently every day, but somewhat more likely to spend class time asking open ended questions, use collaborative group work, and have students present their work to classmates. Less experienced teachers were also significantly more likely work with struggling students every day and use ICT every week, and were significantly less likely to use formal tests to assess learning at least once a week. On the other hand, science teachers with less experience were 11 percentage points less likely to have students conduct laboratory experiments, compared to more experienced science teachers.

Our third subgroup analysis (Table F.3) presents differences between teachers who attended a TEE subject-specific training module (with material tailored to the subjects of science, mathematics, English, or geography) and teachers who attended at least one of the sequence's three core modules without attending a subject module. For practices related to the core training modules, we found no differences between these two groups of trainees. However, we did find one statistically significant difference in usage of a subject-specific teaching practice related to science instruction. Science teachers who attended the science training module were 16 percentage points more likely to have students conduct laboratory experiments at least once per month, compared with teachers who had not attended a subject module.

Table F.1. Comparison of practices one year after training between practitioner and non-practitioner teachers

	Practitioner teachers	Senior, lead, mentor teachers	Difference
Practices related to critical thinking, motivation, and collaboration			
Ask open-ended questions: Every day?	0.44	0.60	-0.16**
Ask open-ended questions: Percentage of class time (p.p.)	25.5	29.1	-3.6**
Collaborative group work: At least three times per week?	0.37	0.40	-0.04
Collaborative group work: Percentage of class time (p.p.)	32.5	40.5	-8.0**
Students present work: At least three times per week?	0.37	0.51	-0.14**
Students present work: Percentage of class time (p.p.)	26.9	30.4	-3.6**
Students work independently: Every day?	0.48	0.57	-0.10**
Practices related to tailoring lessons to student needs			
Lesson plans include differentiated activities: Every day?	0.11	0.13	-0.02
Work with struggling students: Every day?	0.21	0.21	0.00
Practices related to assessing student learning			
Prep lesson plans to achieve specific learning goals: Every day?	0.40	0.41	-0.01
Use formal tests to assess learning: At least once per week?	0.64	0.64	0.00
Use informal tests to assess learning: Every day?	0.41	0.58	-0.17**
Change instruction in response to tests: Every day?	0.19	0.23	-0.04
Practices related to inclusion			
Discuss inclusion of ethnicities/religions/sexual identities: Every month?	0.36	0.29	0.07*
Discuss inclusion of girls: Every month?	0.48	0.36	0.12**
Discuss inclusion of special needs: Every month?	0.48	0.45	0.03
Practices related to ICT use			
Use ICT in instruction: Every week?	0.50	0.45	0.05
Practices related to professional development			
Discuss teaching/professional development with other teachers: At least once per week?	0.85	0.86	-0.01
Attend professional meetings or events: At least once per month?	0.18	0.12	0.06*
Update professional portfolio: At least once per month?	0.53	0.49	0.04
Review professional portfolio: At least once per month?	0.55	0.51	0.03
Practices related to teaching science courses			
Students conduct laboratory experiments: At least once per month?	0.54	0.66	-0.12
Students practice making or testing hypotheses: At least once per month?	0.74	0.82	-0.08
Practices related to teaching mathematics courses			
Students work on math problems or projects: Every day?	0.28	0.36	-0.08
Teach both mathematical theory and work through examples?	0.77	0.83	-0.06
Class time spent teaching mathematical theory: Percentage of class time (p.p.)	23.3	24.4	-1.1

	Practitioner teachers	Senior, lead, mentor teachers	Difference
Practices related to teaching English courses			
Students read authentic English written material?	0.86	0.92	-0.06
Students listen to authentic English audio material?	0.80	0.91	-0.11*
Students discuss materials: Every day?	0.21	0.32	-0.12*
Teacher always provides guidance during discussion of materials?	0.85	0.80	0.05
Practices related to teaching geography courses			
Students collect geographic data: At least once per month?	0.92	0.97	-0.05
Students interpret maps or other geographic materials: Every day?	0.47	0.75	-0.28**

Note: Samples included 815 practitioner teachers and 328 senior, lead, and mentor teachers. We estimated *p*-values of mean differences between practitioner teacher and senior, lead, and mentor teachers using *t*-tests. "p.p." indicates that the reported means and differences were in percentage points, with a range between 0 and 100. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

Table F.2. Comparison of practices one year after training between less and more experienced teachers

	Less than 20 years of experience	20 or more years of experience	Difference
Practices related to critical thinking, motivation, and collaboration			
Ask open-ended questions: Every day?	0.45	0.52	-0.07*
Ask open-ended questions: Percentage of class time (p.p.)	28.0	25.4	2.6*
Collaborative group work: At least three times per week?	0.39	0.37	0.03
Collaborative group work: Percentage of class time (p.p.)	36.7	33.5	3.2*
Students present work: At least three times per week?	0.42	0.40	0.02
Students present work: Percentage of class time (p.p.)	29.9	26.6	3.3*
Students work independently: Every day?	0.46	0.54	-0.08**
Practices related to tailoring lessons to student needs			
Lesson plans include differentiated activities: Every day?	0.12	0.10	0.02
Work with struggling students: Every day?	0.25	0.19	0.06*
Practices related to assessing student learning			
Prep lesson plans to achieve specific learning goals: Every day?	0.38	0.43	-0.05
Use formal tests to assess learning: At least once per week?	0.58	0.68	-0.09**
Use informal tests to assess learning: Every day?	0.47	0.46	0.02
Change instruction in response to tests: Every day?	0.20	0.21	-0.01
Practices related to inclusion			
Discuss inclusion of ethnicities/religions/sexual identities: Every month?	0.35	0.33	0.02
Discuss inclusion of girls: Every month?	0.44	0.46	-0.02
Discuss inclusion of special needs: Every month?	0.48	0.47	0.01
Practices related to ICT use			
Use ICT in instruction: Every week?	0.54	0.46	0.08**
Practices related to professional development			
Discuss teaching/professional development with other teachers: At least once per week?	0.87	0.84	0.03
Attend professional meetings or events: At least once per month?	0.15	0.17	-0.02
Update professional portfolio: At least once per month?	0.50	0.53	-0.03
Review professional portfolio: At least once per month?	0.54	0.54	0.00
Practices related to teaching science courses			
Students conduct laboratory experiments: At least once per month?	0.48	0.59	-0.11*
Students practice making or testing hypotheses: At least once per month?	0.75	0.75	0.00
Practices related to teaching mathematics courses			
Students work on math problems or projects: Every day?	0.30	0.31	-0.01
Teach both mathematical theory and work through examples?	0.77	0.80	-0.04
Class time spent teaching mathematical theory: Percentage of class time (p.p.)	23.1	23.9	-0.8

	Less than 20 years of experience	20 or more years of experience	Difference
Practices related to teaching English courses			
Students read authentic English written material?	0.89	0.86	0.04
Students listen to authentic English audio material?	0.86	0.79	0.07
Students discuss materials: Every day?	0.24	0.31	-0.08
Teacher always provides guidance during discussion of materials?	0.81	0.89	-0.07
Practices related to teaching geography courses			
Students collect geographic data: At least once per month?	0.93	0.93	0.00
Students interpret maps or other geographic materials: Every day?	0.48	0.56	-0.08

Note: Samples included 466 teachers with less than 20 years of experience and 677 teachers with 20 or more years of experience. We estimated p-values of mean differences between teachers with less than 20 years of experience and teachers with 20 or more years of experience using *t*-tests. "p.p." indicates that the reported means and differences were in percentage points, with a range between 0 and 100. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

Table F.3. Comparison of practices one year after training between teachers who had attended subject training modules and those who had not

	Attended subject module	Had not attended subject module	Difference
Practices related to critical thinking, motivation, and collaboration			
Ask open-ended questions: Every day?	0.48	0.54	-0.06
Ask open-ended questions: Percentage of class time (p.p.)	26.2	26.4	-0.2
Collaborative group work: At least three times per week?	0.38	0.38	0.00
Collaborative group work: Percentage of class time (p.p.)	34.1	35.9	-1.8
Students present work: At least three times per week?	0.42	0.39	0.03
Students present work: Percentage of class time (p.p.)	26.5	29.6	-3.1
Students work independently: Every day?	0.51	0.50	0.01
Practices related to tailoring lessons to student needs			
Lesson plans include differentiated activities: Every day?	0.11	0.10	0.01
Work with struggling students: Every day?	0.21	0.23	-0.02
Practices related to assessing student learning			
Prep lesson plans to achieve specific learning goals: Every day?	0.40	0.46	-0.06
Use formal tests to assess learning: At least once per week?	0.65	0.60	0.05
Use informal tests to assess learning: Every day?	0.49	0.46	0.02
Change instruction in response to tests: Every day?	0.22	0.16	0.06
Practices related to inclusion			
Discuss inclusion of ethnicities/religions/sexual identities: Every month?	0.34	0.40	-0.05
Discuss inclusion of girls: Every month?	0.45	0.49	-0.04
Discuss inclusion of special needs: Every month?	0.48	0.51	-0.03
Practices related to ICT use			
Use ICT in instruction: Every week?	0.52	0.44	0.08
Practices related to professional development			
Discuss teaching/professional development with other teachers: At least once per week?	0.86	0.87	-0.01
Attend professional meetings or events: At least once per month?	0.17	0.15	0.03
Update professional portfolio: At least once per month?	0.54	0.52	0.02
Review professional portfolio: At least once per month?	0.55	0.54	0.01
Practices related to teaching science courses			
Students conduct laboratory experiments: At least once per month?	0.59	0.44	0.16*
Students practice making or testing hypotheses: At least once per month?	0.78	0.70	0.08
Practices related to teaching mathematics courses			
Students work on math problems or projects: Every day?	0.32	0.23	0.09
Teach both mathematical theory and work through examples?	0.78	0.82	-0.04
Class time spent teaching mathematical theory: Percentage of class time (p.p.)	22.7	23.7	-0.9

	Attended subject module	Had not attended subject module	Difference
Practices related to teaching English courses			
Students read authentic English written material?	0.92	0.95	-0.03
Students listen to authentic English audio material?	0.89	0.90	-0.01
Students discuss materials: Every day?	0.26	0.21	0.06
Teacher always provides guidance during discussion of materials?	0.83	0.92	-0.09
Practices related to teaching geography courses			
Students collect geographic data: At least once per month?	0.94	0.91	0.03
Students interpret maps or other geographic materials: Every day?	0.53	0.57	-0.04

Note: Samples included 875 teachers who had attended a subject module and 138 teachers who had not attended a subject module but attended at least one core module. We estimated p-values of mean differences between teachers who attended a subject module and those who had not using *t*-tests. "p.p." indicates that the reported means and differences were in percentage points, with a range between 0 and 100. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

This page has been left blank for double-sided copying.

APPENDIX G

TEE MATCHED COMPARISON GROUP ANALYSIS FOR INFREQUENT TEACHING PRACTICES

This page has been left blank for double-sided copying.

To evaluate whether the results of our TEE matched comparison analysis (presented in Table IV.6 in the main report) were influenced by the specific frequency-cutoffs that we chose for each binary teaching-practice outcome, we conducted an additional analysis that examined less frequent cutoffs for each outcome. For example, rather than testing if a practice was used multiple times per week, we examined if it was used at least once per week.[14] The results of this analysis are presented in Table G.1. Note that in cases where the outcomes in the main report were already coded using the least-frequent survey response option, the results are unchanged from the findings in Table IV.6. In general, the findings using these less-frequent cutoffs are very similar to our main results. However, two exceptions are whether teachers prepared lessons plans to achieve specific learning goals every month (using the monthly cutoff reveals a small statistically significant difference, potentially driven by ceiling effects in the survey data) and whether teachers changed instruction in response to tests every month (with this less-frequent cutoff there is a positive and statistically significant effect that did not appear using a more frequent cutoff-value).

Table G.1. Matched comparison group analysis of lowest frequency practices for practitioner teachers

	Cohort 1 teachers	Cohort 2 teachers	
	After training round	Before training round	Difference
Practices related to critical thinking, motivation, and collaboration			
Ask open-ended questions: Every week?	0.95	0.96	-0.02
Collaborative group work: Every week?	0.83	0.86	-0.03
Students present work: Every week?	0.82	0.77	0.06
Students work independently: Every week?	0.92	0.94	-0.02
Practices related to tailoring lessons to student needs			
Lesson plans include differentiated activities: Every week?	0.68	0.68	0.00
Work with struggling students: Every week?	0.78	0.76	0.02
Practices related to assessing student learning			
Prep lesson plans to achieve specific learning goals: Every month?	0.97	1.00	-0.02**
Use formal tests to assess learning: Every month?	0.96	0.95	0.01
Use informal tests to assess learning Every month?	0.97	0.96	0.01
Change instruction in response to tests: Every month?	0.92	0.84	0.07*
Practices related to inclusion			
Discuss inclusion of ethnicities/religions/sexual identities: Every month?	0.39	0.37	0.02
Discuss inclusion of girls: Every month?	0.51	0.52	-0.01
Discuss inclusion of special needs: Every month?	0.50	0.51	-0.01
Practices related to ICT use			
Use ICT in instruction: Every week?	0.51	0.51	0.00

[14] For cases with a "Never" option, we considered the next common frequency to be the "least frequent option".

	Cohort 1 teachers	Cohort 2 teachers	
	After training round	Before training round	Difference
Practices related to professional development			
Discuss teaching/professional development with other teachers: Every week?	0.84	0.82	0.02
Attend professional meetings or events: Every month?	0.19	0.17	0.02
Update professional portfolio: Every month?	0.55	0.45	0.10*
Review professional portfolio: Every month?	0.57	0.48	0.09
Practices related to teaching science courses			
Students conduct laboratory experiments: Every 2–3 months?	0.59	0.70	-0.11
Students practice making or testing hypotheses: Every month?	0.72	0.70	0.02
Practices related to teaching mathematics courses			
Students work on math problems or projects: Every month?	0.97	0.92	0.05
Practices related to teaching English courses			
Students discuss materials: Every week?	0.91	0.80	0.12

Note: Samples included 573 Cohort 1 and 279 Cohort 2 practitioner teachers. Differences between Cohort 1 and Cohort 2 means and *p*-values of those differences were estimated using multivariate ordinary least squares regressions with weights estimated by using propensity score matching. Details of the matching are presented in Chapter II. The regressions included all controls used to conduct the propensity score matching, as well as indicators for region (not reported). Standard errors were robust to heteroscedasticity. The standardized weighted knowledge indices were estimated by using principal components analysis (PCA). We present details of the PCAs in Appendix A. We restricted the matching analyses to outcomes with a comparison sample of at least 25 respondents. The geography measures did not reach this threshold and were therefore excluded from the analysis. “p.p.” indicates that the reported means and differences were in percentage points, with a range between 0 and 100. The reported means and differences without units listed were in percentage points, with a range between 0 and 1.

**/* indicates that differences were significant at the 1/5 percent levels.

APPENDIX H

STAKEHOLDER COMMENTS AND MATHEMATICA RESPONSES

This page has been left blank for double-sided copying.

Table H.1. Responses to stakeholder comments on the draft interim report

Page Number	Comment	Mathematica Response
pg. 19	Mathematica selected 2 schools per region (11 geographic regions) and conducted total of 44 observations (22 teachers were observed twice) with the Stallings method. Observations were conducted right after the trainings. Classroom observation method was used to measure behavior change in the classroom. With the proposed methodology, the method was not used properly, since the timing of the observation was not selected in a meaningful manner. Observations were conducted immediately after the trainings and teachers did not have a chance to digest training materials and apply to the classroom. In addition, sample of 22 teachers is very low and results cannot be generalized. Although, we know that triangulation method was used in the study and various methods adopted to validate data, it is crucial to have relevant sample size for all sources.	We appreciate this feedback about the nature of the classroom observation data collection activity. These observations were conducted for teachers who completed training in fall 2017 (Cohort 1 teachers), and the observations took place at two points in time: (1) spring 2018, approximately six months after training; and (2) fall 2018, approximately one year after completion of the training sequence. In our view, this 6-12 month follow-up period did allow a reasonable amount of time to elapse after training, before observations took place. That said, we fully recognize that the sample size was very small, and is unlikely to be representative of all trained teachers. Accordingly, we have moved the classroom observation findings to the TEE qualitative results appendix. The study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.
pg. 63	On the page 63, report says: "we found evidence that many of the practices that were encouraged by the TEE training sequence were only being applied to a limited extent in classrooms" . Please see comment above. Observations were conducted immediately after the trainings and only 22 teachers were observed. I believe that this statement is too loud to make such a conclusion on such a limited sample	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.
pg. 70	Figure IV.8 gives information in percentages. Please provide information in numbers (disaggregated by Teacher status). Although in the footnote it is indicated that the sample size is 22 teachers, it is better to avoid misunderstanding and put numbers on the figure for more clearance rather than percentages.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.
Pg. 71	Would be better to present pie charts with diverse colours rather using one colour in different tones. It makes the chart difficult to read	We have adjusted the color scheme in the pie chart figures.
pg. 72	Figure IV.11 either should be removed from the report or additional information provided: If schools observed had: labs, computers, internet connectivity, notebooks. Otherwise this information is misleading and draws reader to the wrong conclusions. Those schools might not have labs or computers at all. Reporting that teaches are not using them if schools are not equipped with the technology, won't be fair and correct.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.

Page Number	Comment	Mathematica Response
pg. 73	Report says: "Although it's possible that the small sample of teachers observed by the study team could differ in important ways from the full survey sample, these results suggest that survey findings using teachers' self-reported practices should be interpreted with caution, as they may not correspond strongly with actual practices for all teachers". - with the limited sample size, the conclusion can be an opposite: that current sample size does not give possibility to generalize what teachers actually say with what teachers actually do and not vice versa.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.
Pg. 76	Report says: "However, at least in the initial month after the end of the training sequence, we did not find evidence of training impacts for most of the self-reported teaching practices measured in the survey". Making such a general conclusion in the report without mentioning the sample size and the limitation of such comparison, is not relevant	To clarify the interim-nature of this set of findings, we have revised this paragraph to emphasize the timing of data collection.
pg. 81	report says: "Some teachers in the focus groups stated that the TEE training helped them improve their lesson planning; others noted that the lesson planning approach recommended in the trainings was difficult to implement and time-consuming" please instead of using "some" and "other" could you please specify majority, few? otherwise it is difficult to make conclusions. This comment refers to the whole report	Because these statements refer to qualitative findings, it could be misleading to characterize the results with more precise or quantitative language. Qualitative focus groups are intended to surface information about potential mechanisms or patterns that could be driving the study's quantitative results, and are not intended to represent the population of all trained teachers.
pg. 82	Report states: "However, the quality of the discussions, the participants' enthusiasm, and the level of teacher participation varied widely across study group meetings. For example, in one study group meeting some teachers were passive and seemed unwilling to contribute to the group discussion. This was in sharp contrast to another group meeting in which all teachers were highly engaged with each other's ideas and participated in a dynamic and vivacious discussion" . Drawing conclusions on two study group meetings does not seem valid. Please provide more details about the number of study groups observed, otherwise, reporting and concluding on two study group meetings should be removed from the findings.	We have clarified in the report that the study team observed five study group meetings during the first quarter of 2017. This was part of the study's qualitative data collection effort--we recognize that qualitative observations do not support strong quantitative conclusions about all trained teachers, but we believe the insights from these observations are valuable and help to enrich the study's quantitative findings with additional context about what occurred during these teacher study groups.
pg. 91	Report states: "Instead, the program was designed to produce rapid improvements in teachers' knowledge and their professional development resources (through the use of teacher study groups and other professional networks), which would in turn produce changes in their teaching practices and ultimately improve students' learning outcomes over longer periods of time. To examine whether this pattern is actually occurring, the final evaluation report will include a longer-term follow-up analysis of	Thank you for this suggestion--we agree it could be interesting and useful to add an additional classroom observation data collection component to the evaluation, using a larger and more representative sample. We are open to adding this component to the study, if MCC chooses to do so.

Page Number	Comment	Mathematica Response
pg. 91 (continued)	teachers' and school directors' practices up to three years after the training sequence was completed." - This won't be feasible if in the follow up data collection, the number of classroom observation will not increase and can be considered as representative sample	
pg. 92	Drawing such a conclusion based on the 22 classroom observations after a month of a training does not seem valid "However, outside of professional development activities (where we found a stronger pattern of improvements), the interim analysis did not reveal consistent evidence of short-term changes in teachers' classroom practices. Although school directors reported that they believed the training was improving classroom instruction, we did not observe a quantitative pattern of improvements in teachers self-reported practices in the initial study period, and there is currently substantial room for improvement in teachers' use of the types of practices encouraged in the training sequence."	This finding is based on the study's much larger analysis of survey data from a broadly representative sample of teachers, using a propensity score matching design to compare trained teachers to untrained teachers. We believe the data does support the finding as written.
General/TEE p. 58 - 92	Related to finding that TEE training did not change teaching practice, suggest clearer explanation somewhere in report (one option is in conclusions, p. 91/92) that the study did not include a baseline survey or observation or teaching practices prior to the start of TEE training in Georgia, and that the finding indicating no change on teaching practice is based solely on comparison of Cohort 1 and Cohort 2 after the 1st year of TEE training, with some potential for spillover effects.	We agree that it is important to interpret the interim findings with caution--we have added a clarifying statement in the conclusion acknowledging the descriptive nature of the analysis and the reasons why the matched comparison group study may not have captured all of the effects of TEE training.
p. 79	Note additional potential for spillover effects due to Leadership Academy training. While Cohort 2 teachers were surveyed prior to receiving any direct training from TEE, all or nearly all had spent the previous year teaching in schools where their principals and SPDFs were completing the Leadership Academy training, which included content on student-centered learning and supported school leaders to promote teachers' use of student-centered learning.	We present the evaluation's findings about the potential for spillover effects in the following section of the report after this page (discussing data from school directors). In that section, we discuss the fact that spillover effects could be masking potential impact of the TEE training in our matched comparison group analysis.
General/TEE p. 58 - 92	Re: general structure. Suggest making it clear in the introduction to the TEE data analysis (pg. 58) that the TEE evaluation starts with presentation of the post-treatment Cohort I and Cohort II data, and is then followed by comparison of Cohort I and pre-treatment Cohort II data to explore change in practice. At the moment, it feels like you have to figure that out as you go along and a signpost would have been helpful. In the pre-and post-treatment comparison, also suggest modifying labels	We have added an additional roadmap for the TEE analysis section, and adjusted the column labels for the matched comparison group analysis results table.

Page Number	Comment	Mathematica Response
General/TEE p. 58 – 92 (continued)	because confusing to continue with "Cohort I" and "Cohort II" when you're actually measuring different things.	
p. xvi, 76, A-6, C-5	"Confidence" seems to be used as a proxy for "knowledge" in the principle component analysis PCA (pg. A-6). I worry about the impact of this on the matched comparison with pre-treatment Cohort II, since there is some evidence that people become more critical (less confident), the greater their knowledge. This is corroborated by some of the quotes from the focus groups e.g. "I thought I was doing it well, but after the last training, which was about planning, I saw how to plan a lesson..." and "The formative assessment turned out not to be what we thought it was. We found out that the formative assessment is a verbal, sentence" (pg. C-5). Without a baseline, I have some doubts about the validity of the Cohort II pre-treatment self-reported data. However, I haven't seen the survey questions and it might be that they are so specific you think this risk is mitigated. If not, suggest adding a statement about the risk of overconfidence bias.	While in general we share this concern about the validity of self-reported knowledge measures, in practice this issue does not appear to be affecting the data in this study. Instead, there is a pattern of observed increases in self-reported knowledge after training (when comparing trained teachers to untrained teachers providing data at baseline); in other words, on average the TEE training program does not appear to have had negative effects on the self-reported knowledge constructs used as outcomes in the study.
p. 75, E-4	I'm interested in how the different levels of reported occurrence were selected e.g. Groups of students work together during class: Every day. Students present their work to the rest of class: At least once per week. Teachers lecture without students speaking: At least three times per week. Doesn't this pre-selection potentially excluded smaller movements e.g. are we able to know if a teacher previously had groups of students working together during class at least once per week, and now does this at least three times a week? Are greater effects between Cohort I and the comparison group found if lower frequencies are reported?	We have added a new sensitivity test examining practice rate outcomes defined by a less frequent cutoff value (e.g. an outcome defined by whether teachers use formative assessments on a monthly basis, rather than weekly). There were very few differences between the results of the matched comparison group analysis using our benchmark cutoff values and these alternative cutoffs (see Appendix G).
p. xiv, 77-79	Is it possible to compare TEE impact on practice (especially re: ICT and laboratory use) in schools that were and were not rehabilitated?	Because of the need to limit respondent burden across the two evaluations, we did not include rehabilitated schools in the sample for the TEE evaluation study.
p. xv, 18, 29, 77	Propensity Score Matching: Did the propensity score matching take into account participation in prior teacher professional development initiatives? If not, this could be another indicators with a significant masking effect.	Yes, the treatment and comparison groups had very similar rates of participation in prior professional development activities. We have updated the baseline equivalence table to show that the two groups attended prior professional development activities at very similar rates.
p. xv, 18, 29, 77	Propensity Score Matching: By necessity, the PSM selected practitioner teachers for the matched groups. Our analysis of the PMU training management system shows significant variation in training attendance and achievement of learning outcomes between senior/lead teachers and practitioner teachers. Suggest clearer statement on the potential impact of	Data from senior teachers are included in the outcome reporting data presented at the beginning of the TEE analysis (prior to the presentation of matched comparison group analyses that are limited to practitioner teachers). To explore the pattern for senior teachers in more detail, we have added an exploratory subgroup analysis to the TEE

Page Number	Comment	Mathematica Response
p. xv, 18, 29, 77 (continued)	selecting this group to provide generalized findings across the population.	analysis appendix that compares survey results for senior teachers to the survey results for practitioner teachers (see Appendix F). Senior teachers are more likely to use practices related to critical thinking, increasing motivation, and increasing collaboration, and they are more likely to use formal tests to assess learning. We summarize these patterns briefly in the main report, and the full results are shown in Appendix F.
General/TEE p. 58 - 92	Is it possible to do some comparative analysis of Minority and Georgian schools? Analysis of the PMU training management system hows significant variation in training attendance and achievement of learning outcomes between Georgian and Minority schools.	Because minority-language schools were not included as part of the initial waves of training in 2017 and 2018, data collection at these schools did not align with the study's longitudinal data collection plan and study design (which calls for surveying the same sample of teachers in 2017, 2018, and 2019).
p. xvi	re: 55% attendance rate for Cohort II - suggest adding a clearer statement that the survey/report was conducted mid-way through TEE and so the numbers are not final. The attendance rate is now significantly higher.	We have clarified this point.
General/TEE p. 58 - 92	There is no discussion of the impact of exogenous factors on TEE effectiveness (possibly this was out of scope?). Change of Government is currently having significant impact and suggest included in scope of the final report.	Ahead of the final report, we plan to interview government stakeholders regarding any potential policy changes that may affect teacher professional development or recommended classroom instruction practices in the period after the Compact ends. The current change of government did not coincide with the period of data collection included in this interim report.
p.6	Given the research evidence that subject specific prof. dev. has better results, could the final evaluation compare comparative impacts on teachers who only attended the core, versus those who attended the subject specific trainings?	We have added a new sensitivity test examining outcomes separately for teachers who did and did not attend the subject specific training module (see Appendix F). Generally speaking, the pattern of knowledge and practice outcomes in the two groups are very similar.
p.28, 62	Stallings Protocol: suggest that it is more clearly stated that data from the observation of 22 teachers (an extremely small sample) cannot be used to make generalized statements about the population and that clearer statements about the limitations of this data are made each time a conclusion is drawn.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.
p.28	Stallings Protocol: At the time of design, IREX raised some concerns about an imperfect alignment between the Protocol and TEE target teaching practices, and the application of the Protocol by non-ed experts. Suggest clearer acknowledgement that Stallings is administered by non-ed experts and doesn't have a 1:1 correlation for the intended outcomes of TEE.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.

Page Number	Comment	Mathematica Response
P.58	An additional reason for targeting senior teachers for attendance in the first cohort was the Ministry's decision to help assure senior teachers could benefit early from credit awarded for pay and promotion. Practitioner teachers successfully completing courses will accumulate credit but will not be able to use these credits until they have passed their certifying examination in their subjects. This impacted practitioner teachers' motivation.	Thank you for this additional context--we have added this point to the report.
P.61, Table IV.3	Spelling errors in subsection reasons for not attending any training modules, last five responses.	Addressed.
P.85, Table IV.11	Is the reporting of school directors on their belief that the school is a welcoming and safe environment for all students accurately stated? Strongly agreed 68%, Agreed: 100%	We have clarified the table: 100% of directors either 'agreed' or 'strongly agreed' with this statement.
p.62-64	Would recommend two separate graphs for cohorts 1 and 2, or do small multiples that include an overall category. Current format makes it seem like they're being compared despite the narrative emphasizing the different makeup.	We believe some readers have a strong interest in comparing results for the two cohorts of teachers, in part because their seniority levels are so different. Our preference is to leave the data disaggregated for the two cohorts, so the additional detail is shown for readers who prefer to examine each cohort separately.
p. 70	Are the differences between groups significant? Are there implications if so?	The difference is not statistically significant, due to the small sample sizes in the subject-level subgroups. We have clarified this point in the report.
p. 86	The Y axis is not in proper % format	Addressed.
Figures IV.13; 1V.14; 1V.15	Would recommend making these bar graphs instead of column charts as the column chart is hard to read for this type of data	Column charts provide a compact way to convey data and findings or a wide range of outcomes. When the general pattern of outcomes is very similar across survey items (as is the case here), we believe the format strikes an appropriate balance between clarity and space efficiency. In this format it is easy to see at a glance that the pattern of survey responses is very similar across columns.
p. xvi, parag. 3	Attendance rates indicated (school directors - 93%, cohort 1 teachers - 82%, cohort 2 teachers - 55%) refer to survey data. It should be noted as a footnote or somewhere in the text that it does not reflect the actual percentage.	We have clarified this point.
p. 18, parag. 2 p. 25, parag. 2 p. 29, parag. 2	The comparison matching design eliminated the senior level teachers from the sampling. The main focus of the survey is made on practitioner teachers and senior teachers are not reflected in the report. We do not know if the project effect was the same or in any way differed for senior teachers. I think the report shall reflect senior teachers as well although it would be a limitation not to make any valid judgements since we have no baseline data for them but at least give descriptive report of the survey data.	Data from senior teachers are included in the outcome reporting data presented at the beginning of the TEE analysis (prior to the presentation of matched comparison group analyses that are limited to practitioner teachers). To explore the pattern for senior teachers in more detail, we have added an exploratory subgroup analysis (see Appendix F) that compares survey results for senior teachers to the survey results for practitioner teachers.

Page Number	Comment	Mathematica Response
p. 18, parag. 2 p. 25, parag. 2 p. 29, parag. 2 (continued)	It must be taken into account that senior teachers had better attendance rates than the practitioners in both cohorts. What if they were more motivated to apply TEE gained knowledge into practice? Moreover, there were senior teachers in the second cohort and small sampling could have hopefully been made to make comparison match. If Stallings sampling is ok why small sampling would not make any sense to compare cohort 1 and cohort 2 senior teachers?	
p. 26, parag. 1	TPDC's data list of teachers that was shared with Mathematica did not include Core 3 and Subject teachers from cohort 2. It should be noted that it was not final	We have clarified this point.
p. 28, Stalling classroom observation	No clear correlation between stallings survey questions and TEE learning outcomes. More details are needed to explain in what way stallings observation data reflects what was taught by the TEE modules.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.
p. 28, parag. 4	Sampling of 22 teachers in the Stallings is too small to generalize or even use as the validation for survey data.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.
p. 59, parag. 3	Percentages given shall be footnoted that they represent survey data and not the TEE participants administrative data.	We have clarified this point.
p. 60, parag. 3	"Among teachers who did not attend any TEE trainings... , a quarter of nonattendance appeared to occur because teachers believed they had not been invited to attend the training...." It sound too loud in the text and could be misinterpreted. Table IV.3. note says that it's only out of 33 C1 and 29 C2 nonattendants and giving the percentages (31%, 25% or even 12%) may drive the reader to the wrong translation. 25% of 33 nonattendants is only 8 teachers and 31% of 29 is only 9. at least should be taken out from the major findings.	We agree that the sample of nonattending teachers is relatively small, and we have made edits to the results table and corresponding text to emphasize this point.
p. 63, parag. 1	Second line should be corrected: 65% of cohort 1 teachers were more senior level teachers, not practitioners	We have clarified this statement.

Page Number	Comment	Mathematica Response
p. 63, parag. 2	"...among the teachers in the survey sample and across both cohorts, we found evidence that many of the practices that were encouraged by the TEE training sequence were only being applied to a limited extent in classrooms." It would be interesting to explore this more. what kind of changes were observed even in the limited number of teachers. and whether these teachers differed from others in terms of age, status, cohort.	In the main report we compare the teaching outcomes of Cohort 1 and Cohort 2 teachers in the first month after completing the training sequence. To explore these patterns in more depth for additional subgroups, we have added a new appendix to the report (Appendix F) that disaggregates results by teacher age and teacher seniority status as well.
p. 66, parag. 2	Slightly different percentage numbers in text and in Figure IV.6.	Addressed.
p. 69	Stallings observations findings - it is interesting to have more details about the lessons observed - what kind of lessons were observed. From my experience, teachers change the goal of the lesson when they have an observer without prior notice. It's almost always a revision of previous lessons and very rarely it's an ordinary lesson. The point is that revision lessons differ in nature from a new topic introduction lesson. On a revision lesson they might not need to use IT or labs. So that IT and Lab use data might be irrelevant in these cases.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix, since the study's primary findings are based on the much larger sample of teachers who completed the study's quantitative survey focused on teacher knowledge and practices.
p. 72, Figure IV.11.	Percentages given sound too loud than numbers out of 22 observed teachers would.	As noted above, in response to these concerns we have moved the classroom observation findings to a report appendix.
p. 83, School director knowledge and practices after training	Only the data from school directors survey questionnaires are described in the report - what they perceive has changed in practice of their own or of teachers. Very large percentages show that they overestimate the processes. 50% provide advice on teaching practices once a week and 89% - at least once a month. Number of observations is also quite high, and many issues that have that high percentages - it would be great if teachers reflection is also represented in the report - how they felt the school directors' described behaviors are happening or not and to what extend - the same way as was done for teachers, having triangulation of students survey and focus groups.	We have added a new table to this section of the report summarizing what teachers say about the instructional practices of school directors. Broadly speaking, teachers' survey responses were consistent with what school directors said in the self-reported survey.
multiple	The report lacks any findings about the study group/quarterly meeting experience among the teachers and school directors. Don't remember whether these questions were reflected in the survey design but since study groups were something new for teachers as well as quarterly meeting for directors, it would be interesting to hear what was the experience and whether they had any effect on teachers.	The study's qualitative findings about the teacher study groups are summarized on p. 85 (section 6 of the TEE results chapter).

Page Number	Comment	Mathematica Response
multiple	For policy makers and donors, it would also be interesting to have the findings disaggregated by age groups - to know for example whether over 20-30 year-experienced teachers show better results or, on the contrary, are resistant to changes, while younger generations of teachers are enthusiastic and motivated to adapt changes or not. Possible differences among status levels would also be helpful.	To explore the pattern for teachers with more than 20 years of experience, we have added an exploratory subgroup analysis as an appendix (see Appendix F). We did not find any substantial pattern of differences between teachers with more than 20 years of experience and less-experienced teachers, but there are differences between senior teachers and practitioner teachers.
pg. xi	All the issued comments to the Report draft findings have been based on given stated questions that are directed to evaluate the actual results of the beneficiary schools infrastructure rehabilitation and not just measure perception of the beneficiaries regarding generated results, which is the common explanation by Report authors that the aim of the interim report was not to evaluate reaching infra related objectives but to identify subjective judgment regarding particular infra improvement performed within the Compact II project frames. All the stated questions are more like project implementation efficiency related audit asked questions and not just beneficiaries subjective perception evaluation questions.	The evaluation is designed to measure not only program outputs (the changes in infrastructure occurring at rehabilitated schools) but also the corresponding changes in the learning environment that could ultimately produce changes in student learning and educational attainment. We report on the data collection team's direct measurement of changes in school conditions, but we also believe that survey and qualitative data from students, teachers, and directors provide important insights about whether these infrastructure changes are working as intended.
pg. xii	Here, to our mind, the used questioners are not methodologically and conceptually correct and consistent: it is not clear what exactly under "One problem one" or "two problems" both the author and the beneficiary ment prior rehabilitation and post rehabilitation, since no explanation has been provided in the report text. For instance, let's elaborate a little bit on just cracks (when the provided analysis has also been applicable to other listed factors such as "water damage, mold, chipped or peeling paint, or holes in ceilings and floors") Prior rehabilitation does "one problem" mean "one crack in one classroom" or "multiple cracks in one classroom" or one crack is "one problem" and multiple cracks "more than two problems"? Here also is very important whether the crack is structural or just cosmetic, since in different cases the implications are totally different.	Thank you for this suggestion. The infrastructure assessments (completed by teams of engineering students) did categorize the types of "problems" in more detail, and we have added a new figure (Figure III.3) to the report showing the pre-rehabilitation and post-rehabilitation pattern for each type of observed problem with classroom walls. As you note, the figure shows that there is a more pronounced pattern of improvements for more severe problems, and a less pronounced pattern of improvements for more superficial issues (such as a visible crack in the wall that is not large enough to be classified as a "hole").
pg. xii	In so far as comment #1, under interim findings, one can develop an impression that 19% of the students reported that installed heating system had been a concern because the heating system had not been properly installed or had not been working at all. If, as authors have been stating that, this is just subjective perceptive evaluation by certain fraction of the polled beneficiaries, it should be made clear that despite the modern heating system had been installed in the newly rehabilitated school, some students report it as a concern.	We have revised this statement to clarify that the heating system was installed in all rehabilitated schools.

Page Number	Comment	Mathematica Response
pg. xiii	Proceeding from the comment #1, especially referring to the question: "1. Was the ILEI activity budgeted and planned appropriately, forecasting key risks?", the reader will definitely have formed an opinion that, given stated fact, the ILEI activities had not been properly budgeted and planned, when, all project related papers or relevant technical documentation clearly envisions increase of utility costs for rehabilitated schools, which is natural outcome of the project related activities. If prior rehabilitation school either did not have at least on bulb in the classroom, central heating system, any scientific lab equipment or even water supply, certainly after all of these listed means will be installed/provided, generally maintenance cost, including utility bills definitely would have increased three, four five time as much as it was before rehabilitation. Having expected increases, the MCA fund negotiated with Georgian Government to co-finance newly incurred expenses and has reached agreement on the matter. Thus if this Report evaluates ILEI activity run by MCA fund provided clarification definitely should have been included, not remaining any room for dual interpretation that the ILEI activity implementation team did not properly budget or plan the project and forecast respective risks. Before rehabilitation, school administration was providing heating in limited spaces, using wood stoves for the duration of lessons approx. 30-45% of school space. After rehabilitation all 100% of the building is being heated. As per sq.m heating cost remained or decreased, heating area is now larger than previously, hence the increased cost.	We have clarified that these increases were an expected consequence of rehabilitation investments. However, it remains true that in Phase I schools directors are reporting that they do not have sufficient funds to meeting these increased utility costs, and we believe this is an important findings to state clearly in the evaluation report. In the final evaluation report, we will assess if the government arrangements to offset increased utility costs are helping to make these costs easier to manage in the Phase II and Phase III schools in the evaluation sample.
pg. xiii	Once again, it should be mentioned that provided wording may create wrong expectation that air quality in the classroom has not been improved as a result of the performed rehabilitation. Due to budget limitations in the design criteria elaborated for the purposed of this particular project, mechanical ventilation systems have been provided only in Scientific labs and partly gyms. In all other rehabilitated classroom the natural ventilation - opening the classroom windows - has been envisioned. Thus if former air pollutant means - wood stoves - have been eliminated as envisioned by the framework of the project implementation the objective has been achieved if, again, all newly installed central heating systems operate. As to the air quality subjective assessment by students or by direct air quality measurements by specific instruments, to our mind, that should not be attributed to the rehabilitation results exactly complying with the design requirements but both perceptive assessments and specific instrument	We have revised this section of the report to clarify that other potential sources of air pollution are unrelated to heating systems and the infrastructure investments that were the main focus of the Compact. Our focus here was to examine the extent to which air quality improvements related to removing wood stoves are likely to affect learning outcomes. If there are other sources of indoor or outdoor air quality problems in these schools, it might limit the extent to which students will experience learning (or health) benefits related to air quality changes.

Page Number	Comment	Mathematica Response
pg. xiii (continued)	measurement of air quality outside the school facility provided, since if classrooms are being systematically ventilated by opening the windows, air quality will be the same inside and outside, and if not, this means that classrooms are not being systematically ventilated, which is not within the frames of MCA responsibility.	
pg. xiii	Lighting of the rehabilitated schools facilities had been installed according to the approved design criteria, specifying modern international best practices and scientific requirements toward lighting in the educational services providing facilities, including specific requirements for secondary education facilities - Public Schools. Since all the designs and respectively lighting Luminas quantities were exactly the same across all rehabilitated schools, additional clarification is needed to underline that lighting had been installed according to the set design and requirements, but some fraction of students reported having difficulties reading from the blackboard that also can be caused by other reasons, including health issues, that definitely go beyond rehabilitation project responsibility boundaries.	We have added a clarifying statement here explaining that the remaining issues after rehabilitation were not related to an absence of installed lighting. In Chapter III, we have also included an additional finding (footnote 6) explaining the data collected by infrastructure assessments teams on the actual light-levels (measured in lumens) in each classroom.
pg. ix	All toilets were in working order at the point of rehab completion and was approved by the engineer. During DLP period some of the defects have been identified and addressed as required. It needs to be noted that toilet systems as any other parts of the school are subject to proper operation and maintenance, which is exclusive responsibility of school administration.	We have clarified the distinction between maintenance issues and the initial rehabilitation package.
pp. 55-57	Maintenance is largely ignored in this interim evaluation report. Section III.D.2 is titled "Operations and Maintenance" but it focuses on operation costs (the tripling of utility costs caused by school rehabilitation) and says nothing about maintenance. It is important to remember that MCC was asked to rehabilitate schools in Georgia only because the Georgian government had neglected maintenance for decades, with the result that many (most?) public schools were in extremely dilapidated condition. Rehabilitating schools that will just fall apart again due to lack of maintenance makes no sense and is a waste of taxpayer money. If the schools are not maintained post-compact, the benefits won't be sustained and the project will have failed. MPR does not seem to appreciate the severity of this problem. Although "facility-maintenance funding" is mentioned as one of the evaluation questions, and the first of the four questions on page xii is about maintenance policy and practice, no	This interim evaluation report is focused only on a small subset of ILEI schools, and the timing of data collection for the interim report means that the analysis is almost exclusively limited to data collected in the first year following rehabilitation efforts in treatment schools. The final evaluation report will include two years of follow-up data for each treatment school, including direct measurement of infrastructure conditions in rehabilitated schools two years after the program is complete. If heating systems, electric lighting systems, water and sanitation systems, or classroom walls, ceilings, or floors are not properly maintained during this two-year follow-up period the evaluation will be able to directly measure those outcomes (and quantify the extent to which conditions deteriorated between the first and second follow-up year, if such a pattern occurs). The current evaluation design is limited to two follow-up years of data collection. MCC may consider pursuing longer-term analyses to assess maintenance issues

Page Number	Comment	Mathematica Response
pp. 55-57 (continued)	meaningful approach to assessing maintenance post-compact is presented in the report. Table I.1. (page 8) titled “Evaluation questions for the ILEI activity and approaches to answering them” makes it clear that MPR intends only to “Interview school directors to gather data on operations and maintenance funding and maintenance practices” as part of the planned qualitative survey. This approach will not reveal whether compact-funded school infrastructure is actually being maintained. School directors are unlikely to have a maintenance line item in their budget and might not track maintenance expenditures. Moreover, what the director says and what happens in practice could be very different. The only way to reasonably assess whether schools are being maintained post-compact is to send building engineers to visit the schools several years after rehabilitation work has been completed. According to page 21, “The infrastructure assessment teams were comprised of enumerators with engineering backgrounds who received training on how to consistently measure air quality, building systems, light levels, and temperature.” This makes it clear that MPR can afford to send engineers to the rehabilitated schools, but the current plan is to use engineers only to measure air quality, building systems, light levels and temperature.” It does not appear that MPR has any plan to send engineers to visit schools during the second follow-up survey to assess, through visual inspection, whether or not the compact-funded infrastructure is actually being maintained. Given the critical importance of post-compact maintenance for the sustainability of the benefits and the success of the project, this is a major oversight.	over an extended period, and we would agree that such an analysis could be an interesting and valuable contribution to the study.
page 4	MPR asserts that MCC's ERR calculations for the school rehabilitation activity aim to produce "a 10 percent improvement in the number of students enrolling in upper secondary school and a 10 percent improvement in postsecondary enrollment rates." This is not correct. MCC's ERR calculations assume that student transition rates (from lower secondary to upper secondary and from upper secondary to post-secondary education) will improve by 10% not that enrollment will improve by 10%. Enrollment is not the same thing as the student transition rate.	We have clarified this statement.
Pg. xi - xii	More context is needed in the executive summary to allow the reader to understand the reported results. The bottom paragraph on this page, going into the next page, or that paragraph before the interim findings are provided would benefit from noting what baseline is, when it was done, and when the follow-up was done, a sense for how much time	We have clarified the data collection timeline in this section of the executive summary.

Page Number	Comment	Mathematica Response
Pg. xi - xii	had passed between the two or rather the time since rehabilitated, the seasons the info was collected, etc. The mention of pre-post is insufficient without noting more about what is pre and what is post. There is one finding on pg. xiii that notes 'one-year follow-up surveys', but unclear how that fits in. When was the data collected? There is mention of what the endline will do, but clear timing note would be helpful here too - when data is expected to be collected and/or how much time will have passed for the schools rehabilitated.	
Pg. xiii	During recent mission there was mention that schools with lots of children enrolled should easily be able to cover the increased utility costs, but those with less children may face the greatest challenge because GoG funding is based on $$ per student. Does the data collected to this point shed any light on this possibility? Will the endline be able to examine this potential difference? It was also noted that labs are where the electricity will increase – besides heat - and this would be a year-round cost. Did you explore this in the interim? Do you plan to for the endline?	The endline report will compare the cost-burden of utilities at the smaller-enrollment Phase I schools to the cost-burden of utilities at the larger-enrollment Phase II and Phase III schools. The final report will also feature RCT impact analyses that compare the cost burden of utilities in the treatment group to the cost-burden in the control group of schools that were not rehabilitated.
Pg. xiii	Note that air quality was reported as fair, rather than good, and that this seems to be confirmed by the air tests. What are the other pollutants and their sources? What is causing this? Did we not address a key issue? It seems worth noting upfront.	In the executive summary we have clarified the other potential sources of air pollution (dust, deteriorating paint, or outdoor air pollution). The RCT analysis in the final report will shed light on whether there have in fact been meaningful improvements in air quality in rehabilitated schools, as compared to a well-identified control group.
Pg. xiii	Based on the reported results it is not clear whether there were actual tests completed to measure lumins. Were there? If so, then how does this relate to what was reported by the students.	We did record lumen-levels in classrooms, and we have added information to the report noting that there is a modest pattern of improvement in directly measured classroom light levels.
Pg. xiv	Will the endline ask teachers specifically if they received training in using a science lab? That would seem necessary to better understand whether they are being used. This is also a linkage with the teacher training activity.	Yes, the endline teacher survey does include a question about whether teachers have been trained in how to use a science lab, and we will examine this as part of the analysis of science lab usage rates in the final report.
Executive Summary	I know that this is the executive summary, but it would be useful to provide more information on the objectives of each Activity - rehab and teacher training. What were the outputs? This interim evaluation is mainly focused on whether the inputs to outputs were achieved and maybe a glance at short-term outcomes. It is hard to get a sense for what to compare the summarized results against without telling the reader what the project was aiming to produce within this time period that you are looking at. It seems to jump into the research design and interim	Thank you for this suggestion--we have added additional contextual information about each program in the executive summary.

Page Number	Comment	Mathematica Response
Executive Summary (continued)	findings. For example, how many schools are aimed to be rehabilitated, how many teaching modules, how many hours of training, it is nationwide, etc.	
throughout	Minor, but typically we use Activity as capitalized when referring to the specific program, same with Project and Compact - this can help to distinguish and highlight when we are talking about the specific component vs. the terms in general.	We have reviewed this punctuation pattern and capitalized terms for proper nouns.
general	There are a couple of places where the report notes the final report in 2021, but I had thought that it was expected by the end of 2020. Just wanted to be clear and ensure that all are aware of the timing.	We expect to complete a draft report for stakeholder review by the end of 2020, and finalize it for public release in early 2021.
multiple pages	MPR's interim evaluation finds that, while teachers increased their knowledge of good teaching practices by attending the compact-funding training courses, for the most part teachers did not apply the recommended practices in their classrooms. The conclusion states "…the interim analysis did not reveal consistent evidence of short term changes in teachers' classroom practices" [p. 91]. MPR attributes this lack of uptake to the fact that the main assessment took place shortly after the training finished, and notes that "...the program's theory of change did not predict that teaching practices would change in the immediate aftermath of the training sequence." According to MPR, the program's theory of change "...states that the training will improve teacher knowledge…which will then (over a period of several years) improve teacher's classroom instruction in ways that can ultimately improve students' learning outcomes" [p. 62]. In other words, MPR is asserting that the impact of the training on actual classroom practice will increase over time. This is counterintuitive. One would expect application of newly acquired knowledge to be greatest when memory of the new knowledge is freshest. MPR presents no evidence to support the assertion that teachers will gradually increase their classroom application of the promoted teaching practices over time. Moreover, the assertion of a lag effect is directly contradicted by the evidence in Annex D-1 Table D.1 of their report, which shows that Cohort 1 teachers did not increase their application of the promoted teaching practices over time. MPR presents the data in Annex D-1 but never discusses what the data show.	The program logic developed by MCC, MCA-G, and consultants for the TEE activity repeatedly stated in very clear terms that implementers did not expect to observe changes in teaching practices in the first year after completion of the training sequence. We are not asserting whether or not this pattern is likely to occur. Rather, we designed the evaluation and its data collection schedule to examine if the pattern assumed in the program logic is taking place. This is why the final analysis will include follow-up data collection activities for Cohort 1 teachers over three years (and follow-up data collection activities for Cohort 2 teachers over two years). Because this interim report does not include enough data to conduct the evaluation's complete trend analysis, in our view it would be premature to focus on the partial trend-data for Cohort 1 teachers that are currently shown in Appendix D. The final evaluation report will prominently feature findings from all of the follow-up years in the trend analysis, and more fully assess whether the program logic's assumed pattern of medium-term improvements in teacher-practice outcomes actually occurred.

Page Number	Comment	Mathematica Response
multiple pages (continued)	The interim evaluation needs to incorporate and discuss the findings from Table D.1 into the main body of the report. Since this will be a public report, MPR's assertion of a lag effect, without any evidentiary basis, and their failure to discuss contradictory evidence available in Annex D-1, could raise concerns among readers that the evaluators want to avoid conveying bad news.	
	Directly related to Peter's comment above: What does 'shortly after' mean here? In another place it notes 'initial period'. And, although that is from when they completed the final training module there were 4 modules, so how does this idea link to this theory of waiting longer to see impacts. Were there topics taught in the first module that we would be more likely to see implemented now, but not those in say module 3 or 4? What was assumed to happened pre-evaluation? How does that stack up to what we found?	As noted in the response to the prior comment, the final evaluation report will present results over a three-year follow-up period (assessing if the pattern assumed in the project logic took place over time). In the interim report, our primarily analyses focused only on survey data gathered within one month after teachers completed the training sequence, and we have clarified this is what we mean by the term 'shortly after training' in the report.
pg. xvi & xvii, program logic reference	Also related to above two comments: The timing is not really given in the logic (at least the diagram) other than the distinction between outputs of them being trained and the improvements in teaching. Actually, the elements of teacher study groups and professional networks among directors are missing from the program's logic (again, referring to the diagram). This seems to be an oversight and something that may need to be updated in a final M&E Plan or capture of the program's design. Perhaps useful to emphasize this element within the evaluation and the role that it is believed/expected to play in getting teachers to apply their new knowledge and use the newly acquired resources to improve classroom practices. Why wouldn't we expect 'immediate' changes in teaching practices? What else needs to occur? Is that planned to occur?	As explained in the evaluation design report for the TEE activity, stakeholders and implementers (as shown in GOPA and later IREX program planning documents) believe that the earliest potential timeline for observing changes in teaching practice would be 1-3 years. Implementers hypothesized that teachers would be slow to adopt new practices for multiple reasons. For example, the training sequence concluded in September (the first month of the school year), and implementers thought it would be unreasonable to expect teachers to incorporate new lesson plans as the school year was getting underway. They expected teachers to begin piloting new practices during the first year, and enact changes more consistently over time. We designed the evaluation and its data collection schedule to test this hypothesis. See page 7 of the report for a revised discussion of these issues.
pg. xvi	end of first paragraph on findings summarized: 'both groups attended at least one training session'. Do you mean completion of a module or that they literally just showed up to one training session? How many modules are there? How many training sessions are there within a module? This finding is not particularly helpful without more context.	We have clarified here that the teacher training included 4 modules, where each module was scheduled to take place in a multi-day, in-person training session.

Page Number	Comment	Mathematica Response
pg. xvi, last paragraph	Overall, I don't think that it is particularly clear how an increase in confidence demonstrates knowledge. Are there other measures or questions that we can draw from to get at knowledge, perhaps other data collection methods? Will there be others in the endline? Seems like you should mention here that they were already doing really well on this - at least as self-reported. Does this seem high? It is statistically significant, but is it substantive? Do you think that the non-trainees fully understood what it meant to know how to do this work?	We recognize the limitations of self-reported knowledge measures in survey data, which is why the evaluation complemented this outcome-measurement approach with additional measures (student surveys, director surveys about teachers, classroom observations, and teachers' self-reported use of these practices). In addition, the fact that the evaluation did detect differences between the confidence-levels of trained teachers and untrained teachers suggests that there was room for improvement in the comparison group at baseline (even though baseline confidence levels were relatively high).
xvii, first para. Findings	There is no mention within this paragraph on how this compares to the comparison group, which makes it more difficult to interpret the results fully. (2) For informal assessments, what is the timing asked of the students - per day, week, month, etc.? (3) For classroom observations - this is a small sample size, so caution in interpretation and reporting is needed, as well as noting results relative to comparison.	We have revised this paragraph to clarify that the classroom observations and student surveys did not attempt to compare trained teachers to untrained teachers. Rather, these other data sources cross-check the pattern of self-reported practices among trained teachers. All of these data sources suggest that there is room for improvement in the use of practices targeted by the TEE activity.
xviii, 2nd para. Finding	Given what the focus group notes, it sounds worth exploring further the ability to disaggregate the results - particularly by classroom size, perhaps categorical groups (small, medium, large?!?). Is this possible at this interim point or for the endline?	The evaluation did not collect data linking teachers to classroom rosters or classroom-sizes, so our existing data would not support a subgroup analysis categorizing teachers by class size. It is also likely that class size is highly correlated with confounding variables (such as urban vs. rural locations) that would make it difficult to interpret the results of such an analysis.
xvii, 3rd para.	Portfolio' is unlikely to be understood by the larger audience, so helpful to define.	Revised and clarified.
general	Typically, directly after a training there is a satisfaction survey that asks about pieces they learned, areas for improvement, etc. Where are these types of questions? Who is doing this type of evaluation to ensure that the courses are updated as needed? As it is a common practice it could be good to explicitly state that this was not done and explain why.	The study did include multiple measures of teacher and school director perceptions about the quality of the training sequence. For example, in the analysis of teachers' motivations for attending and completing the training sequence (see Table IV.1) we report that nearly 90% of attending teachers believed attending the training would improve their practice. This is in accordance with findings in teacher focus groups that satisfaction with the training was generally high. Our understanding is that interim satisfaction surveys were also collected by TPDC/IREX as part of program implementation, but that type of implementation-support data is not the primary focus of the evaluation design. Since the evaluation includes more direct measures of the training program's inputs (attendance rates), outputs (changes in teachers and school director knowledge and attitudes), and outcomes (changes in practice), we believe incorporating additional satisfaction survey data would be of limited value to the report.

Page Number	Comment	Mathematica Response
pg. 1	Overview of evaluated activities: Additional information on the project is needed here, particularly the teacher training piece which provides no sense for what the teachers and school directors were trained in, for how long, when, who, etc. There is a little more for school rehab.	We have added more introductory information about the TEE teacher and director training sequence here.
p. 6 and p. 15	In the literature review on page 6, MPR writes that: "…teacher training interventions tailored to specific academic subjects tended to be associated with larger gains in student learning." This implies that interventions that provide training in subject matter will have a bigger impact on student learning than those that focus on pedagogy – a finding consistent with published literature. Despite this acknowledgement, MPR's evaluation questions in Table I.5 (p. 15) focus entirely on pedagogy (e.g., student-centered instruction, formative assessments, and classroom management) with no questions related to subject matter training. The report makes clear that MPR has no plan to assess subject matter training and will only assess the impact of training in pedagogy. If this is because the TEE activity did not provide subject matter training, MPR should say so and should acknowledge explicitly that the design of the activity did not conform to good practice or to available evidence about what works. If the TEE activity did provide subject matter training, then MPR's evaluation questions need to be revised.	The TEE intervention included three training modules focused on general teaching practices, and a fourth module tailored to specific subjects that addressed how these core practices can be applied in the fields of science, mathematics, geography, and English instruction. Since all four modules focused on the same set of pedagogical practices, we believe it is appropriate to give those general practices a prominent place the evaluation's research questions. To further investigate the subject-specific module of the training sequence, we have added an exploratory analysis to the TEE results appendix (Appendix F). This analysis tested whether the outcomes of teachers who completed the subject-specific module of the TEE training program differed from the outcomes of teachers who only completed one or more of the core training modules (without attending the subject-specific module). We did not find any differences in the core teaching practices in these two groups, but we analysis did reveal that science teachers who attended the science training module conducted lab experiments at a greater rate than science teachers who did not attend that module. See Appendix F.
pg. 6-7, lit review	In general, I think that this section could benefit from additional detail as it is currently quite vague using words like largest, wide range of results, improvements in learning, large effects, etc. and it also needs tighter linkages back to this study. Additionally, I think that a major flaw is that the narrative seems to be equating lack of evidence within the literature - due to no one having done this yet - to lack of the project's potential for success - i.e., overall impacts. It seems to be heavily slanted on what is wrong with our program's design, based on almost no evidence or linking to what we know, rather than highlighting the positive potential pieces of the program's designs to reaching results but noting that these have not been evaluated - so we don't know - and then emphasize the large gap in the literature that this can help to fill. (1) For US literature mentioned please provide the SD reported or other relevant outcomes to compare to those assumed for the CBA. Without this it is difficult to understand it's context and comparison;	We have revised and clarified the content in this brief literature review section to address many of these items. In particular, we appreciate that it is possible to clarify the extent to which prior studies examined interventions that are similar to (or different from) the TEE trainings, and have clarified those examples. More generally, in our view this literature review section is intended to provide a brief overview of the relevant literature on the effects of teacher training interventions and assess how relevant that literature may be to the TEE activity. Since this is only an interim report, we do not think readers will be looking for an in-depth or exhaustively detailed literature review or meta-analysis. The program logic assessment and public evaluation design report for this study do include some of the additional requested detail, and we hope those documents will be a useful resource.

Page Number	Comment	Mathematica Response
pg. 6-7, lit review (continued)	(2) You state that Evans and Popova 2015 find greater outcomes for those that include on-going follow-up support for teachers. Based on your program description in the report this seems to be part of the program, as related to teacher study groups and professional networks among directors. So, does this mean that you do not think that this will actually be implemented? If not, then why? (3) Last line of first paragraph on pg. 7 is really unclear on what is intending to be said, is this in reference to the MCC project, what is the main takeaway here?; (4) Regional differences are not just related to teachers, but the students, schooling system and culture overall, expectations of schooling, etc.; (5) Reference to Hill, et al. 2008 is unclear and not compelling with the level of info currently provided. 'Substantial evidence' in the form of what exactly - student learning, what are the differences in SD reported, what types of studies are these and what are 'early' grades vs. 'later' grades? - and what is the context of measuring these outcomes - i.e., a specific program, national testing, etc. (6) There are also potentially positive aspects about TEE Activity being nationwide, but none are mentioned. I have not scoured the literature, but I would think that there this could create a herd mentality to adopt new practices, facilitate discussions on the topics and share experiences. This has the potential to change the culture of the system and what is the idea of expectations of a teacher; (7) There are two other potentially positive aspects about the program's design that I don't think are mentioned - first, if students are getting teachers at all levels (grades) that have the new training and across all subjects then that should help to solidify the gains in learning. This comprehensive approach is not typically taken but has the potential to change their learning environment for 6 years - if results of teacher learning and practices were to be sustained. (8) There is no mention about training of school directors, this also helps to enforce application of teaching practices, better school management of resources, etc. This was also nationwide.	
In general	The component on laboratory training seems to be missing. As this was a large investment in the rehabilitated schools and related to the overall STEM focus for the Compact this seems to be an important omission that should be examined further in the endline data collection, analysis and reporting. This would not only be about laboratory safety, but training to complete the actual lab work.	We have gathered additional information about the laboratory training activity and the timing of the trainings in relation to the interim data collection and analysis. As it happens, much of the laboratory training took place after the interim data collection round for the ILEI study: we have noted in the revised report that we will be able to examine the effects of the laboratory training sequence more fully in the endline data collection and analysis.

Page Number	Comment	Mathematica Response
pg. 14, 15 Evaluation Design	Perhaps there is an easier way to clearly demonstrate the various forms of data collection, the sample, and how these link to answering the evaluation questions. For example, when are you pulling data from the other teachers, are their students included in the student survey? Were they included in the classroom observations?	We have revised Table I.5 to help clarify the links between samples in the TEE analysis
pg. 16-17	As noted in the executive summary, this performance evaluation description only provides dates of when things were done, but not how much time passed from training to surveying or x to data collection. Perhaps a timeline would be useful. As the timing could be quite key to how we interpret the results I think emphasizing them more would be helpful.	We have added additional clarifying information about the timing of data collection here.
In general	MCC has a specific definition that is used for the term 'beneficiary' - I can provide more detail if wanted. However, given this, it would be helpful to use words like participants, students, teachers, directors, etc. or define the term within the report. This will help to avoid any confusion by the reader.	Revised and addressed.
In general	Is there sufficient data to incorporate student survey comparisons for teachers in the treatment and comparison groups? Given the study's design, less rigorous causal claims, having an additional way to triangulate the data could be useful.	The student survey data used in the TEE analysis was derived from a convenience sample of students in the ILEI study's treatment and control group. Since the data did not contain teacher-student links, it is not possible to separate students of teachers in the TEE treatment group from students of teachers in the TEE matched comparison group. The student survey also took place in spring 2018, after the TEE matched comparison group completed the training sequence.
pg. 19	Would seem appropriate to provide more detail on the potential strengths and weaknesses of the design, or lay out a few caveats of how the results can and cannot be interpreted. Particularly, this will be important for the classroom observations, given the small sample size.	We have added an explanation here about the limitations in the classroom observation data. As noted above, in response to these concerns we have also moved the classroom observation findings to a report appendix.
	My biggest question/comment relates to changes in absenteeism due to our interventions (or not). Perhaps I missed this, but how was MPR able to account for the fact that the length of the school day is actually variable, dependent on grade level, e.g. first graders go to school for ~4 hours, vs. 7-th-graders, who may have 7.5 hours of class a day? I think we could obtain official Ministry of Ed requirements on schooling hours, though perhaps MPR has this already.	The evaluation collected a direct measure of student attendance at a consistent point in time, when all enrolled students in a given school-shift were present. The survey-based measures of attendance collected from directors, teachers, and students would not have been affected by differences in class-hours by grade, and the pre-post comparison of school-wide attendance patterns before and after rehabilitation also would be affected by this issue.

Page Number	Comment	Mathematica Response
	Second comment is re: shifts, e.g. do heating costs go up because schools are required to run a second shift or open more sections due to increased student numbers (e.g. students are "migrating" from our rehabilitated schools, away from nearby, still-dilapidated ones)?	In the ILEI chapter section on enrollment outcomes, we report that directors did not find it necessary to add a second shift after rehabilitation (at least in the interim study's sample and follow-up period). We will examine this issue again in the endline report.
	Grade promotion is also a very "political" with a small "p" issue, e.g. not sure we could really make an impact on grade promotion rates with our interventions, but that is perhaps a different story.	We will be able to make a final assessment on whether the ILEI activity impacted grade promotion as an outcome, in the final report.
xiv	Interesting qualitative finding on the impact of improved sanitation facilities on girls. Will be interesting to see if in the long we can correlate these changes with increases in female student performance.	We agree it will be interesting and important to conduct subgroup analyses by gender for the achievement outcomes in the final report.
	I don't know if any baseline was established with regards to SBGBV (school based gender based violence) but it would be interesting to see what impact having indoor bathroom facilities have on girls' perceptions of safety and impact for their ability to concentrate and perform.	As noted in the qualitative findings for the ILEI activity, students do perceive that new bathroom facilities have provided important safety benefits for girls. Our baseline survey did not include a specific item on SBGBV, however.
	The issue of uneven usage of school labs is interesting. For the final report, it would be interesting to see if Mathematica can isolate the major constraints (teacher preparedness, concerns to conserve consumable lab ingredients, etc.)	As noted above, we have added information to the report clarifying that the ILEI activity's science lab training sequence had not been provided to teachers in rehabilitated schools at the time of the interim study's data collection. It will be interesting to see if the science lab usage rate changes in the endline survey data, which is being collected after the training.
XVII	the pedagogical changes that have not been as quickly implemented (differentiated learning, collaborative learning exercises, etc) are not surprising as these sorts of changes require forethought and extensive planning. I like that you cross referenced teachers' perceptions of their improved pedagogy with students' reporting on incidence of new teaching techniques. While both are perception based, it would be valuable to share the students' perceptions with teachers.	Thank you for this feedback.
24	Is it by design that bathrooms and sanitation facilities are not included in the measures?	The summary indices developed by the study are intended to present summary measures for outcomes with multiple underlying components. In our view a single bathroom and sanitation facility measure (presence of flush toilets) could serve as a primary outcome measure, so we did not develop an index of multiple measures for the sanitation outcome.
35	What is the reason for the decreased amount of outdoor recreational space in the rehabilitated schools?	In our view, improvements made to indoor gyms may have made maintenance of an outdoor space relatively less appealing, particularly in the winter month of February when data was collected. We have revised the report to explain this possibility more prominently.

Page Number	Comment	Mathematica Response
43	it would be interesting to unpack what "comfortable" using the lavatories in the treatment schools means to respondents- is it about cleanliness? Safety (stall with doors and working locks)?	Our student focus groups provided additional insights on which aspects of the comfort improvements were most important. These findings are discussed at the end of the chapter section addressing sanitation facilities. In focus groups, students reported that the location of renovated lavatories (inside the building versus outside, as previously), the privacy of the stalls (with doors versus without doors, as previously), the presence of flush toilets using running water, and the availability of sinks with running water for hand washing were critical improvements for students.
65	Disappointing preliminary findings regarding teachers' lackluster engagement and promotion of social inclusion. Is it possible to unpack why this is- is it due to a lack of pedagogical materials or a lack of personal beliefs and commitment? Because these might not be testable subjects? This and gender issues are mentioned almost in passing- I'd like to see a greater analysis of these findings.	One way that the evaluation explored these findings in greater depth was in the matched comparison group analysis (see Table IV.6) which tested whether training improved knowledge of inclusionary teaching practices and the frequency of teaching practices related to inclusion. Training did have an effect on confidence levels, but that did not translate into an effect on improving the frequency with which teachers say they are discussing inclusion issues on a monthly (or greater) basis in their classrooms. However, the interim study results are preliminary and only focus on data collected within a few weeks after the TEE trainings. The evaluation will be testing whether these practices change over a more extended follow-up period as part of the endline report.
89	I'd love to see a more in-depth analysis of the heading regarding gender and social inclusion: "Directors' perspectives on inclusion and diversity were mixed—these issues seem to be somewhat controversial and directors did not always agree with the training content." The ensuing content of the paragraph doesn't really develop this to the degree needed, but rather shares some positive examples (albeit one-offs). Attitudes and perceptions are definitely a major barrier to effective norm changing and so I hope in the final evaluation this is really unpacked.	We have added some additional detail about this finding to the report. We agree that directors' mixed response to the inclusion content in the training sequence is interesting and valuable to consider.
40	Paragraph describing lighting switching from describing baseline responses (I assume from the quantitative survey) to the focus groups. This is a little confusing. Please provide detail regarding which data set you are referring to. Are the baseline responses both treatment and control?	Revised and clarified.

Page Number	Comment	Mathematica Response
48	Make clear you are referring to the lack of food service in the school, not the infrastructure itself	We have clarified this point.
56	Mark the pie chart as after rehabilitation	Addressed.
59	I think we need to be clear on the definition of practitioner teacher. This term is used early on in the report but not defined until later on and it still gives the impression that practitioner teachers are younger, especially to a reader that does not know the context. I think it's best to describe them as uncertified according to the Ministry's professional development scheme. It might even be worthwhile to cite the scheme. I can reach out to Nino to get a copy in English if that exists.	We have provided a more explicit discussion of the definition of practitioner teachers on p. 27 (when the term is introduced) and again here, including a discussion of the fact that practitioner teachers are older, on average, compared to more senior teachers who have passed a certification exam in their subject.
multiple pages	I feel like on the TEE evaluation, they may be an opportunity to more convincingly address the question of "falsification", or more specifically, to make sense of what dimensions are natural changes in knowledge and opinions that are expected out of the TPDC trainings and which pieces shouldn't necessarily change. Ira basically said something like "here, these are the dimensions most linked and these others are less linked", but I sort of feel like there should be a more careful treatment of this dimension, otherwise it really just looks like p-hacking (which I hope it isn't, but it could certainly be accused if it). One way I might approach this (but MPR is perfectly able to disagree) would be to create and index of a number of variable of dimensions that are related, not related and perhaps even a third index of dimensions that are plausibly but not directly related. This may increase the likelihood of not finding impact, but I think we're better off having a more conservative formulation than to have a looser approach which ensure some form of positive impact is detected.	Thank you for this suggestion. We conducted a falsification test examining whether there were differences between trained and untrained teachers in perceptions of how their school director provided instructional leadership. Since teachers in the treatment and matched comparison group were in the same school and shared the same school director, we would not expect the training to impact how often teachers say they are observed by their school director and receive mentorship from their school director. As expected, we did not find any significant differences between the treatment and matched comparison group for these 'falsification test' outcomes. (These results are summarized in a footnote to the TEE matched comparison group analysis section).
multiple pages	The report states that the training sequence consisted of multiple modules (five one-day training sessions for directors, and four one-day sessions for teachers) and was held over the course over about one year for each cohort. This is incorrect – the training sessions were of different duration (from two to five days depending on the modules). Teacher trainings were delivered over the course of one year, however principal trainings were delivered over the course of two years.	Thank you for pointing this out. We have corrected the reported information about the duration of the TEE training modules, throughout the revised report.

Page Number	Comment	Mathematica Response
multiple pages	While report has a section on SPDFs, the executive summary doesn't mention them at all. I'm sure most of the readers will not go further than the summary, so it might be worthwhile to include something about SPDFs there.	Thank you for this suggestion. We have added information about SPDFs in the executive summary of the revised report.
multiple pages	Will it be possible to add one spreadsheet which will give the characteristics of surveyed teachers? Like what percentage of them were practitioners, rural/urban, Georgian/ethnic minority, women/men, and any other criteria that was used for matching or is deemed relevant for findings?	We present the characteristics of surveyed teachers in Table II.5. This table includes the characteristics of the overall sample and the balanced sample after identifying a matched comparison group for the TEE evaluation. In the revised report we also refer readers to this table in the TEE results chapter, as an additional reference.

This page has been left blank for double-sided copying.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ SEATTLE, WA ■ TUCSON, AZ ■ WASHINGTON, DC ■ WOODLAWN, MD

MATHEMATICA
CENTER FOR
INTERNATIONAL POLICY
RESEARCH AND EVALUATION