USAID/GEORGIA EXTERNAL IMPACT EVALUATION OF THE GEORGIAN PRIMARY EDUCATION (G-PRIED) PROJECT

IQC AID-114-I-13-00001
TASK ORDER AID-114-TO-14-00008

ENDLINE IMPACT EVALUATION REPORT (FINAL)

January 2016

This publication was produced at the request of the United States Agency for International Development. It was prepared independently by NORC at the University of Chicago, under a subcontract with Mendez England & Associates. The authors' views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

USAID/GEORGIA EXTERNAL IMPACT EVALUATION OF THE GEORGIAN PRIMARY EDUCATION (G-PRIED) PROJECT

ENDLINE IMPACT EVALUATION REPORT

January 2016

PN 7476
IQC AID-114-I-13-0001
Task Order AID-114-TO-14-00008

CONTENTS

LIST OF TABLES

LIST OF Figures

ACKNOWLEDGEMENTS

The impact evaluation of the Georgia Primary Education Project would not have been possible without the hard work and collaboration of many people. First, we would like to thank USAID/Georgia, in particular Medea Kakachia and Lela Kerashvili, for allowing us to undertake this evaluation and facilitating communications with the G-PriEd team. We would also like to thank the entire team at G-PriEd. Nancy Parks, Indira Amiranashvili and Sophie Malashkhia were absolutely instrumental in developing our understanding of the intervention, from sharing all documents from the baseline data collection, to providing all relevant monitoring data and openly and thoroughly answering our many questions. We are also extremely grateful for their collaboration regarding the training of the enumerators for the endline data collection. Paata Papava, Giorgi Nozadze and Sophie Malashkhia generously provided of their time to lead some of the training sessions which ensured consistency between the baseline and endline data collections.

Our partners at GORBI and in particular Nino Gulashvili provided excellent leadership for the data collection. The data entry process would not have been possible without the strong collaboration between GORBI and G-PriED's IT staff. We are also thankful to Marika Gorgadze for her support at the beginning of the project and during the dissemination workshop.

Finally, we thank the enumerators and the school principals and teachers for supporting us throughout the data collection process and during our field visits.

AUTHORS

This report was produced by Principal Investigator Dr. Alejandro Ome, Project Manager Yvonne Cao, Psychometrician Dr. Michel Rousseau, and Research Analysts Elise Le and Russell Owen, with the input of psychometricians Irina Samsonia (reading specialist) and Ekaterine Kordzaze (math specialist) and education expert Natia Gorgadze. Impact Evaluation Advisor Dr. Alicia Menendez, and Project Director Varuni Dayaratna also provided support.

LIST OF ACRONYMS

ANOVA	Analysis of Variance
DID	Difference-in-Differences
GDA-M	Georgian Diagnostic Assessment in Math
GDA-R	Georgian Diagnostic Assessment in Reading
GDAR-GSL	Georgian Diagnostic Assessment in Reading of Georgian as a Second Language
G-PriEd	USAID/Georgia Primary Education Project
IE	Impact Evaluation
M&E	Monitoring and Evaluation
ME&A	Mendez England & Associates
MES	Ministry of Education and Science
NORC	NORC at the University of Chicago
PLD	Performance Level Descriptors
PMP	Performance Management Plan
TLC	Teacher Learning Circles
TOT	Training of Trainers
TPDC	Teacher's Professional Development Center
USAID	United States Agency For International Development
VAM	Value-Added Model

EXECUTIVE SUMMARY

PROJECT BACKGROUND

The USAID/Georgia Primary Education Project (G-PriEd) is a 5-year (2011-2016) $8.7 million pilot project designed to provide comprehensive assistance to the primary education system to improve reading and math competences of Georgian and ethnic minority students in 122 pilot schools. This project, implemented by Chemonics, is in line with the government's reforms to change the education system from a teacher-centered model to a student-centered model. In collaboration with the Georgian Ministry of Education and Science (MES), G-PriEd aims to strengthen key components of the education system through teacher trainings, in-service professional development, classroom diagnostic assessments, provision of instructional resources and greater accountability and transparency in schools as well as greater community and public engagement.

EVALUATION PURPOSE AND EVALUATION DESIGN

The purpose of the evaluation is to document and measure the impact of the G-PriEd pilot intervention in 122 schools in Georgia in terms of improvement of learning outcomes in math and reading. The results of the baseline and endline studies also contributed to the establishment of national norms and standards for reading and math competences in the primary grades.

The evaluation of G-PriEd seeks to address three research questions:

1. What is the student performance against grade-level norms and standards in reading and math before implementing the project? What are the differences in performance between student sub-groups (ethnicity, gender, region, small/medium/large schools)?
2. Has students' performance in reading and math improved against the initial grade-level standards as a result of the interventions in the pilot schools? What is the extent and magnitude of improvement? What are the differences in performance improvements between student sub-groups (ethnicity, gender, region, small/medium/large schools) as a result of this pilot intervention?
3. What are the changes to the national norms and grade-level standards proposed for reading and math based on the data gathered throughout this project? Do these differ for students in the different sub-groups for which data were collected?

STUDENT PERFORMANCE AT BASELINE (RESEARCH QUESTION 1)

To address Research Question 1 above, we assessed student performance measures against initial grade-level standards in reading and math before implementing the project. This required a descriptive analysis of the baseline data collected in 2013 by G-PriEd. For each grade, we

presented norms and standards for math and reading at baseline[1]. The results of these analyses were presented in the Baseline Report[2].

We calculated norms for the sample as a whole as well as for pilot and control schools separately at baseline. For the standards, we identified four levels of proficiency, determined the level which corresponds to the "minimum grade-level requirement" for math and reading, and provided the proportion of students who fall into each level of proficiency and that reached the minimum requirement. We found that for math the fraction of students that reached the minimum requirement was between 40 and 80 percent depending on the grade. For reading, the fraction of students that reached the minimum requirement varies between 60 and 80 percent, depending on the grade.[3]

UPDATING NORMS AND STANDARDS (RESEARCH QUESTION 3)

The third question requires updating the standards and norms analyses discussed when answering the first research question, using the endline data. This was addressed in a separate companion Norms and Standards Report[4]. Norms and Standards are also discussed in this report. Standards are reviewed throughout the main section of the report. Norms are discussed in Annex VIII.

EVALUATION FINDINGS (RESEARCH QUESTION 2)

The main purpose of this report is to address the second research question, which specifically deals with the causal impact of the G-PriEd intervention on student's performance in reading and math.

The impact evaluation of G-PriEd uses a quasi-experimental methodology whereby pilot and control school students were assessed prior to the start of the intervention in spring 2013 and again two years later, after the intervention ended, in spring 2015.

This was a quasi-experimental design as the process by which G-PriED assigned schools to treatment and comparison groups was not randomized. In effect, schools were invited to apply for the program by the MES through a promotional campaign. Pilot schools were then selected from the pool of applicants on a first-come first-served basis while control schools were selected from the pool of non-applicants. The fact that schools were not randomly assigned into treatment and comparison groups implies that both observable and unobservable characteristics of schools in these two groups may be rather different. To address this challenge, the impact evaluation employs a Difference-in-Differences (DID) model. This method involves comparing the changes between baseline and endline test scores in treatment schools

[1] It is important to distinguish between *norms* and *standards*. Norms are used to situate the performance of a specific student in comparison with the performance of a specific student population. Standards specify what level of performance on a test (i.e. what score) is required for a student to be classified into a given performance category. In other words, the goal of the standard setting is to describe what a student who achieves a given score on a specific test, typically knows.

[2] G-PriEd Baseline Report, final version submitted by ME&A/NORC to USAID in May 2015.

[3] Note that these figures correspond to fractions of students reaching minimum requirements associated to math and reading *Rasch* scores (see G-PriEd Baseline Report for details). An analogous exercise was conducted for each math and reading competence. The results at the competence level are much nosier, so for math the fraction of students reaching the minimum requirement varies between 5 and 86 percent, and for reading this fraction varies between 15 and 87 percent.

[4] G-PriEd Updated Norms and Standards, submitted by ME&A/NORC in October 2015.

to changes between baseline and endline test scores in comparison schools for a given grade level. We also present results using Value-Added Models (VAM), an approach that differs from the DID model in that it focuses on the progress at the student level rather than changes at the school level (in other words it involves following a panel of students over time). In terms of outcomes of interest, the evaluation focuses on reading and math test scores as well as proportions of students who meet minimum grade-level requirements.

We implement the DID and VAM approaches to analyze the effects of the program on math and reading. For reading, two types of exams were fielded, one for Georgian native speakers and another for Georgian as a Second Language in ethnic minority schools (Armenian, Azeri and Russian); therefore we present the analyses for these two groups separately.

For math we analyze the effects on a single math score (Rasch score), as well as proportions of students reaching the minimum requirement for each grade. Using the DID model, we find that for the Rasch score the program has positive and significant effects for grades 3 and 4, positive but only marginally significant effects for grades 1 and 2 (that is, these results are significant at the 10 percent level of confidence, but not the 5 percent), and no significant effects for grades 5 and 6. Regarding the effect on the proportion of students achieving the minimum grade-level requirement, we find positive and significant effects only for students in grade 3.

For reading Georgian as native speakers, we analyze not a single Rasch score, but raw scores for each of the reading competences evaluated in each grade. We also analyze the effect of the program on the proportion of students achieving the minimum requirement for each competence. Using the DID model, for 1st grade we find no effects for any raw score, and for achieving the minimum requirement we find a positive and significant effect only for Phoneme Segmenting. For 2nd grade we find positive and significant effects for two competences, Letter Sound Fluency and Vocabulary, for both achieving the minimum requirement and the raw score. For grades 3 to 5 we find no significant effects for any raw score; for proportions of students meeting minimum thresholds we find positive effects for Comprehension of Narrative Text in 4th grade and Passage Reading Fluency in 5th grade. For 6th grade, we find a detrimental impact for Vocabulary, both in the case of the raw score and the minimum requirement.

For the Georgian as a Second Language (GSL) exams we also analyze the effects of the program by competence. In the case of GSL we present results only for raw scores because the sample of students who took the GSL tests is too small to produce good estimates of the proportion of the population of students that are at each level of proficiency. In effect only 447 students were assessed in GSL (as compared to 2,837 students in Georgian schools). This also implies that finding significant effects is going to be less likely for these tests. In fact, we did not find any significant effect for GSL for any grade level. Other than the small sample sizes, the lack of significant results could be explained by the fact that, according to the G-PriEd Pilot Phase report[5], the training of ethnic minority school teachers proved more challenging than that of Georgian school teachers. Indeed, the project found that it was difficult to identify qualified translators to translate the training materials and supplementary reading materials, and that some teachers from ethnic minority schools did not have a mastery of the Georgian language

[5] G-PriEd Pilot Phase Report, 10 August 2015. Shared by G-PriEd.

that was adequate to understand the trainings. Other mechanisms that may explain these findings are discussed.

We also use the DID model to analyze treatment heterogeneity across student gender, school size, language of the test (for math) and regions. We find evidence of treatment heterogeneity for language of the test, as the effect of the program is positive and significant for students taking the exam in Georgian and Azeri, not significant for students taking the exam in Armenian, and detrimental and significant for students taking the exam in Russian. We also find evidence of treatment heterogeneity for school size, in particular for math. Specifically, we find larger effects for small schools than for large schools. We do not find strong evidence of treatment heterogeneity by gender. By region we find that for math there are positive impacts for Achara, Kvemo Kartli, Imereti, Mtskheta-Mtianeti and Samegrelo & Zemo Svaneti for the Rasch score; for reading we do not find any region having significant effects for more than a couple of competences. No clear pattern is worth highlighting in terms if heterogeneity by region for reading.

Finally, using the VAM we analyze the impact of GPriEd on the same outcomes. We argue that VAM is a preferable specification than the DID model because it focuses on changes at the student level rather than at the school level, and because a growing literature shows that VAM produce relatively similar estimates to those found in schooling interventions where treatment assignment is randomized.

For math Rasch scores we find positive impacts for grades 3-5 at endline. For reading we find positive and significant impacts for three competences for students in grade 3 at endline, one competence in grade 4 and two competences in grade 5. No detectable impacts for students in grade 6 at endline for either math or reading are found. Note that VAM only allows to estimate the effect of the program for students that at endline are in grades 3-6, as students that at endline are in grades 1 and 2 cannot be included in the analysis because they were not observed at baseline.

DISCUSSION AND FINAL COMMENTS

In sum, we found that G-PriEd has had positive and significant impacts on math and reading outcomes, especially when we focus on the VAM results, which is our preferred specification. If the program is going to be expanded, special attention should be placed on two aspects:

- No effects for 6th graders. We did not find evidence that the program affected any outcome for students in 6th grade. This could be because 5th and 6th grade teachers received no training in 2014, but any extension of the program should make sure that the program has the expected effects on this population when teachers receive training in full.

- No effects for GSL and detrimental impact on math for students taking the exam in Russian. There are potentially various explanations for the lack of positive results for minority students. First, the training of ethnic minority school teachers proved more challenging than that of Georgian school teachers. Second, teachers of minority students did not receive training in 2014 due to budget constraints. Third, the sample size for

minority students was perhaps too small to detect reasonable effects. In any case, if the program is going to be extended special attention should be devoted to the effects on minority students.

A. PROJECT BACKGROUND

The USAID/Georgia Primary Education Project (G-PriEd) is a 5-year (2011-2016) $8.7 million pilot project designed to provide comprehensive assistance to the primary education system to improve reading and math competences of Georgian and ethnic minority students in 122 pilot schools (103 Georgian and 19 ethnic minority language instruction schools). This project, implemented by Chemonics, is in line with the government's reforms to change the education system from a teacher-centered model to a student centered-model. In collaboration with the Georgian Ministry of Science and Education, G-PriEd aims to strengthen key components of the education system through activities at different levels[6]:

- **Teacher trainings and in-service support**: principals and teachers are trained in best instructional practices in reading and math. G-PriEd does this by first training national trainers who are then responsible for training teachers from each pilot school. Schools are clustered into groups ("cohorts") and each group of schools is trained by a team of two reading trainers and two math trainers. Furthermore, teachers receive continuous support through school-based Teacher Learning Circles (TLC). In each medium to large-size school, there are two TLCs, one in math and one in reading while small-size schools have one combined math/reading TLC. During TLCs, teachers discuss student progress, test scores and brainstorm solutions to any challenges. TLCs are led by a teacher facilitator trained by G-PriEd.

  As seen in Table 1 below, in spring 2013, reading and math teachers from Grades 1-6 from all 122 pilot schools were trained. In fall 2013-spring 2014, due to budget restrictions, the G-PriEd team trained reading and math teachers from Grades 1-4 and in Georgian schools only, resulting in 103 schools trained. However, G-PriEd continued to train principals as well as the TLC facilitators from all 122 pilot schools. Similarly, all national trainers participated in ToTs, either as trainers for the teachers or classroom observers. In fall 2014-spring 2015, the training resumed with trainings of teachers from all primary grades, G1-6, in both Georgian and ethnic minority schools.[7] Table 1 shows the total number of days of training offered by G-PriEd by grade and subject. One day of training consisted of 6 hours of training.

Table 1. Number of days of training offered per year, grade and subject

	Georgian Schools				Ethnic Minority Schools		
Grades	**G1-4**		**G5-6**		**G1-6**	**G1-4**	**G5-6**
Subject	**Math**	**Reading**	**Math**	**Reading**	**Gsl**	**Math**	**Math**
Spring 2013	3 days	4 days	3 days	3 days	4 days	3 days	3 days
Nov-Dec 2013	2 days	2 days	none	none	none	none	none

[6] Adapted from G-PriEd Project Fact Sheet. Retrieved on 12 November 2014 from http://www.usaid.gov/sites/default/files/documents/1863/G-PriEd%20factsheet.pdf

[7] Source: Email correspondence with G-PriEd Chief of Party, 13 January 2015 and G-PriEd Monitoring Data 2013, 2014, 2015.

	Georgian Schools				Ethnic Minority Schools		
Grades	**G1-4**		**G5-6**		**G1-6**	**G1-4**	**G5-6**
Subject	**Math**	**Reading**	**Math**	**Reading**	**Gsl**	**Math**	**Math**
March-April 2014	2 days	2 days	none	none	none	none	none
Oct 2014-Feb 2015	4 days	3 days	7 days	5 days	5 days	7 days	7 days

Finally, teachers also received support through classroom visits from the national trainers and regional coordinators. The purpose of these classroom visits was to observe first-hand whether teachers had applied what they had been taught in trainings and to give teachers constructive feedback as a result of the observations.

- **Classroom diagnostic (formative) assessment**: G-PriEd designed an assessment tool for Georgian primary students in order to provide teachers with real-time information that they can use to adapt teaching practices

- **Provision of quality instructional resources**: G-PriEd designed and produced supplementary leveled readers for each grade. In addition G-PriEd provided several types of reading and math materials such as math manipulatives (rainbow fraction tiles, decimal blocks, mathematics games and toys, geometry student kits, math activity cards) and student newspapers for grades 3-6 students, as well as educational equipment (projector, CD/DVD players). All 122 pilot schools received educational equipment and math manipulatives in spring 2013, and in October 2013 and March 2014 all 122 pilot schools received the supplementary leveled readers.[8]

In addition, G-PriEd aimed to enhance community and parental engagement, accountability and transparency in all target schools. To do this, G-PriEd created school report cards with information from school observations, training records, parental engagement activities and TLC activities to distribute to each school principal at a principals' workshop held at the end of every school year. Schools with the highest scores received recognition while schools with the lowest scores received additional support from the project. Furthermore, G-PriEd created parental engagement cards which provided parents strategies for them to support their child's reading and math skills development through simple activities. G-PriEd also provided schools with ideas of competitions that they could implement in order to bring parents into the schools.

Mendez England & Associates (ME&A), with its partner NORC at the University of Chicago, were contracted to conduct the impact evaluation of the G-PriEd project in order to assess the impact of the project on learner outcomes. A parallel goal of the evaluation is to establish national norms and standards of reading and math for Grades 1-6.

[8] Source: email correspondence with G-PriEd Chief of Party, 6 November 2014.

B. EVALUATION PURPOSE & EVALUATION DESIGN

B1. EVALUATION PURPOSE

The purpose of the evaluation is to document and measure the impact of the G-PriEd pilot intervention in 122 schools in Georgia in terms of improvement of learning outcomes in math and reading.

The evaluation seeks to measure the improvement towards Goal 1 of USAID's Global Strategy – improving reading skills for 100 million children in primary grades by 2015 worldwide – by measuring the percentage change in proportion of students in primary grades who, after two years of schooling, demonstrate sufficient reading fluency and comprehension to read to learn. Additionally, the evaluation will use the two Performance Management Plan (PMP) indicators of USAID/Georgia to measure the project's achievements including: 1) the change in the proportion of students who, by the end of the primary cycle, are able to read and demonstrate understanding as defined by a country curriculum, standards, or national experts; and 2) the change in the proportion of primary grade students who, by the end of each school year, are meeting math and reading requirements as defined by a country curriculum, standards, or national experts.

B2. EVALUATION QUESTIONS AND INDICATORS

The evaluation of G-PriEd seeks to address three research questions:

1. What is the student performance against grade-level norms and standards in reading and math before implementing the project? What are the differences in performance between student sub-groups (ethnicity, gender, region, small/medium/large schools)?
2. Has students' performance in reading and math improved against the initial grade-level standards as a result of the interventions in the pilot schools? What is the extent and magnitude of improvement? What are the differences in performance improvements between student sub-groups (ethnicity, gender, region, small/medium/large schools) as a result of this pilot intervention?
3. What are the changes to the national norms and grade-level standards proposed for reading and math based on the data gathered throughout this project? Do these differ for students in the different sub-groups for which data were collected?

The first question required a descriptive analysis of the baseline data and was addressed in the Baseline Report[9]. The third question was addressed in a separate companion Norms and Standards Report[10]. Norms and Standards are also discussed in this report. Standards are reviewed throughout the main section of the report. Norms are discussed in Annex VIII. The main purpose of this report is to address the second research question, which specifically deals with the causal impact of the G-PriEd intervention on student's performance in reading and math.

[9] G-PriEd Baseline Report, final version submitted by ME&A/NORC to USAID in May 2015.
[10] G-PriEd Updated Norms and Standards, submitted by ME&A/NORC in October 2015.

B3. EVALUATION DESIGN

The impact evaluation of G-PriEd used a quasi-experimental methodology whereby pilot and control school students were assessed prior to the start of the intervention in spring 2013 and again two years later, after the intervention ended, in spring 2015. Below we describe the process taken for measuring indicators of interest as well as empirical methods used for the impact analysis.

B3.1 Indicator Measurement

The evaluation focuses on two types of outcomes:

1. Student's test scores on reading and math tasks linked to specific reading and math competences
2. Proportions of students who meet minimum grade-level requirements.

In this section, we describe the process taken to measure these two types of indicators.

Prior to the baseline data collection, G-PriEd, in collaboration with a large group of education and child development experts, developed separate assessment tools for math, reading in Georgian, and reading in Georgian as a Second Language (GSL) for each of the primary grades, such that the tools were intended to be leveled appropriately for each grade. For each grade and in each subject, two versions (forms) of the test were created. It is common for different versions of an assessment tool to be created so that students can be re-assessed at different points in time without the familiarity of the tool confounding the actual skill level of the student. The tests assessed students in the skills listed in Table 2 below. The number and complexity of items (questions) for each of these skills differed from grade to grade. More information can be found in Annex II.

Table 2. Evaluated competences by grade

	Math	Reading	GSL
Grade 1	Counting Number identification Comparing numbers Operation on numbers Patterns Geometric figures	Phoneme segmenting Syllable segmenting Letter sound fluency Word reading fluency	Phoneme segmenting Syllable segmenting Letter sound fluency Word reading fluency
Grade 2	Counting Number identification Comparing numbers Operation on numbers Algebra Patterns Geometric figures Data analysis	Letter sound fluency Word reading fluency Passage reading fluency Vocabulary	Phoneme segmenting Syllable segmenting Letter sound fluency Word reading fluency Passage reading fluency
Grade 3	Number identification Comparing numbers Operation on numbers Algebra Patterns Geometric figures Data analysis	Passage reading fluency Vocabulary Comprehension of narrative text Comprehension of informational text	Word reading fluency Passage reading fluency Vocabulary Comprehension of narrative text Comprehension of informational text
Grade 4	Number identification Comparing numbers Operation on numbers Algebra Patterns Geometric figures Data analysis	Passage reading fluency Vocabulary Comprehension of narrative text Comprehension of informational text	Word reading fluency Passage reading fluency Vocabulary Comprehension of narrative text Comprehension of informational text
Grade 5	Number identification Comparing numbers Operation on numbers Algebra Geometric figures Area Data analysis	Passage reading fluency Vocabulary Comprehension of narrative text Comprehension of informational text	Passage reading fluency Vocabulary Comprehension of narrative text Comprehension of informational text
Grade 6	Number identification Comparing numbers Operation on numbers Algebra Relations between quantities Geometric figures Area Data analysis	Passage reading fluency Vocabulary Comprehension of narrative text Comprehension of informational text	Passage reading fluency Vocabulary Comprehension of narrative text Comprehension of informational text

Given that ME&A/NORC was not involved in the development of these tools, the first step in the analysis of the data was to assess the quality of these assessment tools to ensure that all items in the tools were valid from a psychometric standpoint and should be included in the analysis. The second step was to ensure that Form 1 and Form 2 of each test were comparable.[11] As mentioned above, for each grade, two different forms of the test were developed. "Forms" are versions of tests that were

[11] Both forms I and II were used at both baseline and endline.

constructed for a given grade level but are not composed of the same items (they are composed of the same subtasks). Since each form contains items specific to that form, a simple summary score (e.g. sum of all correct answers) could be biased by the level of difficulty of the items contained within the form. This makes it difficult to compare scores between different forms as it is not possible to determine whether a difference in scores stems from a true difference in the performance of students or from a difference in the levels of difficulty in the forms. In other words, we cannot be sure that a student tested with Form 1 would receive the same score had he been tested with Form 2. Therefore, pooling together the results of students from the two forms could be problematic.

In order to align the scores of the two forms on the same scale, we used a Rasch model. Note that it is only the two forms corresponding to a specific grade that are aligned on the same scale and not tests from different grades. No direct comparison should be made between grades using scores of tests from different grades. In parallel, the Rasch model also enables the verification of the quality of the items and ensures that a unidimensional competence is measured in reading and mathematics.

The Rasch model is part of a family of models called *Item Response Theory* (IRT) or latent trait models. These models link the probability of a student giving a correct answer on a specific item to the characteristics of the students and of the item. In an IRT model, the student parameter being considered is his/her ability in the cognitive domain of interest. For example, if a test is designed to measure mathematics achievement, the student's ability level in mathematics is the parameter that would influence his response on any mathematics item. The item parameter of interest is the level of difficulty of the item. If an item really measures ability level in mathematics, only its difficulty can influence the probability that a student gives a correct answer. Therefore, in Rasch analysis, the probability of a student giving a correct answer on a given item is considered to be dependent on the level of difficulty of the items relative to the level of ability of the student. Thus, the model considers that a test measures a given ability on a continuum, ranging from a low level of ability to a high level of ability. The ability of the students and the difficulty of the items are all put on this scale. All items that do not fit the model (based on fit statistics) are removed since those items are viewed as of low quality (for instance, the items are too difficult or too easy, and don't discriminate between different levels of student ability).

After running the Rasch analysis on the G-PriEd Reading and Math tests, we removed 3 problematic items from the Mathematics test (one in grade 2 and two in grade 5[12]). The Rasch model did not identify any item in the reading test as being problematic from a psychometric standpoint. The Math items that were removed will also be removed from the tests that will be used for the endline data collection.

The Rasch analysis transforms the scores such that the continuum of scores is based on a normal distribution with a mean of 500 and a standard deviation of 100. The mean was determined based on the performance of the students on each test. The 500 value represents the mean of scores of students for a specific test. It is important to note that the scales are not comparable across grade levels. In other words, a value of 500 on a grade 1 math test cannot be compared to a value of 500 on a grade 4 math test, each test has its own scale. For more information on the Rasch analysis, please refer to Annex II.

[12] The items that were removed are: Grade 2 (Form 2, item 8), Grade 5 (Form 1, item 1; Form 2, item 27).

In addition to the psychometric analysis, our Georgian math and reading experts also conducted a thorough content analysis of the tests in order to understand the objective of each item in relation to the national curriculum. While the psychometric analysis of the test showed that most items are valid from a psychometric standpoint, the content analysis revealed a few issues with the content of some of the items. First, we note that the G-PriEd tests were developed as rapid diagnostic tools that follow the national curriculum loosely. For instance, the national math curriculum targets more than 20 indicators (competences) while the G-PriEd math test includes only 10; also, the national reading curriculum does not include a standard for passage reading fluency in 4th grade while the G-PriEd 4th grade reading test does include a measure of passage reading fluency[13]. Therefore, the G-Pried tests may not be completely exhaustive in testing students against the national curriculum. Second, some test items were problematic. For math, we found that some items were categorized incorrectly while for reading, it was not clear what some items were intended to measure (see Annex II for more information).

While we have kept all of these items given that they were not problematic from a psychometric standpoint, we have re-classified the math items into the correct content category for the analysis. Finally, we also note that more than 70% of the students gave a correct answer to about 75% of the items in Grade 1 and about 60% of the items in Grade 2, indicating that these tests may be too easy overall for those grades.

Reading and math scores used

The Rasch score allows us to pool the data from students tested with Form 1 and Form 2 ensuring that the results from the two forms are comparable. It also takes into account both the item difficulty level as well as student ability level and puts them on a shared continuum. The model assigns different weights to the items depending on their level of difficulty, and the overall Rasch score is estimated based on students' correct answers on these items. The Rasch score is an overall score for the test. However, following discussions with USAID, we agreed that giving separate scores for each reading competence (e.g. syllable segmenting, phoneme segmenting, word reading fluency, comprehension, etc.) would be more useful to the project, given that these competences are distinct and build on one another. In Math, however, an overall Rasch score is more appropriate given that the number of items in each competence is small, thereby making it difficult to give meaningful scores by competence. As such, we present raw scores by competence for reading[14] and overall Rasch scores for math.

Calculation of proportion of students that meet minimum grade-level requirements

Once reading and math scores were calculated, we needed to determine the threshold scores for each grade that students would need to obtain to be deemed meeting the minimum grade-level requirements. Once that determination was made, it then became possible to calculate the

[13] We understand that this was of interest to G-PriEd from a research standpoint and that the Ministry of Education agreed with it.

[14] While the use of a Rasch score would have been preferable in terms of the comparability between the two test forms, we performed a comparison between raw scores of students who took the test using Form 1 and those of students who took the test using Form 2 and found that the differences were not important between the two forms in most cases. Therefore the analysis uses the raw scores without correction.

proportion of students who meet these minimum requirements. In order to develop these threshold scores, we used a two-step process (for more information on this process, see Annex IV[15]):

- First, we developed grade-level standards. Standards specify what level of performance on a test (i.e. what score) is required for a student to be classified into a given performance category, what we call a "mastery level". For G-PriEd, we developed four mastery levels for each grade. Students were then categorized into a specific mastery level depending on their performance; and each mastery level was described using Performance Level Descriptors, to specify what students are able to achieve at that particular level.
- Second, after Performance Level Descriptors were developed, our team's local psychometricians, who are experts in the Georgian math and reading curricula, determined which of the four levels corresponded to the minimum requirement that students should know at each grade level. In addition, thresholds for each competence in reading were also defined. These thresholds were determined based on an analysis of the items in each level and competence as compared to the national reading and math curriculum. They can, therefore, be used to calculate the main indicators of interest to USAID, i.e. proportion of students who have met the minimum math and reading requirement at each primary grade level, and also specifically at Grade 2[16].

The thresholds for math (Rasch scores only) and reading (Rasch scores and raw scores by competence) are given in the table below. For each reading competence, we give the threshold as well as the maximum score possible (e.g. for syllable segmenting, the threshold score is 37 while the maximum possible score for that competence was 47). The reading thresholds correspond to the number of points scored by the student in a given competence and not to the number of items answered correctly. Phoneme segmenting, syllable segmenting, letter sound fluency, word reading fluency and oral reading fluency (from reading passage) were all timed at 1 minute, except for letter sound fluency in Grade 2 which was timed at 30 seconds.

[15] Also see G-PriEd Baseline Report.
[16] For more information on these thresholds, please refer to the G-PriEd Updated Norms and Standards Report.

Table 3. Math and Reading Thresholds per Grade Level

Grade	1	2	3	4	5	6
Math – Rasch Score	408.01	441.01	524.01	414.01	517.01	457.01
Reading – Rasch Score	430.01	415.01	474.01	455.01	489.01	453.01
Reading – Syllable segmenting	37/47					
Reading – Phoneme segmenting	65/81					
Reading – Letter fluency	52/65	30/65				
Reading – Word fluency	30/60	30/60				
Reading - Vocabulary		8/12	10/15	14/20	14/20	13/20
Reading – Passage fluency		50/90	58/115	75/195	75/233	100/234
Reading – Comprehension narrative			7/9	9/11	9/15	11/15
Reading – Comprehension informational			4/6	4/6	4/6	4/7

B3.2 Sample Description and Data Collection

Sample

To participate in G-PriEd, schools had to apply to the program, and final selection for the intervention was determined on a first come-first serve basis. Comparison schools (control schools) were randomly selected from the pool of schools that did not apply. Before the program started, G-PriEd conducted a baseline data collection in spring 2013 with samples of students from 122 pilot and 119 control schools from grades 1 to 6 using the reading and math assessment tools that the project developed. In each school, a sample of 1 to 6 students per grade was randomly selected depending on school size. With this sampling strategy, the baseline target sample size consisted of 1,665 students from pilot schools and 1,579 students from the control schools for a total of 3,244. The final baseline sample size was 3,244. For more information on the sampling strategy, see Annex 1.

Table 4. Baseline Study Population and Sample Size

Grade	# of students in pilot schools	# of students in control schools	Sample size in pilot schools	Sample size in control schools	Total # of students in pilot and control group
1	2,976	2,975	274	265	539
2	3,446	3,295	276	262	538
3	3,441	3,098	278	263	541
4	2,966	2,793	280	262	542
5	3,105	2,781	279	264	543
6	3,136	3,065	278	263	541
Total	19,070	18,007	1,665	1,579	3,244

In spring 2015, the evaluation team conducted the endline data collection in the same schools from which we collected data at the baseline. To the extent possible, we also attempted to assess the same students at endline and baseline. In other words, we aimed to have a panel of students. Given that two years had elapsed since the baseline in spring 2013, we expected that students who were in Grades 1-4 at baseline would be in Grades 3-6 at endline. On the other hand, Grade 5 and 6 students at baseline would have graduated on to middle school by the endline period, while Grade 1-2 students at endline would be new cohorts of students who had not yet been in primary school during the baseline period.

For the endline sample, a target quota for each grade and in each school was determined based on the baseline sampling distribution. Then, the sampling strategy was the following:

1. First, attempt to re-assess the same students as baseline (Grades 3-6).
2. If target quota is not reached with the panel students, randomly sample new students to reach quota.
3. For Grades 1-2, randomly sample students to reach quota.

At endline, the total number of students tested in Grades 1 through 6 was 3,285 – 1,569 from pilot schools and 1,716 in control schools (against a target of 3,289 students). Table 5 shows the distribution of students by grade and between pilot and control schools

Table 5. Number of students tested at endline by school type and grade

Grade	Sample size in pilot schools	Sample size in control schools	Total # of students in pilot and control group
1	262	297	559
2	260	282	542
3	263	285	548
4	259	282	541
5	264	286	550
6	261	284	545
Total	1,569	1,716	3,285

As for the panel sample, out of a total 2,179 students who were in Grades 1-4, we managed to re-assess 1,735 of them. The attrition rate was therefore 20.42% (445 students out of 2,179).

Data Collection

All assessors and supervisors attended a comprehensive training on how to administer the math and reading assessments. The training was followed by a field practice in schools. Teams of 2 to 3 enumerators were sent to each school (2 enumerators for small and medium size schools and 3 enumerators for large schools). During the field, each assessor was observed at least once a week by the supervisor and by a NORC representative using a standard Assessor Observation Checklist. For data entry, USAID asked that the data entry platform created by G-PriEd be used. For quality assurance purposes, 100% double data entry was completed.[17]

B3.3 Evaluation Approach and Methodology

In any impact evaluation, constructing a valid counterfactual constitutes the main methodological challenge. The ideal comparison group stems from the use of experimental methods in which eligible participants are randomly assigned to receive the intervention or not. The process by which schools were selected for the USAID G-PriEd program *was not* based on random assignment. Schools were invited to apply for the program by the MES through a promotional campaign. Of a total of 817 applications received, 122 pilot schools were then chosen on a first-come first-served basis to participate in G-PriEd. For each pilot school, G-PriEd and the MES then selected a comparison school from the same region from the pool of schools that did not apply for G-PriEd, taking into consideration the school size category and language of instruction.

The fact that schools were not randomly assigned into treatment and comparison groups implies that both observable and unobservable characteristics of schools in these two groups may be rather different. For instance, we know that schools differ in their willingness to participate in the program; schools that were selected for the program were picked because they applied first, indicating that they were more willing and eager to participate. In addition, schools may differ in other ways we cannot observe. For example, schools that wanted to participate (applicant schools) may have more motivated staff, while comparison schools that did not apply may be more isolated and/or may have less motivated staff. Because some characteristics between treatment and comparison schools are

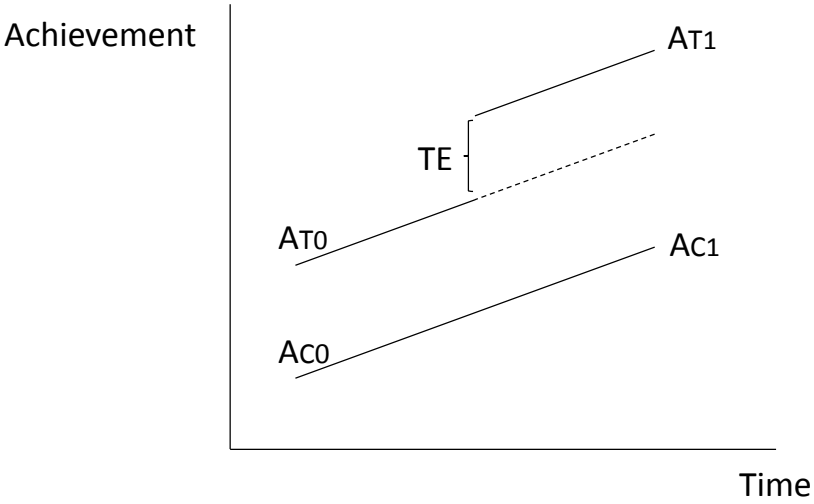
[17] More information can be found in the baseline and endline Administrative Process Reports, written by GORBI.

presumably different, comparing their outcomes after the intervention will most likely lead to biased estimates of the treatment effect. In other words, the difference in their outcomes will reflect not only the impact of the intervention, but also the fact that schools in the two groups would be different even in the absence of the program.

To address this challenge, we conduct a Difference-in-Difference (DiD) analysis. This method involves comparing the changes between baseline and endline test scores in treatment schools to changes between baseline and endline test scores in comparison schools. Baseline data was collected on a sample of students in grades 1-6 in spring 2013 and endline data was collected two years later (see Section B4 and Annex I on data collection and final samples). Therefore, our main analytical approach is to compare the changes in scores between 2013 and 2015, and between treatment and comparison schools. This method allows us to determine the impact of the G-PriEd intervention for each grade level.

A graphical representation of the proposed methodology is depicted by Figure 1:

Figure 1. Difference in difference estimator



Where:

A_{T0} is the average test score for a given grade at baseline in the treatment group
A_{C0} is the average test score for a given grade at baseline in the control group
A_{T1} is the average test score for a given grade at endline in the treatment group
A_{C1} is the average test score for a given grade at endline in the control group
TE is the treatment effect for the corresponding grade

In words, the proposed approach measures the difference between mean test scores for a given grade between baseline and endline, and then compares these differences between treatment and control groups.

Analysis on math and reading scores

Mathematically, we estimate

$$A_{it} = \alpha + \beta D_i + \gamma E_t + \delta D_i E_t + \mu S_i + u_{it} \qquad (1)$$

Where A_{it} measures student achievement for student *i* in period *t*; D_i is a dummy variable for treatment status; E_t is a dummy variable for the endline, S_i indicates the sex of the student *i*; u_{it} is an error term and α, β, γ, δ and μ are parameters to be estimated. The parameter of interest is δ, which captures the effect of the program on students' outcomes at endline.

The identification assumption of this approach is that, in the absence of treatment, students in pilot and control schools would experience the same changes in the outcomes of interest and, therefore, any differential change between the two groups can be attributed to the program.

Analysis on proportion of students who meet minimum requirements

In addition to looking at changes in test scores, we also analyze the effect of the program on the probability that students reach certain thresholds (i.e. meeting math and reading grade-level requirements). For this analysis we implement a model in the same fashion as equation (1), except that the dependent variable is a dummy for reaching a certain achievement threshold. Mathematically, we estimate:

$$\mathbf{I}(A_{it} > A^*) = \alpha + \beta D_i + \gamma E_t + \delta D_i E_t + \mu S_i + u_{it} \qquad (2)$$

Where$\mathbf{I}(\cdot)$ is an indicator variable and A^* is the score needed to achieve the minimum requirement for each grade. This model will allows us to estimate the effect of the program on the probability that students reach a meaningful achievement threshold.

Analysis by subgroup

When sample size allows for it, we conduct separate regressions to analyze heterogeneity, i.e. differential impacts on different subgroups (for example, to study how boys and girls are differently affected by treatment). The approach is a little different when we want to analyze heterogeneity over dimensions where there are more than two or three classes or categories, like when we analyze the results by region. In these cases, instead of splitting the sample, we will include interaction terms between region dummies and all the key parameters of the DID model. This will imply estimating:

$$A_{it} = \sum_l \alpha^l d_l + \sum_l \beta^l D_i d_l + \sum_l \delta^l D_i d_l E_t + \sum_l \gamma^l d_l E_t + \mu S_i + u_{it} \qquad (3)$$

where α^l correspond to region-specific intercepts, β^l reflect the time-invariant effect of being in a pilot school for each region, γ^l are region-specific trend terms and δ^l are region-specific treatment effects. The rest of the terms remain the same as in the original specification.

Attrition

We did not find any correlation between attrition and treatment status, so we do not control for attrition in our main impact evaluation results. However, we did find that attrition was negatively correlated with baseline test scores, which could imply that the estimated results are only valid for the type of students that are less likely to drop out from the sample. In regressions not shown we run the model depicted by equation (1) but weighting the observations using Inverse Probability

Weights, attempting to approximate the parameters if there had been no attrition.[18] The main caveat of this methodology (other than it assumes that attrition is driven by observable characteristics) is that only students that are observed both at baseline and endline can be used; therefore, no estimates can be produced for grades 1 and 2, because students at endline in these grades were not observed at baseline. We do not find differences compared to the original estimates that are worth highlighting using this correction.

Value-Added Models (VAM)

So far we have discussed the empirical approaches where the treatment effect is defined as the difference between the changes in the outcomes of interest between treatment and control groups at the school-grade level. Given that for some students we have data at both baseline and endline, we can also analyze how the program affected test scores at the student level, rather than at the school level. For this analysis, we review the results when we focus on changes at the student level. In other words, we study how the progress of students in pilot schools compares to students' progress in control schools. These models are referred in the literature as Value–Added Models (VAM).

The key assumption underlying the VAM is that baseline test scores are a sufficient statistic to characterize the cognitive ability of students at baseline. Mathematically, for each student *i* we estimate:

$$A_{i1} = \alpha + \beta D_i + \gamma A_{i0} + \mathbf{x_i'}\boldsymbol{\delta} + u_{i1} \qquad (4)$$

where, A_{s1} and A_{s0} are measures for achievement, such as a test score, for student *s* at endline and baseline, respectively; D_s is a dummy variable for treatment status; $\mathbf{x_i}$ is a vector of characteristics at the student level, specifically age and gender, as well as characteristics at higher levels of analysis, in particular geographic region and school size; u_{s1} is an error term and α, β, γ and δ are parameters to be estimated.

It is important to highlight that VAM models require a panel of students; that is, baseline and endline data for each student. Therefore, this approach is only feasible for students that were in grades 1-4 in 2013 and in grades 3-6 in 2015. In other words, no estimates can be calculated for the impact of the program for students that in 2015 are in 1st grade or in 2nd grade.

There are a few reasons why VAM may be preferable to DID to analyze a program like G-PriEd. First, given that we are comparing the progress of each student (looking at changes in achievement between baseline and endline for the same student), we do not need to be worried about changes in school composition because we are looking only at students that are both at baseline and endline.

Second, the identification assumption underlying each model is different, and possibly the one associated with the VAM is more reasonable to believe than the one associated with the DID model. In particular, the identification assumption for the DID model is that, in the absence of treatment, students in pilot and control schools would experience the same changes in the outcomes of interest;

[18] This method consists of modeling the probability that observations attrite from the sample, and then using the predicted probabilities to give more weight to the observations that were more likely to attrite.

for VAM on the other hand, the assumption is that each student's baseline test score is a sufficient statistic to characterize the cognitive ability of students at baseline.

Whether VAM can provide unbiased causal estimates of the impact of a given program is an empirical question. A growing literature shows that VAMs produce relatively similar estimates to those found in interventions where treatment assignment is randomized, which is considered the gold standard in program evaluation.[19]

C. EVALUATION FINDINGS

In this section, we present the main findings for the three subjects that we evaluated, namely math, reading for Georgian native speakers and for Georgian as Second Language (GSL). We first present the results of the Difference-in-Difference methodology. For math and reading we discuss the results for scores (Equation 1), and for the fraction of students reaching the minimum requirement (Equation 2). For GSL we only present results for raw scores as the sample for these exams are too small to produce good estimates of the minimum requirement thresholds. Afterwards we present a summary of the main findings for the heterogeneity analysis (Equation 3). Finally, we describe the results of the Value-Added Models (Equation 4).

C1. DIFFERENCE-IN-DIFFERENCES ANALYSIS

Key Findings

- **Math**
  - For Rasch scores: positively but marginally significant impacts for Grades 1 and 2, and positive and significant impacts for Grades 3 and 4. No detectable impacts in Grades 5-6.
  - For achieving the minimum requirement: positive and significant impacts only for Grade 3.
- **Reading**
  - Positive impacts for two competences for Grade 2 for both the raw score and achieving the minimum requirement (Letter Sound Fluency and Vocabulary).
  - A few additional positive impacts: proportion meeting threshold and raw score for Phoneme Segmenting in Grade 1, proportion meeting threshold in Comprehension of Narrative Text in Grade 4 and in Passage Reading Fluency in Grade 5.
  - Detrimental impact on Vocabulary in Grade 6
- **GSL**
  - No detectable impacts in any grades for any competence.

C1.1 Math

Table 6 presents results for math Rasch scores by grade. Rasch scores are standardized so the mean is 500 and the standard deviation is 100. In the first two columns baseline mean Rasch scores for control and pilot groups are shown and in the third column is the difference between these two

[19] For a review see Thomas J. Kane. "Do Value-Added Estimates Identify Causal Effects of Teachers and Schools?" *The Brown Center Chalkboard*, Brookings Institution, October 30, 2014.g

groups. For all grades but the 4th pilot schools have higher mean test scores than control schools. This suggests that pilot and control schools were different from the beginning, which underscores the importance of using a DID model as opposed to simply comparing the endline averages between pilot and control schools. The same constructs at endline are displayed in columns (3) to (6). Pilot schools have higher test scores at all grades at endline. In column (7) the 'raw' or non-parametric Difference-in-difference is shown. This is simply the difference, between pilot and control groups, of their respective changes in mean scores over time. These estimates of the impact of the program indicate that schools in the pilot group observed greater test score improvements than in the control group at all grades but grade 6. Finally, in columns (8) and (9) the results from the regression analysis described in equation (1) are shown; specifically, column (8) shows the DID estimate of the program effect and column (9) the corresponding standard error. It is clear that the results between the 'raw' DID and the regression DID are not too different, which is perhaps a consequence of the fact that only a few characteristics are being included in the model. For the regression results the main conclusions are:

- For 3rd and 4th grade: The estimates are positive, relatively large and highly significant. Third and fourth graders participating in G-PriEd increased their math scores by 49.8 and 41.4 percent of a standard deviation, respectively, as compared to control students.
- For 1st and 2nd grade: The estimates are positive but only marginally significant (10 percent of confidence). It's possible that we are observing smaller impacts for 1st and 2nd graders than for 3rd and 4th graders because the former have been exposed to G-PriEd less time than the latter.
- For 5th and 6th grade: The effects are small and not statistically significant.

Table 6. Results for math Rasch scores by grade

	Baseline			Endline			'Raw' DID	Regression DID[a]	
	Control	Pilot	Dif	Control	Pilot	Dif		Effect	SE
	(1)	(2)	(3)=(2)-(1)	(4)	(5)	(6)=(5)-(4)	(7)=(6)-(3)	(8)	(9)
1st grade	494.9	504.3	9.4	502.5	542.4	39.9	30.5	28.6	(17.6)
2nd grade	492.5	506.7	14.2	505.4	547.8	42.4	28.2	28.4	(14.9)
3rd grade	496.8	502.8	6	495.5	551.9	56.4	50.4	49.8***	(13.8)
4th grade	505.2	495.2	-10	495.6	526.2	30.6	40.6	41.9**	(14.0)
5th grade	487.0	512.3	25.3	502.9	533.3	30.4	5.1	4.95	(13.8)
6th grade	495.5	504.1	8.6	527.7	533.5	5.8	-2.8	-2.93	(13.0)

[a] All DID regressions include pilot and endline dummies, a sex dummy and fixed effects for language of test, school size at baseline and region. Standard errors are clustered at the school level. Sample sizes are 1,101 for 1st grade, 1,073 for 2nd grade, 1,095 for 3rd grade, 1,082 for 4th grade, 1,084 for 5th grade, and 1,081 for 6th grade.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

Table 7 presents the same results as Table 6 but for proportions of students reaching the minimum requirement for math. For grades 1, 2, 5 and 6, at baseline, pilot schools have higher proportions of students reaching the minimum requirement than control schools, while the opposite pattern is observed for grades 3 and 4. At endline, the fraction of students reaching the minimum requirement

is higher for pilot than for control schools at all grades. For the regression estimates, displayed in column (8), the main conclusions are as follows:

- For 1st, 2nd and 4th grade: the impact is positive but not significant. The result for 4th grade is perhaps surprising, given that we found a significant effect for the 4th grade Rasch score. This suggests that the gains found for the Rasch score are likely observed either far below or far above the minimum requirement and, thus, did not contribute to changing the status of students from not meeting the requirement to meeting the requirement.
- For 3rd grade: the result indicates that the program increased the proportion of students who meet the requirement by 22 percentage points. This is substantial given that the proportion of students achieving the minimum requirement in the control group at baseline was 36.2 percent.
- For 5th and 6th grades: the impacts are negative but very small and not significant.

Table 7. Results for proportions who meet minimum thresholds in math by grade

	Baseline			Endline			'Raw' DID	Regression DID[a]	
	Control	Pilot	Dif	Control	Pilot	Dif		Effect	SE
	(1)	(2)	(3)=(2)-(1)	(4)	(5)	(6)=(5)-(4)	(7)=(6)-(3)	(8)	(9)
1st grade	0.855	0.857	0.002	0.832	0.902	0.07	0.068	0.062	(0.052)
2nd grade	0.719	0.781	0.062	0.704	0.848	0.144	0.082	0.081	(0.055)
3rd grade	0.362	0.338	-0.024	0.361	0.554	0.193	0.217	0.22***	(0.059)
4th grade	0.840	0.827	-0.013	0.780	0.826	0.046	0.059	0.062	(0.051)
5th grade	0.345	0.449	0.104	0.413	0.507	0.094	-0.01	-0.011	(0.063)
6th grade	0.301	0.336	0.035	0.425	0.451	0.026	-0.009	-0.0091	(0.064)

[a] All DID regressions include pilot and endline dummies, a sex dummy and fixed effects for language of test, school size at baseline and region. Standard errors are clustered at the school level. Sample sizes are 1,101 for 1st grade, 1,073 for 2nd grade, 1,095 for 3rd grade, 1,082 for 4th grade, 1,084 for 5th grade, and 1,081 for 6th grade.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd data for 2013 and 2015.

Overall, G-PriEd has a positive effect on math outcomes, particularly for 3rd and 4th grades and less strongly for 1st and 2nd grades. However, no effects are observed for 5th and 6th grades. These results may be explained by the way primary grade teachers are assigned to classes in Georgia. Indeed, it is often the case that teachers in Grades 1 through 4 follow their cohort of students as they move up grades; in other words a teacher starts with a cohort in Grade 1 and follows that cohort all the way to Grade 4 and once they reach Grade 4, they return to Grade 1 to follow the next cohort of students. In Grades 5 and 6, reading and math are taught by different teachers, and these teachers do not necessarily follow students over grades; this is dependent on the school principal and beyond the control of the G-PriEd project. Given this system, students in grades 1 to 4 may have been taught by teachers with more accumulated training than their counterparts in 5th and 6th grades.

Furthermore, as explained in Section A, three G-PriEd training waves were deployed between 2013 and 2015, but 5th and 6th grade teachers were trained in only two of them (2013 and 2015) due to budget constraints. This may have affected the impact of the training on teacher performance, and ultimately on students' outcomes in these grades.

C1.2 Reading (for Georgian native speakers)

The analysis for reading is less straightforward, given that we focus on specific competences rather than on a single score. Table 8 shows results for mean scores for each reading competence by grade. The score for each competence is the summation of correct answers by competence, or 'raw' score; except for vocabulary, comprehension of narrative text and comprehension of informational text, where we display the average of the percent of correct answers instead.[20] At baseline, with only a few exceptions in grades 1 and 2, pilot schools outperform control schools. At endline pilot schools observe higher averages in all competences and all grades, except for 6th grade passage reading fluency and vocabulary. Looking at the DID regression results, the main findings are:

- For grades 1, 3, 4 and 5, most impact estimates for the scores by competence are positive, although none of them is significant.
- For grade 2 we find positive impacts for all four competences, and two of them are statistically significant: letter sounds fluency and vocabulary.
- For 6th grade, on the other hand, we find that the parameters are *negative* for three of the four competences, and for one of them, vocabulary, the coefficient is statistically significant.

[20] Following discussions with USAID, we agreed to present results in terms of percent of correct answers for these three competences as this was more useful to the project.

Table 8. Results for reading scores by grade and competence

	Baseline			Endline			'Raw' DID	Regression DID[a]	
	Control	Pilot	Dif (3)=(2)-(1)	Control	Pilot	Dif (6)=(5)-(4)	(7)=(6)-(3)	Effect	SE
	(1)	(2)		(4)	(5)			(8)	(9)
A. Grade 1									
Phoneme segmenting	51.3	52.4	1.1	57.8	63.5	5.7	4.6	3.88	(3.55)
Syllable segmenting	38.6	37.9	-0.7	38.2	39.8	1.6	2.4	2.10	(1.65)
Letter sounds Fluency	48.9	49.6	0.8	55.0	58.1	3.1	2.3	1.93	(2.00)
Word reading Fluency	20.5	23.4	2.9	25.6	27.8	2.2	-0.6	-0.85	(1.93)
B. Grade 2									
Letter sounds Fluency	43.0	42.5	-0.5	42.4	49.5	7.1	7.5	7.58***	(2.06)
Word reading Fluency	31.8	34.7	2.9	34.1	40.3	6.3	3.4	3.57	(1.86)
Passage reading fluency	35.2	38.0	2.8	40.2	46.4	6.2	3.4	3.57	(1.93)
Vocabulary (% correct)	58.6	58.2	-0.4	58.3	66.8	8.5	8.8	8.86**	(2.94)
C. Grade 3									
Passage reading fluency	44.6	49.8	5.2	51.7	59.9	8.1	3.0	2.83	(2.28)
Vocabulary (% correct)	66.8	70.0	3.2	66.0	73.6	7.6	4.4	4.40	(2.82)
Comp narrative text (% correct)	58.8	63.3	4.5	56.9	65.2	8.4	3.9	3.87	(3.09)
Comp informational text (% correct)	41.2	44.7	3.5	41.8	44.4	2.7	-0.8	-0.76	(3.41)
D. Grade 4									
Passage reading fluency	63.4	65.9	2.6	68.6	73.7	5.1	2.5	2.44	(2.93)
Vocabulary (% correct)	62.3	62.9	0.7	62.5	65.9	3.4	2.7	2.71	(2.51)
Comp narrative text (% correct)	68.0	70.8	2.9	65.5	69.3	3.8	0.9	0.73	(2.52)
Comp informational text (% correct)	51.6	52.4	0.8	55.3	60.8	5.5	4.7	4.61	(3.40)
E. Grade 5									
Passage reading fluency	72.8	74.6	1.8	76.2	83.9	7.6	5.8	5.55	(3.92)
Vocabulary (% correct)	55.2	57.5	2.2	57.4	62.7	5.3	3.1	2.93	(2.46)
Comp narrative text (% correct)	58.7	63.1	4.4	63.8	64.7	1.0	-3.4	-3.58	(2.54)
Comp informational text (% correct)	49.7	50.8	1.0	53.9	57.1	3.2	2.2	2.05	(2.90)
F. Grade 6									
Passage reading fluency	76.4	81.7	5.3	89.4	88.6	-0.8	-6.1	-5.13	(3.81)
Vocabulary (% correct)	62.7	65.6	2.9	69.0	66.3	-2.6	-5.5	-5.17*	(2.04)
Comp narrative text (% correct)	67.0	71.3	4.3	71.7	72.9	1.2	-3.1	-2.55	(2.52)
Comp informational text (% correct)	56.0	57.7	1.7	58.0	60.1	2.1	0.3	0.48	(3.06)

[a] All DID regressions include pilot and endline dummies, a sex dummy and fixed effects for school size at baseline and region. Sample sizes are 940 for 1st grade, 941 for grade 3rd grade and 934 for 4th grade. For grades 2, 5 and 6 sample sizes vary a little due to competence-specific missing data. For grade 2 sample size was 922 for Letter Sounds Fluency, 921 for Passage Reading Fluency, and 923 for Wording Reading Fluency and Vocabulary. For grade 5 sample size was 936 for Passage Reading Fluency and 938 for the other competences. For grade 6 sample size was 929 for Passage Reading Fluency and 930 for the other competences. Standard errors are clustered at the school level.
* p<0.05 ** p<0.01 *** p<0.001

Source: G-PriEd data for 2013 and 2015.

Table 9 shows results for proportions of students reaching the minimum requirement for each reading competence by grade. Focusing on the figures at baseline, pilot schools outperform control schools in the majority of cases, although for some competences the opposite pattern is observed (control schools outperform pilot schools in 1st grade Phoneme and Syllable Segmenting, 2nd grade Letter Sounds Fluency and Vocabulary, 4th grade Comprehension narrative text and 5th grade Passage reading Fluency). At endline pilot schools outperform control schools across all competences and all grades, except for 6th grade, where control schools observe better results for two competences. The main results for the regression DID are:

- For achieving the minimum requirement at 1st grade, we only found a significant effect for Phoneme Segmenting; the impact estimate indicates that G-PriEd increased the proportion of students reaching the minimum requirement for the competence by 15 percentage points.
- For 2nd grade students, all impact estimates are positive, but only two are significant – Letter Sound Fluency and Vocabulary. The results indicate that G-PriEd increased the proportions of students passing the Letter Sound Fluency threshold by 10 percentage points and the Vocabulary threshold by 13 percentage points, compared to their peers in comparison schools.
- For the impact on proportions of students meeting minimum thresholds for grades 3 to 5 there are a couple of significant effects - Comprehension of Narrative Text in 4th grade and Passage Reading Fluency in 5th grade, but there are also a few (not significant) negative coefficients.
- For grade 6, only the effect for Vocabulary is significant, and is negative.

Table 9. Results for proportions who meet minimum thresholds in reading by grade and competence

	Baseline			Endline			'Raw' DID	Regression DID[a]	
	Control	Pilot	Dif (3)=(2)-(1)	Control	Pilot	Dif (6)=(5)-(4)	(7)=(6)-(3)	Effect	SE
	(1)	(2)		(4)	(5)			(8)	(9)
A. Grade 1									
Phoneme segmenting	39.8	38.8	-1.0	46.4	61.6	15.2	16.2	15*	(7.4)
Syllable segmenting	68.5	67.4	-1.2	67.9	74.4	6.6	7.7	6.7	(6.5)
Letter sounds Fluency	47.2	52.9	5.7	70.1	79.8	9.8	4.1	3.1	(6.3)
Word reading Fluency	18.5	27.7	9.2	30.8	38.8	8.0	-1.2	-1.7	(7.1)
B. Grade 2									
Letter sounds Fluency	81.9	80.4	-1.4	83.3	91.8	8.5	10.0	10*	(5.0)
Word reading Fluency	50.5	60.4	10.0	60.4	74.3	13.9	4.0	4.3	(6.3)
Passage reading fluency	15.3	18.9	3.6	23.9	33.1	9.2	5.6	6.1	(5.1)
Vocabulary	32.9	30.4	-2.5	29.7	40.8	11.1	13.5	13*	(6.2)
C. Grade 3									
Passage reading fluency	20.2	31.6	11.4	31.7	47.4	15.7	4.3	4.0	(5.7)
Vocabulary	48.6	56.7	8.1	55.1	61.9	6.8	-1.3	-1.4	(7.2)
Comp narrative text	12.4	21.1	8.7	16.7	24.9	8.2	-0.5	-0.6	(5.0)
Comp informational text	9.6	10.9	1.3	8.4	8.8	0.5	-0.8	-0.9	(3.8)
D. Grade 4									
Passage reading fluency	29.7	33.3	3.7	38.3	47.8	9.5	5.8	5.8	(6.0)
Vocabulary	29.7	33.3	3.7	32.0	42.1	10.1	6.5	6.5	(6.4)
Comp narrative text	23.7	22.4	-1.4	16.2	25.9	9.7	11.1	11*	(4.9)
Comp informational text	22.8	24.0	1.2	20.3	30.8	10.5	9.4	9.2	(6.4)
E. Grade 5									
Passage reading fluency	44.8	41.5	-3.3	48.0	61.9	13.9	17.2	17*	(6.6)
Vocabulary	16.2	23.1	6.9	22.7	32.5	9.9	2.9	2.7	(5.5)
Comp narrative text	41.9	54.1	12.2	53.3	58.6	5.3	-6.9	-7.1	(5.9)
Comp informational text	24.8	28.1	3.3	33.8	37.8	4.0	0.6	0.5	(6.5)
F. Grade 6									
Passage reading fluency	20.6	27.3	6.6	37.4	36.0	-1.4	-8.0	-6.8	(5.8)
Vocabulary	40.4	52.7	12.3	56.3	49.0	-7.3	-19.6	-0.19**	(7.0)
Comp narrative text	42.7	49.8	7.1	46.9	54.3	7.4	0.3	1.2	(7.1)
Comp informational text	39.0	40.7	1.8	41.0	44.1	3.1	1.4	1.7	(7.5)

[a] All DID regressions include pilot and endline dummies, a sex dummy and fixed effects for school size at baseline and region. Sample sizes are 940 for 1st grade, 941 for grade 3rd grade and 934 for 4th grade. For grades 2, 5 and 6 sample sizes vary a little due to competence-specific missing data. For grade 2 sample size was 922 for Letter Sounds Fluency, 921 for Passage Reading Fluency, and 923 for Wording Reading Fluency and Vocabulary. For grade 5, sample size was 936 for Passage Reading Fluency and 938 for the other competences. For grade 6, sample size was 929 for Passage Reading Fluency and 930 for the other competences. Standard errors are clustered at the school level.

* p<0.05 ** p<0.01 *** p<0.001

Source: G-PriEd data for 2013 and 2015.

As we mentioned before, an explanation for the lack of positive results for grade 6 could be that teachers from these grade received less training sessions than their counterparts in grades 1 to 4. However, it is still unexpected that the program would have a significant detrimental effect, even if the teachers did not receive the full training 'dosage'.

C1.3 Georgian as a Second Language (GSL)

Reading abilities for students in minority schools were assessed using a Georgian as a Second Language (GSL) assessment tool. In total 447 students were assessed in GSL (as compared to 2,837 students in Georgian schools). Given the small sample size, it is reasonable to expect that not many significant results will be found for this subpopulation. In the case of GSL we present results only for raw scores because the sample of students who took the GSL tests is too small to produce good estimates of the proportion of the population of students that are at each level of proficiency.

Table 10 shows results for GSL by grade and competence. The differences between pilot and control schools present a very different pattern than what was found for math and reading Georgian as a native language. Indeed, with a few exceptions in 2nd, 5th and 6th grades, control schools outperformed pilot schools at baseline. At endline the pattern is less clear, as pilot schools observe better results than control schools over a few more competences compared to baseline. This suggests that the program may had have a positive effect on some of these competences. In effect, the 'raw' DID is positive for 16 of the 27 grade-competences analyzed. However, when we look at the DID regression results in column (8), we do not find any significant effects for any competence at any grade.

Table 10. Results for GSL raw scores by competence and grade

	Baseline			Endline			'Raw' DID	Regression DID	
	Control	Pilot	Dif	Control	Pilot	Dif		Effect	SE
	(1)	(2)	(3)=(2)-(1)	(4)	(5)	(6)=(5)-(4)	(7)=(6)-(3)	(8)	(9)
A. Grade 1									
Phoneme segmenting	36.6	32.4	-4.1	50.6	55.1	4.5	8.6	7.88	(7.04)
Syllable segmenting	25.3	20.5	-4.8	31.1	32.4	1.3	6.1	5.66	(4.10)
Letter sounds Fluency	24.3	21.1	-3.2	24.9	26.8	1.9	5.0	5.22	(4.59)
Word reading Fluency	11.8	10.1	-1.7	14.4	14.1	-0.3	1.4	1.37	(2.00)
B. Grade 2									
Phoneme segmenting	52.5	59.6	7.1	64.0	66.8	2.8	-4.3	-5.83	(9.07)
Syllable segmenting	27.6	30.6	2.9	34.8	32.9	-1.9	-4.8	-6.11	(5.84)
Letter sounds Fluency	28.6	25.9	-2.8	34.4	35.0	0.6	3.4	1.74	(5.43)
Word reading Fluency	15.3	12.0	-3.3	14.7	15.9	1.3	4.6	2.86	(2.40)
Passage reading fluency	19.2	15.2	-4.0	23.9	24.0	0.1	4.1	2.25	(2.38)
C. Grade 3									
Word reading Fluency	21.8	19.7	-2.1	25.7	24.3	-1.3	0.8	0.40	(2.52)
Passage reading fluency	21.9	17.5	-4.4	29.3	26.7	-2.6	1.7	1.90	(3.87)
Vocabulary (% correct)	49.5	47.3	-2.2	54.4	40.8	-13.6	-11.4	-12.1	(7.00)
Comp narrative text (% correct)	56.3	53.0	-3.3	62.8	52.2	-10.6	-7.2	-7.84	(11.3)
Comp informational text (% correct)	61.8	56.1	-5.8	68.8	48.6	-20.1	-14.4	-15.8	(8.05)
D. Grade 4									
Word reading Fluency	35.2	33.7	-1.5	38.7	34.9	-3.8	-2.3	-0.49	(4.20)
Passage reading fluency	26.9	26.6	-0.3	39.9	36.3	-3.6	-3.3	-0.97	(5.46)
Vocabulary (% correct)	58.2	56.3	-1.8	60.6	53.7	-6.8	-5.0	-3.21	(9.33)
Comp narrative text (% correct)	60.9	50.0	-10.9	56.3	52.2	-4.1	6.8	9.66	(8.07)
Comp informational text (% correct)	70.0	66.3	-3.7	67.2	68.6	1.3	5.0	5.89	(7.21)
E. Grade 5									
Passage reading fluency	39.2	28.7	-10.6	48.5	42.9	-5.6	4.9	4.50	(4.37)
Vocabulary (% correct)	58.4	50.6	-7.7	53.3	51.4	-2.0	5.8	5.38	(5.60)
Comp narrative text (% correct)	49.7	44.9	-4.7	43.9	52.4	8.5	13.2	12.8	(6.58)
Comp informational text (% correct)	55.4	55.9	0.5	59.8	61.3	1.4	1.0	1.25	(12.3)
F. Grade 6									
Passage reading fluency	35.9	37.8	1.9	56.2	50.8	-5.4	-7.3	-7.85	(4.81)
Vocabulary (% correct)	56.0	55.4	-0.6	62.9	60.4	-2.5	-2.0	-2.30	(7.17)
Comp narrative text (% correct)	60.2	59.7	-0.5	64.4	59.8	-4.6	-4.2	-3.94	(6.45)
Comp informational text (% correct)	57.0	50.9	-6.1	51.7	52.3	0.5	6.6	6.10	(9.57)

[a] All DID regressions include pilot and endline dummies, a sex dummy and fixed effects for mother tongue (assumed to be the language of the math test), school size at baseline and region. Sample sizes are 148 for 1st grade, 146 for 2nd grade, 147 for 3rd grade, 147 for 4th grade, 150 for 5th grade, and 150 for 6th grade. Standard errors are clustered at the school level.
* p<0.05 ** p<0.01 *** p<0.001. Source: G-PriEd and EMIS data for 2013 and 2015.

The lack of significant results could be explained by the fact that, according to the G-PriEd Pilot Phase report[21], the training of ethnic minority school teachers proved more challenging than that of Georgian school teachers. Indeed, the project found that it was difficult to identify qualified staff to translate the training materials and supplementary reading materials, and that some teachers from ethnic minority schools did not have a mastery of the Georgian language that was adequate to understand the trainings. One of the lessons learned identified by G-PriEd was that trainings for ethnic minority school teachers needed to be better tailored to their language level and needs.

Furthermore, teachers of minority students did not receive training in 2014 due to budget constraints. The lack of full 'dosage' of treatment on minority teachers may also explain the lack of results for this population.

Finally, these findings may also be a result of the small sample sizes. Increasing the number of students assessed is likely necessary to provide a more reliable evaluation of the program on minority students.

C1.4 Summary

In sum, the DID analysis finds positive effects for grades 1 to 4 for math Rasch scores although the results for 1st and 2nd grade are only marginally significant, and for achieving the minimum requirement for math we find a positive and significant effect only for 3rd grade. For Georgian as a native language, we found positive and significant effects for two reading competences in 2nd grade (Letter Sounds Fluency and Vocabulary), for both the raw score and the minimum requirement. We also found positive effects for one competence in 1st grade (Phoneme Segmenting), one competence in 4th grade (Comprehension of Narrative Text) and one competence in 5th grade (Passage Reading Fluency), but only for the minimum requirement and not the corresponding raw score.

As a robustness check we also conducted an extension of the DID model called DID - Propensity Score Matching. In essence, this approach discards or underweights control schools that are too different in terms of observables characteristics at baseline compared to the pilot schools. The results using this approach are presented in Annex VI, but overall they are similar to those obtained using a simple DID.

Another approach we explored was using the take up rate as the covariate of interest rather than the dummy for participating in the program. We define take up rate as the rate at which teachers participated in training sessions, averaged at the school level. We do not find major differences in the results with respect to the simple DID model, which is not surprising as training take up was relatively high. For this reason we do not present these results in this report but they are available upon request.

[21] G-PriEd Pilot Phase Report, 10 August 2015. Shared by G-PriEd.

C2. ANALYSIS BY SUB-GROUP

Different populations may observe different treatment effects. For example, having a sizeable impact on a very large school might be more difficult than improving the outcomes of students in small schools. We have already discussed how G-PriEd has different effects on students by grades. In this section we focus on additional heterogeneity dimensions, namely sex of the student, language of the test (for the math test), school size, and region. We stratify the sample to analyze the effect of the program over gender, school size and language of test. We also analyze effects for the 12 regions in Georgia. In this case rather than stratifying the sample we incorporate regional dummies interacted with the treatment dummy. We conduct these heterogeneity analyses pooling all grades in each regression and only for math and reading Georgian for native speakers. The sample sizes for GSL are too small for these exercises, and given that we did not find any major effects for the more basic specifications it is unlikely that any analysis on heterogeneity is going to be valuable.

Tables presenting results for these analysis can be found in Annex VII. The tables present results for both scores (Rasch for math and raw for reading) and achieving the minimum requirement. Our main conclusions are:

- **Differences by sex**
  - Math: No major differential impact across female and male students.
  - Reading: While there are a few cases where there seems to be a difference between the effects by gender, we don't think there is strong evidence that there is gender-driven heterogeneity for the impact of G-PriEd.
- **Differences by language of test for math (Georgian, Russian, Azeri and Armenian)**
  - Positive and significant impacts of G-PriEd for students taking the exam in Georgian and Azeri (although for the latter the impact for the Rasch score is significant only at 10 percent).
  - No significant impact detected for students taking the exam in Armenian.
  - Negative impact for students taking the exam in Russian for the Rasch score. Given that the sample size for Russian students is so small (n=92) perhaps not much should be read into this result.
- **Differences by school size**

  To analyze heterogeneity by school size we divide the schools in three categories according to total number of students at baseline: small (less than 300 students), midsize (between 300 and 599 students), and large (600 students or more).
  - Math: Only effects for small schools are found to be significant, effects for midsize and large schools are positive but not significant. Heterogeneity across school size seems to be correlated with baseline mean scores; in effect, small schools had lower mean scores at baseline compared to large schools, which suggests that this intervention has helped small schools to 'catch up' with large schools.
  - Reading: For Letter Sound Fluency, Word Fluency, Vocabulary and Passage Reading Fluency we found significant and positive effects for small schools but not for midsize or large schools. No evidence of treatment heterogeneity is apparent for other competences.
- **Differences by region**

- Math: We find positive impacts (either for the Rasch score or the minimum requirement or both) for Achara, Kvemo Kartli, Imereti, Mtskheta-Mtianeti and Samegrelo & Zemo Svaneti.
- Reading: For ease of exposition only results for raw scores are presented. We do not find any region having significant effects for more than a couple of competences. No clear pattern is worth highlighting in terms if heterogeneity by region.

C3. VALUE-ADDED MODELS

In this section we discuss the results of the Value–Added models. This approach may be preferable to the DID because it analyzes student progress rather than changes at the school-grade level. One limitation of this model is that only results for students that are assessed both at baseline and endline can be produced, which has two main implications. First, we can only estimate the effects of the program for students that are in grades 3 to 6 at endline. Second, given that we have to discard all students that were observed only once, the resulting sample sizes are too small for the GSL analysis, which is why we do not discuss the results for this group using this model.

Key Findings

- **Math**
  - Significant positive impacts for students that are in grades 3-5 at endline.
  - No detectable impacts for students in grade 6 at endline.
- **Reading**
  - Positive impacts for three competences for students in grade 3 at endline, one competence in grade 4 and two competences in grade 5.
  - No detectable impacts for students in grade 6 at endline.

C3.1 Math

Table 11 shows mean math Rasch test scores by treatment group and grade. Note that these figures include only students that were observed both at baseline and endline. Pilot schools observe higher scores than control schools for grades 1 to 3, and the effect for 2nd graders is statically significant. Mean test score is higher for the control group for 4th graders, but the difference is not significant. This suggests that pilot and control schools are not directly comparable as pilot schools observe better results than control schools even before the intervention.

Table 11. Baseline Math Rasch scores

	Control	Pilot	Dif
Grade 1	492.9	506.5	13.67
Grade 2	488.6	512.1	23.50*
Grade 3	493.0	503.4	10.43
Grade 4	507.7	496.7	-10.98

* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

Table 12 presents VAM results for math Rasch scores by grade. In the first row we can see the effect of G-PriEd and on the second we present the coefficient on the key control variable for this model:

the math Rasch score at baseline. As a reminder, the Rasch scores have been developed such that the mean is 500 and the standard deviation is 100. For 3rd graders at endline, the effect of G-PriEd is 41.3% of a standard deviation. For 4th graders it is 27.6% and for 5th graders it is 33.1%. In education studies, effect sizes of 0.2-0.4 of a standard deviation are considered medium size effects. For 6th graders the parameter is positive but small and not significant. We can also see that, with no exception, the baseline math score is positively and significantly correlated with the endline test score. In fact, the correlation seems to be increasing with grade level, which is consistent with a model where human capital formation is described as a cumulative process.

Table 12. Estimated impact of G-PriEd using the VAM model on Math Rasch scores

	Grades			
	3rd	4th	5th	6th
Effect of G-PriEd	41.3***	27.6**	33.1**	8.88
	(11.9)	(9.51)	(10.9)	(10.3)
Coefficient for Baseline math Rasch score	0.53***	0.63***	0.69***	0.69***
	(0.086)	(0.060)	(0.072)	(0.067)
Obs	423	438	430	431

Note: All specifications include students' age and gender, categorized student teacher ratio, fraction of certified teachers and class size, and dummies for language of test, school size at baseline and region.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

C3.2 Reading

Table 13 displays baseline mean raw scores for each competence by grade and treatment status. Similarly to what we see for math, for the most part pilot schools outperform control schools at endline. In all analyzed cases but one, mean scores are higher for pilot schools than for control schools, although the differences are significant only for three grade-competences.

Table 13. Baseline Reading raw scores

	Control	Pilot	Dif
A. Grade 1			
Phoneme Segmenting	38.91	39.06	0.149
Syllable Segmenting	51.62	53.80	2.172
Letter Sound Fluency	48.64	50.79	2.155
Word reading Fluency	20.54	24.09	3.547*
B. Grade 2			
Letter Sound Fluency	43.51	42.52	-0.992
Word reading Fluency	31.27	35.05	3.782*
Passage Reading Fluency	34.83	38.56	3.726
Vocabulary	6.977	7.046	0.0693
C. Grade 3			
Passage Reading Fluency	44.23	50.64	6.408*
Vocabulary	9.986	10.43	0.448
Comprehension narrative txt	5.225	5.698	0.472
Comprehension info txt	2.485	2.612	0.127
D. Grade 4			
Passage Reading Fluency	63.61	66.00	2.382
Vocabulary	12.50	12.61	0.106
Comprehension narrative txt	7.576	7.834	0.257
Comprehension info txt	3.384	3.413	0.0295

* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

Table 14 to 16 present results by grade for reading. As in the previous section, for reading we do not use a single Rasch score but analyze raw scores for each reading competence. Table 14 shows results for the four competences in which 3rd graders were evaluated at endline. Each column gives results for a given regression model. In the first row we can see the effect of G-PriEd for each of the competences listed. There are positive and significant effects for Passage Reading Fluency (4.94 points), Vocabulary (0.66 points) and Comprehension of narrative text (0.47 points). To provide a sense of the relative size of these effects, at the bottom of the table we show the mean of the competence of interest (dependent variable) for the control group. The coefficients for Passage Reading Fluency and Comprehension of narrative text are roughly 10 percent of the mean score for those competences and for Vocabulary the coefficient is 6 percent of the mean score for this competence.

As mentioned previously, in the VAM, we use student scores at baseline to control for cognitive ability at the student level. Therefore, as control variables we are not including one single score (like for math), but all scores for each of the four competences that 3rd graders at endline were evaluated in when they were 1st graders at baseline, namely Phoneme Segmenting, Syllable Segmenting, Letter Sound Fluency and Passage Reading Fluency. This allows us to study which competences evaluated at baseline are (conditionally) correlated with the competences at endline. For example, Passage Reading Fluency at endline is not correlated with Phoneme or Syllable Segmenting at baseline, but it is

correlated with Letter Sound and Word Reading Fluency. Overall, Word Reading Fluency at baseline is positively correlated with all four endline competences, meaning that Word Reading Fluency in 1st grade is a strong predictor of performance in passage reading fluency, vocabulary and comprehension of narrative and informational texts in 3rd grade.

Table 14. Estimated impact of G-PriEd using the VAM model on reading raw scores - 3rd grade

Competence of interest	Passage Reading Fluency	Vocabulary	Comprehension narrative txt	Comprehension info txt
Impact of G-PriEd	4.94**	0.66*	0.47*	-0.040
	(1.89)	(0.32)	(0.23)	(0.14)
Regression Coefficients for Baseline scores:				
Phoneme Segmenting	-0.012	0.016	0.011	0.000015
	(0.10)	(0.014)	(0.011)	(0.0077)
Syllable Segmenting	-0.0027	0.017*	0.011	-0.00057
	(0.045)	(0.0069)	(0.0060)	(0.0037)
Letter Sound Fluency	0.23**	0.045***	0.017	0.0018
	(0.079)	(0.013)	(0.0099)	(0.0069)
Word reading Fluency	0.65***	0.037**	0.037***	0.032***
	(0.094)	(0.012)	(0.011)	(0.0077)
Observations	358	358	358	358
Mean raw score for control group	52.0	10.1	5.22	2.54

Note: All specifications include students' age and gender, categorized student teacher ratio, fraction of certified teachers and class size, and dummies school size at baseline and region.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

Table 15 shows results for students in 4th grade in 2015. All impact estimates for G-PriEd are positive but only the one for Comprehension of Informational Text is significant, an effect of 0.29 points, equivalent to roughly 10 percent of the raw score mean for this competence. The other three coefficients are pretty small, both relative to their standard errors and the mean of the corresponding competence score.

In terms of the correlations between scores at baseline and those at endline, we see that Word Reading Fluency and Passage Reading Fluency in 2nd grade are strong predictors of Passage Reading Fluency in 4th grade, and that Vocabulary in 2nd grade is a strong predictor of all reading competences in 4th grade.

Table 15. Estimated impact of G-PriEd using the VAM model on reading raw scores – 4th grade

	Passage Reading Fluency	Vocabulary	Comprehension narrative txt	Comprehension info txt
Impact of G-PriEd	0.94	0.36	0.19	0.29*
	(1.95)	(0.36)	(0.23)	(0.14)
Regression Coefficients for Baseline scores:				
Letter Sound Fluency	0.13	-0.00042	0.0082	0.00064
	(0.082)	(0.015)	(0.010)	(0.0065)
Word reading Fluency	0.51***	0.020	0.029	0.025*
	(0.13)	(0.022)	(0.015)	(0.011)
Passage Reading Fluency	0.64***	0.042*	0.020	0.0090
	(0.11)	(0.017)	(0.014)	(0.0098)
Vocabulary	0.89*	0.40***	0.20***	0.085*
	(0.40)	(0.074)	(0.047)	(0.036)
Observations	372	372	372	372
Mean raw score for control group	69.4	12.5	7.19	3.22

Note: All specifications include students' age and gender, categorized student teacher ratio, fraction of certified teachers and class size, and dummies school size at baseline and region.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015

In Table 16 the results for students in 5th grade in 2015 are displayed. In this case we find positive and significant effects for Passage Reading Fluency (5.02 points) and Vocabulary (0.73 points), equivalent to approximately 6 percent of the mean of these competences' scores. No effects for either of the two reading comprehension competences are found. Regarding the estimated correlations between baseline and endline scores, we can see that in almost all cases, the estimated coefficients are statistically significant.

Table 16. Estimated impact of G-PriEd using the VAM model on reading raw scores – 5th grade

Competence of interest	Passage Reading Fluency	Vocabulary	Comprehension narrative txt	Comprehension info txt
Impact of G-PriEd	5.02*	0.73*	-0.18	0.28
	(2.04)	(0.37)	(0.31)	(0.16)
Regression Coefficients for Baseline scores:				
Passage Reading Fluency	0.95***	0.043***	0.034***	0.0090
	(0.066)	(0.011)	(0.0092)	(0.0063)
Vocabulary	-0.30	0.31***	0.25***	0.031
	(0.37)	(0.072)	(0.058)	(0.037)
Comprehension narrative txt	1.42*	0.17*	0.35***	0.12**
	(0.61)	(0.085)	(0.071)	(0.046)
Comprehension info txt	1.73*	0.13	0.28**	0.15*
	(0.68)	(0.11)	(0.089)	(0.058)
Observations	366	366	366	366
Mean raw score for control group	74.7	11.4	9.56	3.78

Note: All specifications include students' age and gender, categorized student teacher ratio, fraction of certified teachers and class size, and dummies school size at baseline and region.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015

Finally, Table 17 presents results for students in grade 6 in 2015. In this case none of the coefficients are significant, and in fact three of the parameters are negative, although they are all pretty small.

Table 17. Estimated impact of G-PriEd using the VAM model on reading raw scores – 6th grade

Competence of interest	Passage Reading Fluency	Vocabulary	Comprehension narrative txt	Comprehension info txt
Impact of G-PriEd	-2.05	-0.43	0.019	-0.029
	(2.29)	(0.32)	(0.27)	(0.14)
Regression Coefficients for Baseline scores:				
Passage Reading Fluency	0.69***	0.012	0.023***	0.0022
	(0.063)	(0.0063)	(0.0063)	(0.0032)
Vocabulary	0.47	0.26***	0.13**	0.087***
	(0.37)	(0.047)	(0.047)	(0.023)
Comprehension narrative txt	1.89***	0.31***	0.32***	0.13***
	(0.50)	(0.070)	(0.067)	(0.038)
Comprehension info txt	0.99	0.20	0.38***	0.081
	(0.77)	(0.10)	(0.083)	(0.061)
Observations	377	377	377	377
Mean raw score for control group	88.8	13.8	10.7	4.09

Note: All specifications include students' age and gender, categorized student teacher ratio, fraction of certified teachers and class size, and dummies school size at baseline and region.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015

In sum, ***the estimated effects using the VAM approach are relatively different from the ones using the DID model***. While it is not possible to compare the results one by one because we can't estimate any impact for students who are in 1st or 2nd grades at endline with the VAM while it is possible with DID, we can see that ***for those cases where comparisons can be made, the VAM approach finds more positive and significant results than the DID model***. For math, DID finds positive effects only for 3rd and 4th grade, while VAM finds effects for 3rd, 4th and 5th grades. For reading DID finds almost no positive and significant effect for the raw scores in grades 3 to 6, but VAM finds effects for three competences in grade 3, one competence in grade 4 and two competences in grade 5.

F. DISCUSSION AND FINAL COMMENTS

In this report we presented the results of the impact evaluation of G-PriEd. We first discussed the results of the DID model and document some positive effects for math. In particular, for the Rasch scores we found marginally significant effects for 1st and 2nd graders of roughly a fourth of standard deviation, and strongly significant effects for 3rd and 4th graders of almost half of standard deviation. We did not find effects for 5th or 6th grades for math. For the proportion of students achieving the minimum requirement we found positive and significant effects only for 3rd graders. We also used the DID model to evaluate the impact of G-PriEd on each reading competence. For Georgian as a native language, where four competences in each grade were evaluated, we found positive and significant effects for two reading competences in 2nd grade (Letter Sounds Fluency and Vocabulary), for both the raw score and the minimum requirement. We also found positive effects for one competence in 1st grade (Phoneme Segmenting), one competence in 4th grade (Comprehension of Narrative Text) and one competence in 5th grade (Passage Reading Fluency), but only for the minimum requirement and not the corresponding raw score. We also found a significant decrease for Vocabulary for 6th graders in pilot schools. For GSL we found no significant effects for any competence at any grade.

We also used the DID model to analyze the effects of G-PriEd across different heterogeneity dimensions. The analysis by students' gender didn't show any consistent pattern that are worth highlighting. In terms of school size we found that at, least for math, small schools observed greater improvements than larger schools, which is possibly a consequence of small schools having lower average scores at baseline than large schools. We also found evidence of treatment heterogeneity across language of the test (for math). While there were positive impacts for Azeri and Georgian populations, we found no effects for students taking the exam in Armenian and *negative* effects for students taking the exam in Russian, although the sample size for this particular subsample was small.

In addition to the DID method we also explored VAM, which we argue is perhaps a preferable specification than the DID model because it focuses on changes at the student level rather than at the school level. The main caveat of VAM is that we need to observe the same student over time, which has two main implications. First, we drop from the analysis all the students that were observed only once; and second, and most importantly, we can only produce treatment impacts for students that at baseline were in grades 1-4 (so in 2015 they are in 3rd to 6th grades). Using this method we found positive effects for math for students in grades 3 to 5 at endline, and no effects for students in 6th grade at endline. This differs from the results using the DID model in that no significant effects were found for 5th grade using DID. For reading we found effects using VAM for three competences in grade 3 (Passage Reading Fluency, Vocabulary and Comprehension of Narrative Text), one competence in grade 4 (Comprehension of Informational Text) and two competences in grade 5 (Passage Reading Fluency and Vocabulary), but no effects for 6th graders. Compared to the results for reading using DID, we did not find any effect for the raw scores in grades 3 to 6 (with the exception of a negative effect for Vocabulary for 6th graders). We could use the VAM approach to explore the effects of the program on GSL scores but given the small samples that will result, we do not think this is productive.

There are two subpopulations for which we did not find any positive program effects regardless of the method used: Students in 6th grade in 2015 and GSL. For 6th graders we discussed the possibility that the fact that 5th and 6th grade teachers did not receive the full training 'dosage' maybe the reason why we did not find any effects for 6th graders. This was also the case for teachers of students that do not speak Georgian as their native language. In addition, two other situations made the analyses for students that do not speak Georgian as their native language more complicated. First, training teachers in minority schools proved more challenging than that of Georgian school teachers. Second, sample sizes were perhaps too small for a proper analysis, especially considering that these three minorities are very heterogeneous, and bundling them in one group is perhaps not appropriate (but splitting the sample even further would exacerbate the small sample problem).

In sum, we found that G-PriEd has had positive and significant impacts on math and reading outcomes, especially when we focus on the VAM results, which is our preferred specification. We consider VAM a more appropriate approach to measure the impact of G-PriEd than DID because this methodology allows us to control for cognitive ability at baseline at the student level, while the DID model focuses on changes at the school-grade level over time.

However, if the program is going to be expanded, special attention should be placed on two aspects:

- No effects for 6th graders. We did not find evidence that the program affected any outcome for students in 6th grade. As we have argued this could be because 5th and 6th grade teachers received no training in 2014, but any extension of the program should make sure that the program has the expected effects on this population when teachers receive training in full.

- No effects for GSL and negative effects on math for students taking the exam in Russian. If the program is going to be extended special attention should be devoted to the effects on minority students.

ANNEX I. G-PRIED: SAMPLING STRATEGY

Source: G-PriEd

In order to select an appropriate and meaningful distribution of schools to make up the initial cohort, a decision was made to focus on students as the unit of emphasis. Therefore, the sampling population was defined as Grade 1-6 students in Georgian and ethnic minority public schools. This student-focused rationale reflects the overall goal of the quality improvement initiative as measured in terms of impact on student learning. The student-focused rationale also organizes strategic resources in ways that reflect the demographic distribution of children in Georgia and targets interventions in proportion to school-aged populations.

The sample consisted of students in Grades 1-6.

An initial estimation of G-PriEd resources allowed working with approximately 13,000 students in approximately 110-120 schools. The same schools and students were to be included in the impact study; therefore, **the initial estimation of the pilot impact study population was 13,000**[22]**.**

A Randomized Block Design (otherwise known as multi-stage proportionate random sampling) or strategy was employed for dual purposes. It was important that schools selected for project pilot (hence, for the pilot impact study) were country-representative, or selected randomly; and students selected in such schools were school-representative, selected randomly. The process of applying this sampling strategy is described as the following:

a) Identifying the blocks of schools for pilot intervention

First, the blocks of schools were identified within which the student population was to be homogenous. The blocks were created by the geographic/administrative location of schools, language of their instruction, and size (number of students).

1. Geographic clusters: 12 clusters (11administrative regions of Georgia+ Abkhazeti)
2. Types of schools in each cluster by the language of instruction:
    a. Georgian
    b. Non-Georgian
3. Types of schools in each cluster by the school size:
    a. Small, 1-299 students
    b. Medium-size, 300-599 students
    c. Large schools: over 600 students

The first stage of sampling resulted in the identification of **43 blocks of schools; within each of these blocks, the student population was considered homogenous.**

b) Identifying the number of schools to be selected from each block for private intervention

[22] The total number of 1-6 grade students in Georgia's public schools is 260,060

Secondly, the multi-stage proportionate approach was used to identify how many schools were to be selected from each of the 43 blocks. The calculation of the number of students in schools was as follows:

- Initially, the number of students to be represented from each block of schools was identified in accordance with their ratio to the total number of Grade 1-6 students in Georgia (see table below).

  For instance, if in the **Abkhazeti Block**, there are 1,044 students of grade 1-6, this is approximately 0.40% of all students of grades 1-6 in Georgia's public schools (260,060 students); hence, the 0.4% of the 13,000 students (rough size of the study population), or **52 students of grade 1-6**, had to be from the Abkhazeti Block. The same rationale applied to all blocks.

- Then, it was decided to have at least 200 students in each of 12 regional clusters or at least 2,200 students per grade enrolled in the project implementation. Therefore, where the number of students per region was less than 200, it was *disproportionally* adjusted to 200. This approach was used with two (Abkhazeti and Racha-Lechkhumi) regions. As a result, the total number of students to be enrolled in the schools selected for the pilot was identified as **13,188** (see table below).
- Finally, the number of schools in each block that would comprise the estimated number of students from that block was calculated; the number of students was divided by the block to the average school size in that block; these numbers were then rounded up and adjusted to come up with the discrete numbers of schools.

  In case of Abkhazeti Block, **200 students were divided by 70** (average number of 1-6 students per Abkhazeti schools), and the number of schools (2.8) to be included was **rounded up to three schools**.

The table below provides the results of the multi-stage random sampling; it represents the theoretical framework, based on the average numbers of students in each school by the category (of geographic location, size, and language of instruction). With this theoretical framework, the total number of schools to work with was estimated at 121; and the total number of students to be included in G-PriEd activities from these schools as 13,188.

Table 18. Expected Number of 1-6 Grade Students and Corresponding Schools by the 43 School Blocks

Region		Total # of schools	Total # of 1-6 students	Average # of 1-6 students	% from the total	Initial Stipulation of pilot student	Hand-adjusted # of students	# of school	Hand-adjusted	BLOCK numbers
Abkhazeti	1.	**15**	1 044	70	100%	52	200	2,9	3	**1**

Sub-Total for Apkhazeti		**15**	**1 044**	**70**	**0,40 %**[23]	**52**[24]	**200**	**2,9**	**3**	
Adjara	1.	185	8 988	49	38%		449	9,2	9	**2**
	2.	29	5 529	191	24%		276	1,4	2	**3**
	3.	15	8 966	598	38%		448	0,7	1	**4**
Sub-total for Ajara		**229**	**23 483**	**103**	**9,03 %**	**1 174**	**1 174**	**11,4**	**12**	
Guria	1.	85	4 190	49	60%		209	4,2	4	**5**
	2.	10	2 177	218	31%		200	1,0	1	**6**
	3.	2	658	329	9%		33	0,1	1	**7**
Sub-total for Guria		**97**	**7 025**	**72**	**2,70 %**	**351**	**351**	**5,0**	**6**	
Tbilisi	1.	19	1 767	93	2%		88	0,9	1	**8**
	2.	36	8 026	223	10%		401	1,8	2	**9**
	3.	119	67 224	565	87%		3 360	5,9	6	**10**
Sub-total for Tbilisi		**174**	**77 017**	**443**	**29,62 %**	**3 850**	**3 850**	**9,0**	**9**	
Imereti	1.	299	14 385	48	41%		719	15,0	15	**11**
	2.	45	9 137	203	26%		457	2,3	2	**12**
	3.	24	11 617	484	33%		581	1,2	1	**13**
Sub-total for Imereti		**368**	**35 139**	**95**	**13,51 %**	**1 757**	**1 757**	**18,0**	**18**	
Kakheti	**1.**	**134**	**9 654**	**72**	**44%**	**483**	**483**	**6,7**	**7**	
	Georgian	121	9 351	77	97%		468	6,0	6	**14**
	Ethnic-Minority	13	303	23	3%		15	0,6	1	**15**
	2.	**42**	**8 250**	**196**	**37%**	**412**	**412**	**2,1**	**3**	
	Georgian	36	5 453	151	66%		273	1,8	2	**16**
	Ethnic-Minority	6	2 797	466	34%		140	0,3	1	**17**
	3	**9**	**4 178**	**464**	**19%**	**209**	**209**	0,4	**2**	
	Georgian	7	2 960	423	71%		148	0,3	1	**18**
	Ethnic-Minority	2	1 218	609	29%		61	0,1	1	**19**
Sub-total for Kakheti		**185**	**22 082**	**119**	**8,49 %**	**1 104**	**1 104**	**9,0**	**12**	
Mtskheta-Mtianeti	1.	76	3 121	41	56%		111	2,7	3	**20**
	2.	7	1 449	207	26%		52	0,2	1	**21**
	3	3	1 046	349	19%		37	0,1	1	**22**
Sub-total for Mtianeti		**86**	**5 616**	**65**	**2,16 %**	**281**	**200**	**5,0**	**5**	
Racha-Lechkhumi & Kvemo Svaneti	1.	66	1 416	21	89%		179	8,3	8	**23**
	2.	1	170	170	11%		21	0,1	1	**24**
	3				0%		0			**25**
Sub-total for Racha-Letchkhumi/Kv. Svaneti		**67**	**1 586**	**24**	**0,61 %**	**79**	**200**	**12,0**	**9**	
	1.	207	11 221	54	54%		561	10,3	10	**26**

[23] 0.4 % is the percentage of 1,044 students of 1-6 grade in this block from 260, 060 students of grade 1-6 in the country
[24] 52 is the 0.04% from 13,000, an estimated total number of Grade 1-6 students in all pilot schools

Samegrelo & Zemo Svaneti	2.	20	3 776	189	18%		189	1,0	1	**27**
	3.	13	5 625	433	27%		281	0,6	1	**28**
Sub-total for Samegrelo and Zemo Svaneti		**240**	**20 622**	**86**	**7,93 %**	**1 031**	**1 031**	**12,0**	**12**	
Samtskhe-Javakheti	**1.**	**188**	**8 778**	**47**	**68%**	**439**	**439**	**9,4**	**9**	
	Georgian	83	3 961	48	45%		198	4,2	4	**29**
	Ethnic-Minority	105	4 817	46	55%		241	5,3	5	**30**
	2.	**12**	**2 389**	**199**	**18%**	**119**	**300**	**1,5**	**2**	
	Georgian	5	912	182	38%		115	0,6	1	**31**
	Ethnic-Minority	7	1 477	211	62%		185	0,9	1	**32**
	3.	**4**	**1 789**	**447**	**14%**	**89**	**300**	**0,7**	**2**	
	Georgian	3	1 521	507	85%		255	0,5	1	**33**
	Ethnic-Minority	1	268	268	15%		45	0,2	1	**34**
Sub-total for Samtskhe-Javakheti		**204**	**12 956**	**64**	**4,98 %**	**648**	**648**	**10,0**	**13**	
Kvemo Kartli	**1**	**184**	**11 971**	**65**	**34%**	**598**	**599**	9,2	**9**	
	Georgian	41	4 769	116	40%		238	2,1	2	**34**
	Ethnic-Minority	143	7 202	50	60%		360	7,2	7	**36**
	2.	**43**	**9 747**	**227**	**28%**	**487**	**487**	2,2	**2**	
	Georgian	20	4 959	248	51%		248	1,0	1	**37**
	Ethnic-Minority	23	4 788	208	49%		239	1,2	1	**38**
	3.	**26**	**13 201**	**508**	**38%**	**660**	**660**	1,3	2	
	Georgian	19	10 254	540	78%		513	1,0	1	**39**
	Ethnic-Minority	7	2 947	421	22%		147	0,4	1	**40**
Sub-total for Kvemo Kartli		**253**	**34 919**	**138**	**13,43 %**	**1 746**	**1 746**	**13,0**	**13**	
Shida Kartli	1.	120	7 672	64	41%		383	6,0	6	**41**
	2.	35	7 192	205	39%		359	1,7	2	**42**
	3.	8	3 707	463	20%		185	0,4	1	**43**
Sub-total for Shida Kartli		**163**	**18 571**	**114**	**7,14 %**	**928**	**928**	**8,0**	**9**	
Total		**2 081**	**260 060**	**125**	**100%**	**13 000**	**13 188**	**115, 5**	**121**	

c) <u>Identifying the appropriate numbers of schools randomly from each block for the pilot intervention</u>

Using this theoretical framework, the G-PriEd and USAID has started working with the Ministry of Education and Science to identify the given number of schools randomly from the sub-groups as specified above. For illustrative purpose only, the process of school selection in Ajara is described: 9 small size (less than 300 students), 2 mid-size (between 300 and 600 students), and 1 large size (over 600 students) schools were identified randomly. In case of Samtskhe-Javakheti, where the language of instruction is Georgian as well as Armenian, an illustrative example could be the random selection of 9 small size schools, of which 4 had Georgian language of instruction, and 5- Armenian.

Once the sampling of the schools was put in practice, the numbers of students in the selected schools represented a change from the initial estimation. The real numbers of Grade 1-6 students in the 122 pilot schools is 19,070; and in 121 control schools- 18,007. The proportion of students per these schools did not change significantly. The total number of students per each pilot and control schools, and the number of samples from them is provided in the annexes 2 and 3 to the SOW.

d) Sampling a student from the selected schools

After schools were identified from each block, the proportionate approach was used to calculate the number of student samples per grade in each school. In each types of schools identified, the number of students to be included in the sample size was calculated as the following:

- For schools with less than 100 students- 1 student per grade= 6 students total
- For schools with 100- 200 students - 2 students per grade= 12 students total
- For schools with 200- 300 students - 3 students per grade= 18 students total
- For schools with 300- 400 students - 4 students per grade= 24 students total
- For schools with 400- 500 students between - 5 students per grade= 30 students total
- For schools with 500 and more students – 6 students per grade= 36 students total

Systematic random sampling was used within each school and each grade. When a school had more than one class of the same grade, the joint roster of all students was developed, from which the desired number of students was selected randomly. With systematic random sampling strategy, the **sampling interval** was first calculated by dividing the total number of students of grade 1-6 in particular school (411) by the appropriate number above (30 for a school with more than 400 and less than 500 students). Therefore, the sampling interval for this school was: 411/6/5=14. A random number between 1 and 14 (in this case 6) was selected. The first selected student in the sampling frame (students' list) is #6. Counting down the list, starting with student #6, each 14th student was selected, i.e. students #20, # 34, # 48, etc. The absent student would be replaced with the random one. The detailed instruction about students sampling procedures is described in General Administration Manual of the Impact Study.

I. Study population and Sample Size

By sampling the students this way, 1665 samples were considered in the pilot schools, and 1579 in the control schools. The total number of samples, therefore, was 3,244, with the study population of approximately 37,000 students of grade 1-6 that study in 244 schools of Georgia; of these students, 19,070 study in the 122 pilot schools, and 18,007 in the 121 control schools. To identify the statistically significant sample size, power analysis with the following framework was conducted:

1. The desired precision level of results was determined as 3%; i.e., no more than 3 percent of errors in the results could be attributable to the sampling error (or margin of error).
2. Determined the confidence level at 95%, or only once in 20 times, the sampling in the same population would have the sampling error higher than 3%.
3. Estimated the degree of Variability: the students within 43 blocks are expected to be homogenous, with the low degree of variance, in terms of their competence/skills in reading

Georgian and math; however, the students between the 43 blocks may differ in this regard, based on the region, the size of schools, language of instruction, etc. Therefore, the most conservative estimate, 50% variance was used, which estimated highly heterogeneous groups and the largest sample size.

4. Because the sampling was not simple random, but the Block Randomized, the coefficient of 1.5 was used to account for the multi-stage randomization.

For the power analysis, the following formula was used:

$$n = P(1-P) / [(A2 / Z2 + (P(1-P)/N)]$$

n= statistically significant sample size required
N = number of students in the study population = 37,000[25]
P = estimated variance in population, as a decimal: (0.5 for 50-50% variance)
A = Precision desired, expressed as a decimal (i.e., 0.03, for 3%)
Z = Based on confidence level: 1.6449 for 90 % confidence; 1.96 for 95% confidence; and 2, 5758 for 99 % of confidence

Therefore, n for the simple random sampling is 537 samples; n for the block random design =537 x 1.5 = 804. The impact evaluation sample size is 3,244.

Table 19. Study Population and Sample Size

Grade	# of students in pilot schools	# of students in control schools	Sample size in pilot schools	Sample size in control schools	Total # of students in pilot and control group
1	2,976	2,975	274	265	539
2	3,446	3,295	276	262	538
3	3,441	3,098	278	263	541
4	2,966	2,793	280	262	542
5	3,105	2,781	279	264	543
6	3,136	3,065	278	263	541
	19,070	18,007	1,665	1,579	3,244

[25] This includes students from the control and intervention schools

Table 20. Number of students tested at both baseline and endline (panel sample) by school type and grade

Grade	Panel students in control schools	Panel students in pilot schools	Total # of students in pilot and control group in the panel data
1	211	223	434
2	206	234	440
3	205	225	430
4	202	229	431
Total	824	911	1735

ANNEX II. TEST DESCRIPTION

Source: Annex 1_Study Design and Framework-Reading_Math_Final.docx, G-PriEd

Georgia's Diagnostic Assessment in Reading (GDA-R)
Georgian as a Native Language

The GDA-R methodology is aligned with the following requirements of Georgia's national curriculum. The requirements could be summarized as the following:

> Grade 1: Letter recognition and word segmentation is the focus; students can read "micro-texts". All letters, their sounds, and letter-reading and letter-writing are concurrently taught; the last month of the year is devoted to the reading of micro- sentences.
>
> Grade 2: Reading a connected text is the focus. Emphasis shifts to reading of connected texts, with the teachers having the students take turns reading aloud from their textbooks.
>
> Grades 3-6: Comprehension (of mostly narrative text) is the focus. Fluency and vocabulary are increasingly emphasized as students move up in grade level. Students continue to read aloud in class, but are also expected to read silently.

The GDA-R has a corresponding structure: reading skills are tested at each grade as appropriate; the test items are increasingly complex and intensive. Length of the sentences and words, as well as their complexity, is commensurate with the grade and age level of students. Table #1 illustrates the distribution of the reading skills tested and the increasing concentration of the items on each sub-test.

Table 21. Test Item Summary for GDA-R

Name	Grade I	Grade II	Grade III	Grade IV	Grade V	Grade VI
Phoneme Segmenting	20 items					
Syllable Segmenting	20 items					
Letter Sounds Fluency	65 items	65 items				
Word Reading Fluency	60 items	120 items				
Passage Reading Fluency		90 items	115 items	195 items	233 items	234 items
Vocabulary		12 items	15 items	20 items	20 items	20 items
Comprehension , narrative text			9 items	11 items	15 items	15 items
Comprehension , informational			6 items	6 items	7 items	7 items

Georgia's Diagnostic Assessment in Reading
Georgian as a second language
(GDA-R-GSL)

The national curriculum requirements for reading Georgian as a second language follows patterns similar to reading acquisition in Georgian; however, the intensity and complexity of words and texts progress at a much slower pace. Progress of grade-appropriate skills, as well as the complexity and intensity of the text and vocabulary at each grade level, is significantly lower than those for native speakers of the Georgian language.

> Grade 1: Letter recognition within a word: Letters, their sounds, and letter-reading and writing are taught concurrently. The emphasis is on building oral vocabulary.
> Grade 2: Letter recognition; Reading commonplace words: emphasis is on oral vocabulary building and listening to the text read by a teacher; starting reading "micro-texts"
> Grade 3: Reading "micro-texts"; reading texts that teacher had read and discussed previously
> Grades 4: Reading connected text, narrative; literal Comprehension; vocabulary of commonplace words
> Grades 5-6: Reading connected text, narrative and informational. Vocabulary gradually emphasized as students move up to grades 5 and 6. Both literal and inferential Comprehension questions asked.

In line with these requirements of the curriculum, the GDA-R-GSL displays the following pattern of distribution of competences to be tested. The complexity and intensity of words, sentences, and texts also follows the curriculum of Georgian as a second language.

Table 22. Test Item summary for GDA-R-GSL

Name	Grade I	Grade II	Grade III	Grade IV	Grade V	Grade VI
Phoneme Segmenting	20	20				
Syllable Segmenting	20	20				
Letter Sounds Fluency	65	65				
Word Reading Fluency	40	40	40	60		
Passage Reading Fluency		56	70	78	124	128
Vocabulary			10	10	15	15
Comp, narrative			5	7	8	8
Comprehension, informational			4	5	6	6

Georgia's Diagnostic Assessment in Math (GDA-M)

The national math curriculum encompasses all major math reasoning skills starting with first grade and increasing in complexity and intensity with each grade level. Skills such as data analysis, patterns, and the relationship between quantities are the new skills for Georgia's math curriculum; however, however, many Georgian teachers could benefit from professional development centered on how best to teach these skills.

Table 23. Test Item summary for GDA-M

	Reading Comprehension and math problem solving	Grade I	Grade II	Grade III	Grade IV	Grade V	Grade VI
Number of items	Counting	9	6				
	Number identification	8	6	6	8	10	8
	Comparing numbers	7	6	10	8	10	8
	Operations on numbers	8	6	9	8	9	8
	Algebra		6	8	8	9	8
	Patterns	8	6	8	8		
	Relations between quantities						8
	Geometric figures	8	6	8	8	6	9
	Area					10	8
	Data analysis		6	7	8	10	7

Assessment of Test Item Quality

In addition to the psychometric analysis, our Georgian math and reading experts also conducted a thorough content analysis of the tests in order to understand the objective of each item of the test in relation to the national curriculum. While the psychometric analysis of the test showed that most items are valid from a psychometric standpoint, the content analysis revealed a few issues with the content of some of the items. First, we note that the G-PriEd tests were developed as rapid diagnostic tools that follow the national curriculum loosely. For instance, the national math curriculum targets more than 20 indicators (competences) while the G-PriEd math test includes only 10; also, the national reading curriculum does not include a standard for passage reading fluency in 4th grade while the G-PriEd 4th grade reading test does include a measure of passage reading fluency[26]. Therefore, the G-Pried tests may not be completely exhaustive in testing students against the national curriculum. Second, as mentioned, some test items are problematic. In particular, for math, we found that some items were categorized incorrectly. The table below provides the item numbers that were misclassified along with their associated original and corrected content categories.

[26] We understand that this was of interest to G-PriEd from a research standpoint and that the Ministry of Education agreed with it.

Table 24. Misclassified Math Test Items

Grade	Item # (Form #)	Original Content Category	Corrected Content Category
Grade 1	#11 (F1)	Operations	Comparing Numbers
Grade 3	#15 (F1)	Operations on Numbers	Comparing Numbers
Grade 3	#3 (F1)	Patterns	Data Analysis
Grade 4	#1 (F1), #3 and #23 (F2)	Comparing Numbers	Number Identification
Grade 4	#4, #8 and #18 (F1) #8, #19, #20, #25 (F2)	Patterns	Algebra

For reading, it was not clear what some items were intended to measure. This was the case for the following items.

Table 25. Misclassified Reading Items

Grade	Item # (Form #)	Comment
Grade 5	Vocabulary: #16 (F1), #11, #12, #13, #14 (F2)	These items do not seem to measure vocabulary but rather knowledge of facts from other disciplines.

While we have kept all of these items given that they weren't problematic from a psychometric standpoint, we have re-classified the math items into the correct content category for the analysis, and taken into account the objective of each item in our description of proficiency levels for the standards (Section D3). Finally, we also note that more than 70% of the students gave a correct answer to about 75% of the items in Grade 1 and about 60% of the items in Grade 2, indicating that these tests may be too easy overall for those grades.

ANNEX III. RASCH ANALYSIS

The G-Pried tests for Mathematics and Reading have been designed in a way that makes it somewhat challenging to scale the tests. For each grade, two different forms of the test were developed (Form I and Form II). We call "forms", tests that were constructed for a specific grade but that are not composed of the same items. Since each form contains items specific to that form, a simple summary score (e.g. sum of all correct answers) could be biased by the level of dificulty of the items contaiend within the form. This makes it difficult to compare scores between different forms as it is not possible to determine whether the difference in scores stems from a real difference in the performance of students or from a difference in the level of difficulty in the forms. In other words, we cannot be sure that a student who was tested with Form I would receive the same score if he had been tested with Form II of the test. Therefore, pooling together the results of students from the two forms could be problematic. Thus, it is necessary to use a model for scaling the tests that can align the two forms on the same scale. For this reason, we used a model called a Rasch model for this scaling exercise. Note that it is only the two forms corresponding to a specific grade that are aligned on the same scale and not tests from different grades. No direct comparison should be made between grades using scores of tests from different grades.

The Rasch model is part of a family of models called Item Response Theory (IRT) or latent trait models. These models link the probability of a student giving a correct answer on a specific item to the characteristics of the students and the item. In an IRT model, the student parameter that is taken into account is his/her ability in the cognitive domain of interest. For example, if a test is designed to measure mathematics achievement, the student's ability level in mathematics is the parameter that would influence his response on any mathematics item. The item parameter of interest is the level of difficulty of the item. If an item really measures ability level in mathematics, only its difficulty can influence the probability that a student gives a correct answer. Therefore, in Rasch analysis, the probability of a student giving a correct answer on a given item is considered to be dependent on the level of difficulty of the items relative to the level of ability of the student. Thus, the model considers that a test measures a given ability on a continuum, ranging from a low level of ability to a high level of ability. The ability of the students and the difficulty of the items are all put on this scale. And all items that do not fit the model (based on fit statistics) are removed since those items are viewed as of "low quality".

More specifically, the Rasch model represents the simplest mathematical representation of the link between student and item characteristics. This model represents the probability that a student gives a correct answer on an item as :

$$probabylity(1|\theta) = \frac{e^{(\theta - D)}}{1 + e^{(\theta - D)}}$$

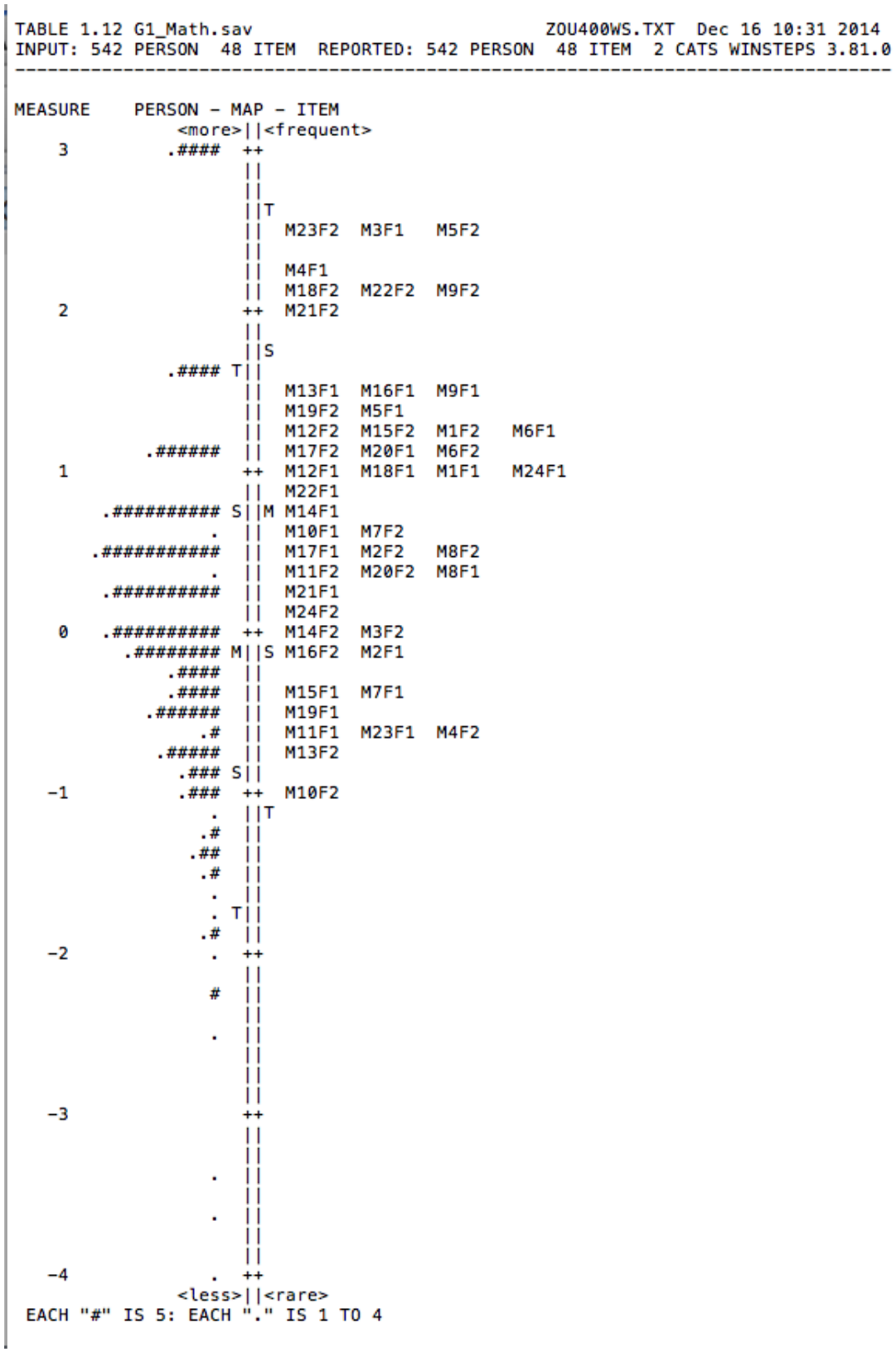
Where θ is the level of ability of the student and D is the level of difficulty of the item. It must be noted that those two parameters are on the same scale, called a logit scale. From this formula, we could state that if the level of ability of a student is greater than the level of difficulty of an item (θ > D) then the most probable outcome is a correct answer while if θ < D, the most probable outcome is an incorrect answer.

The estimation of the model leads to the production of estimates of ability level for each student and estimates of difficulty level for each item. Since those two parameters are on the same scale, it is possible to represent them on a common figure called the Distribution map (see figure 1). In this map, the first column named "Measure" represents the scale of measurement on a normal score scale (mean of 0 and SD of 1). This scale doesn't have an absolute value; we need to fix its mean value 0 to a particular value. In this case, we have fixed the 0 value to the mean ability level of the students. The second column called "Person" represents the distribution of ability level of the students on the measurement scale. The lowest students on this scale (around -3 or -4) represent the less competent students while the highest (around +3) represent the most competent students.

Finally, the last column, called "Item" represents the distribution of difficulty level of the items on the measurement scale. The lowest items (around -1) represent the easiest items while the highest items (around +2.5) represent the most difficult items. From the map, we can see that the less competent students have a low probability of giving a correct answer to the easiest items while the most competent students have a high probability of giving a correct answer to all items in the test. Thus, given the level of ability of a student, we are able to know which items are likely to be answered correctly and which incorrectly.

The Distribution map gives a first idea of the quality of the items that a test is composed of. To be useful, an item must not be too difficult, at least the student with the highest level of ability must be able to give a good answer while the items must also not be too easy, it must represent a challenge even to the student with the lowest level of ability. Items that are too easy or too difficult are not useful items since all or none of the students are able to answer these items correctly. In other words, these items are not able to discriminate the level of ability of the students. In the first Distribution map, no items are identified as too easy or too difficult.

Figure 2 : Distribution Map



```
TABLE 1.12 G1_Math.sav                          ZOU400WS.TXT  Dec 16 10:31 2014
INPUT: 542 PERSON  48 ITEM  REPORTED: 542 PERSON  48 ITEM  2 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>||<frequent>
    3          .####  ++
                      ||
                      ||
                      ||T
                      ||  M23F2  M3F1   M5F2
                      ||
                      ||  M4F1
                      ||  M18F2  M22F2  M9F2
    2                 ++  M21F2
                      ||
                      ||S
              .#### T||
                      ||  M13F1  M16F1  M9F1
                      ||  M19F2  M5F1
                      ||  M12F2  M15F2  M1F2   M6F1
           .######   ||  M17F2  M20F1  M6F2
    1                 ++  M12F1  M18F1  M1F1   M24F1
                      ||  M22F1
        .########## S||M M14F1
                  .   ||  M10F1  M7F2
       .###########  ||  M17F1  M2F2   M8F2
                  .   ||  M11F2  M20F2  M8F1
        .##########  ||  M21F1
                      ||  M24F2
    0    .##########  ++  M14F2  M3F2
          .######## M||S M16F2  M2F1
              .####   ||
              .####   ||  M15F1  M7F1
            .######   ||  M19F1
                 .#   ||  M11F1  M23F1  M4F2
             .#####   ||  M13F2
               .### S||
   -1           .###  ++  M10F2
                  .   ||T
                 .#   ||
                .##   ||
                 .#   ||
                  .   ||
                  . T||
                 .#   ||
   -2             .   ++
                      ||
                     #  ||
                        ||
                  .     ||
                        ||
                        ||
                        ||
   -3                   ++
                        ||
                        ||
                  .     ||
                        ||
                  .     ||
                        ||
                        ||
   -4                .  ++
              <less>||<rare>
 EACH "#" IS 5: EACH "." IS 1 TO 4
```

Other indices of item goodness-of-fit are important in order to verify the quality of the items included in a test. The model states that most students should give a correct answer to the easiest items and that only the students with the highest level of ablity should be able to answer the most difficult items correctly. But real data do not always respect this model. It is possible that a student with a high level of ability gives an incorrect answer to an easy item or that a student with a low level of ability gives a correct answer to a more difficult item. The model is probabilistic and this kind of phenomenon could happen, but if it's too frequent in the data, this could signify that some items are not of good quality. Two indices are used to detect these types of faulty items; those indices are called Infit meansquare and Outfit meansquare.

These two indices are based on the residual values (differences between the observation and the expected values according to the Rasch model). The Outfit is based on the sum of squared standardized residuals. Standardized residuals are modeled to represent a normal distribution while their sums of squares approximate a chi-square distribution. Dividing the Outfit by its degree of freedom will produce the Outfit meansquare with an expected mean of 1 and range from 0 to infinity. A value of 1.0 represents perfect fit while a value that departs too much for 1.0 represents problematic items. The Infit is an information-weighted form of Outfit. The weighting reduces the influence of low variance or off-target response. The interpretation of the Infit is the same as the Outfit. An Infit meansqare or Outfit meansquer value greater than 1.6 or lower than 0.4 represents problematic items that should be removed from the test. Distribution maps and goodness-of-fit statistics of items will be presented in the next section.

One of the main features of the Rasch model is that it is not necessary that a student gives an answer to all items or that an item is administered to all students to produce reliable estimates of student or item parameters. With this feature, we can put all items from the two forms of a test in a given grade level on the same scale and produce a unique scale of scores for all students in a given grade level.

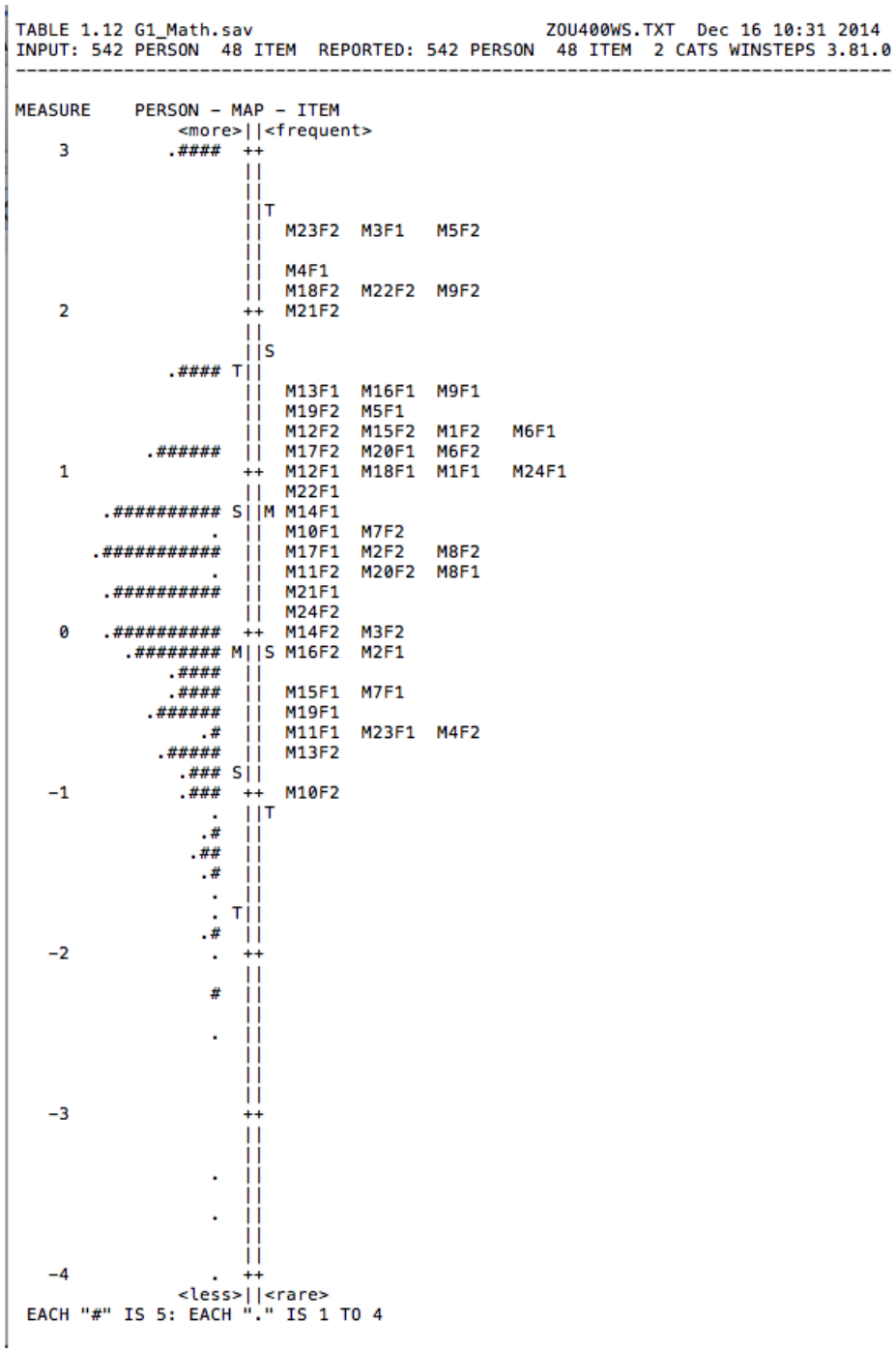
Results from Rasch analysis

For every Distribution map or table with Infit meansquare and Outfit meansquare, a specific coding system was used to identify test items. For mathematics, all items begin with the letter M, the number that follows represents the number of the item in the test and F1 or F2 represents weather the item is in either Form 1 or Form 2 of the test. Thus, item M20F1 represents math item number 20 in Form 1 of the test.

For reading, the coding is a little bit different. Items that begin with the letter R represent one of the first 4 tasks (R1=Phoneme segmenting, R2=Syllable segmenting, R3=Letter sound fluency and R4=Word reading fluency). Letters VOC represent a vocabulary item, NAR a Comprehension question for the narrative text and INF a Comprehension question for the informational text. For those three letter codes, the number represents the position of the items in a specific task; VOC3 is the third word of the vocabulary task. As for mathematics, F1 or F2 identifies whether the items appear on Form 1 or Form 2 of the test. For reading in Georgian as second language, there is only one form; therefore F1/F2 does not appear in the names of the items.

Distribution maps for mathematics tests

Figure 3. Distribution maps for mathematics items – Grade 1



```
TABLE 1.12 G1_Math.sav                          ZOU400WS.TXT  Dec 16 10:31 2014
INPUT: 542 PERSON  48 ITEM  REPORTED: 542 PERSON  48 ITEM  2 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>||<frequent>
    3         .####  ++
                     ||
                     ||
                     ||T
                     ||  M23F2  M3F1   M5F2
                     ||
                     ||  M4F1
                     ||  M18F2  M22F2  M9F2
    2                ++  M21F2
                     ||
                     ||S
              .#### T||
                     ||  M13F1  M16F1  M9F1
                     ||  M19F2  M5F1
                     ||  M12F2  M15F2  M1F2   M6F1
           .######  ||  M17F2  M20F1  M6F2
    1                ++  M12F1  M18F1  M1F1   M24F1
                     ||  M22F1
        .########## S||M M14F1
                  .  ||  M10F1  M7F2
       .###########  ||  M17F1  M2F2   M8F2
                  .  ||  M11F2  M20F2  M8F1
        .##########  ||  M21F1
                     ||  M24F2
    0   .##########  ++  M14F2  M3F2
          .######### M||S M16F2  M2F1
              .####  ||
              .####  ||  M15F1  M7F1
             .######  ||  M19F1
                  .#  ||  M11F1  M23F1  M4F2
              .#####  ||  M13F2
                .### S||
   -1           .###  ++  M10F2
                  .  ||T
                 .#  ||
                .##  ||
                 .#  ||
                  .  ||
                  . T||
                 .#  ||
   -2             .  ++
                     ||
                   #  ||
                      ||
                   .  ||
                      ||
                      ||
                      ||
   -3                 ++
                      ||
                      ||
                   .  ||
                      ||
                   .  ||
                      ||
                      ||
   -4                 ++
              <less>||<rare>
 EACH "#" IS 5: EACH "." IS 1 TO 4
```

Figure 4. Distribution maps for mathematics items – Grade 2

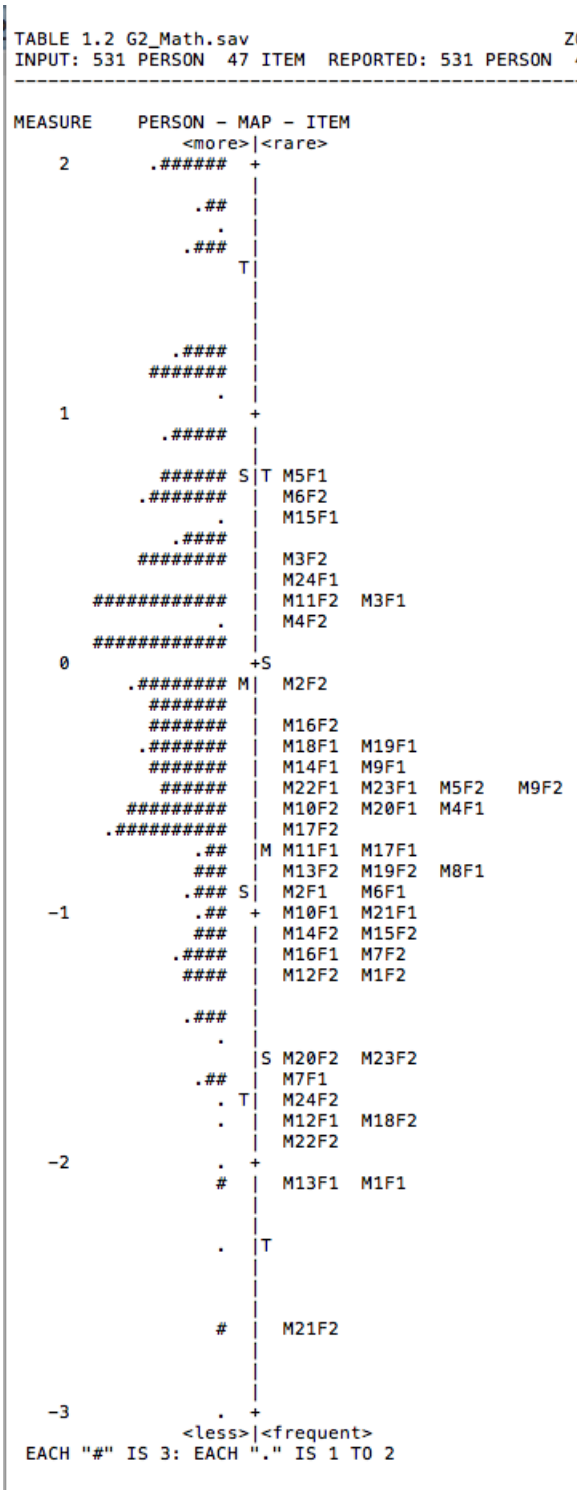
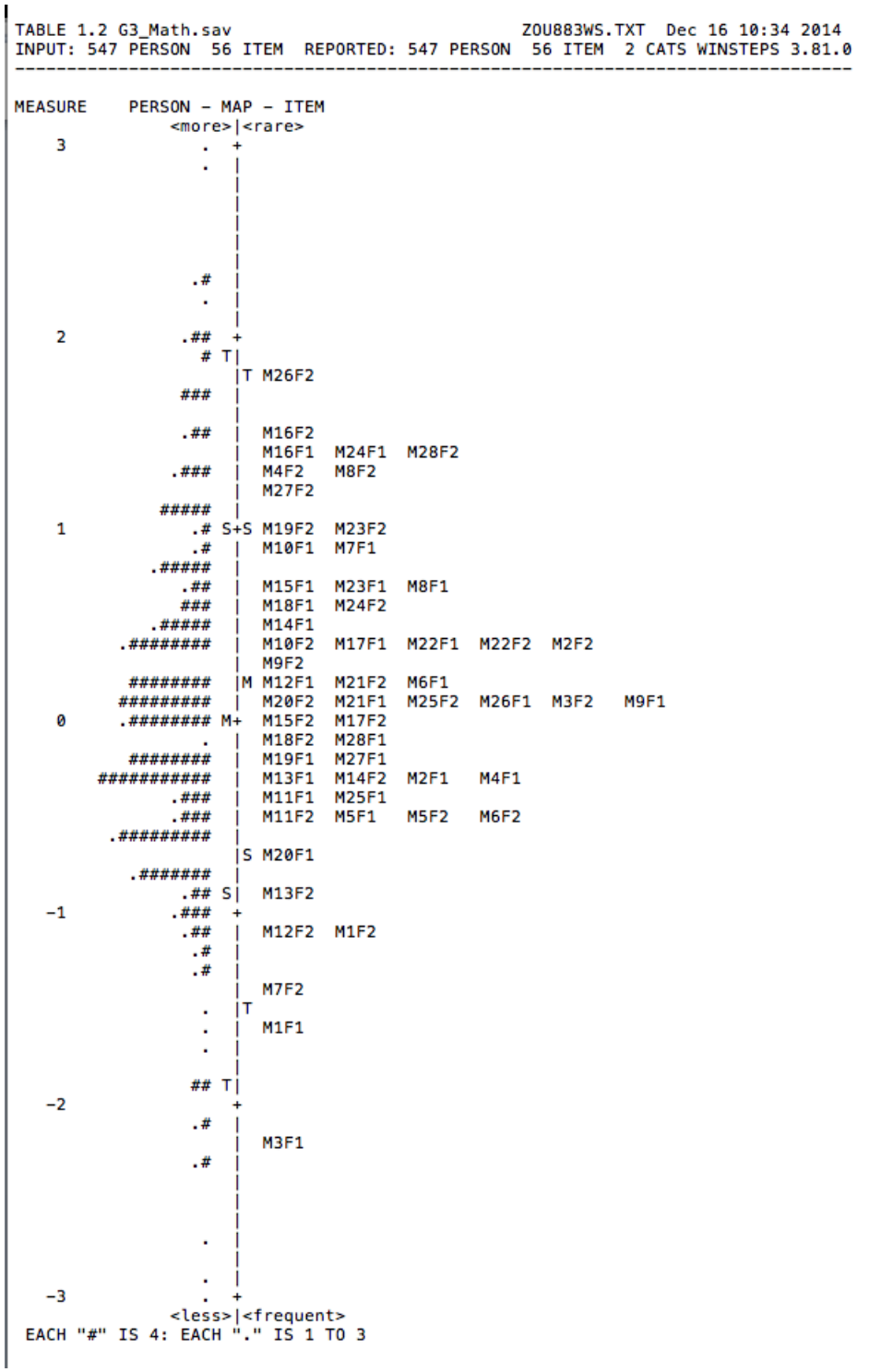


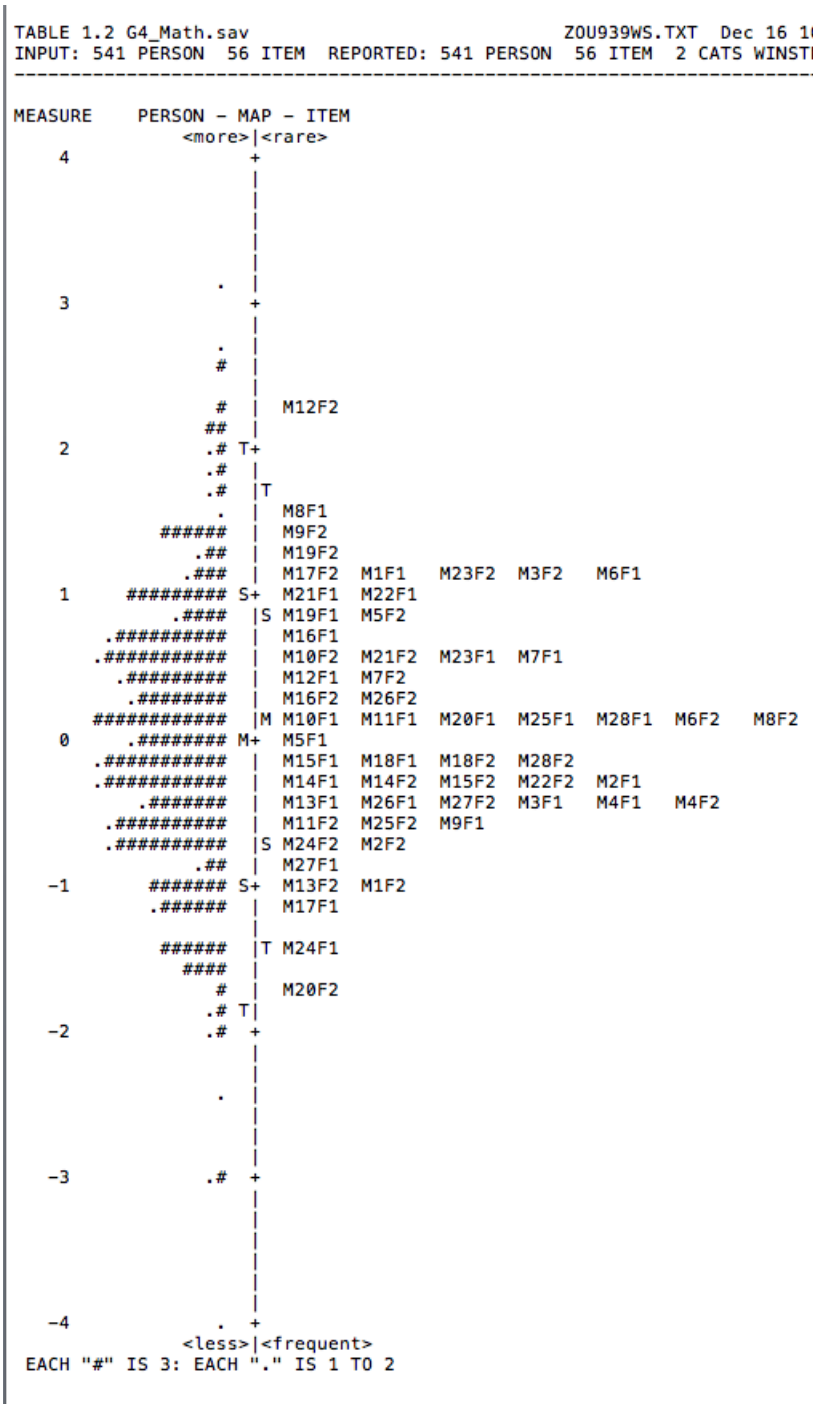
Figure 5. Distribution maps for mathematics items – Grade 3



```
TABLE 1.2 G3_Math.sav                          ZOU883WS.TXT  Dec 16 10:34 2014
INPUT: 547 PERSON  56 ITEM  REPORTED: 547 PERSON  56 ITEM  2 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
    3            .  +
                 .  |
                    |
                    |
                    |
                    |
                    |
                .#  |
                 .  |
                    |
    2           .##  +
                 # T|
                    |T M26F2
                ###  |
                    |
                .##  |  M16F2
                    |  M16F1  M24F1  M28F2
               .###  |  M4F2   M8F2
                    |  M27F2
              #####  |
    1            .# S+S M19F2  M23F2
                 .#  |  M10F1  M7F1
             .#####  |
                .##  |  M15F1  M23F1  M8F1
                ###  |  M18F1  M24F2
             .#####  |  M14F1
          .#########  |  M10F2  M17F1  M22F1  M22F2  M2F2
                    |  M9F2
           #########  |M M12F1  M21F2  M6F1
          ##########  |  M20F2  M21F1  M25F2  M26F1  M3F2   M9F1
    0     .######### M+  M15F2  M17F2
                  .  |  M18F2  M28F1
           #########  |  M19F1  M27F1
         ############  |  M13F1  M14F2  M2F1   M4F1
                .###  |  M11F1  M25F1
                .###  |  M11F2  M5F1   M5F2   M6F2
          .##########  |
                    |S M20F1
           .########  |
                .## S|  M13F2
   -1           .###  +
                 .##  |  M12F2  M1F2
                  .#  |
                  .#  |
                    |  M7F2
                   .  |T
                   .  |  M1F1
                   .  |
                    |
                  ## T|
   -2                +
                 .#  |
                    |  M3F1
                 .#  |
                    |
                    |
                    |
                   .  |
                    |
                   .  |
   -3              .  +
              <less>|<frequent>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

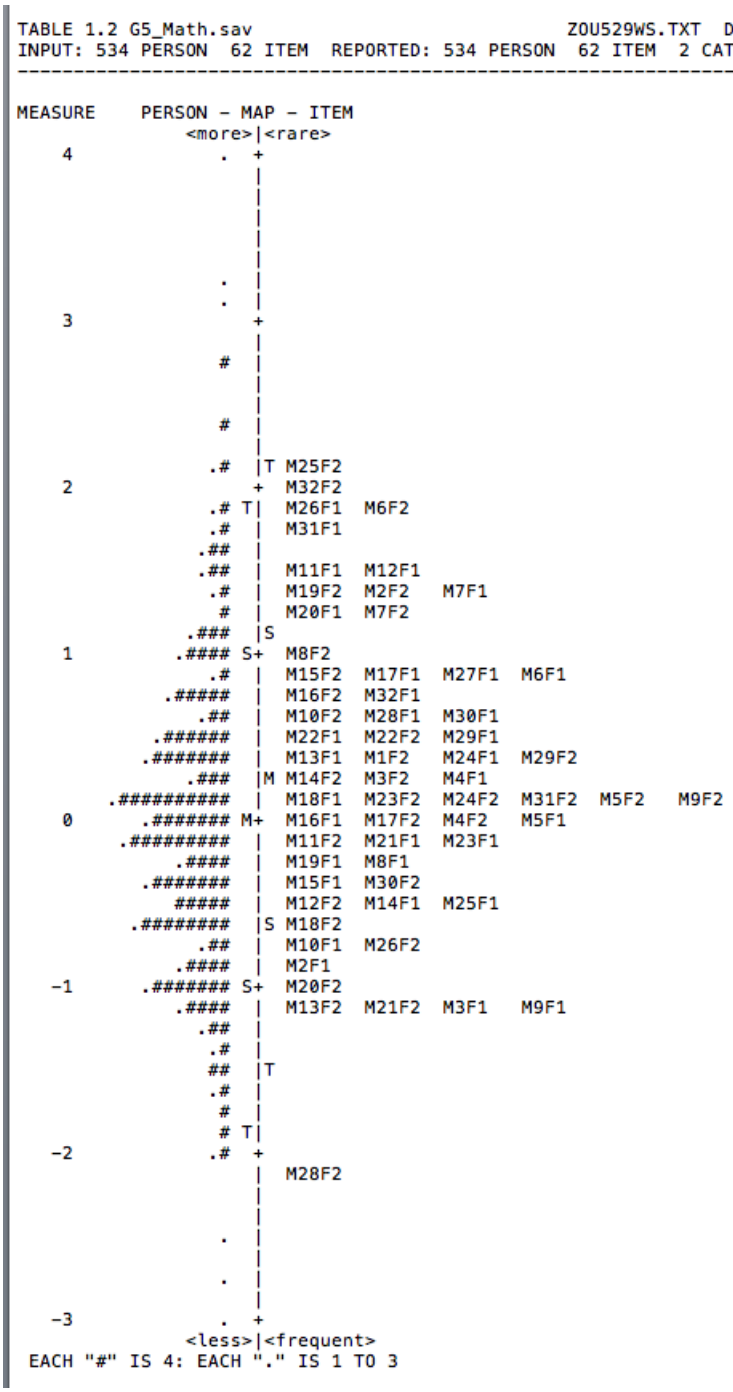
Figure 6. Distribution maps for mathematics items – Grade 4



```
TABLE 1.2 G4_Math.sav                          ZOU939WS.TXT  Dec 16 10:34 2014
INPUT: 541 PERSON  56 ITEM  REPORTED: 541 PERSON  56 ITEM  2 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
    4               +
                    |
                    |
                    |
                    |
                    |
                  . |
    3               +
                    |
                  . |
                  # |
                    |
                  # |  M12F2
                 ## |
    2             .# T+
                  .# |
                  .# |T
                   . |  M8F1
              ######  |  M9F2
                 .##  |  M19F2
                .###  |  M17F2  M1F1   M23F2  M3F2   M6F1
    1     ######### S+  M21F1  M22F1
              .####   |S M19F1  M5F2
         .##########  |  M16F1
       .############  |  M10F2  M21F2  M23F1  M7F1
         .##########  |  M12F1  M7F2
        .#########   |  M16F2  M26F2
       ############  |M M10F1  M11F1  M20F1  M25F1  M28F1  M6F2   M8F2
    0     .######## M+  M5F1
       .############  |  M15F1  M18F1  M18F2  M28F2
       .############  |  M14F1  M14F2  M15F2  M22F2  M2F1
           .########  |  M13F1  M26F1  M27F2  M3F1   M4F1   M4F2
        .##########   |  M11F2  M25F2  M9F1
        .##########   |S M24F2  M2F2
              .##     |  M27F1
   -1       ####### S+  M13F2  M1F2
            .######   |  M17F1
                      |
             ######   |T M24F1
               ####   |
                  #   |  M20F2
                 .# T|
   -2            .#  +
                     |
                     |
                   . |
                     |
                     |
                     |
   -3            .#  +
                     |
                     |
                     |
                     |
                     |
                     |
   -4              .  +
              <less>|<frequent>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

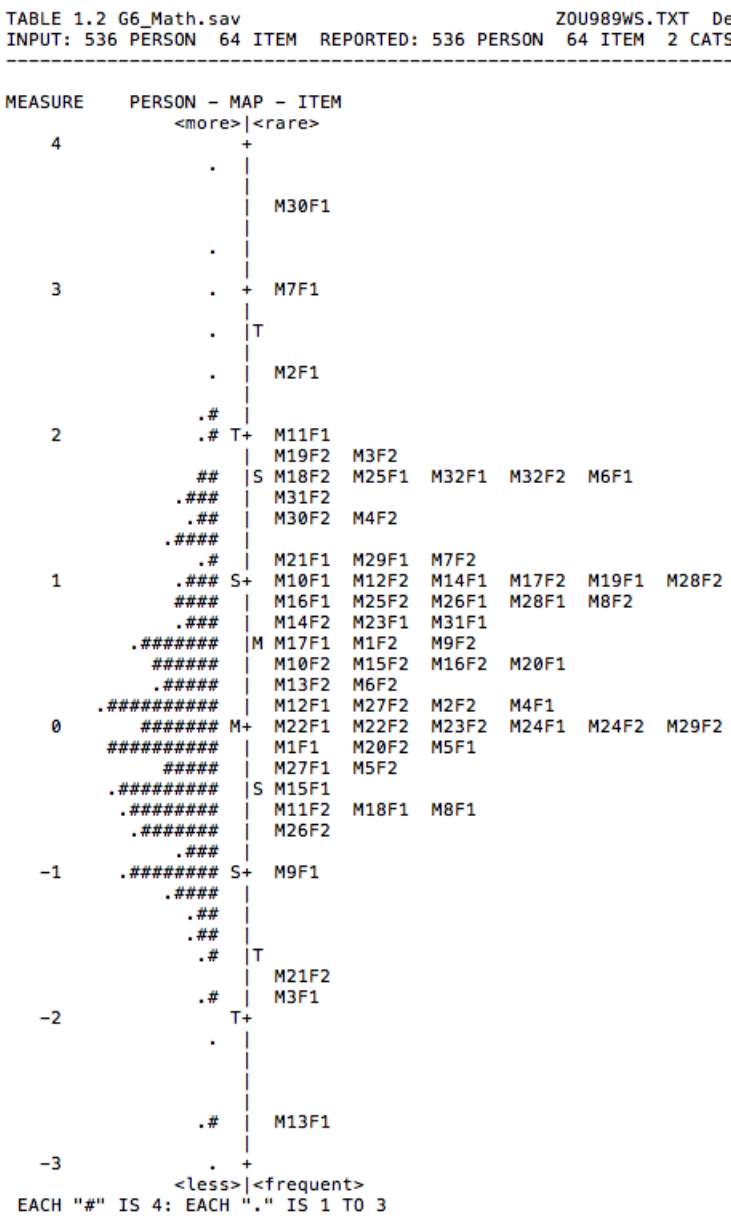
Figure 7. Distribution maps for mathematics items – Grade 5



```
TABLE 1.2 G5_Math.sav                          ZOU529WS.TXT  Dec 16 10:35 2014
INPUT: 534 PERSON  62 ITEM  REPORTED: 534 PERSON  62 ITEM  2 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
             <more>|<rare>
    4            .  +
                    |
                    |
                    |
                    |
                    |
                 .  |
                 .  |
    3               +
                    |
                 #  |
                    |
                    |
                 #  |
                    |
                .#  |T M25F2
    2               +  M32F2
                .# T|  M26F1  M6F2
                .#  |  M31F1
                .##  |
                .##  |  M11F1  M12F1
                 .#  |  M19F2  M2F2   M7F1
                  #  |  M20F1  M7F2
                .###  |S
    1          .#### S+  M8F2
                  .#  |  M15F2  M17F1  M27F1  M6F1
              .#####  |  M16F2  M32F1
                 .##  |  M10F2  M28F1  M30F1
             .######  |  M22F1  M22F2  M29F1
            .########  |  M13F1  M1F2   M24F1  M29F2
                .###  |M M14F2  M3F2   M4F1
         .##########  |  M18F1  M23F2  M24F2  M31F2  M5F2   M9F2
    0      .######## M+  M16F1  M17F2  M4F2   M5F1
          .##########  |  M11F2  M21F1  M23F1
                .####  |  M19F1  M8F1
            .#######  |  M15F1  M30F2
               #####  |  M12F2  M14F1  M25F1
           .#########  |S M18F2
                 .##  |  M10F1  M26F2
                .####  |  M2F1
   -1       .######## S+  M20F2
                .####  |  M13F2  M21F2  M3F1   M9F1
                 .##  |
                  .#  |
                  ##  |T
                  .#  |
                   #  |
                   # T|
   -2              .#  +
                       |  M28F2
                       |
                       |
                    .  |
                       |
                    .  |
                       |
   -3               .  +
                <less>|<frequent>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

Figure 8. Distribution maps for mathematics items – Grade 6



Distribution maps for reading Georgian tests

Figure 9. Distribution maps for reading Georgian items – Grade 1

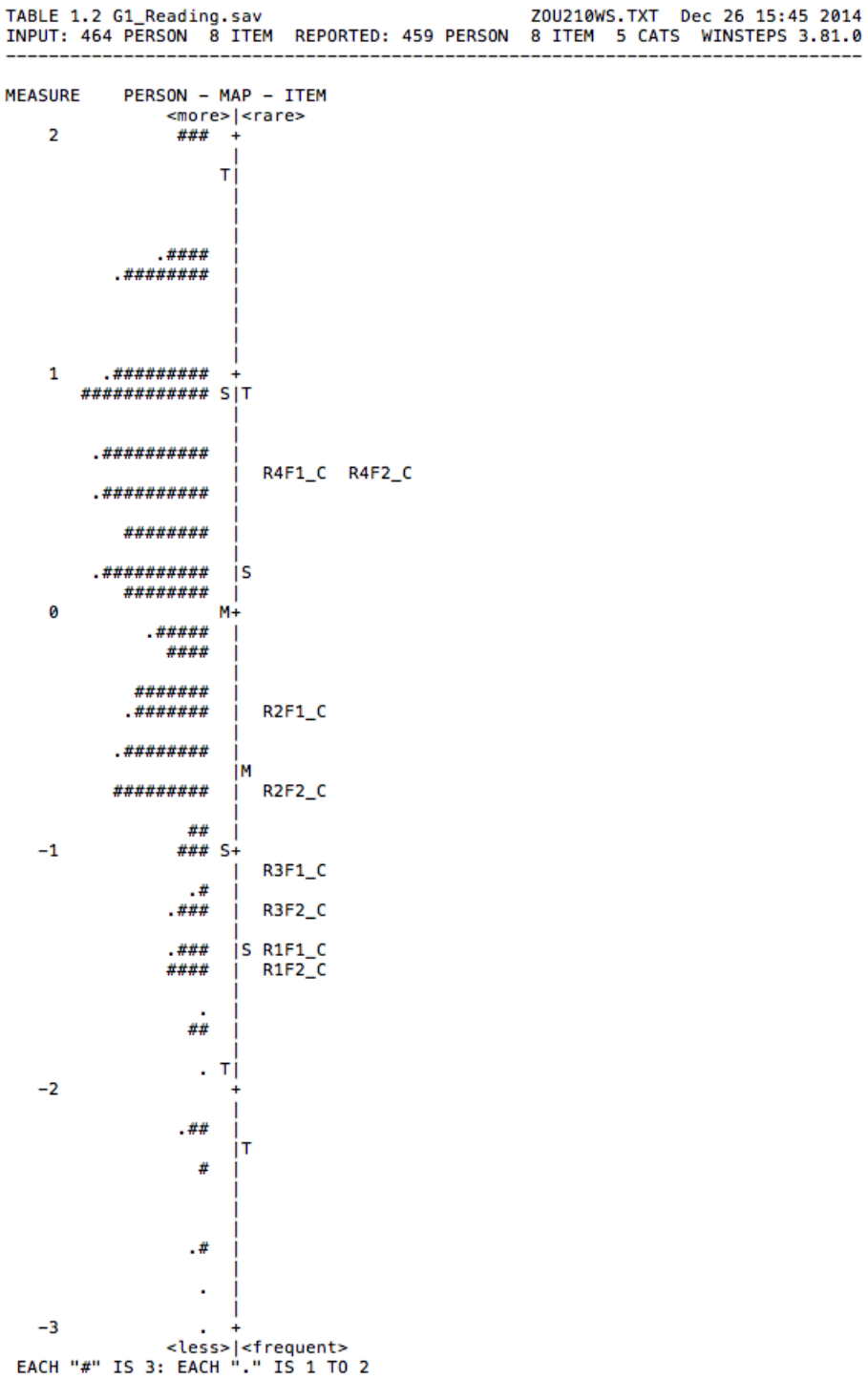


Figure 10. Distribution maps for reading Georgian items – Grade 2

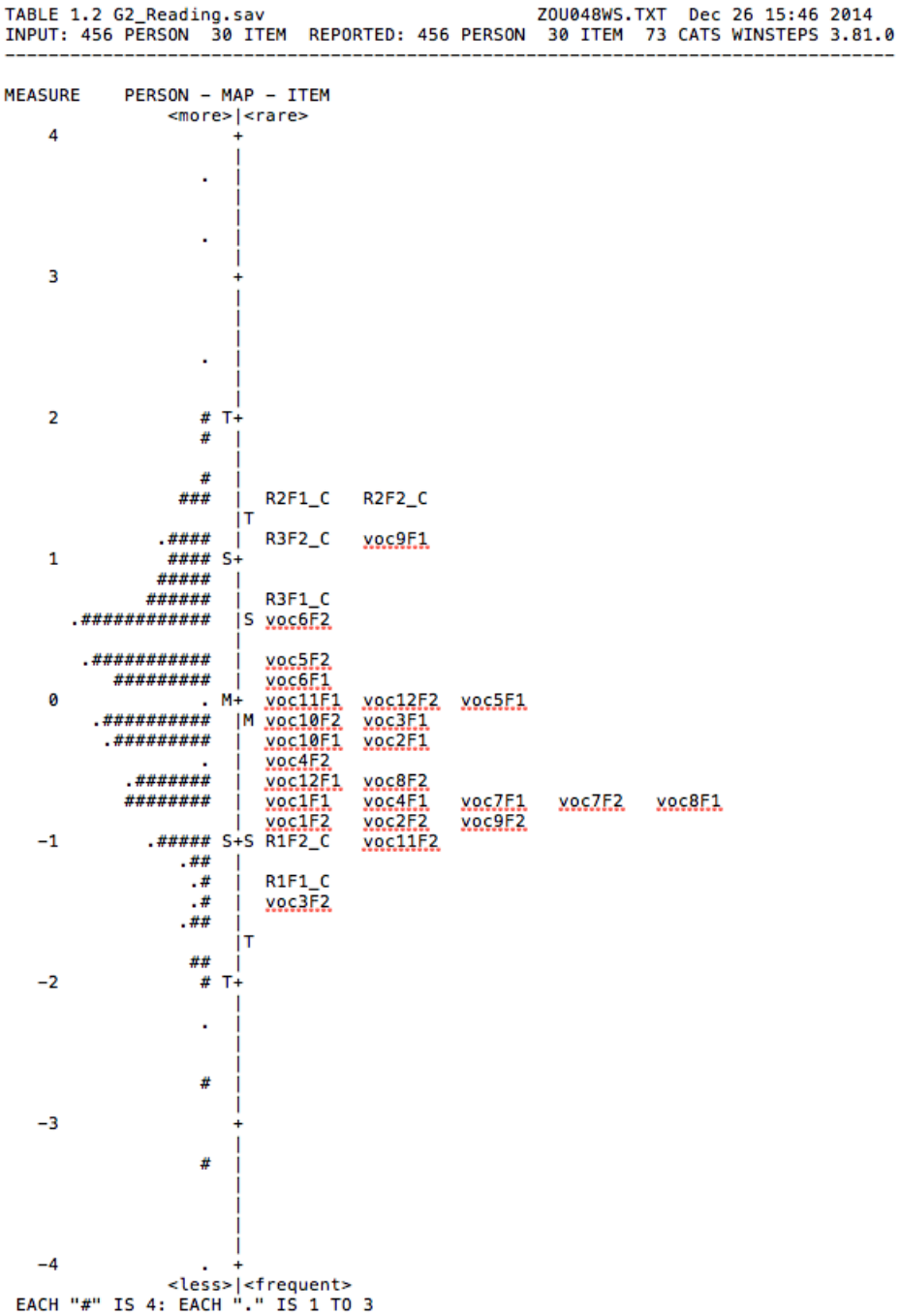
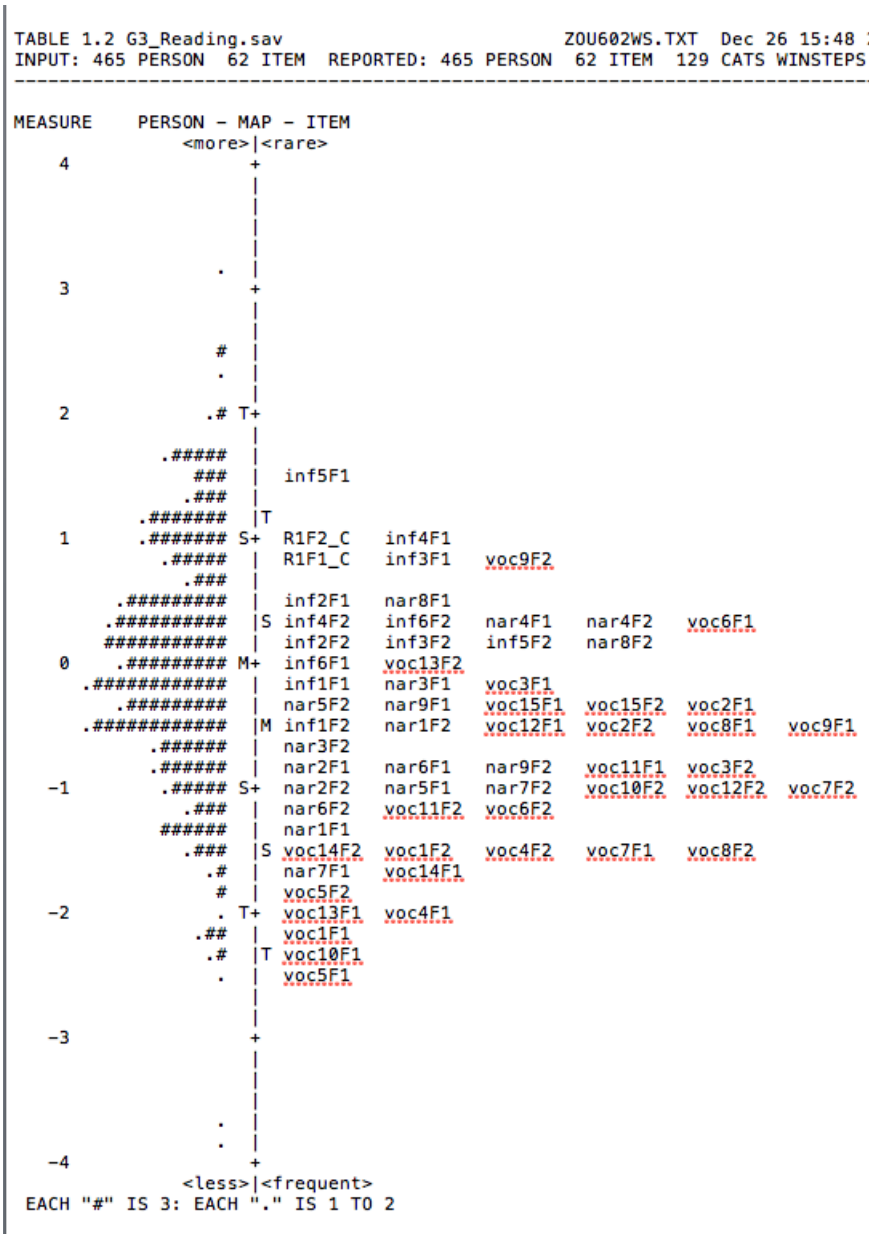


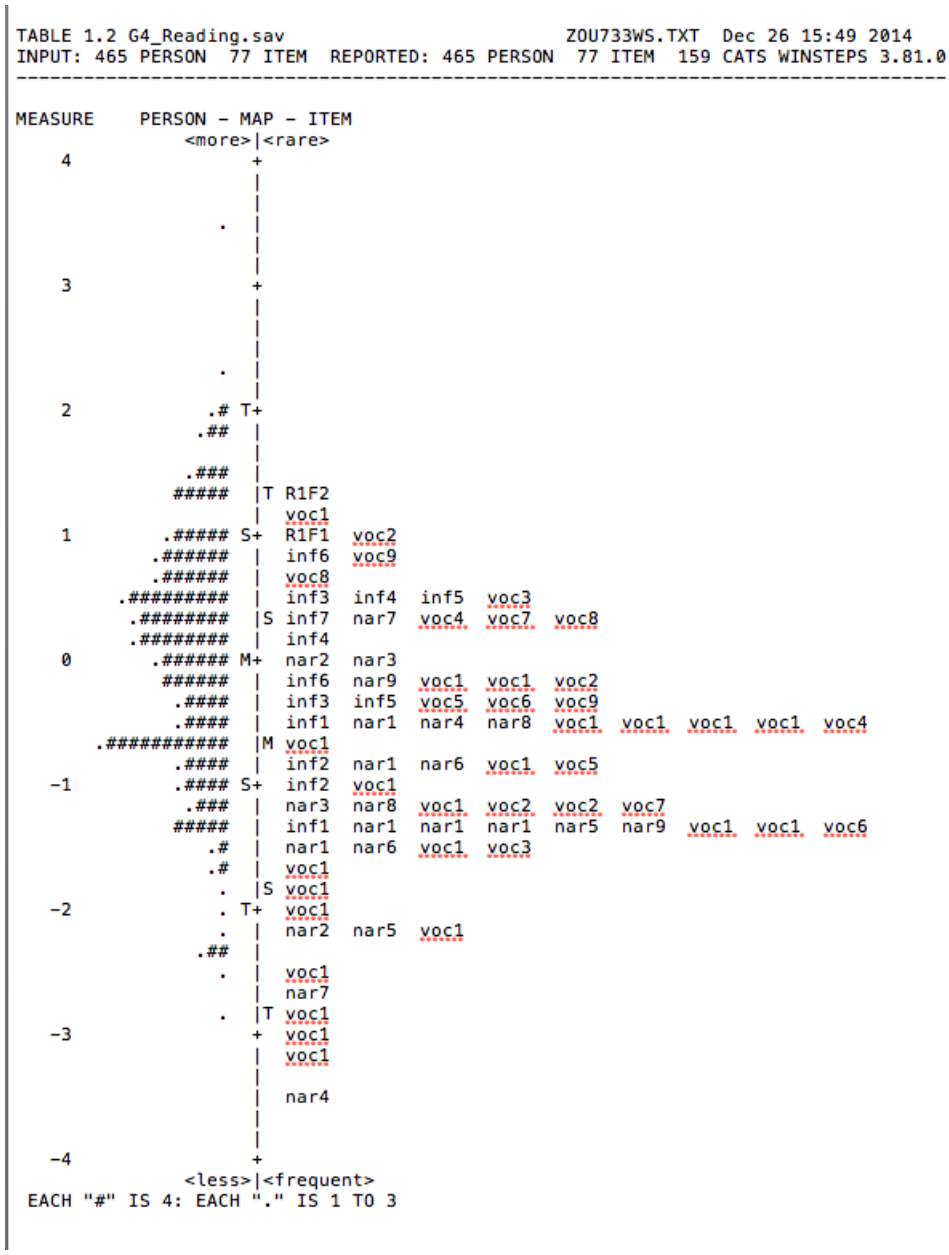
Figure 11. Distribution maps for reading Georgian items – Grade 3



```
TABLE 1.2 G3_Reading.sav                        ZOU602WS.TXT  Dec 26 15:48 2014
INPUT: 465 PERSON  62 ITEM  REPORTED: 465 PERSON  62 ITEM  129 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
    4               +
                    |
                    |
                    |
                    |
                .   |
    3               +
                    |
                    |
                #   |
                .   |
                    |
    2            .# T+
                    |
             .#####  |
                ###  |  inf5F1
               .###  |
           .########|T
    1      .######## S+  R1F2_C   inf4F1
             .#####  |  R1F1_C   inf3F1   voc9F2
               .###  |
         .##########  |  inf2F1   nar8F1
       .###########  |S inf4F2   inf6F2   nar4F1   nar4F2   voc6F1
       ############  |  inf2F2   inf3F2   inf5F2   nar8F2
    0    .########## M+  inf6F1   voc13F2
     .#############  |  inf1F1   nar3F1   voc3F1
        .##########  |  nar5F2   nar9F1   voc15F1  voc15F2  voc2F1
     .#############  |M inf1F2   nar1F2   voc12F1  voc2F2   voc8F1   voc9F1
           .#######  |  nar3F2
           .######  |  nar2F1   nar6F1   nar9F2   voc11F1  voc3F2
   -1       .##### S+  nar2F2   nar5F1   nar7F2   voc10F2  voc12F2  voc7F2
             .###  |  nar6F2   voc11F2  voc6F2
            ######  |  nar1F1
             .###  |S voc14F2  voc1F2   voc4F2   voc7F1   voc8F2
               .#  |  nar7F1   voc14F1
                #  |  voc5F2
   -2            . T+  voc13F1  voc4F1
              .##  |  voc1F1
              .#   |T voc10F1
               .   |  voc5F1
                   |
                   |
   -3              +
                   |
                   |
                   |
                .  |
                .  |
   -4              +
              <less>|<frequent>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

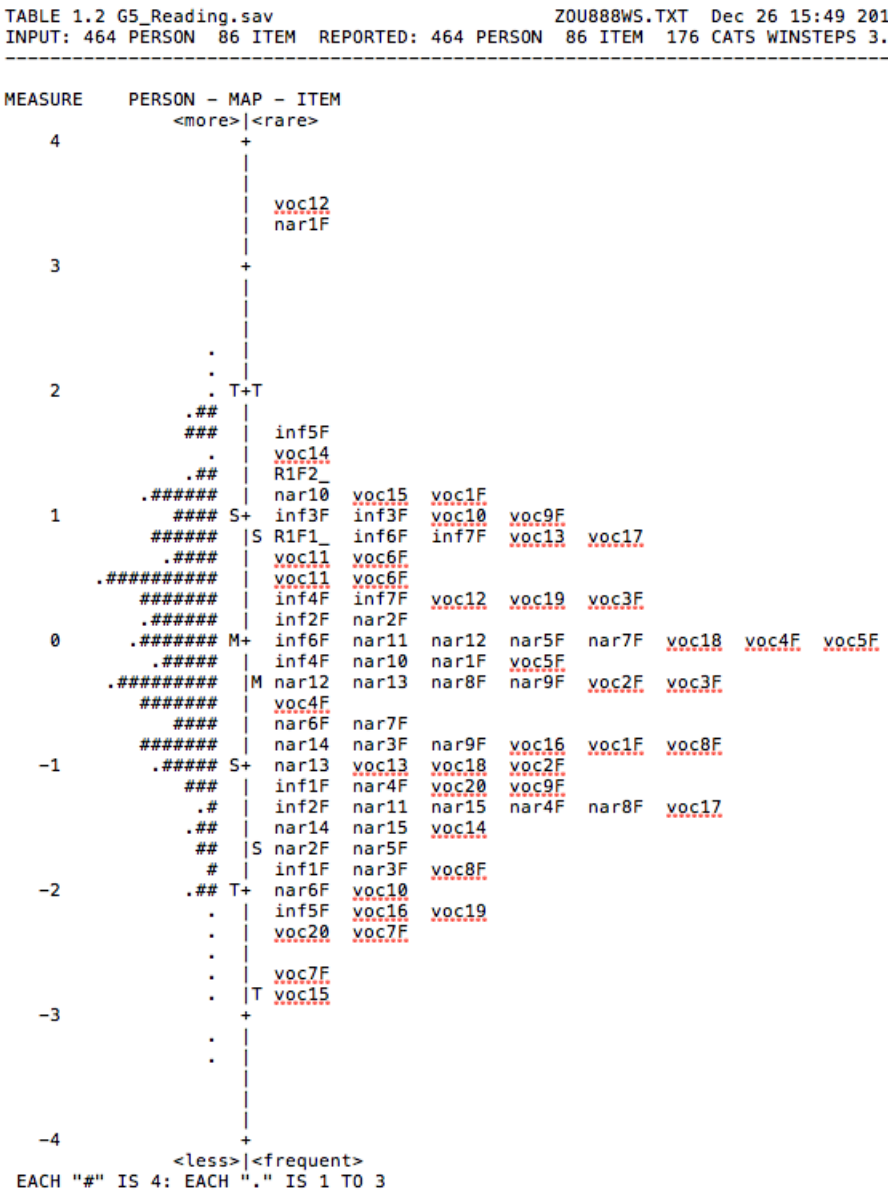
Figure 12. Distribution maps for reading Georgian items – Grade 4



```
TABLE 1.2 G4_Reading.sav                          ZOU733WS.TXT  Dec 26 15:49 2014
INPUT: 465 PERSON  77 ITEM  REPORTED: 465 PERSON  77 ITEM  159 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
    4               +
                    |
                    |
                  . |
                    |
                    |
    3               +
                    |
                    |
                    |
                  . |
                    |
    2             .# T+
                 .##  |
                      |
                .###  |
               #####  |T R1F2
                      |  voc1
    1         .##### S+  R1F1  voc2
             .######  |  inf6  voc9
             .######  |  voc8
           .#########  |  inf3  inf4  inf5  voc3
           .#########  |S inf7  nar7  voc4  voc7  voc8
           .########  |  inf4
    0        .###### M+  nar2  nar3
             ######  |  inf6  nar9  voc1  voc1  voc2
              .####  |  inf3  inf5  voc5  voc6  voc9
              .####  |  inf1  nar1  nar4  nar8  voc1  voc1  voc1  voc1  voc4
         .###########  |M voc1
              .####  |  inf2  nar1  nar6  voc1  voc5
   -1         .#### S+  inf2  voc1
               .###  |  nar3  nar8  voc1  voc2  voc2  voc7
              #####  |  inf1  nar1  nar1  nar1  nar5  nar9  voc1  voc1  voc6
                 .#  |  nar1  nar6  voc1  voc3
                 .#  |  voc1
                  .  |S voc1
   -2             . T+  voc1
                  .  |  nar2  nar5  voc1
                .##  |
                  .  |  voc1
                     |  nar7
                  .  |T voc1
   -3                +  voc1
                     |  voc1
                     |
                     |  nar4
                     |
                     |
   -4                +
              <less>|<frequent>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

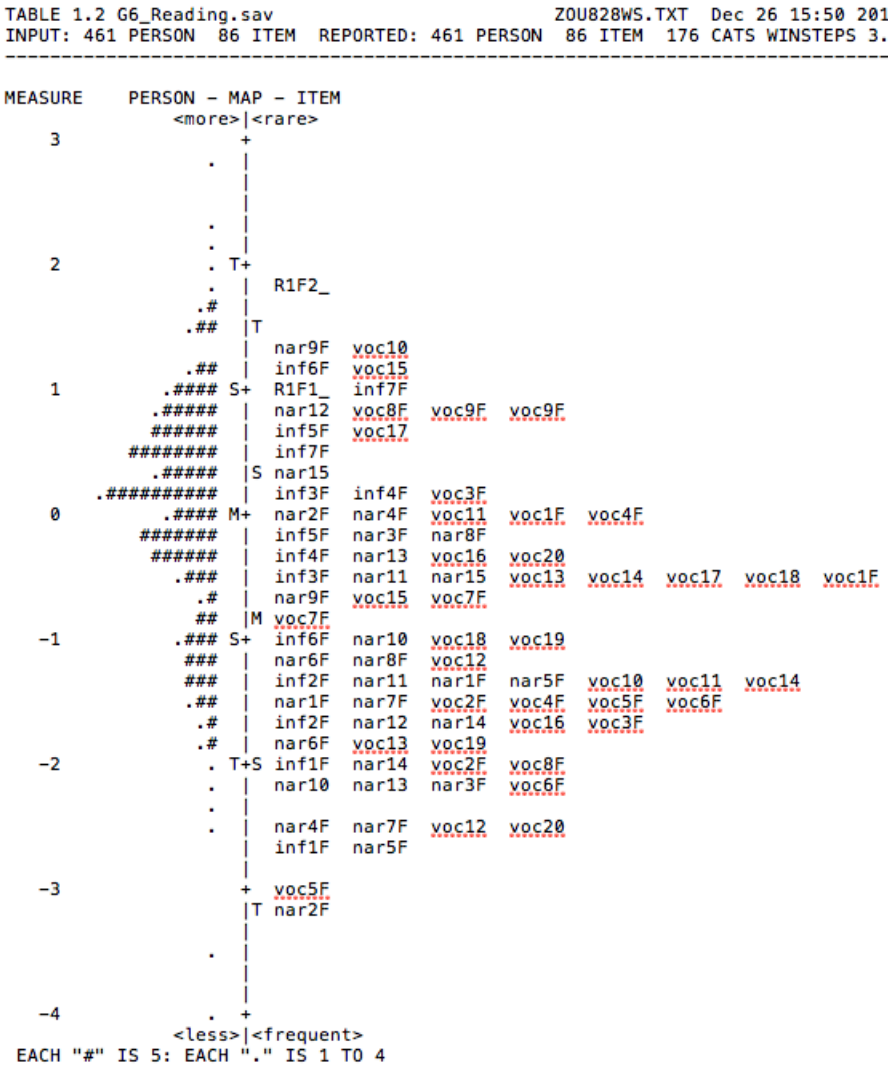
Figure 13. Distribution maps for reading Georgian items – Grade 5



```
TABLE 1.2 G5_Reading.sav                          ZOU888WS.TXT  Dec 26 15:49 2014
INPUT: 464 PERSON  86 ITEM  REPORTED: 464 PERSON  86 ITEM  176 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
    4               +
                    |
                    |
                    |  voc12
                    |  nar1F
                    |
    3               +
                    |
                    |
                    |
                  . |
                  . |
    2             . T+T
                .##   |
                ###   |  inf5F
                  .   |  voc14
                .##   |  R1F2_
             .######   |  nar10  voc15  voc1F
    1           #### S+  inf3F  inf3F  voc10  voc9F
              ######  |S R1F1_  inf6F  inf7F  voc13  voc17
               .####  |  voc11  voc6F
          .##########  |  voc11  voc6F
             #######   |  inf4F  inf7F  voc12  voc19  voc3F
             .######   |  inf2F  nar2F
    0      .######## M+  inf6F  nar11  nar12  nar5F  nar7F  voc18  voc4F  voc5F
              .#####   |  inf4F  nar10  nar1F  voc5F
           .##########|M nar12  nar13  nar8F  nar9F  voc2F  voc3F
             #######   |  voc4F
                ####   |  nar6F  nar7F
             #######   |  nar14  nar3F  nar9F  voc16  voc1F  voc8F
   -1         .##### S+  nar13  voc13  voc18  voc2F
                ###   |  inf1F  nar4F  voc20  voc9F
                 .#   |  inf2F  nar11  nar15  nar4F  nar8F  voc17
                .##   |  nar14  nar15  voc14
                 ##   |S nar2F  nar5F
                  #   |  inf1F  nar3F  voc8F
   -2            .## T+  nar6F  voc10
                   .   |  inf5F  voc16  voc19
                   .   |  voc20  voc7F
                   .   |
                   .   |  voc7F
                   .   |T voc15
   -3                  +
                   .   |
                   .   |
                       |
                       |
                       |
   -4                  +
              <less>|<frequent>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

Figure 14. Distribution maps for reading Georgian items – Grade 6



```
TABLE 1.2 G6_Reading.sav                          ZOU828WS.TXT  Dec 26 15:50 2014
INPUT: 461 PERSON  86 ITEM  REPORTED: 461 PERSON  86 ITEM  176 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
    3               +
                .   |
                    |
                    |
                .   |
                .   |
    2           . T+
                .   |  R1F2_
               .#   |
               .##  |T
                    |  nar9F  voc10
               .##  |  inf6F  voc15
    1         .#### S+  R1F1_  inf7F
             .######  |  nar12  voc8F  voc9F  voc9F
             ######   |  inf5F  voc17
           ########   |  inf7F
             .#####   |S nar15
         .##########  |  inf3F  inf4F  voc3F
    0         .#### M+  nar2F  nar4F  voc11  voc1F  voc4F
            #######   |  inf5F  nar3F  nar8F
            ######    |  inf4F  nar13  voc16  voc20
              .###    |  inf3F  nar11  nar15  voc13  voc14  voc17  voc18  voc1F
                .#    |  nar9F  voc15  voc7F
                ##    |M voc7F
   -1         .### S+  inf6F  nar10  voc18  voc19
               ###   |  nar6F  nar8F  voc12
               ###   |  inf2F  nar11  nar1F  nar5F  voc10  voc11  voc14
               .##   |  nar1F  nar7F  voc2F  voc4F  voc5F  voc6F
                .#   |  inf2F  nar12  nar14  voc16  voc3F
                .#   |  nar6F  voc13  voc19
   -2           . T+S inf1F  nar14  voc2F  voc8F
                .   |  nar10  nar13  nar3F  voc6F
                .   |
                .   |  nar4F  nar7F  voc12  voc20
                    |  inf1F  nar5F
                    |
   -3               +  voc5F
                    |T nar2F
                    |
                .   |
                    |
                    |
   -4               +
              <less>|<frequent>
 EACH "#" IS 5: EACH "." IS 1 TO 4
```

Distribution maps for reading Georgian as second language tests

Figure 15. Distribution maps for reading Georgian as second language items – Grade 1

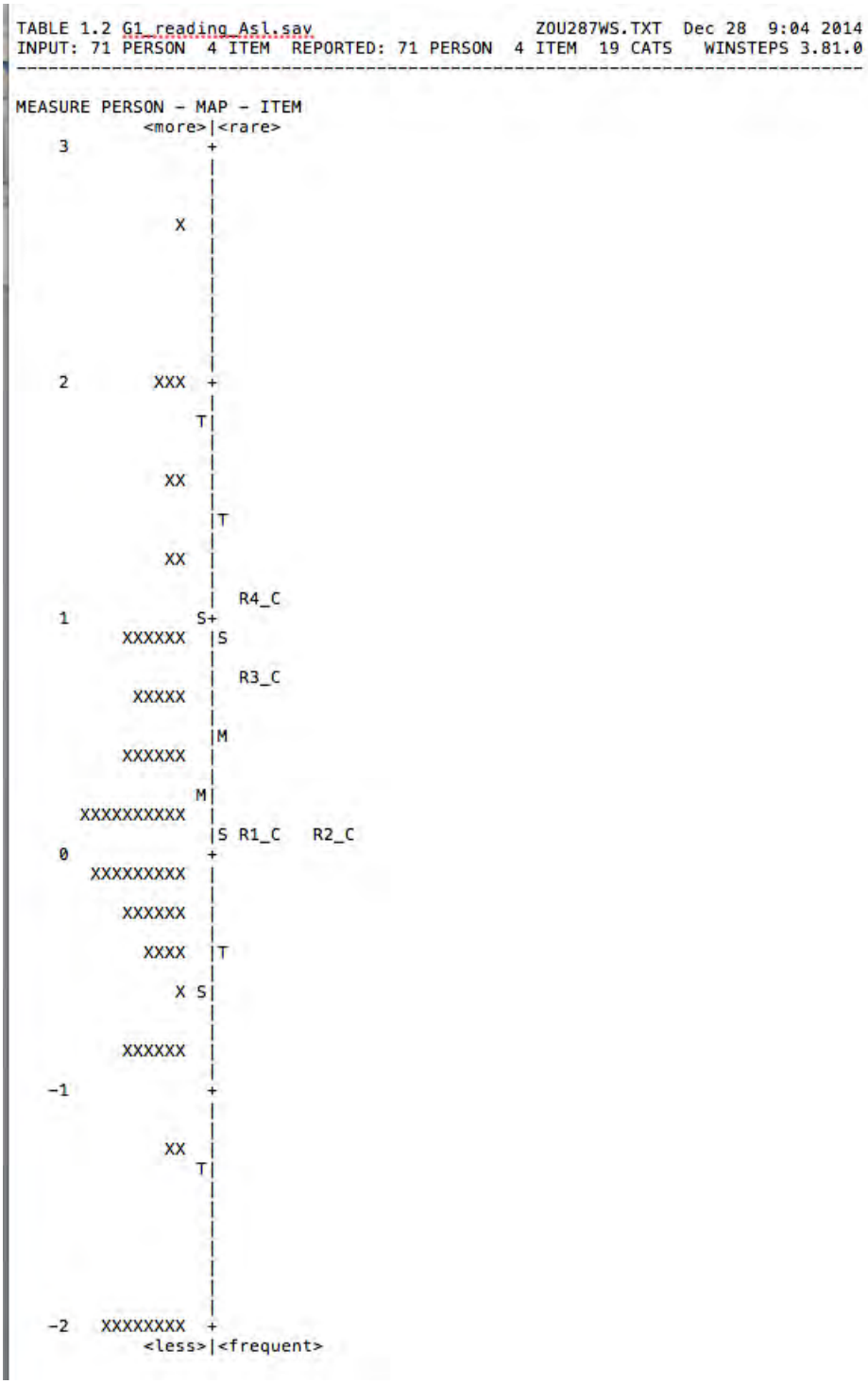


Figure 16. Distribution maps for reading Georgian as second language items – Grade 2

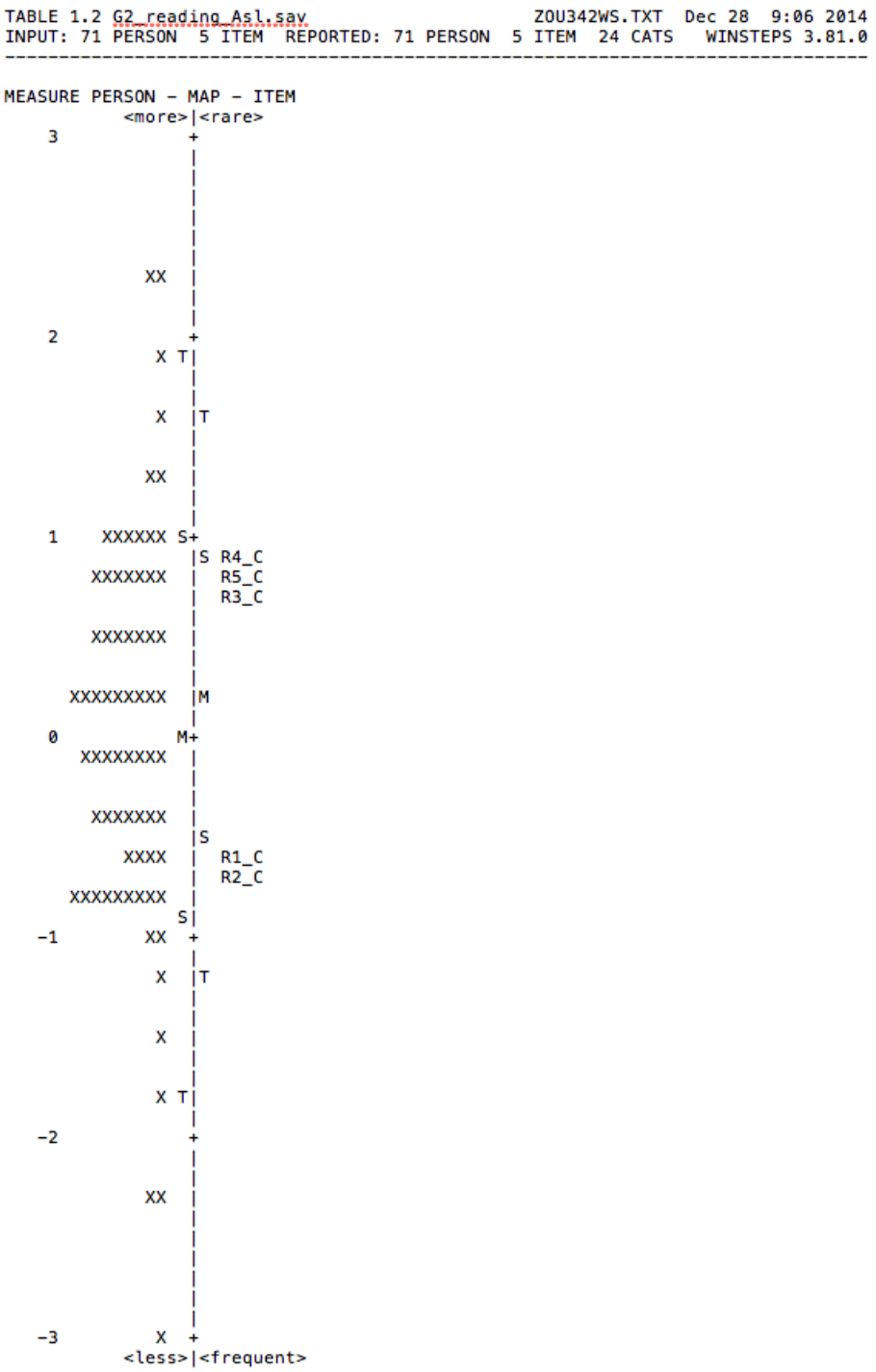


Figure 17. Distribution maps for reading Georgian as second language items – Grade 3

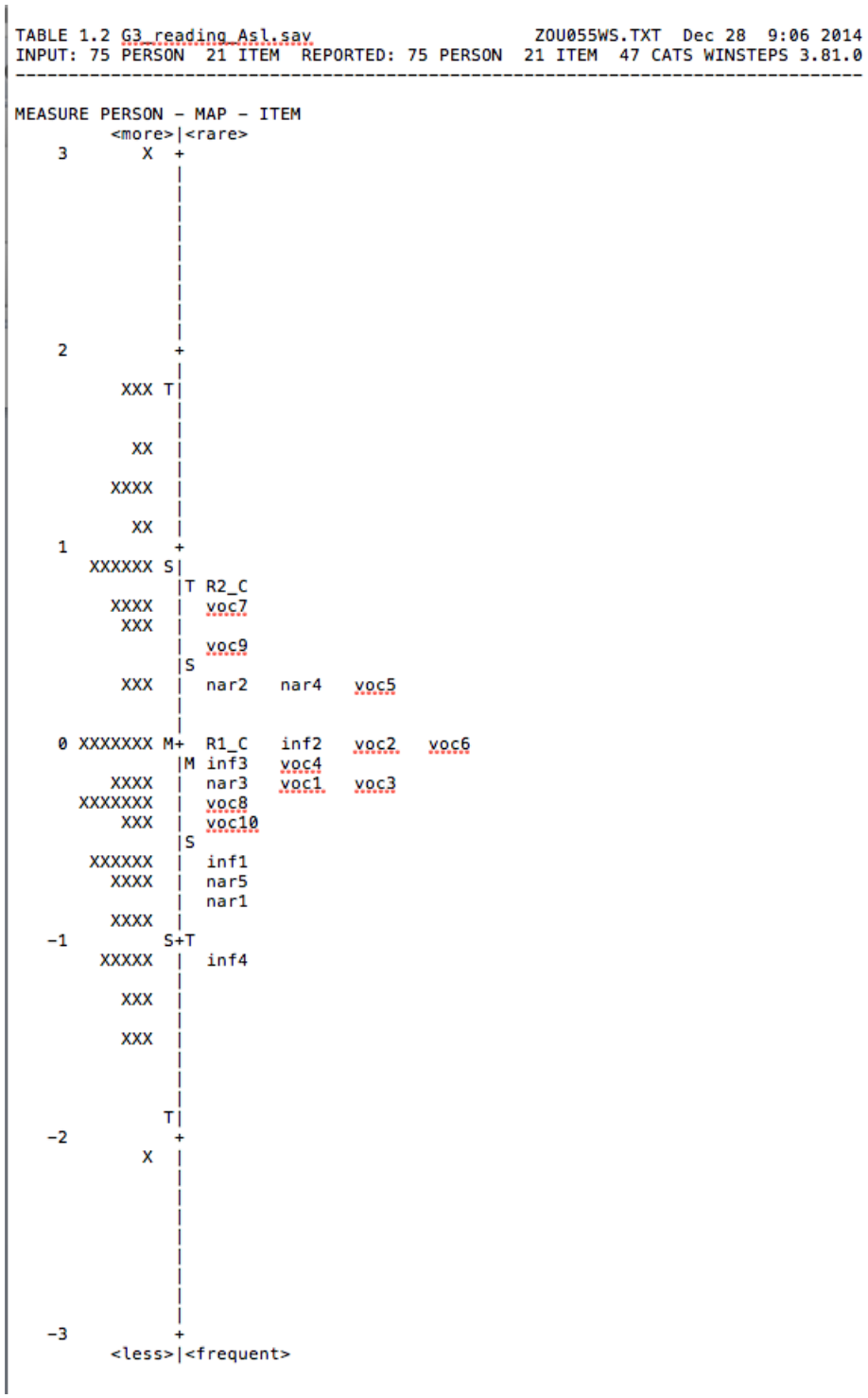


Figure 18. Distribution maps for reading Georgian as second language items –Grade 4

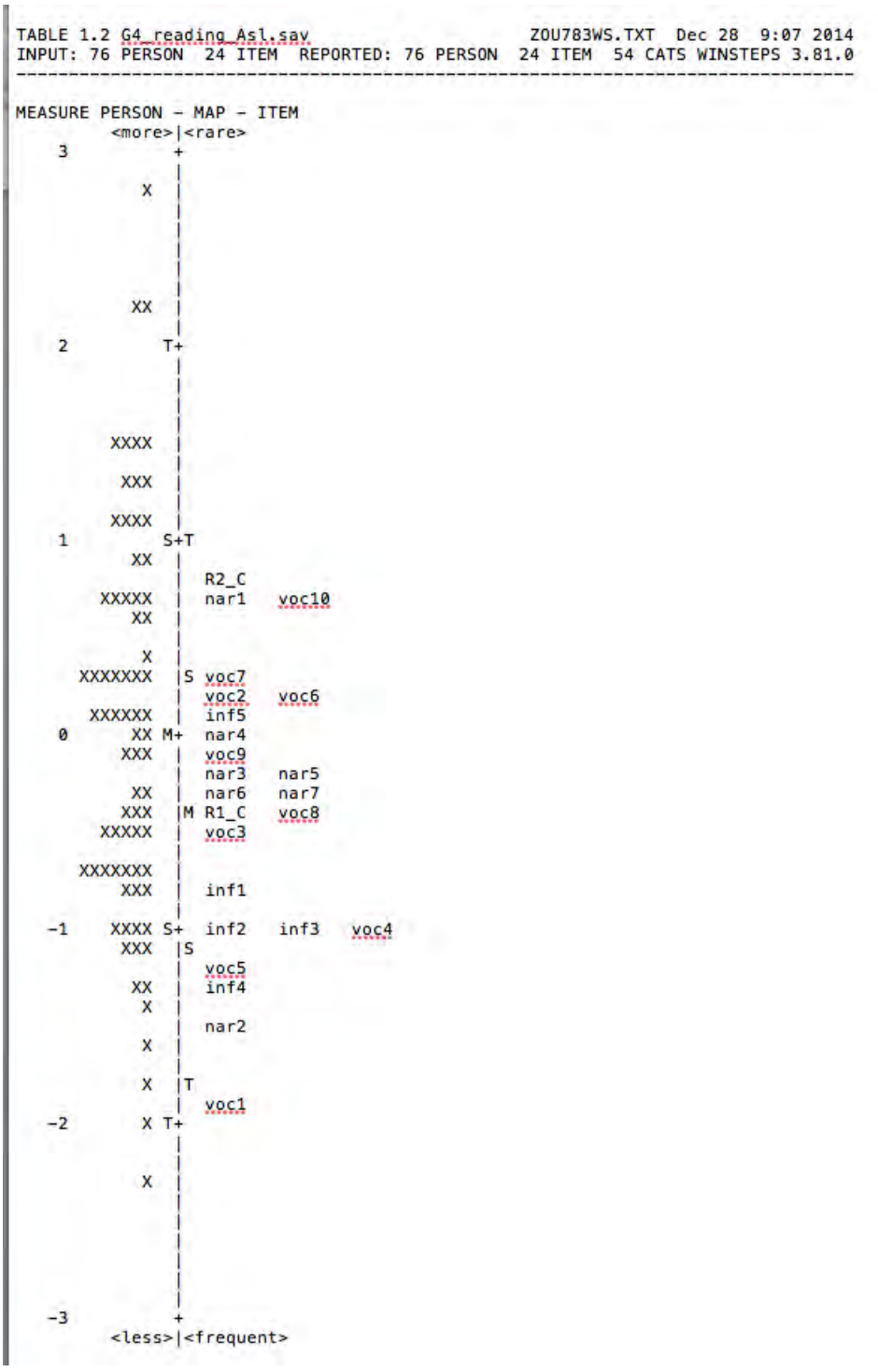


Figure 19. Distribution maps for reading Georgian as second language items – Grade 5

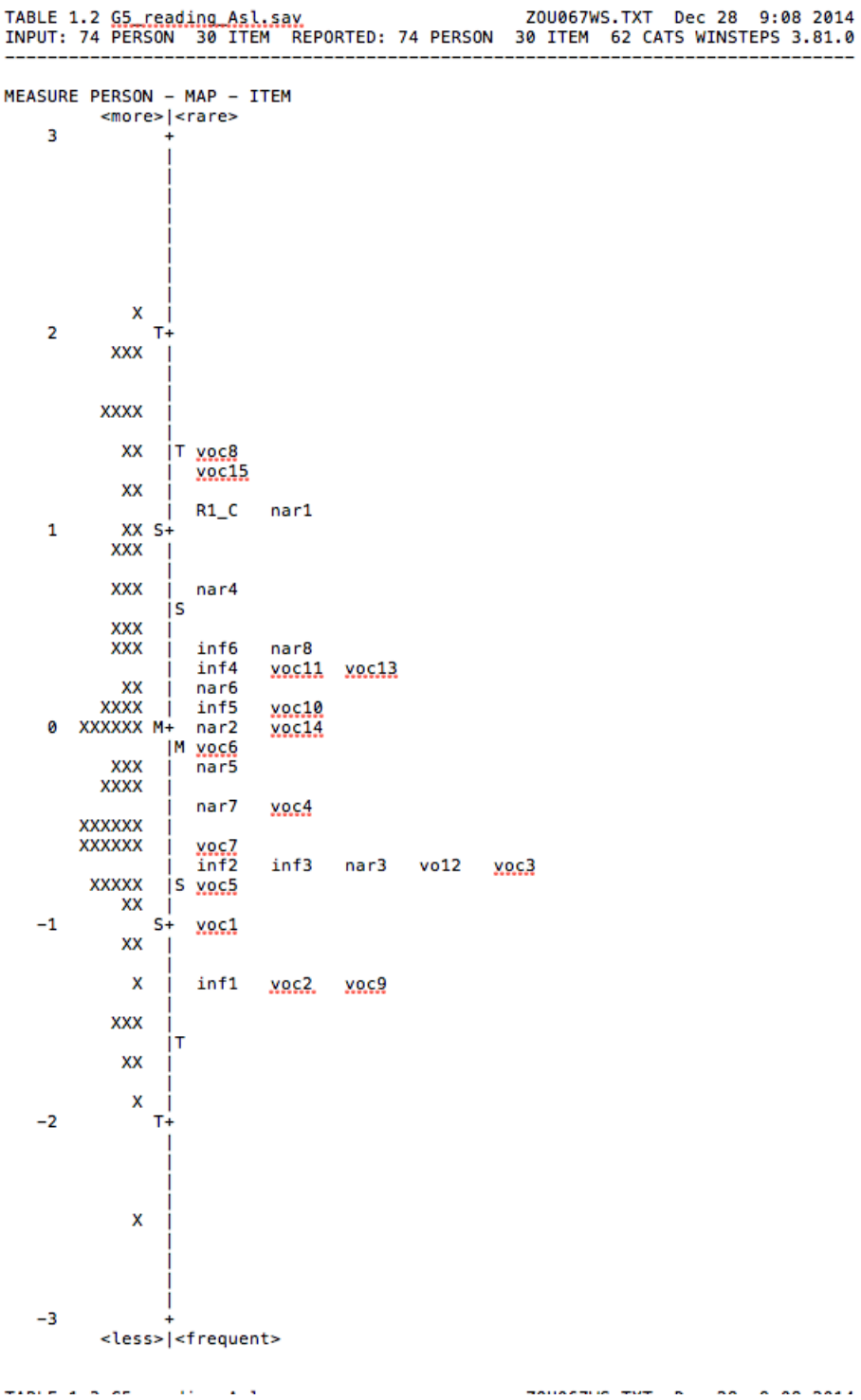
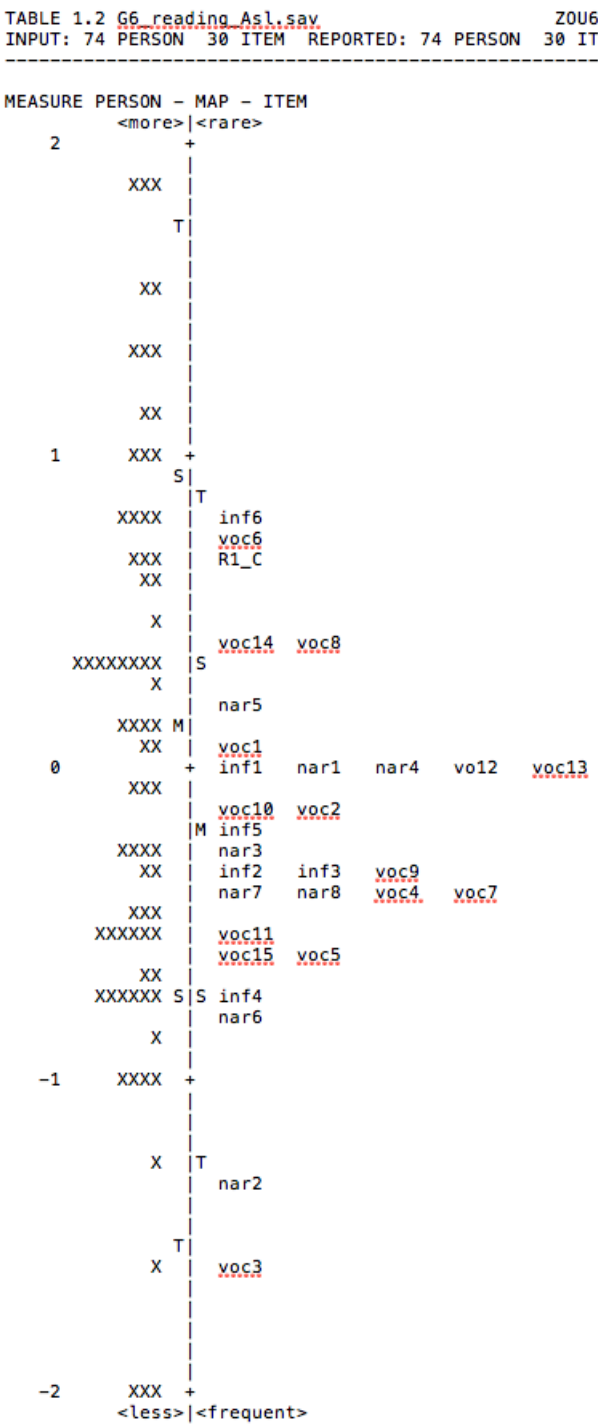


Figure 20. Distribution maps for reading Georgian as second language items – Grade 6



Infit and Outfit statistics

The graphs presented in this section are examples of graphs used to detect if any items are outside the bound of 0.4 or 1.6 Infit or Outfit meansquare. Items that are found to be outside those bounds have been removed from further analysis.

Figure 21. Distribution of infit mean square of mathematics items – Grade 1

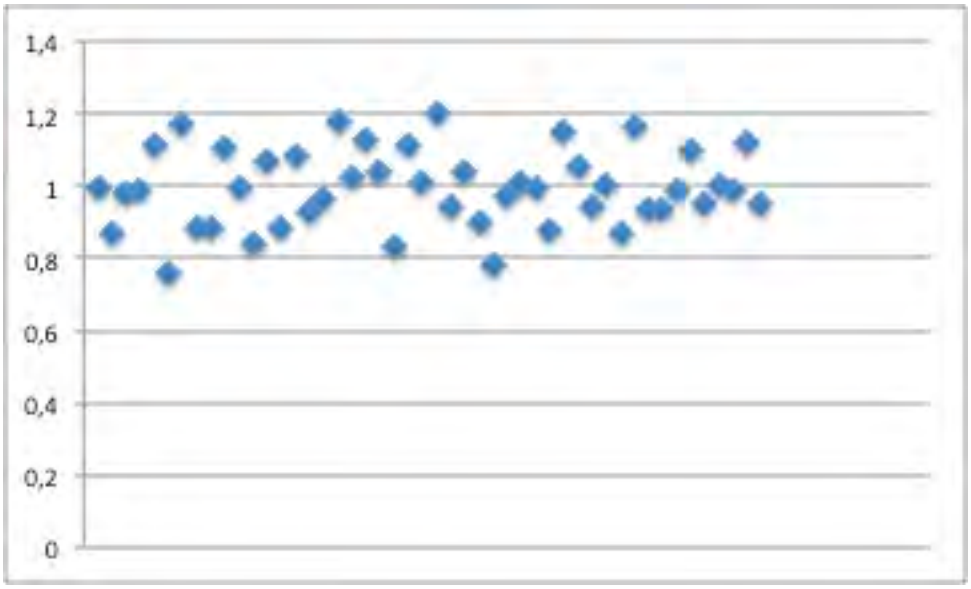
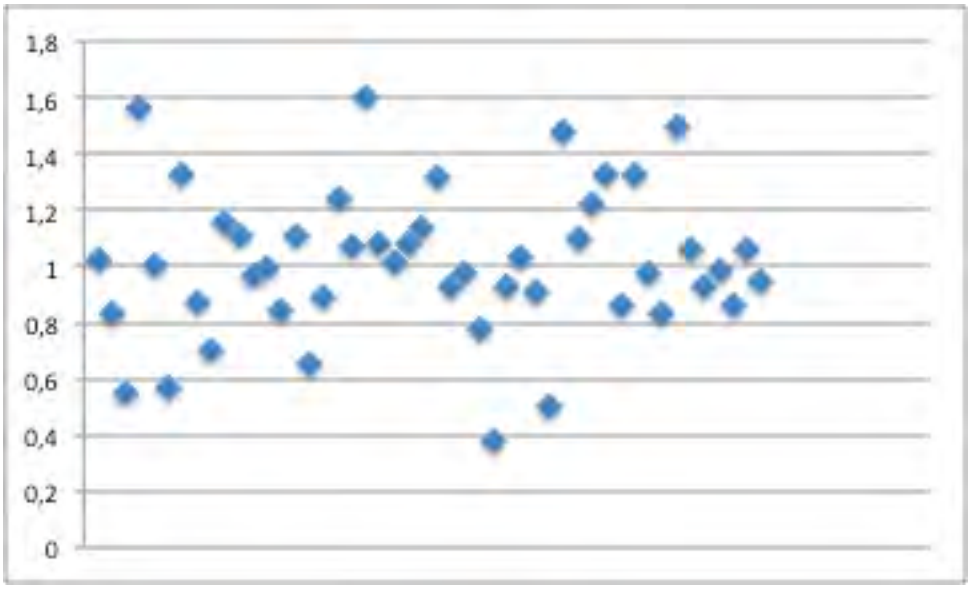


Figure 22. Distribution of infit mean square of mathematics items – Grade 2



Scaling of the Rasch scores

As was stated earlier, the value of the measurement scale must be fixed to specific values. For the analysis, those values were fixed to a mean of 0 and a standard deviation of 1. The main problem with this kind of scale is the presence of negative scores. To address this issue, a linear transformation has been applied to the Rasch scores. This linear transformation results in a scale with a mean of 500 and a standard deviation of 100. Thus, a score lower than 500 means that the student has a performance lower than the mean performance of all students in his grade level in the sample. On the other hand, a score higher than 500 means that he has a better performance than the mean performance of all students in his grade.

The transformation used the following equation:

$$New\ scale = 500 + 100 * \theta$$

where θ is the ability level of the students on the original scale (mean of 0 and SD of 1).

ANNEX IV. STANDARD SETTING METHODOLOGY

While scaling is an important procedure that produces scores which can be correlated with external variable, the interpretation of the scores is not straightforward. We are able to determine whether students have a score that is lower or higher than other students in the sample but this information doesn't provide any information on what this student can actually achieve. Standard setting is an operation that aims to describe what a student is able to perform given his/her score on the test. This section explains how the standards were defined for the tests.

Development of Mastery levels

The measurement scale represents a continuum that goes from lowest levels of ability to highest levels of ability. The items that compose the test determine the width of this continuum. For example, if a mathematics test is formed of items from the Grade 2 curriculum and is administered to Grade 6 students, even if students get a high score on this test, it doesn't mean that they are really proficient at their level. It means that they can easily give correct answers to grade 2 items.

In standard setting, cut scores must be identified to produce a classification system of students into performance categories. The statistical distribution of scores on a test doesn't determine the standard setting operation. Standards setting involves the construction of these mastery levels based on the performance of students on items of each test and on the cognitive demand of the items. The construction of mastery levels is done in two steps. First, decisions about where to set cut-off scores for the different levels and how to associate students with each level are made. Second, an analysis of the items linked to each level is performed in order to develop descriptions for each level. There are no natural cut points to distinguish between stages in the continuum of ability level. Dividing the scale into levels of proficiency is essentially arbitrary. However levels of proficiency can help describe what students in a specific level can typically perform.

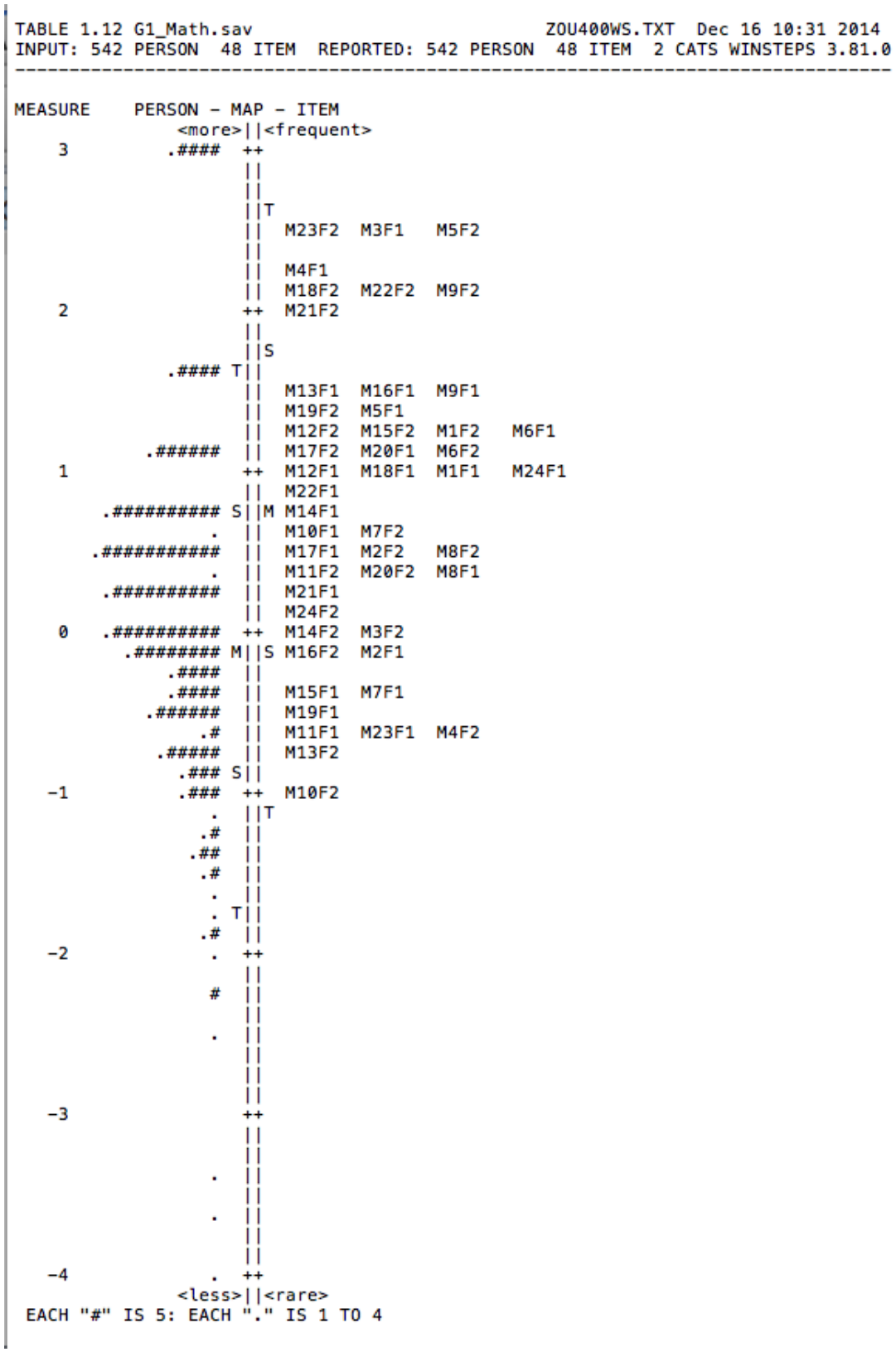
Students are then categorized into a specific level depending on their performance on items linked to each Mastery level. Students at the bottom of a level are able to complete only 50% of the items correctly on the set of items set at the level while students at the middle and top of each level are expected to achieve a much higher success rate, about 80% for the top achievers of the level.

The standards developed for G-PriEd consisted in separating the continuum of scores into four Mastery levels. The lowest and highest levels are unbounded (i.e. they do not have the same width as the other levels) since there are some students who are exceptionally low or high achievers. The middle levels (level 2 and 3) have the same breadth to ensure that the meaning of being at the top or bottom within a given level is more or less the same for each level. The determination of cut-off scores was made for each test depending on the distribution of the scores of students. The distribution of the items is taken into account to ensure that there is sufficient information at each level to develop a meaningful description of what a student at that level can achieve. Given the distribution of the items, a breadth of about 1 standard deviation was chosen for level 2 and 3.

The following Distribution map is used to illustrate how the cut-scores were identified. In this map, we can see that a number of items are grouped together at the lowest end of the continuum. These items are the easiest items in the Mathematics Grade 1 test and constitute Mastery level 1. At the other end of the scale, there is also a group of items that are far from the majority of the items. These items are the harder ones and constitute Mastery level 4. For Mastery levels 2 and 3, the cut-score separating these two levels falls in the middle of the range of values going from Master Level 1 to 4.

For the second step, we describe what a student at a given level can perform based on a content analysis of the items at this level. The content of each set of items linked to a level of proficiency is analyzed to provide Performance Level Descriptors (PLD), i.e. descriptions of what a student should know for different levels of proficiency. PLDs provide a sense of the skills that characterize different levels of performance based on the scores captured by the assessment.

Figure 23. Distribution map



```
TABLE 1.12 G1_Math.sav                          ZOU400WS.TXT  Dec 16 10:31 2014
INPUT: 542 PERSON  48 ITEM  REPORTED: 542 PERSON  48 ITEM  2 CATS WINSTEPS 3.81.0
-------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>||<frequent>
    3          .####  ++
                      ||
                      ||
                      ||T
                      ||  M23F2  M3F1   M5F2
                      ||
                      ||  M4F1
                      ||  M18F2  M22F2  M9F2
    2                 ++  M21F2
                      ||
                      ||S
              .#### T||
                      ||  M13F1  M16F1  M9F1
                      ||  M19F2  M5F1
                      ||  M12F2  M15F2  M1F2   M6F1
           .######   ||  M17F2  M20F1  M6F2
    1                 ++  M12F1  M18F1  M1F1   M24F1
                      ||  M22F1
        .########## S||M M14F1
                  .   ||  M10F1  M7F2
       .###########  ||  M17F1  M2F2   M8F2
                  .   ||  M11F2  M20F2  M8F1
        .##########  ||  M21F1
                      ||  M24F2
    0    .##########  ++  M14F2  M3F2
          .######## M||S M16F2  M2F1
              .####   ||
              .####   ||  M15F1  M7F1
            .######   ||  M19F1
                 .#   ||  M11F1  M23F1  M4F2
             .#####   ||  M13F2
               .### S||
   -1           .###  ++  M10F2
                  .   ||T
                 .#   ||
                .##   ||
                 .#   ||
                  .   ||
                  . T||
                 .#   ||
   -2             .   ++
                      ||
                  #   ||
                      ||
                  .   ||
                      ||
                      ||
                      ||
   -3                 ++
                      ||
                      ||
                  .   ||
                      ||
                  .   ||
                      ||
                      ||
   -4                 .  ++
              <less>||<rare>
 EACH "#" IS 5: EACH "." IS 1 TO 4
```

ANNEX V. MASTERY LEVELS: CUT-OFF SCORES AND DESCRIPTORS

A1. Reading Levels

Table 26. Mastery levels cut-score and description for reading - Grade 1

Mastery level	Cut-scores	Description
4	550.01 and higher	At this level, students can answer all items in the phoneme segmenting, syllable segmenting and letter sound fluency tasks correctly. They are able to read up to 45 words per minute correctly (75% of all items) in the word reading fluency task. Only the most competent students can read all 60 items correctly in one minute in this task.
3	430.01 to 550	At this level, students can answer all items in the phoneme segmenting, syllable segmenting and letter sound fluency tasks correctly. They are able to read up to 30 words per minute correctly (50% of all items) in the word reading fluency task.
2	300.01 to 430	Students at this level can segment 75% of phonemes in the phoneme segmenting task, 50% of syllables in the syllable segmenting task and can identify up to 49 letter sounds per minute (75% of all letter sounds included in the task) in the letter sound fluency task. Also, they can read less than 15 words per minute correctly in the word reading fluency task.
1	300 and lower	Students at this level can segment less than 25% of phonemes in the phoneme segmenting task, less than 25% of syllables in the syllable segmenting task and can identify up to 16 letter sounds per minute (25% of all letter sounds included in the task) in the letter sound fluency task. They are unable to read words correctly in the word reading fluency task.

Table 27. Mastery levels cut-score and description for reading - Grade 2

Mastery level	Cut-scores	Description
4	532.01 and higher	At this level, students can identify all letter sounds and answer all vocabulary items correctly within the alloted time. Only the most competent students can read all 120 items correctly for the word reading fluency task and the 90 words correctly in the passage reading fluency in the alloted time (one minute).
3	487.01 to 532	Students at this level can identify all letter sounds, they can read up to 45 words per minute in the passage reading fluency task and answer 11 out of 12 vocabulary items.
2	415.01 to 487	Students at this level can identify up to 32 letter sounds in the letter sound fluency task. They can also read up to 30 items in the word reading fluency task. They are able to read less than 22 words per minute in the passage reading fluency task. Finally, they are able to answer 8 of the 12 vocabulary items correctly.
1	415 and lower	At this level, students can identify less than 16 letter sounds (25% of the letter sounds) in the letter sound fluency task. They are able to read less than 22 words per minute correctly in the passage reading fluency task. They are unable to read anything in the word reading fluency task and are able to answer 2 out of the 12 vocaulary items correctly.

Table 28. Mastery levels cut-score and description for reading - Grade 3

Mastery level	Cut-scores	Description
4	551.01 to higher	**Fluency:** Students at this level can read 86 words per minute. The most competent students can read 115 words per minute (the entire passage). **Vocabulary:** Students know the meaning of all words. They are able to choose the correct word forms from context (based on grammatical knowledge) and identify words from context. **Comprehension of narrative text:** All students are able to identify the main idea of the story, analyze and explain the motivation behind a character's behavor, identify cause-and-effect relationships between different parts of the text and retrieve explicit information. Students are also able to understand the content of the story in detail, identify relationships between characters and explain the motivation behind their behaviour. They are able to understand the structure of the text. Some students are able to identify the setting of the story. **Comprehension of informational text:** Students are able to understand the text content, identify the timing of events. They are also able to associate specific parts of the text to the illustration, understand the content of what they read thoroughly (including details) and retrieve explicit information from the text. Students at this level are also able to identify a topic of the text, analyze and explain the reason behind an action/event.
3	474.01 to 551	**Fluency:** Students are able to read 57 words per minute. **Vocabulary:** Students are able to answer 95% of the vocabulary items correctly. Students know the meaning of almost all the words. They are able to identify word meaning from context, as well as choose the correct word forms based on context (based on knowledge of grammar). **Comprehension of narrative text:** Students are able to answer all Comprehension questions correctly. All students are able to identify the main idea of the story, analyze and explain the motivation behind a character's behavior, identify cause-and-effect relationships between different parts of the text and retrieve explicit information. Students are also able to understand the content of the story in detail, identify relationships between characters and explain the motivation behind their behavior. They are able to understand the structure of the text. Some students are able to identify the setting of the story. **Comprehension of informational text:** Students are able to answer 3 Comprehension questions correctly. Students are able to identify the main topic of the text. Some students are able to understand some parts of the text, find and retrieve explicit information, as well as analyze and explain the reason behind events (i.e. why is the event occurring?).
2	341.01 to 474	**Fluency:** At this level, students are able to read 29 words per minute. **Vocabulary:** Students know the meaning of most words (12/15 items). Students are also able to identify the meaning of some words from context. They are able to choose the correct word forms from context (using knowledge of grammar).

Mastery level	Cut-scores	Description
		Comprehension of narrative text: All students are able to determine the relationships between characters, identify an action of a character. Some students are able to identify the setting - time and place of events, cause-and-effect relationships between different parts of the text, as well as retrieve explicit information from the text. Some of them are also able to analyze and explain the motivation behind a character's behaviur. **Comprehension of informational text:** Students are able to identify the main topic of the text and can answer one comprehension question correctly.
1	Lower to 341	**Fluency:** Students at this level are able to read 29 words per minute. **Vocabulary:** Students know the meaning of some words (6/15 items), they can identify the meaning of some words from context. **Comprehension of narrative text:** Some students are able to determine the relationships between characters in the story (one Comprehension question correct) **Comprehension of informational text**: None of the students are able to comprehend informational text.

Table 29. Mastery levels cut-score and description for reading - Grade 4

Mastery level	Cut-scores	Description
4	586.01 and higher	**Fluency:** At this level, students are able to read 97 words per minute. **Vocabulary:** Students at this level know the meaning of all words and phrases. They are able to identify the meaning of words from context. **Comprehension of narrative text:** Students are able to find and retrieve explicit information from the text. They are also able to identify characters, specify and draw conclusions about their feelings and purposes, as well as evaluate a character according to his/her behavior, actions and traits. They are able to understand the connection between the title and the text, define the stages of the plot and, identify the chronology of actions and events, and cause-and-effect relations in the story. Students at this level are also able to understand the content of the text so as to differentiate between the author's and characters' words (opinions). Students are also able to identify the main idea of the story. **Comprehension of informational text:** Students are able to identify the main topic of the text and make inferences. Students are also able to find and retrieve explicit information and understand how facts, events and actions are related to each other. They are also able to integrate knowledge/information.
3	455.01 to 586	**Fluency:** Students are able to read 49 words per minute. **Vocabulary:** Students at this level know the meaning of sentences and almost all words (19/20 words). They are able to identify the meaning of all words from context.

Mastery level	Cut-scores	Description
		Comprehension of narrative text: Students are able to find and retrieve some explicit information from the text. They are also able to identify characters, specify and draw conclusions about their feelings and purposes, as well as evaluate the character according to his/her behavior, actions and traits. They are also able to define the stages of the plot and, identify the chronology of actions and events, and cause-and-effect relations in the story. Students at this level are able to understand the content of the text so as to differentiate between the author's and characters' words (opinions). Most students are also able to identify the main idea of the story. They can answer 10 out of 11 narrative questions correctly. **Comprehension of informational text:** Students are able to identify the main topic of the text and make inferences. Students are also able to find and retrieve explicit information and understand how facts, events and actions are related to each other. Some students are able to integrate knowledge/information. They can answer 5 out of 6 informational questions correctly.
2	358.01 to 455	**Fluency:** Students are able to read 49 words per minute. **Vocabulary:** Students at this level know the meaning of some words (12/20 words) and phrases. They are able to identify the meaning of a few words from context. **Comprehension of narrative text**: Students are able to find and retrieve some explicit information from the text. They are also able to identify characters, specify and draw conclusions about their feelings and purposes, as well as evaluate a character according to his/her behavior, actions and traits. They are also able to define the stages of the plot and, identify the chronology of actions and events, and cause-and-effect relations in the story. Students at this level are able to differentiate between the author's and characters' words (opinions). They can answer 7 out of 11 narrative questions correctly. **Comprehension of informational text:** Students are able to identify the main topic of the text, also find and retrieve explicit information from the text. They can answer 2 out of 6 informational questions correctly.
1	358 and lower	**Fluency:** Students are able to read 49 words per minute. **Vocabulary:** Students at this level know the meaning of a few words (5/20 words). They are able to identify the meaning of some words from context. **Comprehension of narrative text:** Students are able to find and retrieve some explicit information from the text. They are also able to specify and draw conclusions about the feelings and purposes of a character, as well as evaluate him/her according to his/her behavior, actions and traits. They can answer 3 out of 11 narrative questions. **Comprehension of informational text:** None of the students are able to comprehend an informational text.

Table 30. Mastery levels cut-score and description for reading – Grade 5

Mastery level	Cut-scores	Description
4	601.01 and higher	**Fluency:** At this level, students are able to read 116 words per minute. **Vocabulary:** Students know the meaning of all words and almost all phrases. They are aware of the semantic correspondence between words and are able to identify almost all words by context. **Comprehension of narrative text:** Students at this level are able to identify the main idea of the story. They are able to understand and retrieve explicit information from the text, also understand content, e.g. identify how and why, what kind of, as well as identify cause-and-effect relationships between different parts of the story. They are also able to differentiate between the author's and characters' words, dialogue and monologue. Students are able to define the stages of the plot. Students are able to define characters' point of view and explain the motive behind his/her behavior, as well as draw conclusions about a character's thoughts, intentions, and feelings based on character behavior, actions, and traits. Some students are able to suggest an alternative title for the story. **Comprehension of informational text:** Students are able to understand the content of the whole text, identify explicit factual information, find and retrieve it from the text. They are also able to identify the topic of the text, relate facts, occasions, and action to each other and make respective inferences on the basis of their understanding of the text. Students at this level are also able to integrate their knowledge.
3	489.01 to 601	**Fluency:** At this level, students can read 58 words per minute. **Vocabulary:** Students know the meaning of almost all words and phrases (18/20). They are aware of the semantic correspondence between words and are able to identify almost all words from context. **Comprehension of narrative text:** Students at this level are able to identify the main idea of the story. They are able to understand and retrieve explicit information from the text, also understand content, e.g. identify how and why, what kind of, as well as identify cause-and-effect relationships between different parts of the story. They are also able to differentiate between the author's and characters' words, dialogue and monologue. Students are able to define the stages of the plot. Students are able to define the characters' point of view and explain the motive behind his/her behavior, while some of them are able to draw conclusions about a character's thoughts, intentions, and feelings based on character behavior, actions, and traits. Some students are able to suggest an alternative title for the story. **Comprehension of informational text:** Students are able to identify explicit factual information, find and retrieve it from the text. They are also able identify the topic of the text, relate facts, occasions, and actions to each other and make respective inferences based on their understanding of the text. Students at this level are also able to integrate their knowledge.
2	393.01 to 489	**Fluency:** At this level, students can read 58 words per minute.

		Vocabulary: Students know the meanings of some words and phrases (12/20). They are aware of the semantic correspondence between words and are able to identify some words from context. **Comprehension of narrative text:** Students at this level are able to identify the main idea of the story. They are able to understand and retrieve explicit information from the text, also understand content, e.g. identify how and why, identify cause-and-effect relations among different parts of the story. They are also able to differentiate between the author's and characters' words, dialogue and monologue. Students are able to define a character's point of view and explain the motive behind his/her behavior, while some of them are able to draw conclusions about a character's thoughts, intentions, and feelings based on character behavior, actions, and traits. Some students are able to define the stages of plot. They can answer 11 out of 15 narrative questions correctly. **Comprehension of informational text:** Students are able to find and retrieve explicit information from the text. They are also able to identify the main topic of the text, relate facts, occasions, and actions to each other and make respective inferences based on what they read. They can answer 2 out of 7 informational questions correctly.
1	393 and lower	**Fluency:** At this level, students can read 58 words per minute. **Vocabulary:** Students at this level know the meaning of some words and phrases (8/20). They are aware of the semantic correspondence between words and are able to identify a few words from context. **Comprehension of narrative text:** Students at this level are able to understand and retrieve some explicit information from the text, also understand some parts content (e.g. identify how and why). Some of them are also able to draw conclusions about a character's thoughts, intentions, and feelings based on the character's behavior, actions, and traits, define a character's point of view and explain the motive behind his/her behavior. They are able to answer 5 out of 15 narrative questions correctly. **Comprehension of informational text:** Students are able to find and retrieve some explicit information from the text. They are also able to identify the main topic of the text, and make inferences based on what they read. They can answer 2 out of 7 informational questions correctly.

Table 31. Mastery levels cut-score and description for reading - Grade 6

Mastery level	Cut-scores	Description
4	524.01 and higher	**Fluency:** Students are able to read 117 words per min. **Vocabulary:** Students know the meaning of all words and sentences. They are aware of the semantic correspondence between words, and are able to identify all words from context. **Comprehension of narrative text:** Students are able to identify the main idea of the text and suggest alternative titles to the story, as well as define the stages of the plot.

Mastery level	Cut-scores	Description
		Students are able to understand the content of the text (explicit factual information and cause-and-effect relations between different parts of the story), distinguish between main and supporting characters, identify chronology of events in the story and make inferences based on understanding the content of the whole text. They are also able to identify creative expressions (figurative language) and understand their purpose in the text, as well as differentiate between the author's and the character's words. Students are able to define the characters' point of view, identify and explain the motive behind a character's behavior, draw conclusions about him/her based on his/her actions and traits, as well as predict what the character is likely to do next (based on his/her traits in the story). **Comprehension of informational text:** Students are able to identify the topic of the text and make inferences on the basis of overall Comprehension of the text. They are also able to understand and identify explicit factual information, as well as relate different parts of the text to each other. Some students are able to integrate their knowledge.
3	453.01 to 524	**Fluency:** Students are able to read 58 words per minute. **Vocabulary:** Students know the meaning of most words and phrases (17/20). They are aware of the semantic correspondence between words, and are able to identify all words from context. **Comprehension of narrative text:** Students at this level are able to identify the main idea and suggest alternative titles to the story, and some of them are able to define the stages of the plot. Students are able to understand the content of the text (explicit factual information and cause-and-effect relations between different parts of the story), distinguish between main and supporting characters, identify the chronology of events in the story and make inferences based on their understanding of the content of the whole text. They are also able to identify creative expressions (figurative language) and understand their purpose in the text, as well as differentiate between the author's and the character's words. Students at this level are able to identify and explain the motive behind the characters' behavior, draw conclusions about him/her based on his/her actions and traits, as well as predict what the character is likely to do next (based on his/her traits in the story). Some students are able to define the character's point of view. **Comprehension of informational text:** Students are able to identify one of the topics of the text and make inferences based on their understanding of the text. They are also able to find and retrieve some explicit information from the text.
2	341.01 to 453	**Fluency:** Students at this level are able to read 58 words per minute. **Vocabulary:** Students know the meaning of some of the words and phrases (14/20). They are aware of the semantic correspondence between words and are able to identify most of the words from context. **Comprehension of narrative text:** Students at this level are able to identify the main idea of the story. They are able to identify and explain the motive behind the characters' behavior, as well as predict what a character is likely to do next (based on his/her traits in the story). Students are also able to understand the content of the text to a certain extent, e.g. they are able to understand explicit factual information and cause-and-effect relations between different parts of the story. Students are able to identify the chronology of events in the story and make inferences based on their understanding of the text as a whole. Some students are also able to distinguish between main and supporting characters, draw conclusions about him/her based on his/her actions and traits, define the stages of the plot,

Mastery level	Cut-scores	Description
		as well as identify creative expressions (figurative language) and understand their purpose in the text. **Comprehension of informational text:** Students at this level are able to identify the topic of the text. Only a few of them are able to find and retrieve some explicit information and make inferences based on what was read.
1	341 and lower	**Fluency**: Students at this level are able to read 58 words per minute. **Vocabulary:** Students know the meaning of a few words and phrases (5/20). They are partly aware of the semantic correspondence between words and are able to identify very few words from context. **Comprehension of narrative text:** Some students are able to identify and explain the motive behind a character's behavior, draw conclusions about him/her based on his/her actions and traits, as well as predict what a character is likely to do next (based on his/her traits in the story). Some students are also able to understand the content of the text (explicit factual information and cause-and-effect relations between different parts of the story), and identify the chronology of events in the story. They can answer 5 out of 15 narrative questions correctly. **Comprehension of informational text:** Students at this level are able to identify the topic of the text. Only a few of them are able to find and retrieve some explicit information from the text. They can answer 2 out of 7 informational questions correctly.

A2. Math Levels

Table 32. Mastery levels cut-score and description for mathematics - Grade 1

Mastery level	Cut-scores	Description
4	496.01 and higher	Students at this level can identify non regular squares, they can group objects by build, they can find missing numbers in more difficult non-continuous series and identify patterns.
3	408.01 to 496	Students at this level can find missing numbers in a simple non-continuous series, they can perform subtractions on objects, apply a similar pattern to numbers and identify octagons.
2	281.01 to 408	Students at this level can find single missing numbers in continuous series. They can count the number of objects in a picture and perform simple operations on those objects. They can also perform additions on objects and locate objects in the front of a picture.
1	281 and lower	Students at this level can identify numbers and identify single locations of objects.

Table 33. Mastery levels cut-score and description for mathematics - Grade 2

Mastery level	Cut-scores	Description
4	513.01 and higher	At this level, students can identify the common tip on a figure, they can group objects by the dozen, identify decades in a given number and transform problems in an equation.
3	441.01 to 513	Students at this level can rank objects based on length, they can find groups of objects in a picture, they can solve simple problems demanding subtractions or more complex problems demanding additions, they can identify errors in non-continuous series of numbers, compare quantities and perform addition operations that lead to similar results as other operations.
2	345.01 to 441	At this level, students can find missing numbers in an addition operation, they can count objects and compare their numbers with another number, they can compare quantities, they can solve problems demanding simple multiplications and fill complex non-continuous series of numbers.
1	Lower to 345	Students at this level can count objects, they can also solve problems demanding additions of number by itself and identify numeric number.

Table 34. Mastery levels cut-score and description for mathematics - Grade 3

Mastery level	Cut-scores	Description
4	622.01 and higher	At this level, pupils have a thorough understanding of the positional system principles of recording numbers. They can apply that knowledge to compare numbers, when the numbers stand in place value models or/and when a digit is missing from a representation. Pupils can assign a number to the amounts given in place value models with verbal explanations (i.e. without numerical calculations). At this level, pupils can fully understand real world situations related to word problem data; separate relevant and irrelevant data from each other to identify the solution of a problem; construct a numerical expression and perform arithmetic operations to solve a real world situation problem. Pupils use properties of operations to obtain values of numerical expressions. Pupils can identify subfigures comprising a compound figure, identify their common sides and vertices.
3	524.01 to 622	At this level pupils can interpret the value of the digits in a particular place value; arrange a series of numbers in increasing and decreasing order; identify the biggest and the smallest number from given digits; find the numbers corresponding to the indicated conditions (the largest two-digit, the smallest three-digit) and perform arithmetic operations on these numbers. They can solve word problems related to calculations and arithmetic operations; select from numerical expressions an expression needed to solve a real world situation related word problem. They can recognize a pattern in a number sequence and in a correspondence represented by a table; find the omitted member in a sequence, find the preimage of an indicated element of a correspondence table. They can measure the side of a figure using a ruler and express the result in standard units; partition a graphical representation of a plane geometric figure to obtain indicated figure/figures. They can extract needed data from a table.

2	431.01 to 524	At this level, pupils can use the properties of operations in calculations and in simplification of numerical expressions; compare numbers and specify the results of the comparison; arrange numbers in increasing and decreasing order; and apply arithmetic operations to distinguish between numbers. They can find the value of the unknown component of an equality containing addition and subtraction; choose a numerical expression to find the unknown member of an equality; choose the expression needed to solve a problem; recognize the pattern in a sequence of numbers, identify the rule of extension of the pattern and find the omitted member of a sequence. They can identify geometrical figures and their elements, including non-convex polygons; create a graphical depiction of a plane figure according to the indicated instructions; partition a drawing of a plane geometric figure to obtain indicated figures. They can extract from a table data needed to solve a problem and group data by a given characteristic.
1	431 and lower	At this level, pupils can read and write three-digit numbers; find a corresponding number with the numeric name; compare three-digit numbers and write results. They can recognize a pattern for a correspondence expressed by a table (directly proportional dependence) and find the preimage of the indicated element. They can enter the data provided as a list into a prepared table.

Table 35. Mastery levels cut-score and description for mathematics - Grade 4

Mastery level	Cut-scores	Description
4	596.01 and higher	At this level, pupils have a thorough understanding of the numerical positional system principles and use that knowledge to compare numbers. They can review different possibilities to solve problems on numbers and conduct simple analysis; find biggest/smallest number from given digits; perform arithmetic operations using written algorithms (including division of a multiple digit number by a two digit number); pupils can identify, name and compare fractions given on a model; add fractions with the like denominators. They can find the value of an algebraic expression, extend the correspondence given in form of a table according to an indicated rule. They can partition a graphical representation of a plane geometric figure to obtain an indicated figure; apply properties of the rectangle and calculate the perimeter of the figure composed from rectangles; use additivity of the distance and calculate the length of a polygonal line; compose a simple algebraic expression/equality to solve a geometric problem.
3	498.01 to 596	At this level, pupils can represent a number in a place value model; assign a numerical representation to a verbally pronounced number (which does not comprise total hundreds or tens and contains several zero digits); determine a unit interval and represent numbers on a number line; use principles of positional number representation to compare numbers; select the biggest/smallest number among given numbers composed from indicated digits; use written algorithm to perform arithmetic operations; execute a written algorithm

		accurately and correctly for numbers not containing complete positional units (having zeros in their representation); solve word problems requiring calculations (arithmetic operations). They can select numerical expressions among a list of algebraic expressions to solve word problems corresponding to a real life situation; they can solve word problems relating to division and interpret value of the obtained residue using the context of the problem; extend a correspondence between two sets given in form of a table (find an omitted element) according to a verbally indicated rule. They can recognize and count faces and edges on a drawing of a spatial geometric figure; on a drawing of intersecting figures, indicate both common points and points belonging to only one of them; grasps the notions – inside, outside, all, each; name points that belong or do not belong to the indicated area. They can organize data into a table; place data in a needed place of an indicated table; extract needed information from a bar chart; solve simple problems related to proportional dependence which require calculation of a number corresponding to several units from the number corresponding to one unit.
2	414.01 to 498	At this level, pupils can: Represent a number in a place value model, interpret the value of the position of a digit in a number and apply this to compare numbers; restore a missing digit in the representation of an inequality; perform arithmetic operations (in particular, multiplication) using written algorithms; select from a list of the written algorithm the correctly executed one for addition of two numbers; perform division, find the quotient and remainder and justify the obtained answer; interpret the value of the remainder obtained from division. Use arithmetic operations on two digit numbers when solving of simple, money related problems; know money units and relationships among them; Find the value of an unknown component of an equality containing addition, subtraction, multiplication and division; select a numeric expression to find an unknown component in an equality containing division; Create the indicated figure/shape from models of plane geometric figures; partition a graphical depiction or a model of a plane geometric figure to obtain the indicated figure; on a depiction of intersecting figures, indicate both common points and points belonging to only one of them; Identify and count on a drawing of a spatial figure its elements – faces and edges and name their total number; describe spatial geometric figures; indicate adjacent/nonadjacent faces and intersecting/nonintersecting edges in a spatial figure. Extract needed information (one component) from data represented by a bar chart; determine the value of the unit interval on a chart.
1	414 and lower	At this level, pupils know and can indicate the numerical value that a digit has by its position in a number; add, subtract and multiply four-digit numbers using a written algorithm; construct an algebraic expression and use it to solve a simple problem.

Table 36. Mastery levels cut-score and description for mathematics - Grade 5

Mastery level	Cut-scores	Description
4	622.01 and higher	At this level, pupils can: Compare natural numbers (up to billion or more) using positional system and identify the result. Identify and name a fraction based on a model, find the requested fraction, use the main properties of fractions. Add/subtract/multiply fractions with like denominators; compare fractions with unlike denominators (including mixed numbers). Collect the needed information from a table, construct a bar chart, and compare two sets of data represented by a table and a chart. Understand the concept of the area of a figure, use additivity of the area, construct and apply algebraic expressions to solve problems with geometric content; classify triangles by their angles: right, acute, obtuse.
3	517.01 to 622	At this level pupils can : Conceive and name a number exceeding a million described verbally (without digits) – how many digits does it have/how many zeroes does it contain; write down a number exceeding a million described verbally (without digits); correctly use terms (numerator, denominator) for fractions; represent fractions on a number line; select a correct one from several versions of the representation of a fraction on a number line; compare and arrange in the increasing/decreasing order proper, improper and mixed fractions with unlike denominators; select from lists of fractions the one arranged in decreasing order; Perform arithmetic operations (addition, multiplication) on fractions; find a number from its fractional part and find fractional part of a number; solve word problem using operations on fractions (2-3 steps). Represent and describe dependence between quantities given on a diagram or an illustration including what influence a change in one of the quantities will have on the second quantity depending on it, use the obtained conclusions to solve problems; Find the value of a symbolic expression containing one variable and insert the obtained value into a corresponding table; choose an algebraic expression for a dependence given by a table. Identify a circle element – sector and indicate it on a drawing; describe/name position of an object on graph paper using coordinates; orient oneself on graph paper, describe – how to reach a square from a given square (e.g. two squares to the left and then one square up); Compare two sets of data represented by charts, identify the resemblances and differences between them. Obtain from a bar chart the data needed for a problem.
2	431.01 to 517	At this level, pupils can: Assign digital representation of a number exceeding a million given in the form of the sum of place values (there are several zeros in the representation of the number); compare numbers; name and record numbers bigger/smaller than an indicated number; represent fractions on a number line, assign a fraction (including mixed fractions) to a point on a number line; use fractions in a real world context, in particular, express a small unit of time

		by a large unit. Use properties of operations and simplify numerical expressions containing fractions. Perform operations on fractions – find a number from its fractional part and find fractional part of a number; solve word problems using operations on fractions (2-3 steps); Construct an algebraic expression (containing one variable) to solve a word problem; use commutativity and associativity of addition and multiplication and distributivity of multiplication over addition to simplify a symbolic expression (containing one variable); convert a number given in one unit into smaller units. Identify elements of a circle/circumference, distinguish between them and show them on a drawing; use correct terminology (center, diameter, radius, chord) related to the circle/ circumference; Use additivity of the area, calculate area of a nonrectangular figure; compare areas of figures composed from identical rectangles. Compare two datasets and select from given statements about these data a correct one; find total numbers from data represented by a bar chart.
1	Lower to 431	At this level pupils can: Classify fractions by proper and improper fractions; add and subtract proper fractions with like denominators; Select from a list of algebraic expressions an algebraic expression to solve a word problem; Indicate an element (namely, a chord) of a circle/circumference on a figure; use additivity of the area to find the area of a plane figure; Extract needed information from data organized into a bar chart; choose from datasets organized into different forms (table and a bar chart) the identical ones.

Table 37. Mastery levels cut-score and description for mathematics - Grade 6

Mastery level	Cut-scores	Description
4	638.01 to higher	At this level, pupils can: Express proper fractions by decimals and vice versa, compare and arrange in increasing/decreasing order numbers represented as proper fractions and decimals; find how the fractional representation will change if in the corresponding decimal representation some (indicated) digit is erased; relate division operation to multiplying a number by a decimal/fraction; Recognize and extend a directly proportional dependence between two quantities given by a table and/or formula; find the unknown member of a proportion; find the value of one quantity by substituting in the formula the value of a second quantity; simplify an algebraic expression containing two variables; solve a fraction related problem corresponding to a real life situation; Find and express by a formula the area of a simple figure obtained by a non-overlapping configuration of rectangles;

		Compute the arithmetic mean of data represented by a bar chart; find the value of an unknown datum from data represented by a pie chart using operations on fractions/decimals.
3	542.01 to 638	At this level, pupils can: Interpret place values of digits by their decimal positions in a decimal number representation; identify change in the value of a decimal when a digit increase or decrease; compare and arrange decimals in the increasing/decreasing order in the context of a problem; choose the valid inequality from given inequalities; perform operations (addition, subtraction, multiplication, division) on mixed fractions and decimals; Extend a given dependence between two quantities by formula and in words according to a given dependence rule; compose and simplify an algebraic expression when solving problems; Name elements of the indicated spatial figure without a drawing of the figure; recognize on a drawing spatial geometric figures by their descriptions; correctly use the terms – "all", "every", "any", "some", "one of"; Calculate the area of a rectangle with indicated side lengths in the context of a problem corresponding to a real life situation; Classify and order qualitative and quantitative data for the solution of a problem.
2	457.01 to 542	At this level, pupils can: Express a fraction by finite decimals; match a digital expression to a verbally named decimal; determine a unit interval of and represent a decimal on a given number axis; indicate decimal places and name place values of digits by their decimal places in a floating point representation of a number; multiply and divide proper fractions; find which digit must stand in place of a missing digit for an indicated inequality to hold; Choose from a list of equations/algebraic expressions the one needed to solve a word problem; solve problems on proportional dependences, find an unknown member of a proportion. Use additivity of area in calculating the area of a rectangle; compute the area of a rectangle with indicated side lengths in the context of a word problem; Calculate the arithmetic mean of three data given by a bar chart;
1	457 and lower	At this level pupils can: Express a proper fraction by a decimal; represent a decimal on a number axis; Identify on a drawing a spatial figure by a given description; indicate elements of a spatial figure on a drawing, name their total number; Calculate the area of a rectangle with given side lengths; Extract needed information from data represented by a bar chart; recognize the identity of the same data given in two different charts (by a table and a pie chart).

ANNEX VI. PROPENSITY SCORE MATCHING RESULTS

Given that selection of schools into treatment was not randomized, school characteristics may be different between treatment and comparison schools. If these characteristics affect the outcomes of interest, simply comparing the endline results between pilot and control schools would produce biased estimates of the impact of the program. Our main approach to deal with this problem is to implement the DID model. However, it is possible that baseline characteristics affect the trajectories or changes in the outcomes of interest that schools would observe in the absence of treatment, in which case the DID model could produce biased estimates of the treatment effect. In this context, it is recommendable that the DID model is modified so only control schools that are very similar at baseline to pilot schools are used as counterfactuals. For this, we implement a technique called Propensity Score Matching (PSM). In this Annex we present the main results using this approach.

The main idea behind PSM is to construct a counterfactual group for each treatment school, using the schools in the control group that are very similar in terms of the observable characteristics at baseline.

The first step that this technique requires is estimating the probability of selection into treatment. Mathematically:

$$\Pr(D = 1) = \Lambda[\alpha + \mathbf{x}'\boldsymbol{\delta} + u] \qquad (4)$$

Where $\Lambda[.]$ is the logistic function; D is a dummy variable for treatment status; $\mathbf{x}$ is a vector of school characteristics taken from EMIS data, namely total number of students (log), student/teacher ratio, percentage of teachers that are certified, class size and a full set of region dummies; u is an error term; and α and $\boldsymbol{\delta}$ are parameters to be estimated.

Once we estimate this model we can calculate the probability of selection of each school $\hat{p}$, which we use to produce weights for each control school. These weights produce pilot school-specific groups of control schools, which have very similar participation probabilities to each pilot school. By using only control schools with very similar participation probabilities as counterfactuals, it can be argued that pilot and (weighted) control schools ended up in different groups just by chance, approximating the analysis to a randomized control trail.

Table 38 presents results for DID-PSM for math Rash scores. The results are relatively similar to those obtained for the simple DID model. In this case, however, we find significant effects for 1st and 2nd graders, while for the DID the effects were positive but significant only at 10 percent. On the other hand there is no significant effect for 4th grade using DID-PSM, while for DID the effect was statistically significant.

Table 38. DID-PSM regressions for math Rasch scores

	Grades					
	1st	2nd	3rd	4th	5th	6th
Rasch score	32.9*	45.9*	41.4**	26.6	26.8	0.48
	(16.7)	(20.2)	(14.4)	(15.9)	(11.8)	(17.9)
Observations	1101	1073	1095	1082	1084	1081

Note: Participation probabilities are modeled using a logit function and as predictors total number of students (log), student/teacher ratio, percentage of teachers that are certified, class size and a full set of region dummies. Weights are calculated using an Epanechnikov kernel function.
Block-bootstrapped standard errors in parentheses (50).
* p<0.05 ** p<0.01 *** p<0.001

In Table 39 the results for PSM-DID for reading competences for grades 1 and 2 are displayed. Again, the results are not so different from the ones estimated using DID. First, no significant effects are found for 1st grade. For second grade we found positive and significant effects for all four competences, while using the DID model we only found significant effects for letter sound fluency and vocabulary.

Table 39. DID-PSM regressions for reading competences – Grades 1 and 2

	Phoneme segmenting	Syllable segmenting	Letter Sounds Fluency	Word Reading Fluency	Passage Reading fluency	Vocabulary (% correct)
A. Grade 1						
Score	4.28	1.18	2.09	2.03	N/A	N/A
	(3.35)	(1.46)	(2.03)	(1.63)		
Observations	940	940	940	940	N/A	N/A
B. Grade 2						
Score	N/A	N/A	7.37***	6.63*	6.20*	7.97*
			(2.21)	(2.65)	(2.66)	(3.61)
Observations			922	923	921	923

Note: Participation probabilities are modeled using a logit function and as predictors total number of students (log), student/teacher ratio, percentage of teachers that are certified, class size and a full set of region dummies. Weights are calculated using an Epanechnikov kernel function.
Block-bootstrapped standard errors in parentheses (50).
* p<0.05 ** p<0.01 *** p<0.001

Finally, results for grade 3 to 6 are displayed in Table 400. We only find significant effects for passage reading fluency for 3rd graders. Notably, we no longer find significantly negative effects for 6th grade vocabulary, as we did for the DID model.

Table 40. DID-PSM regressions for reading competences – Grades 3 to 6

	Passage reading fluency	Vocabulary (% correct)	Comprehension narrative text (% correct)	Comprehension informational text (% correct)
A. Grade 3				
Score	7.08*	4.34	5.44	2.68
	(3.15)	(3.18)	(3.38)	(2.82)
Observations	941	941	941	941
B. Grade 4				
Score	2.97	0.40	0.90	3.73
	(4.35)	(2.93)	(3.01)	(3.33)
Observations	934	934	934	934
C. Grade 5				
Score	5.20	2.72	-0.53	2.85
	(4.13)	(2.03)	(3.25)	(2.45)
Observations	936	938	938	938
D. Grade 6				
Score	1.26	-1.20	0.76	0.87
	(4.51)	(2.67)	(3.13)	(3.06)
Observations	929	930	930	930

Note: Participation probabilities are modeled using a logit function and as predictors total number of students (log), student/teacher ratio, percentage of teachers that are certified, class size and a full set of region dummies. Weights are calculated using an Epanechnikov kernel function.
Block-bootstrapped standard errors in parentheses (50).
* p<0.05 ** p<0.01 *** p<0.001

Overall, there seem to be no major differences between the results for DID and PSM-DID.

ANNEX VII. HETEROGENEITY

Table 41. DID estimates for math by sex, language of test and school size

	Gender		Language of test				School size		
	Males	Females	Armenian	Azeri	Georgian	Russian	1-299	300-599	≥600
Rasch score	25.3**	25.9**	0.0080	37.4	26.2***	-85.2*	33.0***	23.0	17.8
	(8.99)	(8.76)	(44.5)	(18.3)	(7.39)	(28.7)	(8.46)	(13.0)	(17.1)
Dep_var_mean	494.6	496.2	576.4	428.3	495.2	454.1	488.7	487.8	508.5
>Min requirement	0.097**	0.037	-0.091	0.25**	0.060*	-0.15	0.11***	0.044	0.028
	(0.032)	(0.032)	(0.11)	(0.081)	(0.025)	(0.14)	(0.032)	(0.050)	(0.052)
Dep_var_mean	0.57	0.56	0.76	0.38	0.57	0.41	0.53	0.55	0.63
Observations	3400	3116	323	475	5626	92	2895	1375	2246

Note: All specifications include pilot and endline dummies, a sex dummy, fixed effects for language of test region and grade, and dummies that categorize number of students, student/teacher ratio and fraction of certified teachers.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

Table 42. DID regressions for math with treatment and region interaction terms

	Rasch score	>Min Req
Abkhazeti_i	22.30	0.103
	(15.55)	(0.114)
Achara_i	50.32**	0.170*
	(17.56)	(0.0693)
Guria_i	41.69	0.00451
	(34.31)	(0.111)
Imereti_i	19.34*	0.0415
	(9.734)	(0.0659)
Kakheti_i	23.71	0.0827
	(24.62)	(0.0938)
KvemoKartli_i	35.48	0.161**
	(18.59)	(0.0485)
MtskhetaMtianeti_i	46.46***	0.144**
	(9.994)	(0.0510)
RachaLetchkhumiKvemoSvaneti_i	44.42	0.0675
	(29.12)	(0.112)
SamegreloZemoSvaneti_i	26.60*	0.0597
	(11.19)	(0.0517)
SamtskheJavakheti_i	-8.532	-0.00639
	(31.10)	(0.0916)
ShidaKartli_i	39.53	0.0531
	(25.19)	(0.0670)
Tbilisi_i	10.80	0.0180
	(24.73)	(0.0682)
Observations	6516	6516

Note: All specifications include pilot and endline dummies, a sex dummy, fixed effects for language of test, grade and region, and region-specific endline dummies.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

Table 43. DID regressions for reading competences by sex and school size – A

	Gender		School size		
	Male	Female	1-299	300-599	≥600
A. Phoneme Segmenting, Grade 1					
Raw Score	1.18	6.93	4.64	6.19	1.89
	(4.77)	(4.44)	(5.56)	(7.10)	(5.51)
Dep_var_mean	45.4	57.3	48.8	47.4	56.6
>Min requirement	0.074	0.23*	0.11	0.35*	0.091
	(0.089)	(0.10)	(0.11)	(0.16)	(0.13)
Dep_var_mean	0.29	0.51	0.37	0.47	0.40
Observations	492	448	424	191	325
B. Syllable Segmenting, Grade 1					
Raw Score	0.32	3.98	1.59	0.70	3.81
	(2.29)	(2.08)	(2.55)	(3.69)	(2.43)
Dep_var_mean	36.8	40.5	37.1	36.7	41.6
>Min requirement	-0.0054	0.15	0.042	0.13	0.065
	(0.088)	(0.094)	(0.10)	(0.14)	(0.10)
Dep_var_mean	0.63	0.74	0.62	0.67	0.77
Observations	491	449	424	191	325
C. Letter Sounds Fluency, Grades 1 and 2					
Raw Score	5.74**	3.79*	7.28***	2.31	3.06
	(2.12)	(1.81)	(2.07)	(3.07)	(2.84)
Dep_var_mean	44.6	47.3	45.0	42.0	49.3
>Min requirement	0.092	0.039	0.15*	-0.0061	0.0074
	(0.059)	(0.054)	(0.061)	(0.10)	(0.069)
Dep_var_mean	0.62	0.67	0.64	0.56	0.70
Observations	969	893	832	381	649
D. Word Fluency, Grades 1 and 2					
Raw Score	2.32	0.41	4.12*	-2.87	0.15
	(1.87)	(1.89)	(1.94)	(3.92)	(2.01)
Dep_var_mean	25.6	26.7	24.8	23.2	29.4
>Min requirement	0.041	-0.021	0.086	-0.033	-0.058
	(0.061)	(0.069)	(0.061)	(0.12)	(0.096)
Dep_var_mean	0.35	0.34	0.29	0.31	0.43
Observations	970	893	833	381	649

Note: All specifications include pilot and endline dummies, a sex dummy, fixed effects for region and grade, and dummies that categorize number of students, student/teacher ratio and fraction of certified teachers.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

Table 44. DID regressions for reading competences by sex and school size - B

	Gender		School size		
	Male	Female	1-299	300-599	≥600
A. Passage Fluency, Grades 2 - 6					
Raw Score	4.06**	-0.50	2.55	-1.35	2.80
	(1.51)	(1.53)	(1.46)	(2.46)	(2.22)
Dep_var_mean	55.9	61.5	53.8	57.9	64.8
>Min requirement	0.12***	-0.022	0.061*	0.039	0.044
	(0.031)	(0.033)	(0.030)	(0.047)	(0.040)
Dep_var_mean	0.23	0.30	0.18	0.27	0.35
Observations	2435	2226	2056	971	1634
B. Vocabulary, Grades 2 - 6					
Percent Correct	4.09*	1.33	4.21*	-1.09	3.19
	(1.59)	(1.40)	(1.81)	(1.86)	(2.10)
Dep_var_mean	60.3	61.9	58.4	60.5	64.7
>Min requirement	0.053	-0.047	0.057	-0.10*	0.00076
	(0.043)	(0.037)	(0.042)	(0.049)	(0.058)
Dep_var_mean	0.34	0.33	0.28	0.34	0.40
Observations	2437	2229	2059	973	1634
C. Comprehension Narrative Text, Grades 3 - 6					
Percent Correct	-0.40	-0.32	1.29	-3.49	-0.60
	(2.02)	(1.78)	(1.96)	(2.66)	(2.10)
Dep_var_mean	60.5	66.0	60.0	62.8	67.0
>Min requirement	0.041	-0.023	0.032	0.0017	-0.011
	(0.035)	(0.043)	(0.037)	(0.053)	(0.059)
Dep_var_mean	0.27	0.34	0.25	0.30	0.37
Observations	1959	1784	1650	783	1310
D. Comprehension Informational Text, Grades 3 - 6					
Percent Correct	-0.31	3.58	0.17	-1.72	5.35*
	(1.90)	(2.08)	(2.12)	(2.72)	(2.20)
Dep_var_mean	47.9	51.5	47.4	48.1	53.1
>Min requirement	-0.016	0.069	0.013	-0.039	0.081
	(0.035)	(0.044)	(0.039)	(0.059)	(0.054)
Dep_var_mean	0.21	0.27	0.19	0.25	0.30
Observations	1959	1784	1650	783	1310

Note: All specifications include pilot and endline dummies, a sex dummy, fixed effects for region, school size at baseline and grade.
Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001
Source: G-PriEd and EMIS data for 2013 and 2015.

Table 45. DID regressions for reading competences with treatment and region interaction terms

	Phoneme segmenting	Syllable segmenting	Letter sounds Fluency	Word reading Fluency	Passage reading fluency	Vocabulary (% correct)	Comprehension narrative text (% correct)	Comprehension informational text (% correct)
Abkhazeti_i	-23.3	19.4	1.88	15.7**	10.6	15.4*	4.53	-19.0***
	(18.5)	(13.2)	(10.8)	(5.91)	(8.19)	(6.17)	(3.95)	(3.00)
Achara_i	27.9**	3.72	4.74	1.46	4.13	4.57	4.28	3.92
	(8.79)	(4.39)	(3.32)	(3.48)	(3.51)	(3.14)	(2.71)	(4.53)
Guria_i	2.51	1.58	7.89	6.81	-8.37	10.8**	-4.78	-0.73
	(12.4)	(4.02)	(7.49)	(6.56)	(7.76)	(3.60)	(7.40)	(4.49)
Imereti_i	16.4	3.93	10.3**	2.85	0.024	3.75	-2.47	-3.39
	(12.6)	(5.12)	(3.84)	(5.24)	(3.11)	(2.50)	(3.33)	(3.57)
Kakheti_i	-9.73	-4.43	4.33	0.44	3.06	-0.0081	-1.87	2.06
	(14.8)	(6.34)	(7.78)	(4.94)	(2.66)	(3.92)	(3.40)	(4.14)
KvemoKartli_i	-3.93	-5.33	3.66	2.98	6.02*	-0.74	6.51	0.48
	(15.7)	(6.27)	(4.30)	(3.38)	(2.81)	(4.55)	(4.74)	(5.52)
MtskhetaMtianeti_i	-7.36	5.30	-2.39	-5.79	-0.43	1.47	-1.55	1.11
	(11.3)	(6.18)	(4.85)	(5.79)	(5.71)	(4.32)	(2.31)	(6.35)
RachaLetchkhumiKvemoSvaneti_i	2.40	7.82	11.9**	1.00	-0.20	-5.90	-6.21	3.41
	(9.33)	(5.83)	(4.42)	(4.84)	(4.83)	(5.98)	(5.43)	(5.87)
SamegreloZemoSvaneti_i	-7.52	3.52	5.30	0.13	0.39	5.67	4.64	7.80*
	(7.59)	(3.46)	(3.44)	(3.90)	(2.74)	(2.91)	(2.88)	(3.51)
SamtskheJavakheti_i	-2.28	-1.39	1.15	-2.04	2.92	-3.48	-8.25	4.89
	(11.2)	(3.01)	(5.50)	(4.44)	(5.91)	(2.96)	(8.55)	(3.26)
ShidaKartli_i	15.3	8.41	6.44	4.51	2.93	3.01	-1.55	-4.26
	(9.78)	(4.41)	(3.27)	(2.45)	(3.06)	(2.49)	(4.61)	(5.13)
Tbilisi_i	2.17	-1.43	0.29	-0.52	4.33*	4.03*	-0.079	4.06
	(6.20)	(1.75)	(2.33)	(2.36)	(2.04)	(1.92)	(3.04)	(3.31)
Observations	940	940	1862	1863	4661	4666	3743	3743

Note: All specifications include pilot and endline dummies, a sex dummy, fixed effects for grade and region, and region-specific endline dummies. Standard errors clustered at the school level in parentheses.
* p<0.05 ** p<0.01 *** p<0.001

ANNEX VIII. NORMS FOR PILOT AND CONTROL SCHOOLS

Norms, as defined in this report, are used to situate the performance of a specific student in comparison with the performance of a specific student population. We calculate the norms for math and reading.

For Math, we present the Rasch score, which summarizes all measured competences. For reading (both Georgian and GSL), we present raw scores for each competences. All of the figures reflect unweighted scores both baseline and endline. For Vocabulary, Comprehension-narrative, and Comprehension-informational, we present the percentile scores instead of raw scores.

Math

Table 46 shows math test scores percentiles for grades 1 to 6. For the most part, in grades 1, 5 and 6 percentiles are higher at endline than at baseline, for both treatment and control schools. For grades 2, 3 and 4, scores tend to be higher at endline than at baseline only for the upper end of the distribution.

Table 46. Math test scores percentiles by grade

	10th	25th	50th	75th	90th	N
First grade						
Control (baseline)	365	439	501	547	609	255
Pilot (baseline)	382	447	501	559	609	287
Control (endline)	377	449	505	563	630	262
Pilot (endline)	410	482	537	594	679	297
Second grade						
Control (baseline)	387	437	474	543	613	253
Pilot (baseline)	390	445	506	566	628	278
Control (endline)	363	424	508	578	644	260
Pilot (endline)	404	468	532	626	691	282
Third grade						
Control (baseline)	384	435	497	553	613	260
Pilot (baseline)	399	440	496	567	627	287
Control (endline)	363	419	491	575	650	263
Pilot (endline)	406	474	541	630	705	285
Fourth grade						
Control (baseline)	385	450	510	569	625	257
Pilot (baseline)	379	433	483	556	619	284
Control (endline)	369	425	480	577	633	259
Pilot (endline)	369	447	528	598	687	282
Fifth grade						
Control (baseline)	385	423	483	541	600	258
Pilot (baseline)	385	440	512	578	650	276
Control (endline)	380	432	499	556	631	264
Pilot (endline)	400	454	526	602	673	286
Sixth grade						
Control (baseline)	379	424	487	558	640	256
Pilot (baseline)	398	432	492	559	629	280
Control (endline)	414	451	527	601	664	261
Pilot (endline)	397	451	515	601	681	545

Source: Own calculations using GPriEd data.

Reading – Georgian as native language

In Tables 47 through 52, we show results for Grade 1 through Grade 6 by reading competency. It is important to remember that comparing between grades is not appropriate, except for the letter fluency subtask, given that this subtask is the same in all grades in which it appears.

Table 47 shows the unweighted raw score percentiles for the four competences measured in 1st grade. At endline, for phoneme segmenting, the median score is 65 and 74 points for control and pilot schools, respectively, out of a maximum of 81 points. Students at endline perform

significantly better than at endline across the whole distribution for phoneme segmenting. For syllable segmenting, students at the 75th percentile already reach the maximum score of 47 in both control and pilot schools. For letter fluency, scores show improvement from baseline to endline at the lower end of the distribution only. For word reading fluency, at endline the median score is 25 and 28 words read in one minute for students in control and pilot schools respectively; scores are higher at endline across all distribution.

Table 47. Reading measures Percentiles - Grade 1

	10th	25th	50th	75th	90th	N
Phoneme Segmenting						
Baseline (control)	6	35	56	75	81	216
Baseline (pilot)	11	34	56	78	81	242
Endline (control)	22	45	65	77	81	224
Endline (pilot)	26	52	74	80	81	258
Syllable Segmenting						
Baseline (control)	20	35	44	47	47	216
Baseline (pilot)	20	31	44	47	47	242
Endline (control)	22	34	43	47	47	224
Endline (pilot)	24	37	45	47	47	258
Letter Fluency						
Baseline (control)	27	39	51	64	65	216
Baseline (pilot)	25	39	54	65	65	242
Endline (control)	34	50	61	65	65	224
Endline (pilot)	40	57	64	65	65	258
Word Fluency						
Baseline (control)	5	13	20	28	36	216
Baseline (pilot)	9	13	22	32	42	242
Endline (control)	11	18	25	33	42	224
Endline (pilot)	10	21	28	38	45	258

Source: Own calculations using GPriEd data.

Table 48 shows raw score percentiles for 2nd graders. The test for 2nd grade times the letter fluency and word fluency tasks at 30 seconds, so the scores below show the scores obtained by the students after 30 seconds as opposed to 60 seconds (which was the case for the 1st grade test). It's important to note, again, that due to the low number of items for Vocabulary, the scores for this competency may not be highly informative.

Table 48. Reading measures Percentiles - Grade 2

	10th	25th	50th	75th	90th	N
Letter Fluency						
Baseline (control)	26	35	44	52	63	215
Baseline (pilot)	25	33	44	52	60	240
Endline (control)	27	35	42	51	58	222
Endline (pilot)	34	43	50	58	65	245
Word Fluency						
Baseline (control)	15	21	31	41	51	216
Baseline (pilot)	16	24	35	46	54	240
Endline (control)	15	25	34	43	52	222
Endline (pilot)	19	30	40	51	60	245
Passage Fluency						
Baseline (control)	15	25	35	46	55	216
Baseline (pilot)	15	26	39	48	58	238
Endline (control)	21	32	42	50	56	222
Endline (pilot)	28	36	47	55	65	245
Vocabulary						
Baseline (control)	33	42	58	75	83	216
Baseline (pilot)	25	41	58	75	83	240
Endline (control)	33	42	58	75	83	222
Endline (pilot)	41	58	66	75	91	245

Source: Own calculations using GPriEd data.

Table 49 presents the results for 3rd grade. For Passage Reading Fluency, the median score is 50 and 57 words per minute in control and pilot schools respectively.

Table 49. Reading measures Percentiles - Grade 3

	10th	25th	50th	75th	90th	N
Passage Fluency						
Baseline (control)	23	31	46	55	69	218
Baseline (pilot)	26	36	50	62	73	247
Endline (control)	31	39	50	62	78	227
Endline (pilot)	35	47	57	74	87	249
Vocabulary						
Baseline (control)	33	53	67	87	93	218
Baseline (pilot)	40	60	73	87	93	247
Endline (control)	33	53	73	80	87	227
Endline (pilot)	47	60	80	87	93	249
Comprehension, narrative						
Baseline (control)	33	44	56	78	89	218
Baseline (pilot)	33	44	67	78	89	247
Endline (control)	22	33	56	78	89	227
Endline (pilot)	33	44	67	78	100	249
Comprehension, informational						
Baseline (control)	17	17	33	67	67	218
Baseline (pilot)	17	33	50	67	83	247
Endline (control)	17	33	33	50	67	227
Endline (pilot)	17	33	50	67	67	249

Source: Own calculations using GPriEd data.

Table 50 shows raw score percentiles for 4th grade. The median score for Passage Reading Fluency is 65 and 74 words per minute for control and pilot schools respectively at endline.

Table 50. Reading measures Percentiles - Grade 4

	10th	25th	50th	75th	90th	N
Passage Fluency						
Baseline (control)	35	47	61	79	94	219
Baseline (pilot)	35	50	64	83	100	246
Endline (control)	35	51	65	85	105	222
Endline (pilot)	41	57	74	93	109	247
Vocabulary						
Baseline (control)	40	50	65	75	85	219
Baseline (pilot)	40	50	65	75	85	246
Endline (control)	40	50	65	75	85	222
Endline (pilot)	40	55	70	80	85	247
Comprehension, narrative						

Baseline (control)	36	55	73	82	91	219
Baseline (pilot)	45	64	73	82	91	246
Endline (control)	36	55	73	82	91	222
Endline (pilot)	36	55	73	91	91	247
Comprehension, informational						
Baseline (control)	17	33	50	67	83	219
Baseline (pilot)	17	33	50	67	83	246
Endline (control)	17	33	50	67	83	222
Endline (pilot)	33	33	67	83	100	247

Source: Own calculations using GPriEd data.

Table 51 presents the results for 5th grade. At endline, students in control and pilot schools at the 50th percentile score 75 and 83 words per minute in passage reading fluency. Students at the 50th percentile answer two thirds of the narrative comprehension questions and informational text comprehension questions correctly.

Table 51. Reading measures Percentiles - Grade 5

	10th	25th	50th	75th	90th	Max score
Passage Fluency						
Baseline (control)	37	57	77	96	109	218
Baseline (pilot)	40	60	82	105	119	242
Endline (control)	53	69	88	112	125	222
Endline (pilot)	54	69	88	110	123	247
Vocabulary						
Baseline (control)	40	55	65	75	85	218
Baseline (pilot)	40	55	70	80	85	243
Endline (control)	50	60	70	80	85	222
Endline (pilot)	45	55	65	80	85	247
Comprehension, narrative						
Baseline (control)	33	40	60	80	87	222
Baseline (pilot)	33	47	67	80	87	242
Endline (control)	47	60	73	87	93	222
Endline (pilot)	40	60	80	87	93	247
Comprehension, informational						
Baseline (control)	29	43	57	71	86	218
Baseline (pilot)	29	43	57	71	86	243
Endline (control)	29	43	57	71	86	222
Endline (pilot)	29	43	57	71	86	247

Source: Own calculations using GPriEd data.

Finally Table 52 shows percentile scores for the different competences at 6th grade. At endline, in the passage reading fluency task, students at the 50th percentile could read 88 words per minute for both control and pilot schools while the students at the 90th percentile could read 125 and 123 words per minute in control and pilot schools, respectively.

Table 52. Reading measures Percentiles - Grade 6

	10th	25th	50th	75th	90th	Max score
Passage Fluency						
Baseline (control)	37	57	77	96	109	218
Baseline (pilot)	40	60	82	105	119	242
Endline (control)	53	69	88	112	125	222
Endline (pilot)	54	69	88	110	123	247
Vocabulary						
Baseline (control)	40	55	65	75	85	218
Baseline (pilot)	40	55	70	80	85	243
Endline (control)	50	60	70	80	85	222
Endline (pilot)	45	55	65	80	85	247
Comprehension, narrative						
Baseline (control)	33	53	73	80	87	218
Baseline (pilot)	40	60	73	87	93	243
Endline (control)	47	60	73	87	93	222
Endline (pilot)	40	60	80	87	93	247
Comprehension, informational						
Baseline (control)	29	43	57	71	86	218
Baseline (pilot)	29	43	57	71	86	243
Endline (control)	29	43	57	71	86	222
Endline (pilot)	29	43	57	71	86	247

Source: Own calculations using GPriEd data.

Reading – Georgian as a Second Language

Tables 53-58 present percentile scores for Reading in Georgian as a Second Language. It is important to highlight that norms in this case are based on significantly smaller samples. In the context of presenting norms, the effect of having small samples is that results are more unstable, so if we were to resample schools we may found very different test score percentiles. For this reason, we don't find it informative to provide conclusions derived from these results.

Table 53. Reading (GSL) measures Percentiles - Grade 1

	10th	25th	50th	75th	90th	N
Phoneme						
Baseline (control)	0	18	30	34	40	36
Baseline (pilot)	0	3	24	35	38	35
Endline (control)	15	37	57	71	72	38
Endline (pilot)	26	51	62	71	72	39
Syllable Segmenting						
Baseline (control)	0	23	36	57	70	36
Baseline (pilot)	0	0	34	51	67	35
Endline (control)	9	26	36	39	42	38
Endline (pilot)	6	28	38	43	44	39
Letter Fluency						
Baseline (control)	9	15	21	36	44	36
Baseline (pilot)	5	10	17	30	43	35
Endline (control)	7	13	21	33	50	38
Endline (pilot)	5	12	25	41	55	39
Word Fluency						
Baseline (control)	3	7	12	16	20	36
Baseline (pilot)	1	7	11	15	17	35
Endline (control)	4	9	14	20	24	38
Endline (pilot)	4	9	14	18	26	39

Source: Own calculations using GPriEd data.

Table 54. Reading (GSL) measures Percentiles - Grade 2

	10th	25th	50th	75th	90th	N
Phoneme						
Baseline (control)	6	23	29	37	42	36
Baseline (pilot)	16	27	33	38	41	35
Endline (control)	35	56	66	80	82	38
Endline (pilot)	42	60	71	82	82	37
Syllable						
Baseline (control)	10	46	56	69	80	36
Baseline (pilot)	43	55	61	71	82	35
Endline (control)	25	30	37	43	45	38
Endline (pilot)	20	26	36	42	45	37
Letter fluency						
Baseline (control)	13	21	29	35	46	36
Baseline (pilot)	13	20	24	29	42	35
Endline (control)	18	23	33	44	55	38
Endline (pilot)	19	27	31	45	60	37
World fluency						
Baseline (control)	5	8	14	22	28	36
Baseline (pilot)	3	7	10	16	23	35
Endline (control)	5	9	13	19	27	38
Endline (pilot)	7	9	14	22	32	37
Passage fluency						
Baseline (control)	8	14	18	25	31	36
Baseline (pilot)	7	9	15	20	24	35
Endline (control)	14	17	21	30	35	38
Endline (pilot)	15	15	23	30	35	37

Source: Own calculations using GPriEd data.

Table 55. Reading (GSL) measures Percentiles - Grade 3

	10th	25th	50th	75th	90th	N
Word Fluency						
Baseline (control)	11	15	21	28	36	38
Baseline (pilot)	7	14	19	26	30	37
Endline (control)	12	19	26	33	38	36
Endline (pilot)	9	20	26	29	38	36
Passage Fluency						
Baseline (control)	8	12	18	29	53	38
Baseline (pilot)	6	10	17	22	32	37
Endline (control)	17	19	27	37	42	36
Endline (pilot)	14	19	27	31	37	36
Vocabulary						
Baseline (control)	30	40	60	70	80	36
Baseline (pilot)	10	30	40	60	70	36
Endline (control)	30	40	60	70	80	36
Endline (pilot)	10	30	40	60	70	36
Comprehension, narrative						

Baseline (control)	0	40	80	80	100	36
Baseline (pilot)	0	20	60	80	100	36
Endline (control)	0	40	80	80	100	36
Endline (pilot)	0	20	60	80	100	36
Comprehension, informational						
Baseline (control)	25	50	75	100	100	36
Baseline (pilot)	0	25	50	75	75	36
Endline (control)	25	50	75	100	100	36
Endline (pilot)	0	25	50	75	75	36

Source: Own calculations using GPriEd data.

Table 56. Reading (GSL) measures Percentiles - Grade 4

	10th	25th	50th	75th	90th	N
Word Fluency						
Baseline (control)	16	30	36	45	49	38
Baseline (pilot)	16	25	33	45	51	38
Endline (control)	25	29	34	49	77	36
Endline (pilot)	22	27	30	43	66	35
Passage Fluency						
Baseline (control)	13	17	23	36	49	38
Baseline (pilot)	6	15	22	37	56	38
Endline (control)	30	50	65	75	90	36
Endline (pilot)	30	30	50	70	80	35
Vocabulary						
Baseline (control)	20	50	60	80	100	38
Baseline (pilot)	20	40	55	70	90	38
Endline (control)	30	50	65	75	90	36
Endline (pilot)	30	30	50	70	80	35
Comprehension, narrative						
Baseline (control)	29	43	71	86	86	38
Baseline (pilot)	14	29	43	71	100	38
Endline (control)	14	43	57	79	86	36
Endline (pilot)	29	43	57	71	86	35
Comprehension, informational						
Baseline (control)	20	60	80	100	100	38
Baseline (pilot)	20	40	60	100	100	38
Endline (control)	20	40	80	100	100	36
Endline (pilot)	40	60	80	80	100	35

Source: Own calculations using GPriEd data.

Table 57. Reading (GSL) measures Percentiles - Grade 5

	10th	25th	50th	75th	90th	N
Passage Fluency						
Baseline (control)	16	24	34	47	81	37
Baseline (pilot)	8	16	26	40	54	37
Endline (control)	23	28	48	58	79	39
Endline (pilot)	19	32	41	50	58	37
Vocabulary						
Baseline (control)	27	40	60	80	87	37
Baseline (pilot)	20	33	53	67	87	37
Endline (control)	20	40	53	73	80	39
Endline (pilot)	20	33	47	67	87	37
Comprehension, narrative						
Baseline (control)	25	38	50	63	88	37
Baseline (pilot)	25	38	50	50	63	37
Endline (control)	25	25	38	63	75	39
Endline (pilot)	13	38	63	75	88	37
Comprehension, informational						
Baseline (control)	17	33	50	83	100	37
Baseline (pilot)	17	33	50	83	100	37
Endline (control)	33	33	50	83	100	39
Endline (pilot)	17	50	67	83	100	37

Source: Own calculations using GPriEd data.

Table 58. Reading (GSL) measures Percentiles - Grade 6

	10th	25th	50th	75th	90th	N
Passage Fluency						
Baseline (control)	6	16	35	51	65	38
Baseline (pilot)	11	24	34	48	77	36
Endline (control)	32	40	51	65	103	39
Endline (pilot)	26	40	51	62	69	37
Vocabulary						
Baseline (control)	27	33	60	73	87	38
Baseline (pilot)	13	33	53	83	93	36
Endline (control)	33	47	60	80	93	39
Endline (pilot)	33	53	60	73	80	37
Comprehension, narrative						
Baseline (control)	25	38	63	75	100	38
Baseline (pilot)	13	38	63	88	88	36
Endline (control)	38	50	63	88	100	39
Endline (pilot)	25	38	63	75	100	37

Comprehension, informational						
Baseline (control)	17	33	67	83	100	38
Baseline (pilot)	0	33	50	67	83	36
Endline (control)	17	33	50	83	83	39
Endline (pilot)	17	33	50	67	83	37

Source: Own calculations using GPriEd data.

ANNEX IX. ENDLINE SUMMARY STATISTICS BY CATEGORIES OF INTEREST

This section presents endline summary statistics by gender, school size, language of the test (for math), and region. Data for math and reading test scores are presented. Reading is divided between schools where Georgian is the native language and schools where Georgian is the second language (GSL). We present the raw score data for all competencies, except for Vocabulary, Comprehension-Narrative, and Comprehension-Informational, in which the percentage correct is reported. For Math subject, we present a single score, the Rasch score, and for Reading (both Georgian and GSL) we present the results for each outcome measure (competency) separately using raw scores. In all of the tables presented below, we use stars (*) at the top value of each disaggregation to indicate the level of statistical significance of the difference (*** 0.1%, ** 1% and * 5%).

Math

Table 59 presents mean math test scores. The first two columns correspond to first grade: the first column displays the number of observations for each group, and the second column the mean test scores. The rest of the columns show the same data for the remaining five grades. For each grade, we conducted t-tests or ANOVA tests to evaluate the significance of the differences between the different groups.

The data shows certain patterns that are worth highlighting. First, pilot schools have a higher mean test score than control schools at all grades. Differences in mean scores by gender vary across grades. Females obtain higher test scores at 1st, 2nd, 3rd, 5th and 6th grade, but the difference is only significant for 5rd grade. Males have higher mean scores at 4th grade, this difference is not statistically significant.

There seems to be a strong relationship between school size and test scores. For all grades schools with 600 students or more observe higher test scores on average than smaller schools. Moreover, schools with less than 300 students observe the lowest mean test scores at grades 3 to 6. The differences by school size are significant at all grades except for 1st and 2nd grade.

As with school size differences, the ones observed by language of test are apparent. Those answering the exam in Armenian outperform students answering in Georgian at all grades. Students answering the exam in Azeri have higher mean test scores than students answering in Georgian at all but 5th grades. On the other hand, students answering the exam in Russian outperform students answering the exam in Georgian at all grades but 5th grade. According to the ANOVA test, these differences are significant at all but 4th grades. Finally, we find that the differences between regions are also statistically significant, except for 4th grade.

Table 59. Mean Math score by category and t/ANOVA Tests for difference in means

Source: Own calculations using GPriEd data

	First grade		Second grade		Third grade		Fourth grade		Fifth grade		Sixth grade	
	Number of students	Rasch score	Number of students	Rasch score	Number of students	Rasch score	Number of students	Rasch score	Number of students	Rasch score	Number of students	Rasch score
School type												
Control	262	502.5**	260	505.4**	263	495.5**	259	495.6**	264	502.9**	261	527.7
Pilot	297	542.4	282	547.8	285	551.9	282	526.2	286	533.3	284	533.5
Gender												
Female	257	532.7	273	533.1	258	528	246	504.8	266	529.7*	265	532.4
Male	302	516.0	269	521.6	290	522.0	295	517.2	284	508.4	280	529.0
School size												
1-299	256	519.5	240	522.7	243	499.7**	241	489.6**	242	503.9**	239	512.7**
300-599	115	517.5	115	510.4	115	501.2	112	499.3	119	516.6	118	531.4
>=600	188	533.2	187	543.9	190	571.3	188	547.1	189	539.1	188	553.0
Language of test												
Georgian	482	516.3**	467	521.4*	476	523.8**	470	514.5	474	522.9*	469	534.2**
Azeri	40	551.3	39	550.0	40	484.7	40	470.4	41	464.7	41	464.4
Armenian	27	606.3	27	600.6	28	585.1	27	517.9	26	538.2	26	562.9
Russian	10	548.6	9	525.3	4	635.0	4	536.3	9	485.7	9	556.6
Region												
Abkhazeti	11	520**	9	446**	10	443.6**	9	528.4	11	479**	10	499.7**
Achara	48	511.0	47	478.4	47	509.2	47	500.0	46	497.5	47	489.0
Guria	21	561.6	21	549.9	21	529.0	21	509.7	21	528.6	21	536.5
Imereti	68	545.6	66	543.2	67	533.4	66	519.8	66	527.9	67	563.1
Kakheti	66	535.2	65	517.6	67	494.3	65	478.1	67	493.9	66	504.2
Kvemo_Kartli	57	589.6	56	561.8	54	539.5	55	536.6	57	532.2	57	526.4
Mtskheta_Mtianeti	26	444.8	25	508.6	25	504.8	25	431.4	25	476.0	26	536.3
Racha_Letchkhumi_n_Kvemo_Svaneti	17	553.9	19	560.8	18	516.6	18	493.7	19	520.7	18	525.7
Samegrelo_n_Zemo_Svaneti	61	512.4	64	483.6	62	496.0	62	500.5	63	513.8	62	520.7
Samtskhe_Javakheti	49	540.8	49	550.2	52	537.8	49	482.6	50	496.3	48	523.9
Shida_Kartli	36	511.1	34	538.9	36	482.7	36	534.5	36	531.6	36	532.0
Tbilisi	99	480.3	87	543.2	89	585.8	88	560.3	89	558.0	87	562.9

Source: Own calculations using GPriEd data

Reading

Table 60 presents mean reading test scores for different subgroups of interest for 1st grade. In the first column the number of observations corresponding to each group are presented, and the remaining four columns correspond to the four competences measured at this grade. Pilot schools outperform control schools in all competences. The differences are significant for all competences but Syllable Segmenting. With respect to gender we can see that females outperform males in every competency; moreover, all these differences are statistically significant. The differences in means by school size are significant for Syllable Segmenting and Word Fluency. We can also see that schools with 600 students or more outperform the other schools in all competences and that schools with 300-599 students outperform the schools with less than 300 students in Syllable Segmenting, Letter Fluency, and Word Fluency. Regarding differences between regions, these are also significant for all competences.

Table 60: Mean raw scores by category and t/ANOVA tests for difference in means, Grade 1

	Number of students	Significance			
		Phoneme	Syllable	Letter fluency	Word fluency
School type					
Control	224	57.8**	38.2	55**	25.6*
Pilot	258	63.5	39.8	58.1	27.8
Gender					
Female	230	63.6**	40.4**	58*	28.2*
Male	252	58.3	37.9	55.4	25.5
School size					
1-299	221	59.7	37.5*	55.5	25.2*
300-599	97	59.5	40.0	56.8	27.7
>=600	164	63.2	40.6	58.2	28.4
Region					
Abkhazeti	11	64.3**	37.5*	58**	23.5**
Achara	48	57.3	38.8	57.9	29.8
Guria	21	52.5	38.4	53.0	28.9
Imereti	68	63.3	40.5	59.9	28.9
Kakheti	44	46.0	35.2	46.6	21.5
Kvemo_Kartli	27	69.5	44.3	59.8	32.1
Mtskheta_Mtianeti	26	57.9	35.2	53.2	25.4
Racha_Letchkhumi_n_Kvemo_Svaneti	17	71.9	43.6	61.8	28.8
Samegrelo_n_Zemo_Svaneti	61	67.7	39.3	59.0	25.3
Samtskhe_Javakheti	24	57.6	35.1	53.1	25.8
Shida_Kartli	36	67.2	39.5	56.2	30.6
Tbilisi	99	59.5	39.9	57.7	24.4

Source: Own calculations using GPriEd data

Table 61 shows results for 2nd grade. Pilot schools outperform control across all competences. Females outperform males in all competences and the results are statistically significant for Word Fluency and Passage Fluency. For school size, schools with 600 students or more outperform the rest of the schools in all competences; also, schools with 300-599 students

have higher outcomes than schools with less than 300 students except for Letter Fluency. With respect to differences between regions, we can observe that these are statistically significant, except for Passage Fluency.

Table 61. Mean raw scores by category and t/ANOVA tests for difference in means, Grade 2

	Number of students	Significance			
		Letter fluency	Word fluency	Passage fluency	Vocabulary
School type					
Control	222	42.4**	34.1**	40.2**	58.3**
Pilot	245	49.5	40.3	46.4	67
Gender					
Female	235	47	39.6**	46.1**	64.4
Male	232	45.2	35.1	40.7	61
School size					
1-299	207	47.1	35.6	41.5*	61.2
300-599	97	45.2	37.6	42.7	62.7
>=600	163	45.3	39.5	46.4	64.7
Region					
Abkhazeti	9	42**	23.1**	29.3	51.9**
Achara	47	45.7	39.0	42.8	56.6
Guria	21	44.7	36.6	42.0	57.9
Imereti	66	44.9	38.5	43.8	63.6
Kakheti	44	39.1	31.4	41.7	65.9
Kvemo_Kartli	27	51.0	44.8	51.1	71.9
Mtskheta_Mtianeti	25	43.8	35.1	41.9	55.0
Racha_Letchkhumi_n_Kvemo_Svaneti	19	49.4	38.5	46.2	75.0
Samegrelo_n_Zemo_Svaneti	64	51.9	37.0	41.5	58.7
Samtskhe_Javakheti	24	43.6	33.3	40.3	64.6
Shida_Kartli	34	53.7	39.7	45.0	65.2
Tbilisi	87	43.3	39.0	45.4	64.3

Source: Own calculations using GPriEd data

Table 62 presents results for 3rd grade. Pilot schools outperform control schools in all competences and the differences are significant for all but Comprehension-Informational. Also, females outperform males in all competences, and the differences are significant for Passage Fluency and Comprehension-narrative. For school size the differences are significant for all competences: the largest schools have the highest mean test scores on average for all competences. However, the differences between middle-size school and those with fewer than 300 students are more ambiguous. For the analysis by region, again, we can see that the differences are statistically significant.

Table 62. Mean raw scores by category and t/ANOVA tests for difference in means, Grade 3

	Number of students	Significance			
		Passage fluency	Vocabulary	Comp, narrative	Comp, info
School type					

Control	227	51.7**	66**	56.9**	41.8
Pilot	249	59.9	73.6	65.2	44.4
Gender					
Female	222	60.5**	71.6	64.7**	44.5
Male	254	52.1	68.6	58.3	42.0
School size					
1-299	210	53.5**	67.4**	58.4**	40.3**
300-599	100	53.3	65.0	56.6	40.8
>=600	166	60.8	76.2	67.7	48.2
Region					
Abkhazeti	10	32.6**	58.7**	53.3**	41.7**
Achara	47	56.4	69.8	55.6	39.0
Guria	21	54.8	71.1	59.3	33.3
Imereti	67	56.6	70.3	60.2	42.5
Kakheti	46	47.6	67.5	57.7	40.6
Kvemo_Kartli	29	61.8	71.0	62.5	50.0
Mtskheta_Mtianeti	25	50.6	62.1	50.2	38.0
Racha_Letchkhumi_n_Kvemo_Svaneti	18	55.6	71.1	62.3	41.7
Samegrelo_n_Zemo_Svaneti	62	55.3	68.0	61.6	39.0
Samtskhe_Javakheti	26	52.7	65.9	54.3	37.8
Shida_Kartli	36	56.8	63.7	59.0	41.7
Tbilisi	89	63.5	78.8	73.4	54.3

Source: Own calculations using GPriEd data

Table 63 presents mean reading test scores for different subgroups for 4th grade. In this case, pilot schools obtain higher mean test scores than control schools, and the differences are statistically significant in all competences but Comprehension-narrative. Females outperform males also at all grades, but in this case, the difference is only statistically significant for Passage Fluency. The differences by school size are also significant, except for Vocabulary. Schools with 600 students or more outperform the other schools, but it is not always the case that schools with 300-599 students outperform schools with less than 300 students. In other words, it does not seem that the relationship between school size and test scores is monotonic. Regarding differences between regions, these are significant for all competences except for Comprehension-Narrative.

Table 63. Mean raw scores by category and t/ANOVA tests for difference in means, Grade 4

	Number of students	Significance			
		Passage fluency	Vocabulary	Comp, narrative	Comp, info
School type					
Control	222	68.6*	62.5*	65.5	55.3*
Pilot	247	73.6	65.9	69.3	60.8
Gender					
Female	202	76.6**	64.7	69	60.1
Male	267	67.2	63.9	66.4	56.8
School size					

1-299	209	66.5**	63	65.3**	56.6**
300-599	97	68.7	63.0	64.1	52.1
>=600	163	78.8	66.6	72.3	63.9
Region					
Abkhazeti	9	73**	67.8*	72.7	68.5**
Achara	47	69.9	62.3	65.4	53.9
Guria	21	88.3	66.0	71.4	50.0
Imereti	66	67.0	65.5	65.3	54.3
Kakheti	44	62.9	60.8	62.8	58.7
Kvemo_Kartli	30	81.4	68.3	64.8	63.9
Mtskheta_Mtianeti	25	66.2	55.4	62.9	45.3
Racha_Letchkhumi_n_Kvemo_Svaneti	18	75.2	72.5	68.2	61.1
Samegrelo_n_Zemo_Svaneti	62	74.5	62.3	66.6	61.3
Samtskhe_Javakheti	24	60.4	57.7	67.4	40.3
Shida_Kartli	36	69.4	65.3	68.2	57.4
Tbilisi	87	73.6	67.5	73.7	68.4

Source: Own calculations using GPriEd data

Table 64 displays results for 5th grade. Pilot schools obtain higher mean test scores than control schools, although the differences are only significant for Passage-Fluency and Vocabulary. Females statistically and significantly outperform males in all competences. In terms of school size, the results indicate that the bigger the school, the higher the mean score and these differences are significant for all competences but Comprehension-Informational. Schools with less than 300 students obtain the lowest average test scores for all competences but Comprehension-Informational, while schools with 600 students or more outperform all other schools in all competences. Differences by region are significant at this grade for Passage Fluency and Vocabulary only.

Table 64. Mean raw scores by category and t/ANOVA tests for difference in means, Grade 5

	Number of students	Significance			
		Passage fluency	Vocabulary	Comp, narrative	Comp, info
School type					
Control	225	76.2**	57.4**	63.8	53.9
Pilot	249	83.9	62.7	64.7	57.1
Gender					
Female	230	87.3**	63.2**	69**	59.5**
Male	244	73.6	57.3	59.8	51.9
School size					
1-299	209	74.1**	56.6**	60.8**	54.7
300-599	101	81.4	58.9	63.2	54.3
>=600	164	87.3	65.5	69.3	57.5
Region					
Abkhazeti	11	78.4**	51.4*	55.8	50.6
Achara	46	75.2	51.6	61.9	48.1

Guria	21	103.2	65.5	71.1	65.3
Imereti	66	81.2	62.6	60.8	54.5
Kakheti	45	73.7	60.2	63.4	58.7
Kvemo_Kartli	27	86.9	61.5	62.7	56.1
Mtskheta_Mtianeti	25	75.1	54.2	56.8	57.7
Racha_Letchkhumi_n_Kvemo_Svaneti	19	81.5	59.2	58.9	51.1
Samegrelo_n_Zemo_Svaneti	63	77.3	59.7	68.1	56.5
Samtskhe_Javakheti	26	84.8	60.8	62.8	54.9
Shida_Kartli	36	73.8	59.0	64.3	51.6
Tbilisi	89	82.8	64.6	69.4	58.3

Source: Own calculations using GPriEd data

Finally, Table 65 displays results for 6th grade. The differences in mean test scores are mixed between pilot and control schools. While pilot schools outperform control schools in Comprehension-Narrative and Comprehension-Informational, control schools have higher mean scores in Passage Fluency and Vocabulary. However, none of these differences are statistically significant. Females statistically and significantly score higher than males in all competences but Comprehension-Informational. Differences by school size are also significant for all competences, in which the larger the schools the higher the scores. Differences in mean test scores by region are only statistically significant for Comprehension-Narrative.

Table 65. Mean raw scores by category and t/ANOVA tests for difference in means, Grade 6

	Number of students	Significance			
		Passage fluency	Vocabulary	Comp, narrative	Comp, info
School type					
Control	222	89.4	69	71.7	58
Pilot	247	88.6	66.3	72.8	60.1
Gender					
Female	230	96.4**	70.1**	75.9**	59.1
Male	239	81.8	65.1	68.8	59.1
School size					
1-299	206	85**	65.3*	68.8**	55.8**
300-599	99	87.7	68.6	72.2	60.5
>=600	164	94.8	69.8	76.7	62.5
Region					
Abkhazeti	10	75.3	63.5	64.7*	50
Achara	47	88.7	63.6	70.9	53.8
Guria	21	97.7	63.1	70.8	56.5
Imereti	67	87.0	67.1	69.2	62.9
Kakheti	44	84.8	66.5	68.3	55.8
Kvemo_Kartli	27	91.8	69.8	69.9	58.7
Mtskheta_Mtianeti	26	85.2	69.6	74.1	61.0
Racha_Letchkhumi_n_Kvemo_Svaneti	18	88.1	68.1	73.7	61.9
Samegrelo_n_Zemo_Svaneti	62	90.0	67.1	71.9	57.4
Samtskhe_Javakheti	24	79.4	65.8	66.9	57.7
Shida_Kartli	36	85.4	67.5	74.1	62.3
Tbilisi	87	96.1	71.6	79.5	61.7

Source: Own calculations using GPriEd data

Reading Georgian as a Second Language

In Table 66, mean test scores for GSL are displayed for 1st grade. In this case, pilot schools outperform control schools in all competences but for Word Fluency. However the differences are not statistically significant. The lack of statistical significance could be a consequence of the small sample size that we have for these students. Females obtain higher mean test scores than males, but the differences are not significant except for Word Fluency. Differences between school sizes are not significant. Differences between the three regions with schools with GSL are statistically significant with the exception of Phoneme Segmenting.

Table 66. Mean raw scores (GSL) by category and t/ANOVA tests for difference in means, Grade 1

	Number of students	Significance			
		Phoneme	Syllable	Letter fluency	Word fluency
School type					
Control	38	50.6	31.1	24.9	14.4
Pilot	39	55.1	32.4	26.8	14.1
Gender					
Female	27	58.6	33.7	29.1	16.5*
Male	50	49.8	30.7	24.2	13.0
School size					
1-299	35	51.4	31.7	29.7	14.7
300-599	18	54.3	35.3	24.2	14.6
>=600	24	54.0	29.2	21.6	13.3
Region					
Kakheti	22	58.1	26.6**	18.3**	11.5*
Kvemo_Kartli	30	49.6	32.0	26.6	14.7
Samtskhe_Javakheti	25	52.3	35.9	31.7	16.1

Source: Own calculations using GPriEd data

Table 67 presents the analysis for 2nd grade. Pilot schools obtain higher mean test scores than control schools for all competences, except for Phoneme Segmenting; however, none of the differences is statistically significant. Females outperform males in all competences and the results are significant for Letter Fluency, World Fluency, and Passage Fluency. For this grade, schools in the middle-size school (300-599 students) outperform smaller and bigger schools but the differences are only significant for Syllable and Phoneme Segmenting and Letter Sound Fluency. Differences by region are significant for all outcome measures, with the exception of World Fluency.

Table 67. Mean raw scores (GSL) by category and t/ANOVA tests for difference in means, Grade 2

	Number of students	Significance				
		Syllable	Phoneme	Letter Fluency	Word Fluency	Passage Fluency
School type						
Control	38	64	34.8	34.4	14.7	23.9
Pilot	37	66.8	32.9	35.0	15.9	24.0
Gender						
Female	38	66.6	34.4	37.9**	17.7**	26.3*
Male	37	64.1	33.3	31.4	12.8	21.5
School size						
1-299	33	59.6*	28.8***	29.6*	13.2	22.6
300-599	18	71.4	38.9	39.4	18.4	27.8
>=600	24	68.8	37.0	38.1	15.9	23.0
Region						

Kakheti	21	75.4**	38.1**	39.3*	13.8	20.2**
Kvemo_Kartli	29	60.6	30.1	33.0	14.6	25.6
Samtskhe_Javakheti	25	62.6	34.6	32.6	17.3	25.3

Source: Own calculations using GPriEd data

In Table 68, mean test scores for GSL are displayed for 3rd grade. Control schools obtain higher test scores than pilot schools but the differences are statistically significant only for Vocabulary and Comprehension-Informational. Differences by sex are not statistically significant. There is no clear relationship between school size and test scores. Differences by region are significant for three of the five competences.

Table 68. Mean raw scores (GSL) by category and t/ANOVA tests for difference in means, Grade 3

	Number of students	Significance				
		Word fluency	Passage fluency	Vocabulary	Comp, narrative	Comp, info
School type						
Control	36	25.7	29.3	54.4**	62.8	68.8**
Pilot	36	24.3	26.7	40.8	52.2	48.6
Gender						
Female	36	24.6	26.9	47.5	61.1	59.7
Male	36	25.4	29.1	47.8	53.9	57.6
School size						
1-299	33	25.8	28.5	54.2	53.3	59.8
300-599	15	24.9	28.0	44.0	53.3	66.7
>=600	24	24.0	27.3	40.8	65.8	52.1
Region						
Kakheti	21	24	24.3	41.4**	51.4*	48.8**
Kvemo_Kartli	25	24.6	29.7	42.0	51.2	52.0
Samtskhe_Javakheti	26	26.2	29.4	58.1	68.5	73.1

Source: Own calculations using GPriEd data

Table 69 presents the analysis for 4th grade. Control schools outperform pilot schools across all competences but Comprehension-Informational. In all competences, females obtain higher test scores than males, and the differences are significant for Comprehension-Narrative. No clear pattern can be drawn from the results by school size. On the other hand, differences between regions are significant for Word Fluency, Passage Reading Fluency, and Comprehension-Informational.

Table 69. Mean raw scores (GSL) by category and t/ANOVA tests for difference in means, Grade 4

	Number of students	Significance				
		Word fluency	Passage fluency	Vocabulary	Comp, narrative	Comp, info
School type						
Control	36	38.7	39.9	60.6	56.3	67.2
Pilot	35	34.9	36.3	53.7	52.2	68.6
Gender						
Female	43	39	41	60	62.1**	71.2

Male	28	33.5	33.7	52.9	42.3	62.9
School size						
1-299	32	35.5	35.2	59.1	53.6	71.3
300-599	15	39.7	41.5	59.3	58.1	66.7
>=600	24	36.8	40.0	53.3	53.0	64.2
Region						
Kakheti	21	33.2*	33.5*	52.9	46.9	57.1**
Kvemo_Kartli	25	40.6	43.2	62.0	57.7	76.8
Samtskhe_Javakheti	25	36.1	37.0	56.0	57.1	68.0

Source: Own calculations using GPriEd data

Table 70 displays results for 5th grade. Control schools outperform pilot schools in Passage Fluency and Vocabulary. In three if the four comptences, females obtain higher test scores than males, but no difference is statistically significant. No clear pattern in terms of the differences in mean test scores by school size emerges. For Vocabulary, the difference by region is statistically significant.

Table 70. Mean raw scores (GSL) by category and t/ANOVA tests for difference in means, Grade 5

	Number of students	Significance			
		Passage fluency	Vocabulary	Comp, narrative	Comp, info
School type					
Control	39	48.5	53.3	43.9	59.8
Pilot	37	42.9	51.4	52.4	61.3
Gender					
Female	36	45.4	53.5	51.7	65.3
Male	40	46.1	51.3	44.7	56.3
School size					
1-299	33	44.2	55.2	45.8	53.5
300-599	18	50.6	52.2	45.8	68.5
>=600	25	44.4	48.8	52.5	64.0
Region					
Kakheti	22	40.6	42.1**	41.5	60.6
Kvemo_Kartli	30	50.9	54.4	52.5	58.9
Samtskhe_Javakheti	24	44.2	59.2	48.4	62.5

Source: Own calculations using GPriEd data

Finally, results for 6th grade are presented in Table 71. None of the differences between pilot and control schools is significant at this grade. Differences by gender are significant for Passage Fluency and Vocabulary, in which females outperform males. When looking at the mean scores by school size, the differences are not significant in any competency. Difference between regions is significant for Comprehension- Narrative.

Table 71. Mean raw scores (GSL) by category and t/ANOVA tests for difference in means, Grade 6

	Number of students	Significance			
		Passage fluency	Vocabulary	Comp, narrative	Comp, info
School type					
Control	39	56.2	62.9	64.4	51.7
Pilot	37	50.8	60.4	59.8	52.3
Gender					
Female	35	59.1*	67.2*	66.4	55.2
Male	41	49.0	56.9	58.5	49.2
School size					
1-299	33	52.8	63.6	62.5	54
300-599	19	56.6	59.3	60.5	43.9
>=600	24	52.3	60.8	63.0	55.6
Region					
Kakheti	22	47.5	56.7	53.4**	48.5
Kvemo_Kartli	30	57.4	60.9	62.5	52.8
Samtskhe_Javakheti	24	54.5	67.2	69.8	54.2

Source: Own calculations using GPriEd data

ANNEX X. SCOPE OF WORK

T P! Q I CAUCASUS

Issuance Date:	October 30, 2013
RFTOP Clarification Questions Due:	November 12, 2013
Closing Date:	December 13, 2013
Closing Time:	10:00AM, Tbilisi Time

To: Mendez, England & Associates

Reference: Mission Evaluation Mechanism lndefinite Quantity Contract (IQC) AID-114-I-13-00001

Subject: Request for Task Order Proposal (RFTOP) No.SOL-114-14-000002, External Impact Evaluation of the Georgia Primary Education (G-PriEd) Project.

Enclosed is a Request for Proposal for a Task Order to be issued under the referenced IQC to implement the attached Statement of Work (SOW).

The anticipated start date of o/a fall 2013 and end date of fall 2015.

Attached you will find the following documents:

1. Statement of Work (Attachment I)
2. Instructions for Technical and Cost Proposal (Attachment 2)
3. Special Requirements (Attachment 3)
4. Background Information (Attachment 4)

Accurate and Complete Information: The offeror must set forth full, accurate and complete information as required by this RFTOP. The penalty for making false statements to the Government is prescribed in 18 U.S.C. 1001.

Offer Acceptability: The Government may determine an offer to be unacceptable if the offer does not comply with all of the terms and conditions of the RFTOP.

Proposal Preparation Costs: The U.S. Government will not pay for any proposal preparation costs.

Please submit your proposal to implement this activity by the specified closing date of December 13,2013. The proposal must be submitted via e-mail to me at and

- All electronic files containing the proposal must be compatible with MS Word and MS Excel. Pages containing original signatures must be sent via PDF file.

11 George Balanchine Street
Tbilisi 0131, Georgia
Tel: (995 32) 254 4000
Fax: (995 32) 254 4145
georgia.usaid.gov

Please address any questions you may have to Irina Bakradze via email at ibakradze@usaid.gov and _____ no later than November 12, 2013.

Please acknowledge receipt of this e-mail.

, Jncerely,

Sarah R Bueter
Regional Contracting Officer

ATTACHMENT 1

STATEMENT OF WORK
EXTERNAL IMPACT EVALUATINO OF GEORGIA PRIMARY EDUCATION PROJECT
PILOT PHASE
STUDENTS' OUTCOMES IN READING AND MATH

I. Scope

Project Name: Georgia Primary Education Project (GPriEd)
Project Number: AID-114-C-11-00003
Project Dates: September 2011 – September 2016
Project Funding: $ 8,765,635
Implementing organization/s: Chemonics International

The contractor shall ensure that the evaluation team:

- Gets acquainted with the approved design, methodology, instruments, and approaches to the data analysis of the Pilot Impact Study for the G-PriEd project, provided by USAID; gets acquainted with the national norms and standards of students performance, developed on the broad-based consensus of education experts and the MES; builds on these components while developing the work-plans and evaluation plans; proposes clear statistical methods of analysis of the collected data.

- Upon approval of the detailed methods of analysis, conducts analysis of baseline data collected by the G-PriED project in spring 2013 and produces detailed baseline report;

- Conducts follow-up study during the spring session in 2015 with the same methodology and same schools; develops a detailed report of the follow-up study;

- Compares the results of the baseline- and follow-up studies in the target and control schools;

- Produces the Report of the G-PriEd Pilot Impact, which includes the findings of both, baseline and follow-up studies; and

- Proposes the changes to the norms and standards for the national benchmarking of reading and math competences in Georgia.

- Provides all quantitative data collected and analyzed timely and systematically to USAID Georgia's Mission

The contractor shall conduct all studies listed above in agreement with the USAID-approved and the Ministry of Education and Science of Georgia (MES) agreed methodologies. More

specifically, the evaluation has to follow experimental design and the methodology of test administration (sampling, implementation time, resources, data collection instruments and guides), and use the specifically-designed software for data analysis. The analysis of the 2013 and 2015 data shall be conducted to draw conclusions on the G-PriEd achievements in improving learning outcomes; to recommend areas for improvement; and propose the norms and benchmarks for the national benchmarking of reading and math standards in Georgia.

The contractor shall conduct two in- and out-briefs with USAID/Georgia Mission and MES during each year of study. During the in-briefs the contractor shall present the draft detailed evaluation design, detailed plan for statistical methods of data analysis, and the work plan.

The contractor shall provide two dissemination & capacity building workshops (one during each study) to the Ministry of Education and Science (MES) staff on the impact evaluation, its plan, methodology, and later its findings.

II. The Purpose of the Impact Evaluation

The purpose of this evaluation is to document and measure the impact of the G-PriEd pilot intervention in 122 schools of Georgia in terms of improvement learning outcomes in math and reading.

The results of the evaluation will be used by both USAID and MES and its affiliated agencies primarily for determining whether the project activities should be modified or adjusted before or during their roll-out to other schools of Georgia, as planned.

The results of the baseline and follow-up studies will also contribute to the establishment of national norms and benchmarks for reading and math competencies in the primary grades.

III. Rationale for the independent evaluation of the project pilot impact

The G-PriEd project aims to improve learning outcomes - reading and math skills - of approximately 40,000 primary-grades students of Georgia's schools. This goal is in line with the Goal 1 of the USAID's Global Strategy, which aims at improving reading skills for 100 million children in primary grades by 2015 worldwide. One of the major indicators to measure the improvement towards the Goal 1 is: "Percentage change in proportion of students in primary grades who, after two years of schooling, demonstrate sufficient reading fluency and comprehension to read to learn." Sufficient reading fluency and comprehension is defined as the reading norms *vis-à-vis* the standards of the national curricula, or as set by national experts. USAID/Georgia will report towards this indicator.
In addition, the two PMP indicators of USAID/Georgia will measure the achievements of the G-PriEd project: " The change in the proportion of students who, by the end of the primary cycle,

are able to read and demonstrate understanding as defined by a country curriculum, standards, or national experts"; and "The change in the proportion of primary grade students who by the end of each school year are meeting math and reading requirements as defined by a country curriculum, standards, or national experts".

USAID policies mandate the independent evaluations of the project impact. According to the USAID Evaluation Policy, evaluations should be conducted for all "pilot" projects; and evaluations conducted internally within the project will not be considered as impartial and unbiased. Evaluations commissioned to meet the Policy requirements should be conducted by an external (to the project) team. USAID Education Strategy, 2011-2015 (p. 16), also mandates a quality evaluation with the credible evidence: "Evaluations will use methods that generate the highest quality and most credible evidence that corresponds to the questions being asked, taking into consideration time, budget and other practical considerations."

In order to comply with the policy requirements about impartial and unbiased evaluation of USAID projects; and to report indicators listed above by complying with the credibility requirements, USAID/Georgia is conducting this independent evaluation of the G-PriEd pilot impact.

IV. Methodology of the proposed external evaluation

USAID Evaluation policy mandates the early design and baseline data collection/ maintainance, empirical strength of study design, and use of host country systems and local experts. USAID Education Strategy 2011-2016 (p.p. 15-16) also highlights that: "Per the Evaluation Policy, consideration must be given during the design phase of education programs to the types of evaluation to be undertaken… Identifying key evaluation questions at the outset of a program will guide the actions taken during implementation to capture relevant data. At the initiation of a program, baseline data, including for variables that correspond to key outcomes and impacts, will be collected…Program managers will maintain data and documentation that may eventually be made available to independent evaluation teams."

In compliance with the directives of Evaluation Policy and Education Strategy, USAID/Georgia developed the experimental design of the Pilot Impact Study in parallel with designing and implementing the G-PriEd Project; USAID mandated G-PriEd to create data collection instruments and the guides for administration; and collect baseline data in spring 2013. The approved design of the G-PriEd program impact Study with the use of Georgian Diagnostic Assessment in Reading (GDA-R) and Georgian Diagnostic Assessment in Math (GDA-R) is included as an Annex 1, Research Methodology to the SOW; the baseline data set will be made available to the independent contractor upon signing the contract. The data will be provided in the language in which the data was collected, specifically, in Georgian, Azeri, Armenian, and Russian. The contractor, an independent evaluator, shall accomplish the major tasks of the Pilot Impact Study, building on the approved design and methodology.

a. Research Objective

The evaluation aims to document and measure changes in learning outcomes, attributable to the G-Pried interventions, in primary (1-6) grade students from the G-Pried pilot phase schools. The evaluation findings will allow the G-Pried to modify or adjust interventions, if necessary, before or during their role-out to all targeted schools; and advise the MES on the national norms and benchmarks for reading and math attainment.

b. Research Questions

1. What is the student performance against grade-level norms and standards in reading and math before implementing the project? What are the differences in performance between student sub groups (ethnicity, gender, region, small/medium/large schools)?

2. Have students' performance in reading and math improved against the initial grade-level standards as a result of the interventions in the pilot schools? What is the extent and magnitude of improvement? What are the differences in performance improvements between student sub-groups (ethnicity, gender, region, small/medium/large school students) as a result of the pilot intervention?

3. What are the changes to the national norms and grade-level standards proposed for reading and math based on the data gathered throughout this project? Do these differ for students in the different sub-groups for which data were collected?

c. Methodology and Sampling Strategy

Based on the currently approved components of the research, and those components that are planned to be approved as part of this contract, the contractor shall develop a complete, detailed evaluation plan, which integrates its design, sampling, methodology, data collection, entry, analysis, and quality control arrangements at each stage. The data collection in the follow-up study should be performed by local, field independent interviewers, specifically trained for assessment of students' attainment. Independent specialists should be recruited as supervisors in each region and trained to monitor the data collection process and ensure its quality.

Data analysis shall be performed in the software developed by the G-PriEd project. USAID will maintain the availability of the final version of the software; as well as the double-entered baseline data. The Contractor shall process the data and generate reports within three months of obtaining the baseline data and/or completion of the data collection work. The contractor should provide interpretation of the data and write the report, addressing the major research questions. The Contractor should also submit the raw data file and a code book along with the draft report. In addition to baseline and end-line data comparisons, the experimental comparisons (between the intervention and control schools) and analysis should be conducted and incorporated in the report.

The evaluation plan will be presented to the USAID/Caucasus Mission during the in-brief in more detail and adjusted later based on the Mission's comments. The evaluation plan should include the evaluation matrix (an illustrative evaluation matrix is included). The contractor should also explain in detail the limitations and weaknesses of the proposed methodology.

d. Evaluation Design

In order to produce reliable conclusions about the attribution of changes in the learning outcomes to the G-PriEd interventions, it's not sufficient to compare the follow-up impact evaluation findings with the baseline data only. Comparison with the similar control group students from the similar schools is needed. Such an experimental design will allow for evaluation of student achievements in the "intervention" and "control" schools through the baseline and follow-on studies conducted in 2013 and 2015. "Intervention" and "control" schools have the similar features of their size, resources, geographic location, language of instruction, etc. These variables will be part of the group of independent variables of this evaluation. The contractor should consult the Annex 1 with the description of dependent variables, such as phonemic awareness, comprehension of narrative text, comprehension of the narrative text, number awareness, geometry, in reading and math that shall be included in the design.

e. Sampling Strategy, Study Population, and Sample Size

The sampling strategy has already been developed and applied during the baseline data collection. The contractor shall follow this strategy for the end-line evaluation. The sampling strategy is as the following:

Students of grade 1-6 make the study population.

Sample frame is 1-6 grade students enrolled in the 122 pilot and 122 control schools as of September 30, 2012.

Sampling unit is a student.

A Randomized Block Design Strategy applied in three stages:

- creation of blocks of schools, within each of which students are expected to be homogenous
- random selection of schools from each block for the pilot intervention (and for controlling); and
- random selection of students from each selected schools.

Blocks: G-PriEd "created" 43 blocks from all 2,080 public schools of Georgia by the geographic/administrative location, language of instruction, and size (number of students). Within each of these blocks, student population was considered homogenous.

1. Geographic clusters: 11 clusters (10 regions + Abkhazeti)

2. Types of schools in each cluster by the language of instruction:
    a. Georgian
    b. Ethnic minority
3. Types of schools in each cluster by the school size:
    a. Small-size school, 1-99 students
    b. Medium-size school, 300-599 students
    c. Large schools: over 600 students

Sampling of schools in each block: The G-PriEd resources could allow for pilot activities with approximately 13,000 students. The proportionate number of students and schools was included from each block (see Annex 2 for details). In total, 122 schools for pilot and 122 schools for control groups were to be selected from 43 blocks. USAID and G-PriEd chose schools from each block randomly, from the program applicant schools. The identical principle applied to the sampling of 122 pilot schools.

All students of grade 1-6 in 122 pilot schools have been receiving the same effort and amount of resources from the G-PriEd intervention. Students in 122 control schools have been excluded from the G-PriEd pilot interventions.

Sampling of a student from each school: A proportionate approach was used for identifying the number of students to be sampled from each of the pilot (and control) schools. The minimum number of students sampled from a school was 6 students (one per grade) in the schools with less than 100 students; and the maximum number was 36 in the schools with more than 500 students (see Annex 2). Samples were selected with random numbers from the roster provided by school principals (Annex 5)

Sample size: The selected 122 pilot schools enroll 18,802 1-6 grade students. The 122 control schools enroll 18,068 students.

To identify the representative sample size for the impact study, USAID conducted a power analysis described in detail in the Annex 2. The statistically significant sample size was identified as 1,495 in each of the pilot and control schools (2,990 in total). The final sample size of the study was identified as 3,244 (1,665 in the pilot schools, and 1,579 in the control schools) to be able to account for potentially discarded tests

f. Instruments of Data Collection

The data collection instruments, Diagnostic Classroom Assessment Methodology in Reading and Math, were created by the USAID and G-PrIEd through a consultative process with the MES, national experts, and educational agencies. Primarily, this is a tool for teachers' use in the classroom to identify the learning problems and respond with early intervention will put the child back on track to success. The Diagnostic Classroom Assessment methodology was tested

through the validation study in 10 schools, with 2000 students. The MES has approved the method as a diagnostic instrument.

The Diagnostic Classroom Assessment methodology also measures how a student is doing in reading and math and allows tracking the pace of children's progress towards the grade-level standards set by the MES' national curriculum. Therefore, USAID approved it as a tool for measuring the G-PriEd's achievements in making the impact on reading and math skills towards achieving objectives of national curricula.

g. Data entry and analysis

Data entry shall be conducted in the software developed by G-PriEd. The double-entered baseline data set will be provided to the Contractor by USAID. The contractor shall use these double-entered data, maintain, clean, and analyze in lieu with the research questions. In the follow-up assessment in 2015, the contractor shall conduct the data collection, double-entry, maintenance, cleaning, and analysis.

The following represents the illustrative evaluation matrix:

Research Question	Data Source	Methodology
What is the student performance against initial grade-level standards in reading and math before implementing the project? What are the differences in performance between student sub groups (ethnicity, gender, region, small/medium/large schools)?	Baseline assessments on the paper and in the software	Calculating the ratio; variances of groups.
Have students' performance in reading and math improved against the initial grade-level standards as a result of the interventions in the pilot schools? What is the extent and magnitude of improvement? What are the differences in performance improvements between student sub-groups (ethnicity, gender, region, small/medium/large school students) as a result of the pilot intervention?	Baseline and end-line assessments on the paper and in the software	Calculating the ratio; variances of groups.
What are the final national norms and grade-level standards proposed for reading and math as indicated by the data	Baseline and end-line assessments	Means and standard deviations, bivariate correlation, multiple

gathered thought this project? Do these differ for students in the different sub-groups for which data were collected?	on the paper and in the software	regressions, and percentile points

V. Cooperation with the Ministry of Education and Science (MES)

Per USAID's policy on evaluation, and to support the host country capacity building, USAID has included MES and partners in reviewing the design of evaluations. The MES will be invited to observe the evaluation process; this will provide an ongoing capacity building opportunity to the local counterparts, in line with USAD Forward initiative. Upon completion of the baseline and end-line data analysis, the independent evaluation will have a specific component of the technical assistance to the MES to strengthen its ability to conduct similar studies independently in the future. USAID will share the software and data-base with the MES; and will transfer the property to the Ministry upon completion of the project.

The contractor will also provide two dissemination workshops to the MES on impact evaluation in particular. One workshop will be implemented during each study.

VI. Evaluation Team

The contractor is required to propose a strong team of education evaluation experts. While the contractor can propose the composition of the team, the following are the skills required for the completion of the evaluation:

- Ability to provide strategic management of the project, to manage the evaluation team, cooperate with USAID; a demonstrated strong background in the education evaluations and assessments. Fluency in English language is required. The knowledge of the education system of the region is a plus.
- Demonstrated experience in reading and math learning outcomes, including the analysis of norms and setting of benchmarks. Ability to manage data collection and analysis.
- Demonstrated skills and knowledge of the local context and system very closely; knowledge of knowledge of the reading and math learning outcomes and Georgian national standards. Ability to support international expert/s in developing and implementing the analysis of the data.
- Ability to clean provided and collected data on students' outcomes in reading and math. Ability to supervise and assure quality of the data. Ability and skills to provide a double-entry of all students' data. Ability, skills, and knowledge to administer student

assessments in reading and math based on the methodology developed by the G-PriEd, and using the G-PriEd's experience.

USAID may decide to interview the proposed key members of the team prior to final approval of their candidacy.

VII. Estimated timeframe

Phase 1:

- Fall 2013: Get acquainted with the approved design, methodology, instruments, and approaches to the data analysis of the Pilot Impact Study for the G-PriEd project; and build on these components while developing the work-plans and evaluation plans;

- Fall 2013: Conduct analysis of baseline data in diagnostic assessment of reading and math for grades 1 through 6 collected by the G-PriEd project in spring 2013. And produces baseline report

Phase 2:

- Spring 2015: Conduct follow-up study during the spring session in 2015 with the same methodology and same schools; and

- Fall 2015: Compare the results of the baseline- and follow-up studies in the target and control schools and produces the Report of the G-PriEd Pilot Impact

VIII. Logistics

USAID Mission will not be responsible for arranging logistics for the evaluation team, however it will advise on the fieldwork plan prior to the start of the fieldwork. The evaluation team will also receive all relevant reports and documentation in advance furnished by the mission. These documents are:

Annex 1-Study Design and Framework Reading Math
Annex 2- Test Administration Guidelines
Annex 3- Pilot and Control School Data
Annex 4- Sample Test in Reading for grade 3
Annex 5 - Sample test in math for grade 3
Annex 6- Baseline Impact Assessment Report
Annex 7- GPried SOW

USAID/Georgia will also place the team in contact with the staff of G-Pried program. The contractor should suggest how they plan to arrange translation, transportation and

logistical support to the evaluation team. While in Georgia, the team will meet with the MES, and relevant USAID partners. The field work will include intensive data collection in the treatment and control schools.

ATTACHMENT 2

INSTRUCTIONS FOR TECHNICAL AND COST PROPOSAL

I. Technical Proposal

Detailed research (evaluation and/or assessment) design and the work plan:
The research design must explain in details methodologies that will be used to collect required information. The design must outline in details:
- what methods the contractor will use to get answers for each evaluation question.
- The evaluation design must include:
  - a detailed evaluation matrix (including the key questions, methods and data sources used to address each question and the data analysis plan for each question)
  - draft questionnaires and other data collection instruments or their main features,
  - known limitations to the evaluation design,
  - a work plan, and
  - information dissemination plan

Proposed Evaluation Team:
The contractor is required to propose a strong team of education evaluation experts. While the contractor can propose the composition of the team, the following are the skills required for the completion of the evaluation:

- Ability to provide strategic management of the project, to manage the evaluation team, cooperate with USAID; a demonstrated strong background in the education evaluations and assessments. Fluency in English language is required. The knowledge of the education system of the region is a plus.

- Demonstrated experience in reading and math learning outcomes, including the analysis of norms and setting of benchmarks. Ability to manage data collection and analysis.

- Demonstrated skills and knowledge of the local context and system very closely; knowledge of knowledge of the reading and math learning outcomes and Georgian national standards. Ability to support international expert/s in developing and implementing the analysis of the data.

- Ability to clean provided and collected data on students' outcomes in reading and math. Ability to supervise and assure quality of the data. Ability and skills to provide a double-entry of all students' data. Ability, skills, and knowledge to administer student assessments in reading and math based on the methodology developed by the G-PriEd, and using the G-PriEd's experience.

II. Instructions on Preparation of Branding Implementation Plan and Marking Plan

As part of the proposal submission, the Contractor will develop a Branding Implementation Plan (BIP) and a Marking Plan in accordance with the policies found at Automated Directive System (ADS) Chapter 320, revised on May 5, 2009, or any successor branding policy, and with the "USAID Graphics Standard Manual" that is available at www.usaid.gov/branding. Among other provisions, ADS 320 states that:

1. Contractors and subcontractors' corporate identities or logos must not be used on USAID-funded program materials.
2. Marking is not required on contractor vehicles, offices, office supplies or other commodities used solely for administration of the USAID-funded program.
3. Marking is not permitted on any communications that are strictly administrative, rather than programmatic, in nature. USAID identity is also prohibited on contractor and recipient communications related to award administration, such as hiring/firing of staff or renting office space and/or equipment.

The Contractor must also develop a *Marking Plan* for public communications, commodities, program materials, deliverables, and other items that visibly bear or will be marked with the USAID identity. The marking plan may include requests for exceptions to marking requirements, to be approved by the Contracting Officer. Contract deliverables to be marked with the USAID identity must follow design guidance for color, type, and layout in the Graphic Standards Manual (available at www.usaid.gov/branding) or any successor branding policy.

With respect to this Task Order, the Contractor should develop a BIP and Marking Plan bearing in mind the following branding strategy:

1. Program Name: Mid-Term Performance Evaluation of the Good Governance in Georgia (G3) Program.
2. This task order is funded through the USAID/Caucasus Mission. Materials and communications must be positioned as from the American People, using the USAID Identity.
3. Outreach to Beneficiaries and Host-Country Citizens: No special outreach efforts to beneficiaries and host-country citizens are planned under this Task Order.
4. Level of Visibility: The findings of the final evaluation report will be used by USAID in its implementation and further planning its activities. The report will be submitted to USAID's Development Experience Clearinghouse for wider access.
5. Other Organizations to be Acknowledged: No other organizations are required to be acknowledged.
6. Specific branding issues: The only branding issue expected to arise as a result of implementing this Task Order is the proper use of graphics standards for all reports and other printed or electronically distributed information.

II. Cost Proposal

The cost proposal must include detailed budget schedules and a budget narrative.
The schedules must support and explain proposed costs with breakdowns on direct labor, fringe benefits, supplies and equipment, travel and per diem amounts, other direct costs, and indirect costs; personnel costs, allowances and benefits; travel and transportation costs, including airfares (destinations and number of trips), per diems amounts, taxis, and car rentals; other direct costs such as rent, equipment, supplies, domestic, and international communications and indirect costs. Cost proposal must also include Contractor Biographical Data Sheets (Form 1420-17) for all proposed staff.

ATTACHMENT 3

SPECIAL REQUIREMENTS

1. ENVIRONMENTAL COMPLIANCE

1a) The Foreign Assistance Act of 1961, as amended, Section 117 requires that the impact of USAID's activities on the environment be considered and that USAID include environmental sustainability as a central consideration in designing and carrying out its development programs. This mandate is codified in Federal Regulations (22 CFR 216) and in USAID's Automated Directives System (ADS) Parts 201.5.10g and 204 (http://www.usaid.gov/policy/ads/200/), which, in part, require that the potential environmental impacts of USAID-financed activities are identified prior to a final decision to proceed and that appropriate environmental safeguards are adopted for all activities.

1b) In addition, the contractor/recipient must comply with host country environmental regulations unless otherwise directed in writing by USAID. In case of conflict between host country and USAID regulations, the latter shall govern.

1c) No activity funded under the contract resulting from this RFTOP will be implemented unless an environmental threshold determination, as defined by 22 CFR 216, has been reached for that activity, as documented in a Request for Categorical Exclusion (RCE), Initial Environmental Examination (IEE), or Environmental Assessment (EA) duly signed by the Bureau Environmental Officer (BEO). (Hereinafter, such documents are described as "approved Regulation 216 environmental documentation.")

4a) As part of its initial Work Plan, and all Annual Work Plans thereafter, the contractor, in collaboration with the USAID Contracting Officer's Technical Representative and Mission Environmental Officer or Bureau Environmental Officer, as appropriate, shall review all ongoing and planned activities under this contract to determine if they are within the scope of the approved Regulation 216 environmental documentation.

4b) If the contractor plans any new activities outside the scope of the approved Regulation 216 environmental documentation, it shall prepare an amendment to the documentation for USAID review and approval. No such new activities shall be undertaken prior to receiving written USAID approval of environmental documentation amendments.

4c) Any ongoing activities found to be outside the scope of the approved Regulation 216 environmental documentation shall be halted until an amendment to the documentation is submitted and written approval is received from USAID.

2. PROHIBITION AGAINST DISCRIMINATION (Oct 2011)

FAR Part 27 and the clauses prescribed in that part prohibit contractors performing in or recruiting from the U.S. from engaging in certain discriminatory practices.

USAID is committed to achieving and maintaining a diverse and representative workforce and a workplace free of discrimination. Based on law, Executive Order, and Agency policy, USAID prohibits discrimination in its own workplace on the basis of race, color, religion, sex (including pregnancy and gender identity), national origin, disability, age, veteran's status, sexual orientation, genetic information, marital status, parental status, political affiliation, and any other conduct that does not adversely affect the performance of the employee. USAID does not tolerate any type of harassment, either sexual or nonsexual, of any employee or applicant for employment. Contractors are required to comply with the nondiscrimination requirements of the FAR and in addition, the Agency strongly encourages all its contractors (at all tiers) to develop and enforce comprehensive nondiscrimination policies for their workplaces that include protection on these expanded bases.

3. 752.225-70 SOURCE AND NATIONALITY REQUIREMENTS (FEB 2012)

(a) Except as may be specifically approved by the Contracting Officer, the Contractor must procure all commodities (e.g., equipment, materials, vehicles, supplies) and services (including commodity transportation services) in accordance with the requirements at 22 CFR Part 228 "Rules on Procurement of Commodities and Services Financed by USAID Federal Program Funds." The authorized source for procurement is Geographic Code 937 unless otherwise specified in the schedule of this contract. Guidance on eligibility of specific goods or services may be obtained from the Contracting Officer.

(b) Ineligible goods and services. The Contractor must not procure any of the following goods or services under this contract:

(1) Military equipment
(2) Surveillance equipment
(3) Commodities and services for support of police and other law enforcement activities
(4) Abortion equipment and services
(5) Luxury goods and gambling equipment, or
(6) Weather modification equipment.

(c) Restricted goods. The Contractor must obtain prior written approval of the Contracting Officer or comply with required procedures under an applicable waiver as provided by the Contracting Officer when procuring any of the following goods or services:

(1) Agricultural commodities,
(2) Motor vehicles,
(3) Pharmaceuticals and contraceptive items
(4) Pesticides,
(5) Fertilizer,
(6) Used equipment, or
(7) U.S. government-owned excess property.

If USAID determines that the Contractor has procured any of these specific restricted goods under this contract without the prior written authorization of the Contracting Officer or fails to comply with required procedures under an applicable waiver as provided by the Contracting Officer, and has received payment for such purposes, the Contracting Officer may require the contractor to refund the entire amount of the purchase.

4. NONDISCRIMINATION (JUNE 2012)

No U.S. citizen or legal resident shall be excluded from participation in, be denied the benefits of, or be otherwise subjected to discrimination on the basis of race, color, national origin, age, disability, or sex under any program or activity funded by this award when work under the grant is performed in the U.S. or when employees are recruited from the U.S.

Additionally, USAID is committed to achieving and maintaining a diverse and representative workforce and a workplace free of discrimination. Based on law, Executive Order, and Agency policy, USAID prohibits discrimination, including harassment, in its own workplace on the basis of race, color, religion, sex (including pregnancy and gender identity), national origin, disability, age, veteran's status, sexual orientation, genetic information, marital status, parental status, political affiliation, and any other conduct that does not adversely affect the performance of the employee.

In addition, the Agency strongly encourages its recipients and their subrecipients and vendors (at all tiers), performing both in the U.S. and overseas, to develop and enforce comprehensive nondiscrimination policies for their workplaces that include protection for all their employees on these expanded bases, subject to applicable law.

5. ORGANIATIONAL CONFLICTS OF INTEREST: PRECLUSION FROM IMPLEMENTATION CONTRACT

This task order calls for the Contractor to furnish important services in support of the design of _____ [specify activity] (the "Activity"). In accordance with the principles of FAR Subpart 9.5 and USAID policy, THE CONTRACTOR SHALL BE INELIGIBLE TO FURNISH, AS A PRIME OR SUBCONTRACTOR OR OTHERWISE, THE IMPLEMENTATION SERVICES FOR THE ACTIVITY, EXCEPT FOR SUCH SERVICES THAT MAY BE FURNISHED UNDER A SEPARATE TASK ORDER ISSUED UNDER THIS CONTRACT, unless the Head of the Contracting Activity, in consultation with USAID's Competition Advocate, authorizes a waiver (in accordance FAR 9.503 and AIDAR 709.503) determining that preclusion of the Contractor from the implementation contract would not be in the Government's interest.

6. BRANDING AND MARKING POLICY

Where applicable, the Contractor shall comply with the requirements of the policy directives and required procedures outlined in USAID Automated Directive System (ADS) 320.3.2 "Branding and Marking in USAID Direct Contracting" (version from January 8, 2007) at http://www.usaid.gov/policy/ads/300/320.pdf; and USAID "Graphic Standards Manual" available at www.usaid.gov/branding, or any successor branding policy.

7. AUTHORIZED GEOGRAPHIC CODE

The geographic code applicable to the procurement of goods and service under this task order is 110 AND 937.

ATTACHMENT 4

Background information about the project

USAID G-PriEd Project is dedicated to the improvement of reading and math skills of primary grade (1-6) students in Georgia. Activities include development of teacher training materials in contemporary methods of teaching reading and math skills, conducting teacher training, developing and publishing age-appropriate books and additional learning resources, and assisting in classroom application of the teacher training. Diagnostic classroom assessments and a differentiated teaching approach for students with different learning progression are being introduced as innovative components of the learning process in Georgia.

G-PriEd interventions cover schools from all regions, of different sizes, and different languages of instruction. Approximately 10 percent of all students targeted by the Program are those in the schools with ethnic minority language of instruction; with these students, the G-PriEd works on improvement of their reading skills of Georgian as a Second Language; their math skills improvement efforts are conducted in their native language.

The pilot activities of the G-PriEd project are being implemented in 122 schools with about 14,000 primary grade (1-6) students. The major part of the project activities in these schools will be conducted during April 2013- March 2015. This includes training of teachers, provision of learning materials, school visits and consultations, teacher circles and peer-learning activities, and engagement of parents in the learning process. By the end of this pilot cycle, the intervention should be sustainable and activities and methodologies must-be carried-out by the schools independently.

In October 2014, G-PriEd will start expanding its activities to additional schools. Findings and lessons learned through the implementation and monitoring process, as well as through the independent evaluation of the project pilot impact, the G-PriEd activities will be adjusted in the expansion schools in 2014-2016.

G-PriEd will implement its activities in approximately 300 Georgia's schools (of which 122 are the pilot schools), with about 40,000 students to be targeted through this five-year project.

G-PriEd had conducted the baseline data collection in spring 2013. The detailed report on the baseline data collection will be provided separately. Along with the detailed implementation report, G-PriEd, through USAID, will provide the baseline data set to the contractor; as well as the national norms and standards of reading and math competencies, as a result of a broad-base consensus of stakeholders.