The Introduction of Large-scale Computer Adaptive Testing in Georgia

Political context, capacity building, implementation, and lessons learned

Steven Bakker

Dr. Steven Bakker
DutchTest, the Netherlands
March 2014
Contents

List of Acronyms .......................................................................................................................... 5
Executive Summary ......................................................................................................................... 6
Introduction .................................................................................................................................. 8
Political Context and Decision Making ....................................................................................... 8
Planning the CAT ......................................................................................................................... 10
  Business considerations ........................................................................................................... 10
    Human and Material Resources ............................................................................................ 11
    Hardware .............................................................................................................................. 11
    Item banking software ......................................................................................................... 12
    CAT software modules ....................................................................................................... 13
    Connectivity ......................................................................................................................... 13
  Capacity building .................................................................................................................... 15
    Psychometricians ............................................................................................................... 15
    Proctors .............................................................................................................................. 16
    Test developers .................................................................................................................. 16
    System administrator ......................................................................................................... 16
    Helpdesk staff .................................................................................................................... 16
    Local IT school support staff .............................................................................................. 16
  Item development and calibration of item pools .................................................................... 16
Information and Advocacy ........................................................................................................... 17
  First announcement ............................................................................................................... 17
  NAEC’s campaign .................................................................................................................. 18
  Last press conference .......................................................................................................... 21
  School principals’ perception ............................................................................................... 21
  Teachers’ perception ............................................................................................................. 22
  Students’ perception ............................................................................................................. 22
  Stakeholders’ understanding of the CAT concept ............................................................... 22
Implementation of CAT ............................................................................................................... 23
  Testing centers ..................................................................................................................... 23
  Computers and internet connections ................................................................................... 24
  Registration and administration ............................................................................................ 24
  Proctoring .............................................................................................................................. 25
  Testing program .................................................................................................................... 26
## List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT</td>
<td>Computer Adaptive Testing</td>
</tr>
<tr>
<td>CBT</td>
<td>Computer Based Testing</td>
</tr>
<tr>
<td>CITO</td>
<td>Dutch Institute for Educational Measurement</td>
</tr>
<tr>
<td>EMIS</td>
<td>Educational Management and Information Service</td>
</tr>
<tr>
<td>ESIDA</td>
<td>Educational Scientific Infrastructure Development Agency</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>MoES</td>
<td>Ministry of Education and Science</td>
</tr>
<tr>
<td>NAEC</td>
<td>National Assessment and Examinations Center</td>
</tr>
<tr>
<td>SGE</td>
<td>School Graduation Exam</td>
</tr>
<tr>
<td>SSID</td>
<td>Social Security ID</td>
</tr>
</tbody>
</table>
Executive Summary

After three rounds of computer-adaptive School Graduation Exams (SGEs), stakeholders in Georgia and international experts agree that the instrument was successfully launched, that it is an efficient, fair and objective way of student assessment, that it helped to realize a number of issues that are high on the policy agenda of the Ministry, but also that further development is needed to compensate for the limitations that are inherent to Computer Adaptive testing (CAT).

The SGEs had been re-introduced in 2010/2011 in an effort to fight student absenteeism in grade 12 and increase school accountability. Computer Adaptive Testing (CAT) was chosen as the delivery mode for efficiency and security reasons. The human and material resources needed for developing and implementing CAT were present: (i) a strong testing institution (National Assessment and Examinations Center, NAEC); (ii) a well-trained and motivated pool of proctors; (iii) well-developed internet infrastructure; (iv) sufficient resources to set up testing centers in schools; (v) and local IT support. NAEC designed, administered and coordinated the testing, and NAEC psychometricians received training from Dutch and US CAT experts for developing CAT algorithms. The NAEC IT department used these to develop the CAT software.

The need for a well-planned and coordinated information and advocacy campaign was seriously underestimated by the Ministry of Education, which resulted in a variety of rumors and fears of massive percentages of students failing and punitive measures being taken against under-achieving schools. Eventually, NAEC, using its image as a reliable institution and its good relations with schools, managed to convince the educational community that the test results would not be used against them, but would support them in achieving their goals.

NAEC worked closely with the Educational Management and Information Service (EMIS) in establishing testing centers in schools. Two providers of internet services, Delta Comm and MAGTI, made sure that these test centers had sufficient connectivity to allow for smooth down- and uploading during the CAT exams. A pool of 2500 proctors was created and trained to guarantee secure testing conditions.

The costs of developing and implementing the CAT SGE are estimated by MoES at about USD 2.5 M, which include the purchase of surveillance cameras (USD 160,000) and routers for testing centers, but not computers. NAEC estimates the costs of producing the items and administering the tests at about USD 1.4 M. Proctor’s fees and reimbursables make up the larger part of this budget.

Stakeholders are generally positive about the CAT SGE and secure administration of the tests. So far, there haven’t been any major technical problems that would cause student data getting lost. The concerns among stakeholders about the difficulty grade of the tests were eased by using easy items and applying low cut scores. Some note the limited validity inherent to the machine-scorable item format needed for CAT, which does not allow the assessment of certain skills such as speaking or writing. Negative aspects of CAT include the fact that test items are not released and that, during the testing, no corrections in previously given answers may be made. However, receiving scores immediately at the end of the test is much valued.

Main factors leading to the successful implementation of CAT in Georgia include the following:

1. **Strong government commitment.** Deciding to introduce a national test and to use a technologically advanced delivery mode brings along an immediate commitment to fund
the necessary initial investments, a long-term commitment to fund the recurrent costs and to guarantee continuity in all operations concerned.

2. **NAEC’s leadership and stakeholders’ confidence in NAEC’s competence**. Schools that do not buy into measures that are implemented from the top down would be a recipe for failure. The decision to develop and implement the CAT SGE within a year would have desperately failed if NAEC would not have been able to get schools on board. Using its strong public relations capacity and reputation as a reliable institution for educational assessment, NAEC managed to convince schools that using the CAT was in their interest and that results would not have negative consequences for the schools.

3. **NAEC’s strong psychometric and ICT competence**. International psychometric experts note that what has been achieved in Georgia in terms of implementation of a large-scale, high-stakes computer adaptive testing effort within a very short time is unique in the world.

4. **NAEC’s experience in large scale secure testing**. The presence of a large pool of well-managed, trained and motivated proctors can hardly be underestimated.

5. **Smart test design avoiding network overloads and student data getting lost**. The web-based CAT application doesn’t generate much internet traffic, and sessions are scheduled in such a way that a minimum amount of students log in at the same time. The program saves all keystrokes made by students on the central server, which facilitates resuming testing after a power break or computer crash without loss of data.

6. **Full scale pretest under realistic conditions shortly before the real tests** to understand the effects of scaling up and familiarize all involved with the nature and setting of the test.

Important caveats for future use of CAT for School Graduation Exams or other forms of large-scale high-stakes testing including the following:

1. **Doubts about the validity of the tests among stakeholders** due to the use of simple multiple choice item formats and the increase of coaching practices.

2. **Reliability of the ability estimates, both psychometrically and at face value**. NAEC psychometricians point at the wide confidence intervals in estimates of item difficulty and discrimination due to unreliable pretest outcomes. For stakeholders, results based on a relatively short test, while psychometrically sound, may look unreliable.

3. **Negative backlash effects caused by applying low cut scores**. While these may be needed to avoid massive percentages of failing students, they also encourage minimalistic behavior.

4. **Security of items and right to appeal**. To make CAT work, items cannot be released. At the same time this makes it very difficult for candidates to appeal results and argue that the test was flawed.
Introduction

This paper was prepared as a background paper to inform Armenia’s consideration of adopting computer-adaptive testing (CAT) or computer-based testing (CBT) based on Georgia’s experience with design and implementation of CAT-based school leaving examinations.

In September 2010, the Georgian Ministry of Education and Science decided to use computer adaptive testing (CAT) as the delivery mode for the re-introduced external school graduation exams and to conduct the first administration in May 2011. The international experts hired to advise and train the National Assessment and Examinations Centre (NAEC), the institution charged with the development and implementation of these tests, were rather skeptical about the feasibility of a nation-wide rollout of such a logistically and technologically complex measurement instrument as a large scale CAT. Now, after three rounds of computer-adaptive graduation exams, stakeholders in Georgia and international experts observing the process agree that the instrument was successfully launched, that it is an efficient, fair and objective way of student assessment and that it helped to realize a number of issues that are high on the policy agenda of the Ministry, but also that further development is needed to compensate for the limitations that are inherent to CAT. This report aims to describe and evaluate the introduction of large-scale high stakes computer adaptive testing in Georgia in order to provide lessons for Armenia in the following areas: (i) the political context; (ii) the human and material resources that were needed; (iii) the process of capacity building; (iv) how stakeholders were informed; (v) the implementation of CAT in schools; and (vi) the impact it had on all involved and on the educational system in Georgia.

A number of interviews informed this report. The interviews were guided by questionnaires that covered all relevant topics for describing the perception of the different actors and stakeholders of how the introduction of CAT had taken place, how this specific way of testing had been implemented, the impact it had had on education in Georgia in general, and their personal role in it. Interviews were held with the following actors and stakeholders: the Minister of Education, NAEC staff, test center staff, network service providers, school principals, teachers, students who sat one of the CAT exams, and TV and newspaper journalists. With the exception of the Minister, all of the interviewees in their capacity had first-hand experience with one or more of the aspects of CAT. A full list of interviews is in Annex 1. Others sources of information included documents prepared by NAEC and a selection of articles from newspapers, magazines and internet journals (also included in Annex 1).

Political Context and Decision Making

Till the end of the 20th century school graduation exams in Georgia basically followed the Soviet-Russian tradition of ‘biljets’ and test questions taken from existing collections published by the Ministry. While the content of these books, the ‘biljets’ were decided by the Ministry, these exams were essentially school exams that were administered and scored by teachers. In 2002 and 2003, under the Minister of Education Alexander Kartosia, national written grade IX exams were introduced for Mother Tongue, Modern Foreign Language and Mathematics. These exams were produced by the

---

1 ‘Biljet’ system: a student is presented a box or a table full of tickets (‘biljets’) mentioning a certain topic, for instance “Mendeleev’s Periodic System of the Elements”. A student draws a ticket and, after some time for preparation, is invited to the testing room and there demonstrates his or her understanding of the given topic to the examiners.
NAEC (established in 2002) but administered in and by the schools. After the Rose Revolution, all school graduation exams were abolished by the new Minister of Education, Kacha Lomaia. Ministry Kacha Lomaia introduced the national University Entrance Exams to replace the entrance tests that had been the responsibility of universities. NAEC was entrusted with developing, administering and scoring these university entrance tests under strict security conditions. In 2010, Lomaia’s successor Dima Shashkin decided to re-introduce national school graduation exams for grade 12. The main reasons for this decision were the following:

- **The failing introduction of grade 12.** Before 2008, compulsory education had eleven grades. In 2003, a national curriculum innovation plan included the gradual introduction of an additional grade and in 2008, the first cohort that had had 12 years of compulsory education graduated. It soon became clear that the majority of grade 12 students were skipping school to spend time with tutors that prepared them for the university entrance exams. Passing school graduation exams and having a school graduation certificate was a prerequisite for registering for the University Entrance Exams. However, most schools arranged the final exams in such a way that almost all students would pass, and gave out certificates to students, independent of their attendance rate in the last year. It was hoped that taking the exams out of the hands of the schools would ‘bring the students back to school’.

- **Increasing school accountability.** The Minister decided to also use the outcomes of the national school graduation exams as an instrument to identify weakly performing schools and take ‘appropriate measures’. Indeed, in 2011, after the first administration of the national school graduation exams, a couple of school managers were fired and about 200 schools (out of the about 2,000 schools in Georgia with grade 12) received some kind of punishment, allegedly because of the outcomes of students falling below standards.²

The decision to use CAT for the school graduation exams was taken in September 2010. The initial idea was to have a nation-wide test run of the system in 2011, without any consequences for individual students or schools and a full implementation in 2012. However, the final decision was to have a fully operational system in 2011 so that all grade 12 students would take computer-adaptive final exams in eight subjects in May.

As soon as the initial decision to re-introduce national school graduation exams had been taken, NAEC was invited to advise on a format that would enable secure delivery of these tests, and on the feasibility of combining the UEE and SGE. On the feasibility of combining the UEE and SGE, NAEC advised negatively. The Georgian law stipulates that compulsory education is free of charge and therefore taking part in school graduation tests, which are an inseparable part of compulsory education, should be free as well. However, charges do apply for the UEE. The logistical and financial consequences of combining the SGE and UEE would have required two full weeks or more to administer, bringing 45,000 students to testing centers, and housing and feeding a large number of them for the duration of the testing. Therefore, the combination was deemed not to be feasible. The MoES adopted NAEC’s advice not to combine the UEE and SGE³.

² In the following year (2012) there was a slight decrease of students taking part in the national school graduation exams. It is said that this was the result of schools trying to increase their average score on the tests by withholding weak students from the exams, out of fear of punitive measures that might be imposed by the Ministry.

³ From NAEC’s point of view, another important reason not to combine the two tests (into one test) is the incompatibility of the different purposes of certification and entrance exams, requiring assessment of different types of knowledge and skills. But the practical constraints were decisive.
The next step was to propose a format for secure testing of around 45,000 students in eight subjects, in their own schools or a close-by facility that would assess minimal mastery of the national curriculum content. MoES agreed that computerized testing for this purpose would probably offer an adequate solution. At that stage, NAEC suggested to consider CAT, an approach that NAEC was also considering at that time for another test, the Georgian Graduate Record Exams.

NAEC advised to use CAT⁴ rather than Linear Computer Based Testing ⁵ (CBT) for several reasons. CAT is known to make more efficient use of an item bank because fewer items are needed to estimate the ability levels of individual students in CAT than in linear tests. The fact that each student is challenged at his/her own ability level definitely adds to the reliability of the measurement. In addition, CAT can be set up in a way to allow for a flexible, continuous process that puts less demand on available testing facilities than a linear test with the same reliability would do. For the Minister, the decisive argument to go for CAT was the security that could be achieved by each student having his/her own test. In addition, the idea that the implementation of a technologically advanced instrument as CAT would possibly boost Georgia’s image as a knowledge and technology economy appealed to the Minister, who then charged NAEC with the task to explore business issues connected to CAT.

It was decided to use the infrastructure developed by the Educational Scientific Infrastructure Development Agency (ESIDA) for connecting testing centers with the servers on which the CAT application would be running. Also these servers would be located on the ESIDA premises. ESIDA merged in 2011 with the MoES’s Statistics Information Department to become the Educational Management and Information System (EMIS)⁶.

Planning the CAT

Business considerations

Usually the first stage in CAT development is to ask why exactly one wants to move from a fixed testing format to CAT, and, in view of the answer(s) to that question, to determine whether a CAT approach is even feasible for the testing program. Therefore, practical and business considerations of the CAT approach should be researched first. CAT may be introduced for a variety of reasons, such as to increase security by producing student-specific tests and to save on production and administration costs. Before moving forward with the CAT, several questions should be answered. Is it sufficiently known if CAT delivery in a computerized testing center will indeed be more secure than paper and pencil test in a local gym? Will converting the test to CAT likely bring the expected reduction in test length? Does the reduction in test length translate to enough saved examinee seat time – which can be costly – to translate into actual monetary savings? Even if CAT costs more and does not substantially decrease seat time, are these disadvantages sufficiently offset by the increase in

⁴ Computer Adaptive Testing (CAT) is a way to adapt tests to the proficiency level of individual examinees during the administration. This is basically achieved by selecting a more difficult task after a correct response or an easier task after an incorrect one and repeating this till a precise as possible estimation of the student’s ability level has been achieved. More information on CAT may be found in Report 2 of the Feasibility Study on Introducing CAT for the Unified Exams in Armenia, Computer Adaptive Testing: Definition, Use and Implementation, Steven Bakker, December 2012.

⁵ Linear testing is presenting students with a fixed series of test items.

⁶ In 2011 EMIS started as a department of MoES. In 2012 it became a legal entity on its own, by a charter of MoES. It is in charge of a Virtual Private School Network connecting all Georgian public schools. This network is used by schools to upload information to the national school database. EMIS also employs regional IT school support staff. These employees, about 220 across Georgia, are part of the regional educational support centers but on the EMIS pay roll. Schools are charged by EMIS for their services.
precision and security to make it worthwhile for the organization? Does the organization have the psychometric expertise, or is it able to afford it if an external consultant is used? Does the organization have the capacity to develop extensive item banks? Is an affordable CAT delivery engine available for use, or does the organization have the resources to develop its own?

In the case of Georgia, there was another issue of whether the CAT would be legally allowed for school graduation purposes? CAT very much relies on large, secure item banks, from which only small proportions of items are released after administration for information purposes or because they have been over-exposed, published on the internet or magazines, etc. In some countries, according to the law, public exams must be published after administration in order to give students and other stakeholders the opportunity to scrutinize and discuss the tests, and if needed file appeals against defective or otherwise improper items.

However, the decision to re-introduce school graduation exams and use CAT as the delivery mode was very much politically-driven. The government’s wish to make a statement was the major driving force. At the same time, NAEC was confident that it would be able to successfully introduce CAT, be it not within the very short span of time that was ordered by the Ministry when preparations already had started for a nation-wide pilot. In addition, the Ministry guaranteed financing for the CAT regardless of the budget and was not too concerned with possible legal complications. Therefore in Georgia, researching the practical and business considerations did not take place as a commercial test publisher probably would have done in a similar situation. Rather than a careful analysis of all the risk and cost factors, NAEC produced a paper listing the major issues that needed to be addressed, such as developing CAT algorithms, test logistics and establishing and connecting testing centers. This paper indicated that logistics of student registration and test surveillance would not provide any major challenges, as the numbers of students involved would be similar to those of the University Entrance Exams that NAEC had been successfully administering across Georgia since 2005. NAEC was confident that developing and implementing the CAT tool could be done by its existing staff, with some training by international consultants. The Ministry already had announced that schools would receive netbooks for all students in the early grades that could also be used for testing. Exploratory talks with network service providers suggested that sufficient connectivity could be achieved for all testing centers.

**Human and Material Resources**

To implement CAT, NAEC did not need to hire any new institute staff: development and operations could be run with existing staff while technical support for testing centers and ensuring adequate connectivity was taken care of by EMIS. However, new proctors had to be hired and trained, in addition to the existing pool of proctors who were already active proctoring for the UEE and other NAEC secure testing efforts. Proctors were only hired after they have successfully passed a test. This test, administered at the NAEC offices, is actually a simulation of the present CAT procedures that includes entering student identification information and other tasks. Proctors are tested again after the training on their knowledge of procedures and rules.

**Hardware**

During the first year of CAT (2011) the servers on which the item bank and CAT application were running were located on the premises of ESIDA, which controlled the on-line delivery of the tests to the centers. In the

<table>
<thead>
<tr>
<th>Hardware purchased by NAEC for CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Six HP Blade (G7 generation) servers; each server 24 core, 96 GB Ram.</td>
</tr>
<tr>
<td>One Raid Array</td>
</tr>
<tr>
<td>Two Database servers (Mirrored)</td>
</tr>
<tr>
<td>Four web servers (NLB)</td>
</tr>
<tr>
<td>Two redundant Cisco routers</td>
</tr>
</tbody>
</table>

---

7 In the pool are about 2000 trained and active proctors and 300
following year this task moved to the newly established EMIS, using its own servers. In 2013 the item bank and CAT software package were installed on servers located in the NAEC building, and from that time onwards the entire operation (item generation, calibration, development and implementation of CAT algorithms, test delivery, proctoring and scoring) was in NAEC’s hands. For the online delivery of CAT, NAEC purchased the equipment listed in Table 1. NAEC also bought 1,800 surveillance cameras at GEL 150 (equivalent to USD 100) a piece. These were left at the testing centers in the schools, but remained NAEC’s property.

In 2011, students used netbooks for taking the tests. These had been purchased for all grade 1 pupils but were first used for administering the School Graduation exams. These netbooks provided the uniformity needed for an uncomplicated CAT delivery.

- Internet connectivity 32Kbps or more
- Computers should have minimal 512 MB operational memory and a processor speed of 1000 GHz or more
- Windows XP OS or higher.

| Table 2 General criteria required for computers and connectivity testing rooms |

In 2012, MoES purchased 11,000 new computers that were placed in 1,500 schools along with new routers. This was still not enough for securing uniformity in platforms for CAT delivery in the testing centers. Therefore, the EMIS-managed local IT school support staff inspected rooms that would serve as a testing room and the available computers and decided whether these could be used as testing stations. General criteria are listed in Table 2.

**Item banking software**

Item banking software was developed in-house to be used in combination with the CAT modules, which allowed the use of graphics, diagrams and formulas (the use of multi-media is also supported, but not yet used). The item banking software is geared towards entering multiple choice items with four alternatives only. It does not yet support the use of items that allow students to interact, for instance by clicking on objects, drag and drop objects, manipulate experimental settings, etc.
**CAT software modules**

The CAT software consists of a number of cooperating modules. NAEC’s IT department prepared these modules by programming the algorithms that were developed by NAEC’s psychometricians using C# as the programming language.

The starting algorithm randomly draws some 3 or 4 items of medium difficulty from the bank. Based upon the student’s answers, an initial estimate of his/her ability is made. The selection algorithm then chooses the next item, matching as closely as possible the estimated ability, and repeats this procedure till the stopping criterion is fulfilled. This occurs either after the first 24 items have been answered correctly, or if the estimation of the student’s ability has a standard error of less than 0.3 on a scale ranging from -10 to +10 (see page 16 for details). Due to the lack of relatively easy, well discriminating items, especially items with a difficulty around the cut score, most students receive the maximum number of items allowed, usually 40-50 items, in order to get an acceptable estimate of their ability. The selection algorithm rules out ‘similar items’ and makes sure that there is an acceptable spread over subject content categories in each individual test. Alongside these algorithms, an exposure control algorithm is running, which prevents items to be included in tests once they have been used for a specified number of times. Because the most informative items are the first to be ruled out in this way, the operation of this algorithm may eventually increase test length and the standard error in the ability estimates.

The IT department programmed a platform that manages the session of each student, based on observed ability, a stopping rule, and storage of the student responses. The platform also controls the exposure time of each item. In absence of a response after two minutes, the item will disappear from the screen and it will be assumed that the student gave a wrong answer. If a student needs less than 2 minutes to answer an item, the ‘unused time’ is saved in a ‘time bank’ and may be used for later items that may need more than 2 minutes to be answered. The interfaces with radio buttons for choosing options and an OK button for confirming the choice are a self-standing module but integrated in the platform.

The NAEC psychometricians monitored the pre-pilot with students and checked whether theoretical expectations were met in practice. Most algorithms needed some fine-tuning. For example, the selection algorithm worked in such a way that made it very difficult for a student to reach higher levels of true ability if a student, due to a simple mistake in one of the first items, found himself or herself in a low ability category. The algorithm was improved to minimize this effect.

Between 2012 and 2013, the efficiency of the platform was improved to accommodate more candidates taking part at the same time, using the same servers at NAEC. Another innovation was the introduction of ‘seeding’ (using live tests for pretesting items, see page 22 for details) for which the platform had to be adapted. A next improvement would be to allow students to interact with items while deciding on the correct answer.

**Connectivity**

The Virtual Private School Network, established and maintained by EMIS, was used to connect schools to the servers on which the CAT application was running. EMIS uses the services of Caucasus Online as a primary provider of internet services. The physical networks are owned and managed by Delta (glass fiber) and MAGTI (wireless connections). Delta services 570 schools with an internet speed of 50 Mbps, and MAGTI services 1,600 schools with a speed of minimal 1Mbps (see Figure 4 for an illustration of the Delta fiber network). As one student taking a CAT needs about 32 Kbps, the wireless connections provided by MAGTI should, in principle, still allow 30 students taking a CAT at the same time in one center (see Figure 3 for the network infrastructure for delivering CAT).
Delta reports that to support CAT, no specific additions had to be made to the network and services they were already providing to EMIS, and that the implementation of CAT did not increase the fee EMIS is paying to Delta. Delta’s CEO, George Jaliashvili, noted that Delta just signed a contract with MoES to connect more schools to the fiber-optic network, and that facilitating on-line examinations obviously was a driving force behind this.

For MAGTI the situation is different. Depending on the geographical location of the school and the number of students taking a test at one time, MAGTI has two options. For smaller centers, it uses a CDMA/EVDO connection (comparable to 3G, and used in the US for mobile networks) that provides a download speed of 3 Megabit per second (Mbps) and an upload speed of 1 Mbps maximum. For testing centers that need higher internet speed, MAGTI implements point-to-point wireless technology. In this case, the signal is sent directly from a MAGTI tower to an antenna at the testing center, with the tower and antenna in sight of each other. Distances between towers and antennae usually are three to five kilometers, but up to 25 to 30 kilometers is possible. The point-to-point wireless technology uses the IEEE802.11 standard (same as WiFi). According to MAGTI, protocols used make it very unlikely that the signal may be intercepted and decoded by non-authorized third persons.

The CDMA signal is transmitted over the existing MAGTI network. In principle, the common GSM signal (world standard for digital mobile phoning) could be used, but CDMA works better in remote locations and has fewer users than GSM, which diminishes chances of overloading the network. MAGTI towers transmitting the CDMA and/or WiFi signals are connected to the fiber-optic cables of the Delta-network.

EMIS pays a fee to MAGTI, which is related to the number of test takers and speed provided.
Initially the EMIS VPSN was connected to Caucasus Online through a one Gigabit per second (Gbps) G-link. The CAT-related and other traffic has increased to such an extent that in 2014 this connection will be upgraded to 10 Gbps.

**Capacity building**

**Psychometricians**

CAT algorithms are based on Item Response Theory (IRT). NAEC psychometricians had a good understanding of the theoretical aspects of IRT modeling, but little experience with putting these models into practice. It was for instance not clear how calibration of an item bank would work in practice. To address this issue, three staff members received two weeks of training conducted by CITO, the Dutch Institute for Educational Measurement (one week at CITO and one week in-house). Main topics included building CAT algorithms, the use of the One Parameter Logistic Modeling (OPLM) software, and the use of an Item Response Modeling software package for item calibration and test analysis. An additional training was done by two US scholars to familiarize the NAEC psychometricians with a similar software package (BiLog). The head of the IT department also participated in this training, as understanding of the underlying concepts was thought to be essential for adequately fulfilling the task of programming the algorithms.

Psychometric capacity building was completed by self-study. The three psychometricians, actually mathematicians by training, were highly motivated, fluent in English, and well-connected to the world of applied mathematical research. These may be seen as the main conditions of their eventual
success in developing the required algorithms independently. The international consultant hired to train the NAEC psychometricians confirmed that the knowledge and understanding they brought to the training was more than adequate and helped them master the content quickly. Complicated issues, especially in relation to the calibration of items in the bank, were successfully addressed. It is this very aspect (calibration of items in the bank) that is crucial to the validity and reliability of CAT. The international consultant advises that while there is a deep understanding of this issue, continuing efforts should be made to optimize the calibration.

**Proctors**

NAEC trains all new proctors. They undergo two sessions of about three hours to prepare and familiarize them with their tasks. The first part is theoretical, addressing responsibilities and procedures, including how to act in emergency situations. The second part is hands on, dealing with hardware and software. Existing proctors receive retraining (two sessions of 1.5 hour each) at the NAEC offices shortly before the testing period.

**Test developers**

For the test developers, three days of intensive training addressing all aspects of item writing were conducted by an international consultant.

**System administrator**

In 2011, EMIS hired and trained a new system administrator, 40 percent of whose tasks were CAT-related. In 2012, this function moved to NAEC.

**Helpdesk staff**

NAEC trained their helpdesk staff on answering frequently asked questions. There are two helpdesks operational during the CAT: one operated by EMIS for connectivity problems, and one operated by NAEC for problems related to the administration of the tests. In practice, most calls, including those that have to do with connectivity, are received by the NAEC helpdesk.

**Local IT school support staff**

Local school support staff provided various services to schools and were managed by EMIS. They received training to become familiar with the procedure of checking school facilities to serve as a testing center, and to prepare these centers shortly before the actual testing campaign starts. This training was provided by EMIS and took three days.

**Item development and calibration of item pools**

The items were developed by NAEC’s subject specialists. No existing items were converted, so all items were created especially for the CAT. Item writers were instructed to cover all content areas and difficulty levels. During calibrations of the produced item pools from pretest data, however, it became clear that there was a general lack of items at the lower ability levels, and more importantly, at the intended cut score.

After discussion and revision of draft items, the final versions were pretested in schools. Some 200 schools and 5,000 to 6,000 students took part in this pretesting, during which the items were administered as a short, linear computerized test. The pretests were observed by NAEC staff.

---

8 An offer from CITO to use a beta-version of its institutional integrated package for managing large scale CAT was declined, as the Georgian psychometricians and IT staff felt that they could develop their own package and wanted to avoid becoming dependent of a ‘black box’, having to rely on external services.

9 Prof. Dr. Theo Eggen, University of Twente and CITO and at that time chairman of the International Association for Computer Adaptive Testing
general finding was that students spent less time on items than had been anticipated, and that their motivation was low, casting doubt on the reliability of the obtained item parameter estimates.

For calibration of the pools of items, the NAEC psychometricians had a choice of two options: using the BILOG software package or OPLM (another software package). The former is a commercial package offered by SSI (Scientific Software International, US), the latter a product of the Dutch National Institute for Educational Measurement (CITO) with which the psychometricians familiarized themselves during their training at this institute. OPLM usually requires manual non-standardized adjustments during the last phase of calibration. BILOG's procedures are more standardized, but also while working with this package some arbitrary choices need to be made. In the end BILOG, came out as the preferred package. After it had become clear that too many items had to be rejected because they did not fit the model, the model was changed by moving from a 3-parameter to a 2-parameter one, letting go of the 'guessing parameter'.

During the calibration, the psychometricians worked closely with the subject experts who had developed the items. Even information on the behavior of distractors (incorrect options in a multiple choice item) was helpful for understanding the effect of the item and how to improve it. In cases that items had to be rejected because of a poor model fit while the item writer believed that it was a perfectly valid item, the final decision was left to the author.

BILOG generates a scale for item difficulties and student abilities, usually indicated with the symbol θ, ranging from -5 to +5. Student scores were converted to a scale of 5 to 10, with the cut score being 5.5, indicating the border between insufficient and minimal competence.

**Information and Advocacy**

**First announcement**

The first announcement of the reintroduction of the SGE happened quite unexpectedly. In early 2010, President Saakashvili unfolded the plan during a meeting with teachers. He was very critical about the state of education in Georgia and policies implemented by previous Ministers of Education, among which the cancellation of [national] graduation exams at secondary schools. According to the President, these had led to ‘school education losing its importance, especially in upper grades’. He announced that pass scores for three subjects would no longer be enough to be admitted to university, and that students would now have to pass exams in 10 subjects. In an interview with one of the national newspapers, the Ministry of Education confirmed that the school graduation exams and university admission exams would be unified and that ‘demonstration of minimum knowledge would be sufficient to graduate, but for admission high scores on subjects required by universities would be needed’.

---

10 Resonance, 13 January 2010. The full article is added to this report as Annex 4
Later, during a school visit, the Minister of Education Shashkin revealed that CAT would be the way to administer these new exams. Both announcements caused a lot of unrest and left stakeholders, especially students and teachers, with many questions as the Ministry had not been especially pro-active in setting up an information campaign. In fact, the Ministry did not even see the need for an information campaign and believed that a simple press conference would meet all information needs. NAEC offered that its public relations department would set up a coordinated campaign, but the Ministry declined this offer. There were some concerns at the Ministry that giving too much exposure to the plans would increase the commotion that had been caused by the first announcements and that the Ministry might even be forced to revoke the measure, something that Shashkin wanted to avoid at all costs. Then, in September 2010, Shashkin announced that the first nation-wide CAT SGE would be in May 2011. His announcement also seemed to suggest that the graduation and admission exams would not be merged. The poor public relations policy of the Ministry indeed resulted in students, teachers, and parents coming to the Ministry in protest, which demonstrations were eagerly covered by the media. ‘According to the forecasts of competent specialists working in the education sector, the majority of students will not be able to pass the minimum required thresholds in physics, chemistry, biology and math, because of the low level of performance, and therefore the majority of graduates (70 to 80 percent) will be left without certificates,’ one weekly newspaper wrote. Media complained about lack of information from the side of the Ministry and criticized the impromptu decision making.

NAEC’s campaign

\footnote{All News, 12-18 May 2011}
After this chaotic beginning the NAEC management, and especially its director (see Figure 6), actively started to work on closing the information gap, visiting all regions and seeing representatives of virtually all schools involved. Twelve major regional meetings were organized to explain how students should register for the exams and how they would receive their scores immediately after closure of a session. A working prototype of the test delivery interface was demonstrated and subject specialists showed sample items (Figure 7). For further information NAEC attendants were referred to the NAEC website, which also had a Frequently Asked Questions section.

At the same time the EMIS IT specialists started to inspect the school computers and connectivity. Schools reacted positively to this special and unusual attention paid to them, feeling that their concerns were heard and listened to.
NAEC produced a description of the testing process and sample materials, which was used as a base for many more informational brochures and articles. It then opened a special web page and started actively to talk to the media. At a later stage, a web-based practice test was launched and readily visited by students. Many schools would use their test center facilities to give students access to this linear mock test, which had about 50 items that were spread over all eight subjects. All this helped to take away some of students’ fears and emotions. However, the talks and explanations given by Maia Miminoshvili played a major role in eventually getting the schools on board. A publically known, respected and trusted figure, she managed to convince the schools that these tests would not be used against them and that the outcomes would be not as disastrous as many feared. A couple of days before the actual testing started, all 1,600 testing centers conducted a series of linear computerized tests with items from all subjects. In this way, all 45,000 registered students had an opportunity to familiarize themselves hands-on before the real testing began.

The role NAEC had to play was rather special, as it was the messenger in the first place and not the politically responsible entity. However, NAEC took responsibility for CAT and explained that the difficulty level of the tests would be such that no special preparations other than the training given in regular classes would be needed to pass. In addition, NAEC explained that the CAT would not result in major percentages of failing students.

In early 2011, NAEC started to invite media to come to their offices and attend the workshops that it organized for its own staff, with contributions of international experts. NAEC also prepared a more detailed explanation for the media of how the CAT algorithms worked. This open-door policy as part of NAEC’s mission to be as transparent as possible was highly appreciated. Media confirmed that for

---

12 See for instance 24 Hours, a national daily newspaper, of 9 May 2011, in which NAEC’s head of the Logistics Department explains in detail the principle and procedures of the computer-adaptive graduation exams (Annex 5)
information they relied on NAEC’s public relations department, website and publications instead of the official announcements from MoES. In advance of the first CAT, Maia Miminoshvili’s face was on the cover page of many a national and regional newspaper, and she appeared frequently on TV news broadcasts and talk shows. All this definitely helped to gain the confidence of the audience at large. Audiences expressed their happiness with ending the corruption that riddled the school-set exit exams. But many students were seriously worried by the rather late announcement of the re-introduction of the SGE. As the SGE were so close to the university entrance exams, many students feared that they could not prepare properly for either of them. Criticism focused on the timing and anticipated difficulty grade, though, not about the concept. People seemed to rely on the experience of NAEC to do a proper job. The confidence that NAEC had built up over the years with impeccable administrations of the university entrance exams was useful.

NAEC also started to make use of social media to inform test takers. A popular section of the NAEC Facebook page is the Question and Answer column. A list of Frequently Asked Questions is added as Annex 3. The NAEC Facebook page soon received 60,000 likes. In addition, Maia Miminoshvili’s personal page, where she gives updates on CAT and other NAEC assessment efforts, quickly became very popular with over 5,000 followers.

Last press conference

Just two weeks before the start of the National School Graduation Exams on May 24, 2011, the MoES and NAEC jointly organized a seminar for journalists on the objectives, expectations, and administrative procedures of the computer adaptive graduation exams. The minister announced that 50,000 students reporting at 1,500 test centers (schools) would sit the computer adaptive school graduation exams, being observed through 2,200 video cameras recording images and sound, and that with the help of 500 IT specialists a sophisticated and strong network had been created that would allow 15,000 students to log in simultaneously. The head of the NAEC IT department explained how security would be guaranteed, and why the much-feared computer breakdowns would not lead to the loss of any student answers and not cause major interruptions of the testing process. The head of NAEC’s logistic department announced that all students would get at least 15 items to answer, and no more than 45 items, for which they would have 100 minutes maximum. It was once again emphasized that the test questions would be on the easy side, and, moreover, that cut scores (minimum scores for passing) would only be decided after all students had taken the test, with the purpose of avoiding high failure rates.

School principals’ perception

School principals confirmed that in autumn 2010, after rumors had started to spread about the CAT, the Ministry called them to a meeting where they received more information. During this meeting the Minister, Shashkin, announced that the student outcomes would not only be used for certification, but also for accountability purposes. Principals and teachers whose students would perform below standards would be held accountable and in some way be punished. Indeed, after the 2011 test, some 200 teachers and a number of school principals were fired, allegedly for such reasons.

After this first briefing by the Ministry, the principals mainly relied on information coming directly from NAEC. Tbilisi-based headmasters found it easy to approach NAEC staff and went to ‘open house meetings’. Those outside Tbilisi went to regional meetings organized by NAEC. School principals who were interviewed all spoke very highly about these meetings and felt that NAEC had provided all necessary information and listened to them. When the principals informed their staff about the re-introduction of external graduation exams, reactions were mixed. In those schools where attendance
in grade 12 was a serious problem, teachers predominantly welcomed the measure. In schools where attendance in grade 12 was not a problem, such as in specialized and private schools, teachers felt that they did not need any external assessment. In addition, Shashkin’s threatening with punitive measures caused a lot of fear among teachers. Some principals even recommended that staff try to keep weak students away from the test.

Teachers’ perception

Teachers had to cope with the lack of information and rumors about the Ministry wanting to put schools under pressure on one hand, and with students – aware as they were about the general lack of preparation, especially in subjects they would not take for the university entrance exams – who were fearing the upcoming exams. Teachers say that it was useful to inform students about the duration of the test, the expected number of items and their format, the low level of computer skills required, and low difficulty of test items. It should be noted that this information was not clear at the time the Ministry made its first announcements, and such information only became available once NAEC started its information and advocacy campaign. Teachers really appreciated NAEC’s outreach efforts and say that after having attended a NAEC meeting they understood the exact nature and procedures of the testing, and felt confident that they could prepare their students adequately. They say that in later years, when information was timely and complete, students were much more at ease, and felt confident that they would manage.

Students’ perception

Students report that basic principles of CAT were explained to them by their teachers, but that brochures or flyers – if any – did not reach them. The NAEC web site, where more explanation and sample items were available, was frequently visited by students (as was NAEC’s Facebook page), but there are reasons to believe that students in more remote areas were either not aware of the existence of the site, or lacked opportunities to visit it. Some teachers printed the sample items from the NAEC web site and gave them to students during classes. For the students who had not been part of the mock exam one month ahead of the real test (10,000 students at 700 schools), the practice days offered on May 18 to 21, just before the start of the exams on May 24, was the first time they actually practiced CAT. For the vast majority this seemed to be sufficient preparation, although a few cases of ‘first-time mouse users’ presented themselves in rural areas.

Stakeholders’ understanding of the CAT concept

In spite of efforts made by NAEC to explain the principles of computer-adaptive testing, reactions from students, teachers and school principals, and also from media representatives, indicate that the details are still not widely understood. Usually it is clear that tests are generated from a large collection of items, and that each student gets a different test, which may not be of the same length. Many also understand that the difficulty of items varies depending to previous responses. The concept, however is only roughly understood, and its consequences are not well understood. The fact that an incorrect answer may be compensated by correct answers to following questions and should not necessarily affect the outcome wasn’t known by any of the interviewed students or teachers. If students would be aware of this, it might take the concern away that mistakes once made cannot be corrected later on. Also the fact that the final score is not directly proportional to the number of correctly answered items, and that for instance two students with the same number of correct responses may have very different scores, is mostly unknown, with students, teachers and principals alike. As yet this hasn’t been a problem, and no efforts have been made to appeal against a score. Whether in future information campaigns this issue should explicitly be highlighted is a bit of a dilemma for NAEC, as it may easily be misunderstood as an unfair aspect of the procedure.
Implementation of CAT

Testing centers

NAEC liaised with EMIS to establish testing centers in schools. The latter organization employs regional IT specialists, usually managed by the Regional Educational Resource Centers. Together with NAEC IT specialists these drafted quality criteria for the fitness of schools to serve as a testing center, the main ones being the stability and the speed of the internet connection. EMIS teams visited schools. Once a school had proved to be fit to serve as a CAT center, all IP addresses of the computers in the testing room were noted and a Google Chrome based software application was installed which – during the testing – would prevent making screen prints or dumps, any copying of texts or graphics, the use of external drives or other peripherals, and access to any site other than the NAEC CAT website (also see below). It was also checked whether all internet access outside the testing room could be switched off during the testing.

Figure 8. Testing room in 2013
One month prior to the first administration in 2011, 10,000 students from 700 schools took part in a large-scale test run of the system.

Each testing room has twice as many testing stations as the number of students that is admitted at the start of a testing period. The students are seated with one station in between them (see Figure 8). The formal test length is one hour and thirty minutes, but most students take less time, about 20 to 40 minutes. Sessions are scheduled for each hour and students are seated at one of the available testing stations. Only occasionally students have to wait a couple of minutes. There are no sessions scheduled between 2 and 3 pm, to serve as a buffer period in case waiting times have increased beyond expectations. In practice, this approach has made optimal use of the available facilities and proctoring capacity, at the same time guaranteeing maximal security.

Computers and internet connections

The first time CAT was administered (2011), brand-new netbooks were used in the testing centers. There was no danger of malware installed to breach security. CAT was running as a web application so the only security risk was at the server side. There, the usual precautions were taken, including firewalls, IP-filtering\textsuperscript{13}, and other standard methods to avoid malware entering the server.

In 2012, school-owned personal computers were used. A Windows shell application was designed for installation on these PCs that would deny access to the standard applications running under Windows, and replace the standard Windows interface. This application would connect to the web-based CAT application and present the items on the screen of the students taking the test. Installing this application was one of the tasks of the EMIS teams servicing the test centers.

During the testing all ports on the school router\textsuperscript{14} to which computers outside the testing room were connected, were closed. For instance, not even the school director could use his computer to go on the internet while the CAT was going on.

There is no chance of intercepting a signal between the server and testing centers and decoding it, according to the providers of network services. Some leaking of items after the test, due to students remembering and publishing them, is hard to avoid, though. NAEC actively searches the web and monitors social media for this. In the few cases encountered so far, the items were rather incorrectly reproduced and NAEC decided not to take further measures then to remove these items from the bank and inform the item writers.

Registration and administration.

Normally 85 percent of all Georgian grade 12 students take the computer-adaptive school graduation exams. The 15 percent of grade 12 students that does not take these exams are mainly students from minority groups\textsuperscript{15} (Azeri, Armenian), girls who marry before and leave school, and students who prefer to take the Russian Unified State Exam (ЕГЭ), which would give them access to places in Russian universities.

\textsuperscript{13} The server used a database of IP addresses of all computers in the testing centers. Only these computers, and only via the Virtual Private School Network could approach the web application running on the server. Other computers would not be given access.
\textsuperscript{14} Computers in schools are wired to a router. This router is part of a VPN that connects to the EMIS router, which in turn is connected to the server that has the item bank and CAT applications.
\textsuperscript{15} Of the Azeri and Armenian minorities only 15% of the grade XII students take the national school graduation exams. Computerized on-line school graduation exams are offered in Russian, Armenian and Azeri, be it that these are linear and not adaptive.
In 2011, 47,000 students registered, and in 2012, 45,000 registered. Registration takes place two months ahead of the testing and is done by the school principals by filling out an on-line form with details of the students who will take part. The language in which the student was instructed should be noted and which foreign language test the student will take. There is also a field for entering information about disabilities. NAEC can make provisions for some disabilities, such as more time to take the test for students with dyslexia and extra proctors to read out test questions for visually impaired students.

In many respects planning the administration of CAT is similar to that of any large-scale national test. Anticipating technical problems that might cause the loss of testing time, catch-up days were built into the administration schedule.

Proctoring
In each testing room was a surveillance camera, which recorded the full process of testing and which would provide evidence of any misbehavior of students, proctors, or other staff. Such incidents were very rare, though. Only one case was brought to light in this way, in which school staff pressed a proctor to help test takers.

For every testing room, there is one dedicated proctor and in centers with many students usually an extra proctor outside the testing room for keeping the order and do a first identification check. In each region, there is a set of back-up proctors to stand in if necessary. Proctors usually stay at the testing sites for two full weeks, going home for the weekend only if they live far away. One day before the actual testing starts, proctors assemble at the test centers they are assigned to and check if the internet is working properly and if the CAT applications are properly installed on the computers. There they also receive the list of test takers from the school director and are informed about the presence of students with special needs. Students should sign this list to confirm that they have seen and understood the procedures and rules of the school graduation test.

On the days of testing, the proctor is supposed to be present one hour ahead of time, check the room, and switch on the CCTV surveillance camera. Ten minutes before the test he or she is to start the registration of students, take any personal belongings that are not allowed in the testing room and sit students at the computers. He or she then helps the students to log in. This requires both a proctor code and student code. The latter is in a list the proctor receives just before the start of the testing, together with other documents such as the protocols the proctor needs to fill out at the end of every session. Five minutes before the start of the test the proctor logs in, using his Social Security ID (SSID) and the code of the testing center (school) he is proctoring. He then receives a 4-digit code which only remains valid for 30 minutes. The next step is entering a student SSID, upon which the system replies with showing the student name and a field in which the proctor enters this code in combination with the 4-digit student code from the list. This code is specific for the subject and time slot (morning or afternoon) allocated to the student in question. Then the testing starts. In principle, simultaneous testing of different subjects (the SGE requires each student to take 8 subjects) is possible and this would increase test security and help prevent leaking of items. At the moment only one subject is administered at the time. Teachers report that indeed efforts are made by students to recall items and make these available to students in later sessions, but did not seem to believe that this would seriously affect the testing. Also, so far, security has not been threatened by releases of items on the internet.

16 See Annex 2 for the Student ID card and protocols
As soon as the student receives a score, which signifies that the test is over, he or she should raise a hand. The proctor copies the score on the registration list, and the student signs in agreement. The proctor takes all scratch paper, on which students should have written their first and last name. Students are not allowed to take any papers home. At the end of the day, proctors are supposed to put all papers in sealed envelopes and sign them. At the end of the testing period, these envelopes should be taken to NAEC.

Testing program

In 2011, CAT started on Tuesday, May 24, with Georgian Language and Literature. The next days of that week were Modern Foreign Languages, History and Geography. Testing was continued on Monday (Chemistry), then Biology and Physics, ending on Thursday, June 2, with Mathematics. In 2012, the schedule was more or less the same. In 2013, the national SGEs were cancelled (see page 29 for details). In the 2013/2014 school year, the science tests were administered in October 2013.

The remaining tests are scheduled for May 2014.

During the testing the NAEC management had hot lines for all Ministries responsible for one of the aspects of large-scale secure testing: MoES; the Ministry of Health, which was responsible for maintaining sufficient medical services at the testing locations; the Ministry of Energy and Natural Resources that had to be contacted in cases of power breakdowns; and the Ministry of Internal Affairs that looked after security around the testing centers.
Helpdesk

Before and during the school graduation exams two helpdesks were operational: one operated by EMIS and the other by NAEC. The first time CAT was run on the existing school computers (May 2012) EMIS saw a steep increase in incoming calls. In March, the month preceding the CAT test runs in April, and May to June 2012, during the CAT delivery, the number of calls went up to about 8,500 a month, a 40 percent increase compared to February the same year (6,000 calls). In September and October 2013, before and during the CAT delivery of Science subjects the number of calls handled by EMIS was far less (1,900 calls). Still, this was 600 more than during the summer months and 400 more than in November 2013.

Monitoring and feed-back

NAEC

The NAEC management monitored the introduction of CAT by way of direct observations from teams of NAEC staff members visiting test centers. Virtually all staff were involved in this activity. In addition, protocols filled out by proctors after sessions played an important role in monitoring. On these forms proctors had to report technical problems, any irregularities caused by individual candidates, and measures taken.

At the national center where the CAT application servers were located, staff continuously monitored the access that individual centers had to the system. In case of any irregularities, they would immediately contact the center to find out what was wrong and if necessary take measures.

School Principals

During the first CAT delivery in 2011, it was difficult to get direct feedback on CAT from school principals while the testing was going on. Seeing the CAT introduction in the first place as a measure to hold them accountable for the low attendance rate of students in grade 12, and fearing measures of retribution, principals were quite reluctant to share any opinions at all. Once the student outcomes turned out to correlate well with the school grade point averages and low cut scores resulted in the low failing rates promised by MoES, principals started to speak up. Some even felt that the high pass rates undermined the objective to stimulate teaching and learning in Georgia, and also suggested that while top scores (8 to 10) definitely identified top students, the middle scores (6 to 7) did not really distinguish between middle-range students. Apart from several power cuts that were solved by using generators, principals did not report any major technical problems occurring at their schools.

Teachers

Teachers were not allowed to be present during the testing, and thus for evaluation of the tests, teachers had to rely on information students shared with them. It seemed that the student outcomes on the exams were in line with school results, although some report that weaker students performed better than expected. Hardly any computer problems were mentioned, apart from a few cases where computers seemed to be rather slow.

Students

Students also say that the procedures at the testing site were simple, clear, and straightforward. They knew when and where to report for the testing. The logging-in procedure was executed by the proctor without problems and few instructions were needed. The testing facilities were fine, not overcrowded, and spare computers were there if needed – which indeed happened when a few computers crashed during the log-in. There were two proctors, one of whom in the room and the other in the hallway checking ID’s and referring students to testing stations. Once logged-in, the
testing started smoothly. Some were disappointed that they did not get feedback telling them if they had answered an item correctly and that they could not go back to correct given answers. Students found it easy to read items on the screen as no scrolling is needed. Few problems were reported, only one case of pictures not opening in two chemistry items that forced the student to guess the answer.

Media
Newspaper journalists and TV presenters say that after the first successful administration of the CAT SGE the interest in the topic gradually faded away and the initial concerns died. The media still pays attention to the exams in short reports, but it is not front-page news anymore. Listeners calling in usually do this to inquire if there are any changes to the CAT SGE in a particular year. In 2013, when the security of the item bank could not any longer be guaranteed, the CAT SGE had to be cancelled (see page 29) and schools once again had to set and administer their own exit exams. Many people called the media and NAEC to urge NAEC to do everything possible to prevent the return of the corrupted school exams.

Costs
It is difficult to provide a comprehensive picture of the costs of introducing and maintaining CAT for the school graduation exams in Georgia. As new additional tasks are absorbed by existing staff and existing machinery are used for new, additional operations, many of the costs associated with introducing CAT are hidden. According to MoES, all national exams together (the University Entrance Exams and the School Graduation Exams) roughly cost 13 million GEL (USD 7.5 million), 30 percent of which is SGE and 70 percent of which is UEE17. Another source mentions 20 million GEL (USD 11.5 million) for the school graduation exams alone18, which may include the costs of purchasing computers.

Once CAT is up and running, the operational costs are usually lower than a comparable paper and pencil campaign. Main savings are in the ‘paperless’ administration and the proctoring costs. The way CAT is administered by NAEC brings about a saving of 50 percent in proctoring costs in comparison to paper and pencil testing and even to linear CBT, as only one proctor for each testing room is needed. The Ministry of Education and Science arrived at the conclusion that CAT would be more cost-effective than paper-based testing, if existing computer equipment at schools could be used.

Main cost items for the computer-adaptive delivery of the Georgian School Graduation exams are the following:

- Computers for testing centers; in the first year 21 million GEL were spent on buying the netbooks (‘bukis’) that were first used in the testing centers and then given to all grade 1 students. In 2012, computers were used. The Ministry had invested a large sum for equipping the majority of Georgian schools with computers, which also could be used in testing centers;
- Video cameras;
- Item writers
- Test administration costs (registration, test center management, NAEC office costs, transportation, accommodation and subsistence); and

17 Personal communication Mrs. Tamar Sanikidze, Minister of Education and Science
18 Kvela Siakhle (All news), 15-21 June 2011
• Proctors (the largest continuous cost item).

NAEC estimates the costs for item writing, test administration, proctor fees and providing transportation, accommodation and subsistence for proctors and their own staff during the testing campaign in October 2013 to have amounted to GEL 2,380,000. (USD 1,374,000). A breakdown of these costs is in Table 3.

The amounts in Table 3 are for producing and administering four subject tests. If all eight subjects would be administered in one period of time, as was the case in May and June of 2011 and 2012, the expenditures would increase by 80 percent, due to savings on transportation and accommodation costs.

The NAEC IT department spent about USD 80,000 for updating its equipment for running the CAT platform and delivering test items to the testing centers, when this task was moved to NAEC in 2013. In the years before, similar costs were incurred by EMIS and its forerunner ESIDA.

NAEC also bought 1,800 surveillance cameras at GEL 150 (USD 90) a piece. These were left at the testing centers in the schools, but remained NAEC’s property.

Another non-trivial cost factor is the need to equip all testing centers with a power generator to ensure the continuity of testing during electricity cuts. The Ministry ruled that schools had to buy these from their own budget. Some schools purchased generators, while others borrowed appliances for the time of testing. A minor part of the budget is needed for setting up the connections of schools with the national CAT server, which was carried out by the local IT support staff supervised by EMIS.

### Current issues and future development

#### Continuity of management

A special problem during the first CAT years were the management changes in 2012. On 28 May 2012, when the school graduation exams had just begun, the NAEC director was fired by the Minister of Education for political reasons. Almost all of NAEC’s professional staff left in sympathy with the director. Later that year, after parliamentary elections had resulted in a regime change, the former director was brought back to her old position and most of the former staff was re-hired. Unfortunately, it turned out that security measures had not been enforced under the new director. As a result, items had been exposed, computers with confidential files had disappeared, and archives had been deleted. NAEC was forced to rebuild item banks almost from scratch, and had to cancel the 2013 CAT. In that year, schools had to set and administer their own tests once again.

#### Curriculum structure

An issue in timing the school graduation exams is the way the school curriculum is organized. In most schools, the science curriculum (Biology, Chemistry, Physics and Geography) is completed in grade 11.
and not taught in grade 12. In the 2013/2014 school year, these subjects were tested in October, and
the remaining four will be administered to the 12th graders in spring 2014. Also 11th graders will then
be taking biology, chemistry, physics and geography. This scenario will be continued in the following
years: all subjects will be administered in spring, be it sciences to 11th graders and Georgian
language & literature, math, modern foreign languages and history to 12th graders.

Item bank quality
The quality of the item bank was not optimal. There were not enough items, especially those
measuring at the lower end of the ability scale, and some items, in spite of their acceptable
psychometric quality, lacked face validity. Item writers need more training in authoring items. Also,
better communication between the item writers and psychometricians would help. This could be
brought about by allocating dedicated psychometricians to subject groups, with a good
understanding of the subject content.

Another factor affecting the quality of the item bank in the first two years was the reliability of item
pretests. Due to lack of motivation of students taking part the item parameters sometimes differed
considerably from those obtained in real testing. This caused serious problems during the item
calibration, the process of estimating the psychometric parameters such as difficulty and
discriminatory power of all items on one and the same scale. The estimated parameters showed
undesirably large standard errors, which again would cause large standard errors in the ability
estimates of individual candidates, and, even worse, especially around the cut score. Once the CAT
had been introduced NAEC started to ‘seed’ new items in the live testing sessions and used the
results for establishing their parameters, which provided much more reliable item parameters.

A third factor is item exposure. There were eight sessions per day for each subject and as a result
several items were repeated a couple of times during the day. Students tried to remember items and
put them up on social media, e.g. Facebook, after the session they sat. Although there weren’t that
many items that made it to Facebook, NAEC carefully monitored this phenomenon, and concluded
that the effect on the reliability of the testing would be minimal. Statistical analyses comparing
results of students from earlier and later sessions confirmed this. In 2013, an exposure control
algorithm was implemented in the CAT software (see CAT software modules, page 13).

Test Validity
Test validity has been a concern from the start. The strict use of multiple choice items dictated by the
use of CAT brings along certain limitations to what can be reliably assessed. Another validity issue is
the preciseness of the ability estimates produced by the system. With the help of simulations the
NAEC psychometricians adapted the CAT algorithms in order to minimize errors in these estimates as
much as possible.

Quality of testing stations
The state of the computers in schools caused many problems. In the first year, when the brand-new
netbooks were used, this was not an issue. But once the existing computers had to be used, about
half of them were not in a condition to be used for CAT. In addition, internet connections were slow,
in some cases too slow to run tests for more than one student at the same time. Students of schools
that for such reasons could not serve as a testing center, usually small rural schools in remote,
mountainous areas, were allocated to the nearest testing center. Most of these students needed a
place to stay for the entire testing period, as the distance was usually too big to go home for the day.

19 Seeding: including non-calibrated items in live tests. The student answers are excluded from determining
their ability, but used to establish the item parameters for future use in CAT efforts.
Many stayed with relatives, and for some paid accommodation was provided by MoES. In one case students were even flown in by helicopter, from Shatili in Chevsureti to a testing center in Kutaisi. Nevertheless for most of these students having to take the test in another school than their own constitutes an extra physical and emotional burden. For that reason, efforts will be made to also create testing centers in small and remote schools.

**Future developments**

No important changes are foreseen for the school graduation exams in 2014. The policy agenda of the Ministry is still ‘under construction’. Most likely, new item types will be introduced in the coming years.

**Evaluation and lessons learnt**

**Evaluation by MoES and NAEC**

No strict protocol is maintained for evaluating the quality and effects of the introduction of the computer-adaptive school graduation exams, nor at the Ministry nor at NAEC. At NAEC the evaluation of the CAT operations is the subject of a continuous discussion between the institutional management and specialists involved. So far, technical and logistical issues seem to be well under control, given experience NAEC acquired in other large-scale high stakes testing efforts. A major issue now is the reliability of the testing, or more precisely the precision of the ability estimates that determine the selection of items and the stopping rule. While the users are generally satisfied, and some ‘concurrent validity’ of the computer-adaptive tests might be suggested by the fact that teachers find the scores correlating highly with the outcomes of school-based tests, psychometricians suggest that additional fine-tuning of the existing algorithms is needed to minimize the error in ability estimates.

Another issue is the reliability of the psychometric item characteristics obtained by pre-testing. Recalibration of items confirmed that the estimation of theta’s (ability indices) was generally too low. The psychometricians considered the introduction of CAT as an opportunity to break new and exciting grounds, but at the same time they were struggling with doubts on the validity of the theoretical models they were using. In the end they agreed that ‘the perfect model’ did not exist, and that they were using the best of the available models.

The Ministry of Education and Science receives a report from NAEC with basic statistics (participation, average scores, pass rates) and protocols from observers. From these Mrs. Sanikidze, the Minister who assumed this position in September 2013, concluded that the current school graduation exams provide a level playing field to all students and are definitely a better tool than the previous teacher-made assessments. She was happy to note that the exams did not disrupt the learning process and were not affected by technical problems. In addition, the ‘4x4’ model (the 4 sciences at the end of grade 11, the other 4 at the end of grade 12) was positively evaluated.

MoES conducted an analysis of the costs of CAT as compared to administering the same tests by paper&pencil, and arrived at the conclusion that CAT was cost-effective, if at least existing computers at schools could be used.

The Minister of Education notes that while the introduction of CAT SGE has met its goals, it cannot cover all relevant outcomes of education. It did away with the subjective assessment of students at the end of grade 12 and it did help increasing attendance rates in the last year of general education,
but certain skills such as communication, presentation and higher-order problem solving are not assessed. In combination with school grades it is a useful tool for certification, but not for selection of students for academic studies, which is precisely the reason why the two exams at this point in time cannot be integrated (which is on the wish list of MoES).

The CAT SGE outcomes also revealed once again the problematic situation of teaching and learning in Georgia, as reflected by the low cut scores that were applied to have politically satisfactory pass rates. The average achievement in Georgian schools turned out to be far below expectations.

Stakeholder opinions
School Principals
Schools principals generally evaluate the computerized grade 12 school graduation exams positively. The fact that the outcomes of the SGE CATs correlate well with the school grade point averages of their students adds to their positive involvement. Principals seem to like computer based testing, as is also proven by the fact that NAEC is now successfully selling linear on-line tests for grade 6 and grade 9 school exams.

School principals believe that the re-introduction of the school graduation exams is helping to improve the quality of teaching and learning in Georgia. Suggestions are made to introduce similar exams as well at other interfaces in education, e.g. at the end of grade 9. CAT technology is seen as a fair way to deal with assessing what can be assessed with multiple choice type items. However, they regret that this makes it impossible to see the actual items and receive feedback on the individual item level. They also wonder if and how students may appeal against a score, if the actual items and responses cannot be seen. They do appreciate the feed-back they receive from NAEC, which consists of the scores of all of their students, and an indication how the school did in specific fields of learning.

Figure 10. Web page with CAT feed-back at school level
Teachers
Teachers (and students) confirm that initially, the chaotic and scarce information caused a lot of uncertainty and many questions were raised. Once details of the CAT approach became clear, they started to wonder how the obvious difference in difficulty level between the individual tests could lead to fair scores. Also, the fact that once an item had been answered a student could not go back to change the answer, as is the case in pencil and paper testing and most linear computerized tests, was an issue of much concern. In the end, the detailed explanations given by NAEC helped stakeholders understand how CAT operates to arrive at fair assessments of students’ abilities and accept the approach. Teachers mention the positive effect of the external testing SGE on student motivation. Some say that this has especially given a boost to the science subjects, which were neglected before 2011. Teachers also appreciate the objective, merit-based (meaning non-corrupt) character of the CAT. They agree, however, that the assessment of skills that cannot be tested with multiple choice items, such as speaking and writing, or presentation skills, should become part of the national SGE. Last but not least, they feel relieved that the Ministry does not seem to use the student outcomes for accountability purposes, and that the threatening measures announced by Shashkin have not been adopted by the Ministers who came in after him.

Students
The interviewed students experience the CAT SGE as fair tests, in that they are objective and not too difficult, but they were aware of the limited validity, in that important skills (for instance speaking and writing for modern foreign languages) are not assessed. They acknowledge that due to the re-introduction of national school graduation tests students once again started to study all subjects, and not just the four needed for university entrance, and that numbers of students attending classes in grade 12 have increased. They doubt, however that the exams have lowered the amount of extra-curricular tutoring, and, on the contrary, it is now expanding to tutoring for CAT SGE.

Service providers
Delta nor MAGTI experienced any technological problems during the CAT administrations, and did not expect the network to become overloaded as CAT is based on web technology and does not generate a lot of traffic, even not when many students log in at the same time and graphics are used in test items. However, MAGTI’s CEO, David Lee, emphasized that while on paper this may all seem quite straightforward, an experienced company must set up the necessary infrastructure and implement the technology. There should be an existing network, as building extra towers for CAT (at a cost of about USD 100,000 each) might be too expensive, and the provider should have proven capacity and be able to guarantee continuity.

Comments in the media
Comments in the media after the first administration in 2011 were generally positive, which definitely was helped by the fact that cut scores were low and pass rates relatively high. While at first the diploma awarding rule was ‘8 subjects with a score of 5.5 or higher’, it was decided to also award certificates to students whose score was 5.2 in three subjects and 5.5 in the remaining five. The minimum score for Georgian Language at non-Georgian schools was set at 5.1. Usually, out of the 50,000 graduates in a year, about 35,000 apply for a university place. The graduation exams proved to feasible for almost all of them.

Newspapers reported positively on the technical and security aspects of CAT, but did not pay much attention to the impact these had on education. There were a few articles on the limited validity of CAT, pointing out the restrictions to the use of multiple choice type of questions. When at one point
the Minister of Education (Shashkin) announced that, after the successful introduction of CAT for SGE, he was now considering to introduce it for the university entrance exams as well, this raised a lot of concern and negative commentaries in the media as most experts did not believe that such tests would properly fulfill the selection function of these exams.

The weekly magazine 'All News' interviewed a representative of the oppositional New Rights Party. The representative points at the fact that 13 percent of the students failed,²⁰ which in her opinion is not a small number. And that this is the result of applying low competence thresholds, which in fact should have been higher. To her this merely reflects the low level of teaching and learning at Georgian schools. ‘The fact is that our current schools do not give education to students. In itself the idea of exams is not bad, but it would have been better to do them two years later, giving students more time to study the required subject and avoiding all this stress. Presumably the project cost was GEL 20 million or more. This money could be better spent for the improvement of teachers’ qualifications.’

Also in 2013, after the administration of the science tests in October, the focus of the media was on pass rates. For Chemistry, Biology and Geology only 5 to 6 percent of the students scored below 5.5 on a scale of 1 to 10²¹. For Physics, 15 percent of the students failed. Educational experts interviewed by the daily newspaper Resonance comment that this is not a true picture of the level of achievement, and that especially in Physics the situation is far more alarming than outcomes suggest. Some condemn the low cut scores as a way to cover this up and point out that the cut score is close to the guessing score.

At the last testing day in October 2013, journalists working for the web magazine from Edu.Aris.ge visited 51 schools and noted that ‘Students with sparkling eyes and happy faces were leaving the exam room one after another’. They had conversations with some of the students, who said to have been nervous before and during the tests, but found that their scores were ‘normal’. A parent, however, was concerned about the rise in testing. She wondered why next to the university admission tests school graduation exams were needed and why these were not taken into account for university entrance purposes. ‘Then what are they required for? Could we not use the money for other purposes and needs of our children? Are we making tutors richer?’ she complained.

Lessons learnt and caveats

Many international assessment experts who followed the initiative to introduce CAT in Georgia were quite skeptical about chances of success. Taking into account the huge investments in human and material resources that had to be made in similar efforts elsewhere in the world, in countries with a long-standing tradition in educational assessment and an established technical and scientific infrastructure, it seemed impossible to launch a technologically and logistically complicated instrument as a large-scale computer adaptive test within the very short span of time imposed by the Ministry of Education and Science. ‘Lessons learnt’ therefore is in the first place about the critical success factors in this seemingly impossible but obviously successful mission to use computer adaptive testing for the national school graduation exams. But also about a number of caveats that are important for the sustainability of this effort.

²⁰ All grade 12 students take the CAT school leaving exams, but not all grade 12 students take the university admissions exams. The 13 percent of students failing the CAT are mostly those who do not aspire to go to university.
²¹ In practice only the 5-10 part of the scale was used. All scores below 5 were reported as 5.
Success factor 1: Strong commitment by the government

Deciding to introduce a national test and to use a technologically advanced delivery mode brings along an immediate commitment to fund the necessary initial investments, a long-term commitment to fund the recurrent costs and to guarantee continuity in all operations concerned. The Georgian government has a strong record of investing in fair assessment of student outcomes on a continuing basis and was ready to shoulder these obligations. This, and the way NAEC operationalized the governmental policy to free education from corrupt access mechanisms, have done a lot to convince the audience at large of the use of educational testing.

Success factor 2: NEAC’s leadership and trust among stakeholders

Successful implementation of large-scale high stakes testing efforts often is a matter of informing and motivating all involved, next to bringing the right technical expertise. While of course without the latter a campaign is doomed to fail, it is the former that makes the difference. All stakeholders and experts involved in the Georgian School Graduation Exams agree that it was the way NAEC – and more specifically its director Mrs. Maia Miminoshvili – managed to get the schools on board that made the short term nation-wide rollout of computer adaptive testing possible. A publically known, respected and trusted figure she managed to convince the schools that these tests would not be used against them and that the outcomes would be not as disastrous as many feared. The effect her personal appearance in many meetings with school leaders and teachers, in press conferences, TV shows and seminars has had can hardly be underestimated. The solid reputation NAEC has as a producer and administrator of external tests helped a lot, of course. But this reputation is also built on the intensive and frequent personal interaction of NAEC’s management with stakeholders. Accessibility and transparency, listening to stakeholders and proving that they are heard are core elements of NAEC’s approach that motivated all involved to cooperate in the daunting CAT expedition. The introduction of any new, large-scale testing effort needs careful piloting. The immediate nation-wide introduction of computer-adaptive school graduation exams in 2011 was a difficult scenario. The success was due to the readiness at all levels to make things happen, the flexibility and willingness to improvise where needed and to accept hick-ups, a willingness that stakeholders in more developed systems would probably not demonstrate.

Success factor 3: NAEC’s psychometric and ICT competence

International psychometrics consultants state that what has been achieved in Georgia in terms of implementation of a large-scale, high-stakes computer adaptive testing effort within a very short time is unique in the world. According to them this could only happen thanks to the high-level expertise in the field of psychometrics and ICT at NAEC and dedication and flexibility of all involved. NAEC has managed to attract and retain these experts and support them in their development as a team. The way these experts cooperate in developing assessment instruments is indeed remarkable and has been crucial in producing the CAT software.

Success factor 4: NAEC’s experience in large scale secure testing

NAEC has 10 years of experience in designing and executing large-scale high-stakes tests, and has always been keenly aware of the importance of careful management of the proctors they use in these campaigns. NAEC puts a lot of effort in selecting and training them. Proctoring is a valued position, as many proctors have been in place since 2005, the year the first UEE was administered. They are proud of their job, they get proper recognition and remuneration for it, and they feel that they make an important contribution to the struggle for eradicating corruption at all levels, which is successfully fought in Georgia.
Success factor 5: Anticipating and avoiding network problems

A serious risk of any on-line testing effort is an overloaded network and breakdowns causing student data getting lost. The smart scheduling of individual sessions helped to keep the number of students logging in simultaneously as low as possible. The way the CAT algorithms were programmed as a web-based application helps to keep the amount of internet traffic low. The program saves all keystrokes made by students on the central server, which facilitates resuming testing after a power break or computer crash without loss of data. Indeed the number of problems caused by network breakdowns has been minimal and never led to the loss of student data. Students who experienced computer failures could log in on another machine and pick up their session at the point where it had been interrupted without losing any of the responses they had entered before the breakdown.

Success factor 6: Understanding the effects of scaling up

Still, some technical problems only become manifest once the system becomes operational on a full scale. The human factor also plays an important role. The only way to familiarize all involved with the system and the setting and make them act appropriately and effectively is to have a full-scale pretest under realistic conditions. For this reason, NAEC organized such a test run shortly prior to the real tests. For schools and students this was an ideal and much appreciated opportunity to fully understand the test conditions and what was expected of them.

Caveat 1: Test validity

Comments from teachers and students indicate that on the long run the validity of the school graduation exam may become a problem. A solution must be found to assess those skills that cannot be tested in a computer-adaptive way and make the outcomes part of the final scores. But also efforts should be made to develop more sophisticated CAT items. Till now, only the simplest item format has been used (stem+four alternatives, one correct). Efforts should be made to include more item formats allowing for interaction of the candidate with the item, such as drag and drop or simulation of scientific experiments. It is often believed that, due to the necessity of using machine-scorable items, higher order skills cannot be assessed in a computer adaptive testing setting. While indeed certain productive skills such as writing and speaking need open-ended formats, this is not generally true for all higher order skills. Many machine-scorable items have been developed that reliably and validly assess a variety of problem solving skills.

Another factor affecting the validity of the tests is the increasing amount of coaching that is going on. As with all testing, the SGE may lead teachers to training students for these exams rather than educating them. But on the other side, the introduction of national SGE’s and the strict security of the computerized administration gave back value to the school graduation certificate, which it had lost entirely when the school were running their own assessments. Most grade 12 students returned to the classrooms from which they had been absent, spending time with their university entrance exam tutors. However, tutoring may move to the SGE’s, which seems to be on the rise.

Caveat 2: Test reliability

Till now, the CAT testing is precise enough in a small band of the ability scale, but certainly not at all levels. There are simply not enough items at the higher end of the scale, and especially not at the lower end of it, around the cut score. This creates uncertainty in pass/fail decisions. It takes time to properly populate the bank. Regular pretests proved to be unreliable due to lack of motivation of students. Seeding produces satisfactory results, but is a relatively slow process. NAEC psychometricians point out that calibration is still a problematic operation as many items seem not fit the theoretical model used.
More items in the bank will also facilitate longer testing times. It may not look good to estimate student’s abilities developed over several years with a test of just 30 minutes. While psychometrically spoken the estimate may be adequate, the ‘face reliability’ may not be. The stopping rule should therefore be adapted.

Caveat 3: Test standard and negative backlash effects
A special issue in the Georgian CAT SGE’s are the low cuts scores. While politically expedient for the high pass rates (95%), they seem to encourage minimalistic behavior in the classroom. They are criticized by educational experts who say the tests are far too easy. Others say it is a reflection of the low level of teaching and learning in Georgia, and that cut scores should gradually rise with improving education. This is also the opinion of MoES, which is currently investing in upgrading teacher certification and classroom assessment.

Caveat 4: Test publicity
Comments from stakeholders indicate that the security of test items may become a problem, in the sense that it makes it very difficult to file an appeal against an obtained score on the basis of allegedly flawed items or other test defects. Proper legislature should be introduced to protect NAEC from being forced to release and publish items. At the same time the rights of candidates who are put at a disadvantage due to flaws in the test should be protected, for instance by creating the opportunity to retake the test when the outcome is unexpectedly lower than a student’s grade point average.
## Annex 1 Persons interviewed and documents consulted

### Interviews

<table>
<thead>
<tr>
<th>Institution, company</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ministry of Education</td>
<td>Tamar Sanikidze, Minister</td>
</tr>
<tr>
<td>World Bank</td>
<td>Nino Kutateladze, Task Team Leader</td>
</tr>
<tr>
<td>Delta Comm</td>
<td>George Jaliashvili, Executive Director</td>
</tr>
</tbody>
</table>
| MAGTI | David Lee, President  
David Mujirishvili, Director IT Operations  
Nino Avaliani, Chief Customer Care and Sales Officer |
| EMIS | Lasha Verulava, General Director  
David Saghinadze Director Statistics Department |
| Media | Mari Otarashvili, Resonance and TV Channel 9  
Nino Datukishvili, Public Radio&TV  
Keta Tsvitsinadze, Real TV and Prime Time (newspaper and press agency) |
| NAEC | Maia Miminoshvili, General Director  
Iwa Mindadze, Deputy Director  
Dato Chankotadze, Director IT Department  
Merap Topuria, Director Logistics Department  
Maia Gabunia, PR Department  
David Gabelaia, psychometrician  
Misha Mania, psychometrician  
Mamuka Jibladze, psychometrician  
3 proctors |
| Schools | 5 teachers/school leaders from Telavi (Kacheti)  
4 students from Tbilisi  
2 school leaders from Tbilisi  
4 teachers from Tbilisi |
## Documents

<table>
<thead>
<tr>
<th>Source</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resonance (daily newspaper)</td>
<td>13-01-2010</td>
</tr>
<tr>
<td></td>
<td>20-05-2011</td>
</tr>
<tr>
<td></td>
<td>02-06-2011</td>
</tr>
<tr>
<td></td>
<td>19-10-2013</td>
</tr>
<tr>
<td>All News (Kvela Siakhle) (weekly newspaper)</td>
<td>12_18 -05-2011</td>
</tr>
<tr>
<td></td>
<td>15_21 -06-2011</td>
</tr>
<tr>
<td>24 hours (daily newspaper)</td>
<td>09-05-2011</td>
</tr>
<tr>
<td></td>
<td>06-06-2011</td>
</tr>
<tr>
<td>Netgazeti</td>
<td>14-10-2013</td>
</tr>
<tr>
<td>Edu.Aris.ge (web magazine)</td>
<td>16-10-2013</td>
</tr>
<tr>
<td>CAT FAQ (as published on the NAEC web site)</td>
<td>2013</td>
</tr>
<tr>
<td>GeoCAT Action Plan (NAEC)</td>
<td>2010-2011</td>
</tr>
<tr>
<td>GeoCAT Guide for School principals and IT staff</td>
<td>2013</td>
</tr>
</tbody>
</table>
Annex 2 Student Identification and Testing Protocols

The student identification card is checked before entering the testing room, using the School Graduation Exams Protocol. Once the student is seated, the proctor enters the session password upon which a window pops up in which he enters the student 4-digit password. Then the testing starts.

**Student Identification card, front and back side**

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

**School Graduation Exams Protocol**

<table>
<thead>
<tr>
<th>School Graduation Exams 2013</th>
<th>Abkhazian Resource Center</th>
<th>Public School Nr. 2 Abkhazia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name Candidate</td>
<td></td>
<td>Location Test Centre Public School Nr. 2 Abkhazia</td>
</tr>
</tbody>
</table>

Date, and time of the subject tests. In last column: Practice test, Geography, Physics, Biology, Chemistry

**School Graduation Exams Testing Room Protocol**

<table>
<thead>
<tr>
<th>Public School Nr. 2 Abkhazia</th>
<th>37 students</th>
<th>5 students</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Graduation Exams Testing Room Protocol</td>
<td>Proctor's signature</td>
<td>11.10.2013 9:00 Practice Test Session: 1</td>
</tr>
</tbody>
</table>

Personal Number, Last Name, Name, Password, Score, Signature, Warning/Disqualification

5 students
Annex 3 Frequently Asked Questions
The list below is a compilation of the question and answers from NAEC’s Facebook page

- **How many exams will the student pass for graduation exams?**
  A student will pass 8 exams for graduation exams, these are: Georgian language and literature, foreign language (one), mathematics, history, geography, physics, chemistry and biology.

- **Could the threshold be changed?**
  Defined threshold shall not be changed.

- **Will the student see whether answer is right or wrong, while answering the question?**
  The student will not see whether answer is right or wrong, the program will simply go to the next question.

- **Will the score of graduation exams have influence on awarding the medal?**
  The golden medal is award given by the school and the school issues it to the student which had the best results during learning period. The students which get golden medal normally will participate in graduation exams. Naturally, if the candidate of medal fails in an exam, s/he will not get the medal.

- **How complicated will be the tests in biology, chemistry and physics?**
  In 2011 the tasks for graduation exams (among them in natural sciences) basically will be selected from the simple block – the student will be able to overcome threshold based on simple tasks.

- **If the computer comes out of order or internet connection fails during the exam, what will happen with the student?**
  If the process of exam is impeded the student will not have problem because of this, all his/ her answers will be kept and after elimination of the problem he/she will continue the work, if the interruption lasts for long time the student will pass the exam the other day.

- **Will the number of questions be defined?**
  Number of questions for graduation exams is not defined. The advantage of such exam is that the program selects the next question for the student according to the answer on particular question, until determines his/her capacity.
  In case of one student 30 questions can be asked, while in case of another -18.
  Maximal number of the questions will be determined in advance.

- **Will the graduation exams be in Azeri and Armenian languages?**
  The graduation exams will be conducted in Azeri and Armenian languages, in the same format.

- **If any student fails to overcome the threshold, will she / he have possibility to go to national exams?**
  The student which failed to overcome threshold on the exam and could not take diploma will not be able to participate in national exams.

- **Is number of students in exams sector defined in advance?**
  Number of the students in the sector is determined and will not be more than 15.

- **Who will supervise the diploma exams of 12th grade?**
The supervisors of diploma exams are selected in advance by national exams centre, like this was done for national exams. After the training they went through testing, according to the results of which the bests will participate in the diploma exams.

- **Will the student be able to use the periodicity system and solubility table at the exam in chemistry?**
  - The student will have possibility for using periodicity system and solubility table at the exam in chemistry.

- **Will the student have possibility to use the map at the exam in geography?**
  - At the exam in geography the student will have possibility to use the map.

- **Will the student have possibility to correct mistakes at the graduation exams?**
  - Cat System selects the next question for the students based on previous answer and correction of the answer is not possible with this system.

- **Are graduation exams mandatory for the students of private schools?**
  - The students of private school will also pass the diploma exams and as a result of passing these exams they will receive the diploma issued by the Ministry of Education and Science.

**Will the time for answering each question be limited at the diploma exams?**
- The student will have 2 minutes for answering each question, besides he/ she will have so called time limit.

- **Will the student that wins the Olympiad have any advantage?**
  - Usually the winners of Olympiad will pass the diploma exams.
“Revolution in Educational” - 10 exams required from 2011

Saakashvili starts “Great Activities” together with Shashkin and states that the previous reforms lead by Lomaia, Nodia and Gvaramia done in an “ignorant” way

Mari Otarashvili
(13.01.2010)

Alexander (Kakha) Lomaia, Gia Nodia and Nick Gvaramia are former ministers of education who implemented the education sector reform after the Rose Revolution. Yesterday in his speech President Saakashvili criticized these reforms and called the persons who had implemented them “ignorant”.

The President said that the new Minister of Education and Science Dimitri Shashkin is going to start an absolutely new reform, opposite to the previous one, and the most important element of this reform will be the requirement to pass 10 exams for the university admission.

Yesterday during his meeting with teachers of the 3rd public school Saakashvili talked about mistakes made in the education system and called the initiators of so called “seasonal teaching” “ignorant”. He also criticized cancellation of graduation exams at secondary schools and change of school textbooks every year.

Saakashvili said that many families could not afford to buy a new textbook every year. He believed that a standard must be developed for textbooks that would be used from year to year.

The President also stated that efforts must be made to interest children in sports. For this purpose schools will cooperate with different sports schools and children will have an opportunity to go there.

The President also stressed the issue of safety and catering at schools. According to him providers of catering service were mostly selected based on nepotism. Big companies must start service provision at schools. They will be responsible for the food quality; besides children will be able to buy food by means of credit cards. As for safety, Saakashvili believes that elder and younger children must be separated and pupils of upper grades must be more closely controlled.

Saakashvili said that starting from 2011 admission to universities will depend on the results of graduation exams. According to his explanations passing of three exams will no longer be enough for university enrollment. School graduates will have to pass exams in 10 subjects be admitted to universities based on the results. These subjects will include math, physics, chemistry and computer skills. Saakashvili explained that the reason for this change was that the school education had lost its importance, especially in the upper grades. Pupils missed classes and had private tutors only to pass the exam in skills which was required for entering universities.

"The main subjects are no longer necessary; as a result we received an amateurish situation when students can enter universities only based on the result of the skills test. This must end once and for all. We must restore the prestige of school graduation certificates and main subjects," – said Saakashvili. The President also talked about the necessity of introducing new subjects such as military/patriotic education and study of culture. Saakashvili said that Georgia’s future is in the education and improvement of the education system.
It must be noted that the constantly changing textbooks, introduction of trimesters, closing of special schools for talented students i.e. everything criticized by Saakashvili in his yesterday’s speech was a result of the education reform implemented by his own government after the Rose Revolution. It is also important to note that at that time professors of Tbilisi State University were actively protesting against the decision to neglect natural sciences at schools and to give the utmost importance to the skills exam. At that time their opinion was disregarded by the President and officials from the Ministry of Education.

“Resonance” interviewed the Head of the PR Department of the Ministry of Education Ms. Nino Potrzhelskaia about the format of these 10 exams which the graduates would have to pass next year.

Nino Potrzhelskaia: “Implementation of the education sector reform discussed by the President has started today. In 2011 graduates will have to pass 10 mandatory exams for admission to universities. These 10 exams will be organized by the National Examination Center which means that they will not be conducted by schools. In reality these exams will be simultaneously school graduation exams and university entrance exams. The exact list of the 10 subjects will be finalized jointly by Ministry of Education and Science and the National Examination Center. At this moment I cannot tell you which subjects will be selected, but I can tell you for sure that IT technologies, math, Georgian, physics will definitely be in the list … in reality these will be school graduation exams and university entrance exams at the same time. Those graduates who want to receive the school graduation certificate will have to pass these 10 exams.”

“Resonance”: “How will students be able to prepare 10 subjects in this situation when the education process at schools has completely failed? Is it not a little early to make this decision for 2011?”

The Ministry of Education and Science: "It should not be difficult for pupils to get ready for these 10 exams. These are the same subjects that they have been studying at school. Besides we are asking them to demonstrate at least the minimum knowledge. It means that graduates will have to pass several main subjects. The graduates will need to receive high scores in the subjects required by their selected universities. In other subjects the graduates will just have to pass the minimal required competence threshold."

At this stage the Ministry of Education and Science does not talk about any other details.
Like many new ideas, the idea of holding high school graduation exams has been advanced by President Michael Saakashvili and as always the Georgian government has made haste to put it into practice this year. The proposal raised a wave of protests among students and their parents, however, the government has responded quite strictly: “We will not let D students to make a revolution”. Nobody planned to make a revolution anyway. The problem is that the quality of learning at schools has significantly dropped for the past 20 years. Moreover, the differentiated learning system has been promoted in the country and general education has been almost labeled as a soviet heritage. Therefore most of the students have studied only those subjects they need to enroll in higher educational institutions.

Nobody argues about the need to improve the quality of education in Georgian schools, but when it comes to the future of young people we should be more cautious. As for examinations, we already know that they are computer-based tests. All Georgian schools have been already connected to the network. The exams will be based on the so-called CAT principle, which means Computerized Adaptive Testing. Education Minister Dimitri Shashkin said that they had brought strongest servers in the South Caucasus to the country and installed them. Moreover, they equipped schools with special computers. However, the system encountered some problems when a computer-based test was held in test mode. They say these exams cost the country a lot of money, however the National Examinations Center officials say they are not aware of the amount.

Students have 100 minutes to answer test questions – conditionally 2 minutes per question. The remaining 10 minutes serve as a reserve and can be used by a student if s/he needs this time. If a student answers questions earlier than 2 minutes, the remaining time adds to those 10 minutes.

Merab Topuria, head of the Logistics Department of the National Examinations Center, has answered All News’ questions regarding the issue.

- Could you please explain the CAT principle of exams and its advantages?
  - The abbreviation CAT means “Computerized Adaptive Testing”. This is the examination system successfully practiced in the USA and European states. It helps accurately identify the level of knowledge of a student in a specific discipline. Unlike Unified National Examinations, which belong to the selective type of exams, graduation examinations are considered to be certification exams. The CAT is deemed most efficient for the certification exams. The computer software identifies accurately whether a student knows a subject. It does not consider whether one student is better than the other by the level of knowledge.

- Could you please tell us the order, in which examinations will be held and the time interval between the examinations?
  - The examinations will start on 23 May with the Georgian language and literature exam and will end on 1 June with the math exam. One examination will be held each day. On 26 May and 29 May students will have a day off.

- It is known that a number of schools, especially in regions, encountered technical problems when the CAT was run in a test mode. What kind of difficulties were there? Is it possible to resolve them before the examinations start?
The goal of holding the exam in test mode was to resolve technical problems. Some schools in regions had a low-speed Internet connection, Magti (a telecommunication company) and the Ministry of Education are now taking measures to deal with this problem.

How should a student act in case of a technical problem? This may become an additional stress factor, which will hamper a student’s ability to continue the test successfully.

If the examination stops because of some technical problem, all answers and the time left for the examination of a student will be saved. The students will not face problems because of technical malfunctions. I want to assure that they will not suffer because of that.

If a student becomes sick and is not able to take any graduation exam, will s/he be given a chance take the exam this year?

A student will be able to take an additional exam this year only in exceptional cases, i.e. s/he will present a legal document that proves that a student was not able to attend the exam because of a force majeure circumstance or health conditions etc., each such case will be considered by a special commission and the students will be allowed to take an additional exam only in special cases. Otherwise, a student will have to sit for an examination next year (the same rule applies to the Unified National Examinations).

What items are students allowed to take to an examination room?

Students are required to have only their ID cards, an invitation to an examination and a pen. If necessary, students may also take water and items used for personal hygiene.

They say that even if a student fails two exams, s/he will be allowed to take the Unified National Examinations anyway, however s/he will have to pass these exams next year. Is that really so?

This is false information. A student will be given a high school diploma only when s/he passes all exams. Only those students that have high school diplomas are allowed to take the Unified National Examinations.

It is known that questions are divided into easy and medium questions. Are the questions that differ in complexity evaluated in the same way?

A student is evaluated by the computer software that uses special algorithms. The principle of work of the algorithms is published on the webpage of the National Examinations Center (www.naec.ge) and all interested persons may refer to it.

What is the minimum competency level? What is the number of questions students have to answer correctly to pass an exam? How and when will students know whether they passed an exam or not?

The number of questions a student must answer correctly to overcome the hurdle is not defined in advance. The software gives a new question to a student based on the answer of a previous question. Some students may answer around 40 questions, while others 20. The software stops giving questions, when it identifies the level of knowledge of a student. However, whether this is going to be enough to overcome the hurdle is a different question. A child will be able to see the number of points s/he received upon the end of the test, as for the hurdle, the minimum competency level will be identified after the examinations are over, following the analysis of the results.

Many people think that many students this year will not be able to get high school diplomas, where will they be able to continue their studies?

Nobody knows how many students will fail to get high school diplomas until the examinations are over. This is going to be the first year of holding exams in this way, therefore we will take into account the gaps in the education process at schools. This is why students this year will have rather simple examination programs. Briefly, we will try our best to make it possible for most of the students to get high school diplomas.