MATHEMATICA
CENTER FOR
INTERNATIONAL POLICY
RESEARCH AND EVALUATION

Evaluation Design for the Georgia Improving General Education Quality Project's Training Educators for Excellence Activity

Revised Report

January 6, 2017

Ira Nichols-Barrer
Nicholas Ingwersen
Elena Moroz
Matt Sloan

MATHEMATICA
Policy Research

MATHEMATICA
CENTER FOR
INTERNATIONAL POLICY
RESEARCH AND EVALUATION

Contract Number:
MCC-13-BPA-0040 (CL-002)

Mathematica Reference Number:
40306.200

Submitted to:
Millennium Challenge Corporation
875 15th Street, NW
Washington, DC, 20005
Project Officer: Ryan Moore

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

Evaluation Design Report for the Georgia Improving General Education Quality Project's Training Educators for Excellence Activity

Revised Report

January 6, 2017

Ira Nichols-Barrer
Nicholas Ingwersen
Elena Moroz
Matt Sloan

MATHEMATICA
Policy Research

CONTENTS

TABLES

FIGURES

1. Introduction

The Millennium Challenge Corporation (MCC) seeks to support Georgia's efforts to improve educational outcomes by sponsoring the Improving General Education Quality (IGEQ) Project, which includes three components. The Improved Learning Environment Infrastructure activity invests in school rehabilitation to provide safe learning environments that include adequate facilities and heating. The Training Educators for Excellence (TEE) activity supports professional development by training and mentoring teachers to improve competencies in subjects related to science and math, and by training principals to strengthen school management. Finally, the Education Assessment Support (EAS) activity supports Georgia's ongoing efforts to improve educational outcomes by encouraging use of assessment data and fostering a results-oriented education system. Mathematica Policy Research is designing and implementing a rigorous evaluation of these components to determine their ultimate impact on both intermediate and long-term outcomes.

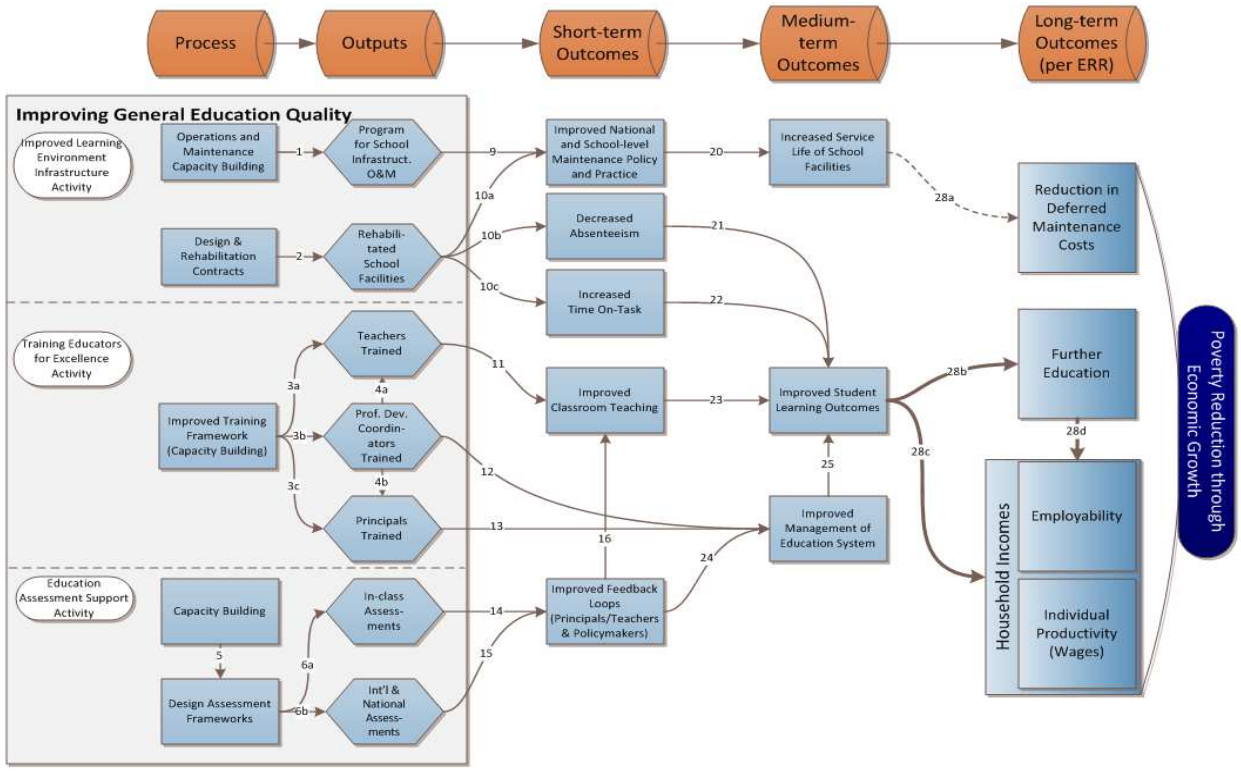
We have previously presented the evaluation approach for the Improved Learning Environment Infrastructure activity in a separate design report (Nichols-Barrer et al. 2016). The current report provides a detailed explanation of the evaluation design chosen for the TEE activity. It begins by presenting an overview of the IGEQ program logic and briefly reviews the existing literature examining the impacts of similar interventions in other countries. Next, we explain the TEE evaluation design and discuss key evaluation questions, methods, and data sources.

2. Overview of the Training Educators for Excellence activity

The TEE activity aims to improve classroom instruction in the subjects of science, technology, English, geography, and math in grades 7–12, through a combination of professional development activities for teachers and school directors. The Georgian government's Teacher Professional Development Center (TPDC), an agency charged with administering training interventions for staff throughout the national education system, will manage these activities. Examples of these activities include an initial core set of teacher training modules related to general pedagogy, student-centered learning approaches, and formative assessment techniques; a second subject-specific set of training modules for teachers to adapt material from the core modules to specific academic subjects; and a series of training modules for school directors focused on school management techniques, including structured approaches to teacher observation. The TEE activity plans to operate on a nationwide basis, including both Georgian-language schools and minority-language schools and reaching up to 18,000 Georgian-language teachers, 2,085 school directors, and 2,085 School-based Professional Development Facilitators (SPDFs) during the rest of the Georgia II Compact (with trainings occurring mainly during the 2016–2017 and 2017–2018 school years).

According to the logic model developed by MCC and Millennium Challenge Account-Georgia (MCA-G) staff (Figure 2.1), the TEE inputs aim to improve the quality of classroom teaching and management of schools throughout the education system, leading to improvements in students' learning and higher educational attainment outcomes. The program logic 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 poverty reduction through economic growth. Each link in the program logic model represents an assumption by IGEQ program designers about how the activities will affect the compact's beneficiaries and 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 cost benefit analysis for the project, informing economic rate of return (ERR) calculations for each activity.

Figure 2.1. The IGEQ program logic



Source: MCC Georgia II Compact investment memo.

Note: Arrows with dotted lines refer to links that MCC does not expect to be evaluable or measurable. O&M refers to operations and maintenance expenses.

To assess the IGEQ program logic and associated ERR calculations, we reviewed the available evidence on the impacts of similar program designs in other contexts and held detailed discussions with local education experts and IGEQ stakeholders during the Mathematica team's initial trip to Georgia in November 2013. These discussions included MCA-G staff, stakeholders in relevant GoG centers and ministries, and site visits to schools to meet teachers and school directors we will invite to participate in training activities. We examined the program logic for each of the three components of the IGEQ separately, noting potential concerns when applicable in a logic assessment report (Nichols-Barrer et al. 2013). The following section summarizes our review of the relevant literature.

3. Literature review

An extensive academic literature investigates the relationship between educational inputs and measures of student learning, educational attainment, and employment outcomes. However, the literature related to the impacts of education interventions in developing countries is scant, and little empirical work focuses on the education system in Georgia.

Based on our initial review of the evidence and discussions with program stakeholders, we conclude in general that the program logic for the TEE activity represents a plausible set of assumptions regarding how improved classroom teaching practices and school management lead to improved student outcomes. According to MCC's cost benefit analysis model for the TEE activity, as a medium-term outcome MCC expects this intervention to produce a 0.18 standard deviation improvement in student learning (particularly in mathematics), ultimately resulting in a 2 percent improvement in annual earnings from employment (in the long term). However, the existing evidence

base does not support strong predictions about whether the size of this projected impact is reasonable. An overview of the relevant literature follows.

Prior studies have shown an uncertain relationship between training inputs for teachers and school directors and each of 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 tend to be in studies in which the measured learning outcome aligned specifically with training materials (Yoon et al. 2007). However, the TEE activity will encompass a much wider range of grade levels and academic skills and are therefore likely to produce more diffuse impacts on general learning outcomes.

Evidence also exists from studies in developing countries that teacher training interventions can improve students' learning. Evans and Popova (2015) conducted a systematic review of reviews, analyzing findings from six evidence reviews focused on education programs in developing countries (these evidence reviews summarized results from 226 separate studies, in total). 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 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 an 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 student learning gains. 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 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, along with 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 produce initially 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.

However, the existing literature examining the effects of teacher training programs 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 in grades 7 through 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 tend 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 will assess a nationwide program that aims 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. It is also possible that such scaling issues could make it more difficult to produce substantial changes in the practices of teachers and school directors.

4. Evaluation design for the TEE activity

This section describes our evaluation design for assessing implementation of the TEE activity and estimating the impacts of these activities on targeted outcomes including the school-management skills of school directors and teachers' pedagogical and classroom management practices.

4.1. Evaluation type

Evaluation studies generally fall into one of two categories: performance evaluations, which use survey data and qualitative methods to measure the key outcomes of beneficiaries in the absence of a counterfactual, and impact evaluations, which measure beneficiaries relative to a comparison or control group to estimate a program's causal effects. For the TEE evaluation, we propose a mixed-methods study design with two components: (1) a performance evaluation to assess the possible effects of the TEE activity on school management and classroom instructional practices, and (2) a matched comparison group design to assess the initial impacts of the activity's teacher training modules. The design also includes an optional third component, which stakeholders could choose to exercise prior to the 2017-2018 school year, which would use a cluster randomized controlled trial (RCT) design to compare two different approaches to managing teacher training or study group activities (study groups are small gatherings of teachers assigned to work together after each training module to discuss, practice and prepare to apply recommended practices).

The performance evaluation and the matched comparison group analysis are designed to answer research questions about the program's implementation and initial outcomes; we will 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.

4.2. Evaluation questions

Table 4.1 presents the research questions that each component of the TEE evaluation will investigate.

Table 4.1. 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 trainers • Conduct a survey of 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? To what extent have school directors' instructional leadership and school management practices improved? To what extent have teachers' pedagogical practices (for example, student-centered instruction, matching practice to subject-matter, formative assessment use) and classroom management (for example, affirmative teaching, eliminating gender bias, time management) 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)?	• Analyze survey data collected from teachers and school directors • Analyze survey data collected from students • 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 to describe pedagogical practices • Triangulate observational data on teachers' practices with self-reported teacher survey data
Effects of training on teachers	**Impact 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)

4.3. Methods

This section explains the methods associated with each component of our evaluation for the TEE activity.

Performance evaluation describing program implementation and outcomes

The performance evaluation will collect information about how the TEE activity was implemented, test whether program activities were implemented as designed, and assess if 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 will analyze several different types of data, including program documentation, survey data, and qualitative research. The performance evaluation will use several key data sources:

- Project reports will 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 in schools, we will collect survey data from a representative sample of teachers and school directors trained by the program. We will collect survey data at two points in time: September 2017 (after the first cohort of teachers has completed its sequence of four training modules) and September 2018 (after the second cohort has completed its full sequence of TEE trainings).
- We will use qualitative data to understand how the program was implemented and how the program might have changed participants' practices.
    - In-depth interviews with implementing staff and stakeholders involved with the project will help us understand project design and implementation. These interviews will take place in the program's second year (during the 2017-2018 school year), after the first cohort of teachers and school directors has completed the full sequence of TEE trainings.
    - Observation and monitoring of the teacher study groups during the program's first implementation year (the 2016-2017 school year). In coordination with TPDC these monitoring activities will measure the number of study group meetings that have taken place, and teacher attendance rates at those study group meetings. In addition, we will carry out qualitative observations of study group meetings to assess whether study group activities align with TEE training modules and support targeted practices.
    - 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 has completed the activity's full course of four training modules—will gather more information about how the training was implemented and will identify possible relationships between training activities (including differences between core training modules and subject-specific modules), study group participation, and changes in the practices of school directors and teachers.
    - The study will also directly observe classrooms of a sample of trained teachers delivering lessons during a regular school day. These observations will occur

during the program's second implementation year (the 2017-2018 school year). Trained observers will visit a sample of schools to conduct a structured observation to measure teachers' use of instructional time, their use of materials (including information and communications technology), and core pedagogical practices.

- The performance evaluation does not include student learning assessments or student exams, and as a result the evaluation will not directly measure student learning outcomes. However, due to concurrent data collection activities related to the evaluation of school rehabilitation activities, it is possible to collect descriptive data from students about their perceptions of teaching practices (using a convenience sample of students who will be surveyed in spring 2018 as part of the school rehabilitation study) at little additional cost. We plan to use this student survey to measure student perceptions related to teachers' use of student-centered instruction, formative assessments, and positive classroom management practices. In addition, such a survey would seek to measure student perceptions related to school culture, gender bias, and knowledge and attitudes toward career options in STEM-related fields.

The performance evaluation will identify implementation successes and challenges and document key lessons learned about implementation of national-scale training programs in Georgia, as well as 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 to the intended set of practices the program expects to encourage, the study will assess whether it is plausible that the TEE training model could ultimately affect students' learning outcomes.

Impact evaluation of teacher training, applying a matched comparison group design

We also will implement an impact evaluation design to directly measure the TEE program's impacts on participants. An impact evaluation involves comparing outcomes for a group of beneficiaries to outcomes for a comparison or control group that does not receive the same activity in a given time period. We assume that the outcomes for the comparison group represent those that the beneficiaries would have realized had there not been a program.

To measure the impacts of the training program on teachers' knowledge and attitudes, the evaluation will apply a matched comparison group design. This design compares a group of teachers who will be trained during the 2016–2017 school year (Cohort 1) with a group of teachers who will not be trained until the 2017–2018 school year (Cohort 2). Specifically, by September 2017 the first cohort of trained teachers will have completed the four TEE training modules. At that time, the second cohort of teachers will not yet have received any training. This provides an opportunity to estimate the impacts of participation in the training modules by comparing teachers in Cohort 1 with those in Cohort 2. This design is well suited to estimate the initial impacts of the program on teachers' knowledge about the types of practices covered in the training intervention, along with 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 assignment to cohorts was not random (for example, Cohort 1 prioritizes teachers with higher certification levels), a matching design is necessary 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 will endeavor 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 will account 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 all of the teacher characteristics that would otherwise generate selection bias in the impact analysis), and the treatment and comparison groups are balanced in their propensity scores, the design will produce an unbiased estimate of the program's impact.

It is important to note that the impact evaluation design also relies on a logistical assumption that there will be a sufficient amount of time to complete a full round of the teacher survey between the end of Cohort 1 training activities and the beginning of Cohort 2 training activities. Under the program's current implementation schedule this appears to be feasible, but any changes to the implementation plan (such as a delay in the rollout of Cohort 1 trainings or any acceleration in the rollout of Cohort 2 trainings) could have important implications for the study's design. Mathematica will closely monitor these implementation plans as they develop, and if changes in the rollout schedule occur we will develop and discuss any necessary evaluation-design adjustments with MCC, MCA-G, and other key stakeholders.

Optional study component: RCT comparing two implementation approaches in 2017-2018

The evaluation design also has the option of incorporating an RCT comparing two different approaches to the management of teacher training or teacher study groups in the 2017-2018 school year. The design could accommodate comparisons of many different possible pairs of implementation strategies, such as the use of professional study group meeting facilitators, or different approaches to teacher-level recruitment and training delivery. For example, if TPDC and other stakeholders choose to do so, the evaluation could assess different approaches to managing teacher study groups (these are activities in which teachers will convene in small groups to discuss the training material and practice applying key concepts and teaching practices). In this example, the study could compare the outcomes of one set of study groups receiving a high level of centralized coordination and guidance from TPDC to a second set of study groups receiving less formal guidance and support.

We developed this optional component of the evaluation in response to discussions with TPDC regarding potential ways in which the TEE evaluation could provide information that would be of greatest use to practitioners and policymakers in Georgia. If exercised, the goal would be to provide evidence to TPDC regarding the difference in impacts between two different implementation approaches that could improve the implementation of future teacher training initiatives in Georgia.

If the option is exercised, Mathematica staff will implement a clustered, district-level random assignment protocol to randomly assign teachers into each tested implementation approach. Random assignment provides an ideal way to define comparable treatment and control groups, because

randomization ensures the two groups will be similar at baseline on both observed an unobserved characteristics. To facilitate implementation logistics, for this optional study the random assign process would be clustered at the district-level: in other words, all teachers in the same district would be assigned to the same implementation approach. Clustering assignment at the district level simplifies the administrative complexity of the study, and minimizes the risk of spillover effects (because teachers in the same area will receive a uniform implementation approach). The random assignment protocol would also stratifies the districts in the study by region: that is, within each region we would assign an equal number of districts to each tested implementation approach.

4.4. Study population

The TEE evaluation will focus on describing the outcomes of training activities delivered to school directors and teachers. Under the current design, the performance evaluation will focus 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 prioritize staff at Georgian-language schools. Thus, the study population will include all Georgian-language school directors and teachers in Georgian-language schools. The study's impact evaluation design will estimate the impacts of teacher training for the subset of Cohort 1 teachers who can be adequately matched to Cohort 2 teachers. Since the TEE activity has prioritized training more senior and experienced teachers in the first cohort, we anticipate that the impact evaluation will be limited to a population of more junior and less-experienced teachers, since these teachers are more likely to be matched successfully to teachers in Cohort 2.

The evaluation will seek to identify a relevant sample of respondents that is geographically representative of this overall population, including a representative sample of school directors across relevant districts and a representative sample of teachers across the TEE activity's targeted set of grade levels and academic subjects. Mathematica will coordinate with MCC and MCA-G staff to develop terms of reference for the data collection that elaborate this sampling protocol in greater detail.

4.5. Study sample and power calculations

The TEE evaluation consists of two separate components: (1) a performance evaluation using descriptive data qualitative methods to assess the possible effects of the TEE activity on school management and classroom instructional practices, and (2) a matched comparison group design focused on assessing the effectiveness of the teacher training modules. Our design also includes an optional third component: a cluster-randomized design focused on estimating the impact of two approaches to managing teacher training or study group activities in the 2017-2018 school year. The data collection plan for the evaluation will realize efficiencies by gathering data for multiple study components in each survey round. The study team conducted power analyses for each of these components separately, as presented next.

To conduct the matched comparison impact analysis of the teacher training modules, we will compare a geographically representative sample of Cohort 1 teachers who have recently completed the training sequence to a matched sample of teachers who are not yet eligible to begin the training (but will receive it eventually). For the comparison group sample, we will target data collection to all teachers in the second TEE cohort (those who are not eligible to receive the core teacher training until October 2017) who teach in the same set of schools as teachers in the treatment sample. To conduct the analysis, we will match teachers in each treatment sample on a number of characteristics, including their level of teaching certification. The design must account for the fact that the second

cohort will consist exclusively of "practitioner-level" teachers (practitioner-level teachers represent approximately two-thirds of the teaching workforce and have yet to pass the government's certification exam needed to obtain senior status), whereas only about a third of the Cohort 1 teachers are practitioner-level.[1] To ensure the treatment and comparison groups are equivalent, the matching study will be limited to practitioner-level teachers. To increase the sample of Cohort 1 teachers in the matching study, we propose to *oversample* practitioner-level teachers in the first cohort: specifically, two-thirds of the surveyed teachers in the first cohort will be practitioner-level teachers (representing strong potential matches for the practitioner-level teachers in the second cohort) and the remaining third of the Cohort 1 survey sample will be senior teachers whose outcome data will be used for the performance evaluation.

Table 4.2 presents power calculations for the matched comparison group design and illustrative power calculations for the study's optional random assignment design in the 2017-2018 school year. The table shows the statistical precision provided by different illustrative sample configurations: a benchmark scenario and a scenario with a 50 percent reduction in sample size, which simulates a subgroup analysis (for instance, outcomes among teachers of science subjects or teachers of upper-secondary grade levels). The power calculations assume an additional 10 percent nonresponse rate during data collection for the estimation of minimum detectable effects (MDEs) in each scenario. We estimate that the optional random assignment evaluation would be able to detect statistically significant teacher-level impacts as small as 0.28 standard deviations in the benchmark scenario and 0.30 standard deviations with a 50 percent smaller sample of teachers within each cluster. For the matched comparison evaluation, we estimate that the evaluation will be able to detect statistically significant teacher-level impacts as small as 0.14 standard deviations in the benchmark scenario and 0.20 standard deviations with a 50 percent smaller sample. The relatively small differences in estimated MDEs between the benchmark and 50 percent subgroup scenarios suggest that the study designs are robust to scenarios in which the uptake rates for project activities are very low. For a desired pedagogical practice used by 50 percent of teachers at baseline, an impact of 0.14 standard deviations represents a gain of approximately seven percentage points (to 57 percent) and an impact of 0.20 standard deviations represents a gain of approximately 10 percentage points (to 60 percent).

Based on our review of other teacher training evaluations in developing countries, we believe that the range of detectable effects shown in the matched comparison scenarios represents a level of statistical precision that is adequate to detect impacts comparable to those reported for teacher training in certain other contexts. Previous studies have found impacts of 0.12 to 0.25 standard deviations on students' learning (Evans and Popova 2015), and the matched comparison group analysis is powered sufficiently to detect an impact of 0.18 standard deviations (the effect size assumed in MCC's ERR analysis for the TEE activity). Although we will not observe student-level outcomes in the analysis, we believe that it is reasonable to assume that a change in student learning would require a similar if not larger change in proximate teacher-level outcomes (for example, beneficial teaching practices). As a result, we believe that the magnitude of detectable effects for teacher-level outcomes presented in Table 4.2 is consistent with the magnitudes of impacts found in studies focused on student learning.

[1] The outcomes of practitioner-level teachers might differ from the outcomes of more senior teachers; the study design will not estimate the effects of TEE training on more senior teachers. However, it is important to note that most of the teachers in Georgia are at the practitioner level, so the impact estimates will be relevant for most of the current teacher workforce. We believe the loss of external validity due to this design decision presents a reasonable tradeoff, because using a matched sample of teachers with an equal level of seniority in the treatment and comparison groups greatly improves the internal validity of the study's quasi-experimental approach.

Table 4.2. TEE minimum detectable effects for different sample sizes

	Optional comparison of two implementation strategies *Random assignment design*		Evaluation of teacher training activities *Matched comparison design*	
	Full sample	50% subgroup sample	Full sample	50% subgroup sample
Level of cluster	District	District	No clustering	No clustering
Number of clusters	34 treatment 34 control	34 treatment 34 control	NA	NA
Number of teachers per cluster (approximate)	36	18	NA	NA
Total teacher sample	2,500	1,250	1,200 treatment 1,200 comparison	600 treatment 600 comparison
MDE	0.28	0.30	0.14	0.20

Notes: MDE calculations assume a two-tailed test with a 5 percent significance level and 80 percent power. We assume a district-level intraclass correlation (ICC) of 0.17, a district-level R-squared of 0.18, and a teacher-level R-squared of 0.04. The ICC and R-squared assumptions are based on a measure of whether students of science teachers frequently conduct experiments that was collected at baseline for the Georgia IGEQ project's Improved Learning Environment Infrastructure activity evaluation (data available from the authors upon request). All power calculations also assume a 10 percent rate of survey nonresponse. For the matched comparison design, we further assume that it will be possible to successfully match two-thirds of the surveyed treatment teachers (that is, those who at the practitioner level) to comparison teachers, using a 1-to-1 matching design. NA = not applicable.

If the option to pursue a comparison of implementation strategies is exercised following the 2016-2017 school year, we will assess whether the range of detectable effects shown in the cluster-randomized scenarios is adequate to detect the relative impacts of the two approaches that will be tested. In particular, we will assess the range of potential effects on proximate outcomes (such as attendance rates in the case of a study focused on study group meetings) where larger effect sizes are more likely to occur. For example, the MDEs of 0.28 and 0.30 standard deviations translate to changes of 14 and 15 percentage points, respectively, for a binary variable with a mean of 0.50.

For the 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. The projected sample sizes for these sources are as follows. To gather information from implementers, we plan to conduct a series of in-depth interviews targeting three groups of respondents: two or three TPDC staff with a prominent TEE management role, two or three interviews with school director trainers, two or three interviews with teacher trainers, and one or two interviews of MCC/MCA-G staff involved in implementation and oversight of TEE activities (the final number of interviews could change following review of the implementers' final set of program documentation reports and training materials). By collecting information from the respondents across various levels of activity planning, management, and implementation, we believe this approach will provide a full picture of the activity's planned implementation, actual implementation, and the reasons for any differences between the planned and actual implementations.

Survey data for the performance evaluation will come from school director, SPDF and teacher surveys. To realize efficiencies in the data collection effort, the study will draw one sample of schools

and teachers for all study components (see the sample sizes described in Table 4.4). That is, in September 2017 (after Cohort 1 teachers have received all training modules) we will administer the performance evaluation survey to the exact same sample of Cohort 1 and Cohort 2 teachers that comprise the evaluation sample for the matched comparison group analysis. We will survey this same sample of teachers a second time in September 2018 to gather descriptive information about longer-term knowledge and practice outcomes for the performance evaluation. As part of these two survey rounds, the study will also gather survey data from school directors and SPDFs located in the same schools as surveyed teachers, providing descriptive information about their knowledge and practices related to the TEE training modules.

Finally, the study team plans to add a small student survey module to the existing sample of students who will be surveyed as part of the IGEQ school rehabilitation evaluation in spring 2018. This survey will include 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 will provide an opportunity to gather data about students' perceptions regarding the presence of targeted teaching practices in their classrooms, at minimal cost. A description of the sampling plan for this school rehabilitation survey is available in Nichols-Barrer et al. (2016).

We will obtain qualitative data for the performance evaluation from three main sources: structured classroom observations, focus groups with teachers, and in-depth interviews with school directors in treatment schools. Classroom observations will enable us to investigate how training has affected classroom teaching practices. High quality classroom observations are resource- and time-intensive; therefore, we will draw a subsample of teachers from each of the program's 11 geographic regions. We plan to collect data in the activity's second year of implementation (the 2017–2018 school year), after the first cohort of school directors and teachers have completed the full set of TEE training activities. For qualitative data collection, we will include at least two schools in each region, for a total of 22 schools. Across regions, we will purposively select schools to include a range of school characteristics, such as school size and urbanicity. In each of these schools, the local data collection firm will conduct an in-depth interview with the school director, an in-depth interview with the school's professional development facilitator, one focus group (with about eight Cohort 1 teachers in each focus group, a majority of whom will be science or mathematics teachers), and classroom observations of two Cohort 1 teachers (observing each teacher twice). To minimize burden on teachers, we will organize focus group recruitment around teacher study groups. That is, we will recruit teachers who participated in the same post-training study group to attend the focus groups. Focus groups may include teachers from the same school or neighboring schools. When study groups are large, we will consider recruiting members of the group by stratifying invited teachers according to grade level (upper versus lower secondary grades) or by academic subject to limit the focus group size.

We believe these samples will produce meaningful descriptive data for qualitative analyses of teachers' and school directors' practices; the purposive sampling plan described earlier will enable the qualitative study to document the presence or absence of targeted school director and teacher practices at a geographically varied subsample of schools. However, this subsample is not sufficient to support quantitative hypothesis testing and, as a result, we do not show power calculations for this portion of the performance study.

4.6. Time frame

We will collect 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). The local data collection firm will also conduct 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 4.3).

Prior to the first round of data collection, we will hold an instrument-development workshop in Tbilisi with key project stakeholders. The purpose of this workshop will be to solicit feedback on draft versions of the teacher survey, school director survey, SPDF survey, student survey, and classroom observation protocols to ensure the data collection instruments capture the activity's full range of near-term outcomes and intermediate outcomes.

Table 4.3. 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	Final outcome survey	Initial outcome survey	Final outcome survey
Qualitative data			
2017–2018 school year	Teacher focus groups Classroom observations		In-depth interviews

Note: The data collection will also include a convenience sample of students to be surveyed in March 2017.

If the TEE implementation plan changes, the study team will consider appropriate revisions to the data collection schedule. For example, if there are any delays in implementation of the training program, the study would consider revising the survey schedule to ensure that data are collected when it is still possible to compare Cohort 1 and 2 teachers before training has begun for the second cohort. Likewise, we will consider alternative or extended data collection schedules as the program develops and in consultation with MCA-G and MCC, particularly if there is interest in assessing the outcomes of later cohorts of trainees (such as teachers and school directors in minority-language schools). The study team recommends the use of a year-by-year contract with the local survey firm. This approach will provide an opportunity to assess whether the existing data collection plan is still advisable following each data collection round, because the contract structure facilitates making adjustments on an annual basis. For example, after the 2017 round we could consider adjusting the timing of the second follow-up data collection round, adding qualitative research activities during the 2018-2019 school year, or adding additional data collection rounds to the study if those additional data collection efforts appear to be merited following review of the quantitative and qualitative data obtained in the first data collection round. The study team will maintain a flexible approach and will discuss the merits of changes to the study design and data collection plan with MCC, MCA-G, and other stakeholders as needed throughout the life of the study.

5. Data sources and outcome definitions

Our study design calls for collecting survey data from teachers and school directors. A combination of administrative data from program implementers, in-depth qualitative interviews and focus groups, and observations of classroom teaching practices will complement the survey data. An MCA-procured local data collection firm will collect survey data, qualitative data from school directors

and teachers, and classroom observations. Mathematica will obtain administrative data from program implementation records and conduct interviews with key project management staff. Table 5.1 summarizes the data sources for each of the study's key outcomes.

Table 5.1. Data sources and study outcomes for the TEE evaluation

Component	Description	Outcome
Data to be collected directly by Mathematica		
TEE design and implementation records	To document the design and implementation of the TEE activity, Mathematica will obtain any available training design reports, training materials, implementation records, and program cost data.	Alignment between training activities and improvement goals (for example, pedagogical weaknesses and school management issues) identified in the program's original design Number of teachers and school directors trained Percentage of teachers receiving credit for completing the TEE training sequence
In-depth interviews with implementers	For the study's performance evaluation, Mathematica will conduct qualitative, in-depth interviews with implementers, including key TPDC program managers, training providers, and MCA-G staff.	Description of barriers to implementation successes
Data to be collected by local survey firm procured by MCA-G		
Teacher survey	A local survey firm will collect survey data on teachers' participation in training and study group activities, knowledge of and attitudes toward targeted pedagogical practices, perceptions regarding the value of training and study group activities, and self-reported pedagogical practices.	Participation in training modules and post-training follow-up activities Knowledge of and attitudes toward key practices (for example, student-centered learning, use of formative assessment techniques, use of group learning methods, and adapting instruction to students' abilities) Perceptions of how training affected teaching practices
School director survey	A local survey firm will collect survey data on school directors' participation in training activities, knowledge of and attitudes toward targeted school management practices, perceptions regarding the value of training activities, and self-reported school management practices.	Participation in training modules and post-training follow-up activities Knowledge of and attitudes toward key practices (for example, attitudes toward student-centered learning, use of student learning data, and provision of instructional leadership to teachers) Perceptions of how training affected school management practices

Component	Description	Outcome
Qualitative data: Teacher focus groups School director qualitative interviews Professional development facilitator qualitative interviews Classroom observations	A local survey firm will collect qualitative data in a subsample of about 22 schools. The sample will consist of teachers, professional development facilitators and school directors in the first TEE cohort (all of whom are scheduled to complete the full training sequence in summer 2017). Qualitative data collection will include in-depth interviews with school directors to investigate how training has affected school management practices; in-depth interviews with SPDFs to assess whether they are actively observing classrooms and providing feedback; focus groups with teachers to investigate how training has affected classroom instruction practices; and classroom observations designed to gather descriptive data on teachers' use of instructional time, use of materials, and core pedagogical practices, and to triangulate findings on self-reported pedagogical practices from the teacher survey.	Relationship between training activities and practices that might ultimately affect students' learning outcomes

6. Analysis plan

To estimate the impacts of the core teaching training activities we will adopt a matched comparison design. We will first conduct a one-to-one matching of practitioner-level teachers who received training with practitioner-level teachers who have not. The teachers will be matched on the following characteristics: school location, grade level, subject area, teaching experience, and educational background. This will ensure that we only compare the outcomes of teachers who are observably similar to one another. After the matching is complete, we will estimate the impacts of the TEE activity using the following ordinary least squares regression:

$$(1) \qquad Y_i = \alpha + \beta * TREAT_i + \boldsymbol{X_i} * \boldsymbol{\gamma} + \varepsilon_i$$

where Y_i is the outcome of interest (for example, attitudes toward student-centered learning) for teacher i; $TREAT_i$ is the treatment dummy variable indicating whether a teacher received the core teacher training; X_i includes a set of teacher-level demographic characteristics; and ε_i is the random error. The estimated value of the coefficient β represents the impact of the core teacher training on the outcome of interest. This model will use robust standard errors calculated at the level of individual teachers (without clustering).

The evaluation will also include subgroup analyses designed to measure whether there are statistically significant differences between the magnitudes of programmatic impacts for key subgroups of teachers (relative to the impacts of the activities among teachers outside each subgroup). Subgroup analyses will include disaggregated impact estimates based on grade levels, subject areas, teachers' characteristics (experience and demographic attributes), and school characteristics (e.g., urbanicity, school size, facilities).

According to documentation MCC provided to Mathematica, MCC produced an ex ante cost benefit analysis model with an estimated an ERR of 18 percent for the overall TEE activity. The ERR is a summary statistic that reflects the economic merits of the investment. Conceptually, it is the discount rate at which the cumulative benefits of an intervention over time are exactly equal to its

costs; a higher (positive) ERR represents higher benefits and lower costs. MCC produced an ex ante ERR estimate of 18 percent based on expected costs and benefits of the teacher training in terms of student learning and, ultimately students' labor market outcomes. Because we will not collect student-level outcomes for this evaluation, we will not be able to estimate a comparable ex post ERR estimate. However, our evaluation can shed light on whether students' learning has likely improved based on the estimated impact on teachers' practices in the classroom, an important factor in students' learning and, therefore, on the plausibility of the MCC's ex ante ERR estimate. Our working assumption is that the TEE activity would have to achieve, at minimum, an impact of 0.18 standard deviations on targeted practices to ultimately produce an effect of 0.18 standard deviations on students' learning (the effect size that the ERR estimate projected).

For analyses of qualitative data, Mathematica will use qualitative transcript-coding software to organize and synthesize the key themes that emerge from document reviews, in-depth interviews, and focus groups. When appropriate, we will compare and contrast information from these data sources with descriptive data available in the study's surveys of teachers and school directors. Data from classroom observations will be a particular focus for comparisons across data sources. By comparing classroom observation data with the survey responses of observed teachers, the evaluation will be able to generate insights about how best to interpret self-reported data from teachers about their classroom practices. More generally, analyses of qualitative data will focus on insights and themes that might explain findings from the study's impact analyses. For example, if the impact analysis uncovered evidence of positive program impacts on some outcomes but not others, we would examine the study's qualitative data to develop a deeper understanding of the relationship between training activities (such as participation in study groups) and the program's impacts.

7. Evaluation risks and monitoring plan

Several risks to the study's internal and external validity will require careful monitoring and management throughout the evaluation period. At every stage of the study period, Mathematica will remain in close contact with the MCA-G to monitor implementation of the TEE activity.

The primary risk to the current evaluation design is that the program's implementation schedule could change. Due to the large-scale, nationwide rollout plan for training activities, it is possible that logistical considerations or implementation constraints could lead to changes in the timing of training delivery to teachers and school directors. This could affect the evaluation's impact study design if (for example) the timing of training activities for Cohort 1 teachers alters to overlap with the timing of training activities for Cohort 2 teachers. If necessary, the study team will consider alternations to the evaluation's scope and schedule to accommodate any major changes to the TEE schedule.

Another potential threat to the internal validity of the matched comparison group study is the risk of within-school spillover effects across teacher cohorts. Because Cohort 2 teachers will be drawn from the same schools as Cohort 1 teachers, it is possible that teachers in the first cohort could influence the knowledge and practices of teachers in the second cohort during the 2016–2017 school year (before formal training begins for Cohort 2). This would reduce observed differences between the two cohorts and consequently reduce the ability of the study to detect the full impact of training activities. To address this issue, the study team will design survey questions to collect data on the amount of interaction between teachers in the two cohorts—for example, the survey will ask Cohort 2 teachers questions about the extent to which they have learned about the trainings given to teachers in Cohort 1. Using these data, in the analysis phase of the study the team will conduct sensitivity analyses to investigate whether the impact findings differ in schools in which the amount of reported interaction is especially high or low.

A final risk to the evaluation pertains to the cost and complexity of classroom observation work. Because large-scale observation of teachers is currently a major area of policy interest in Georgia, there is a particularly strong need to involve stakeholders in the design of the study's classroom observation rubric to ensure the study's approach aligns with existing government-sponsored observation efforts. Coordinating this approach to classroom observations will also help to avoid the risk of doing design work that duplicates other efforts taking place elsewhere in the Georgian education system. Thus, the study team plans to use existing classroom observation tools developed in Georgia as a starting point (including observation rubrics established by the GoG on a nationwide basis and observation tools provided to school directors as part of the TEE program). In addition, the study will consider elements of other widely used and validated classroom observation tools that other developing countries have applied successfully, such as the Stallings classroom snapshot observation system (World Bank 2015).

8. Administrative considerations

8.1. Institutional review board requirements and clearances

Mathematica will prepare and submit an institutional review board (IRB) application for approval of the research and data collection plans. The application materials include three sets of documents: (1) a research protocol, which will draw heavily on the present design report and adds more information about plans for protecting study participants' confidentiality and human rights; (2) copies of all data collection instruments; and (3) a completed IRB questionnaire that summarizes the key elements of the research protocol, plans for protecting participants' human rights, and possible threats to participants if their confidentiality were compromised. Based on prior experiences, we expect that the study will qualify for expedited review because it presents minimal risk to participants. If so, the IRB can typically review the application within one week of its submission.

IRB approval is valid for one year from the date of approval and must be renewed on an annual basis. We expect that the annual renewals will require minimal updates to the core application materials. In addition, if data collection instruments change substantially from those that the IRB approved, then we must reapply for approval. Small changes to the instruments (such as rewording or reordering of questions or editing changes) do not require reapplication, but the finalized instruments must be submitted to the IRB for documentation.

After Mathematica drafts the IRB research protocol, we will coordinate with MCA-G to ensure the data collector and local stakeholders agree on the data collection protocol. Because Mathematica does not have a contractual relationship with the data collector, the data collector's contract with MCA-G must specify that it will abide by the IRB's recommendations. The data collector and Mathematica must also sign an IRB authorization agreement stating that the data collector will adhere to the IRB-approved data collection procedures and protocols.

8.2. Data access, privacy, and documentation

After producing each of the baseline, interim, and final reports, we will prepare corresponding de-identified data files and codebooks that can be made available to the public. These data files, user manuals, and codebooks will be de-identified according to the most recent guidelines set forth by MCC. The public use data files will be free of personal or geographic identifiers that would permit unassisted identification of individual respondents or their households, and we will remove or adjust variables that introduce reasonable risks of deductive disclosure of the identity of individual participants. Mathematica will remove all individual identifiers, including names, addresses, telephone numbers, government-issued identification numbers, and any other similar variables. We will also

remove unique and rare data using local suppression, replacing these observations with missing values instead. If necessary, we will also use top and bottom coding, setting upper and lower bounds to remove outliers and collapse any variables that make an individual highly visible depending on geographic or other factors (such as ethnic classifications or languages spoken) into less easily identifiable categories. Finally, we will introduce random errors into any gathered geographic data (for example, global positioning system or geographic information system coordinates), displacing urban points 0 to 2 kilometers and rural points 0 to 5 kilometers, and additional 1 percent of rural points 0 to 10 kilometers. Data perturbation will take place in a manner that will not significantly degrade the data.

8.3. Dissemination plan

Mathematica will present interim and final evaluation findings in person to MCC and to stakeholders in Georgia. The interim analysis will occur after data collection is completed in September 2017; this analysis will produce preliminary results from the evaluation's matched comparison group analysis examining the initial effects of the activity's training modules for teachers. We plan to summarize these interim findings in a concise issue brief format, which will make the findings more readily accessible and usable to stakeholders and program planners during the program's second year of implementation. Following the 2017–2018 school year, we will analyze descriptive and qualitative data and produce a final evaluation report for the TEE activity.

We will work with MCC to increase the visibility of the study's findings, particularly among education policymakers and development practitioners. We will collaborate with MCC and stakeholders to identify a variety of forums—including conferences, workshops, and publications—to share results and encourage donors, implementers, and policymakers to integrate the findings into future programming. For example, in addition to the project's full impact report, we will develop issue briefs summarizing and visualizing key findings from the final impact report for a broader audience of readers and stakeholders. Potential conferences for presenting evaluation findings will include forums hosted by the Comparative International Education Society, the American Evaluation Association, or the Association for Public Policy Analysis and Management. We will also seek to publish a peer-reviewed article disseminating the study's results in academic or sector-specific journals focused on education systems in developing countries.

8.4. Evaluation team roles and responsibilities

Mathematica's project team has extensive experience conducting mixed-methods, multicomponent, large-scale evaluations in the field of education. **Mr. Matt Sloan** will serve as the program manager, acting as the primary point of contact for MCC. He will manage the relationships with government agencies and other local entities and contractors, while supervising the evaluation design and implementation process and ensuring high data quality. **Mr. Ira Nichols-Barrer** is the principal investigator for this evaluation, providing methodological and technical oversight and serving as a senior analyst supporting the project team. **Dr. Nicholas Ingwersen** will oversee the study's quantitative data collection and analyses, and **Dr. Camila Fernandez** will oversee the qualitative data collection and analysis process. Dr.**. Natia Gorgadze** will serve as the project's in-country consultant, providing substantive knowledge of Georgia's education system and assisting with the study's data collection and other local evaluation management tasks.

8.5. Budget

At this time, Mathematica does not anticipate that the TEE evaluation design and data analysis plans described in this report will require changes to the total evaluation budget figure presented in the study's original proposal. Mathematica will work closely with MCC and MCA-G to ensure data collection is feasible within the compact's budget parameters.

REFERENCES

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.

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.

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.

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.

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." Draft report submitted to the Millennium Challenge Corporation. Washington, DC: Mathematica Policy Research, December 2013.

Nichols-Barrer, Ira, Matt Sloan, Ken Fortson, and Leigh Linden. "Evaluation Design Report for the Georgia Improving General Education Project's School Rehabilitation Activity." Report submitted to the Millennium Challenge Corporation. Washington, DC: Mathematica Policy Research, March 2016.

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*, 2010, pp. 21–41.

World Bank Group. "User Guide: Conducting Classroom Observations – Analyzing Classroom Dynamics and Instructional Time Using the Stallings 'Classroom Snapshot' Observation System." World Bank Education Global Practice. Washington, DC: World Bank, 2015.

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.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

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

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC