The Introduction of Large Scale Computer Adaptive Testing in Georgia

Introduction

In the late 90ties of the previous century an article appeared in a Dutch national newspaper calling for teachers to join efforts in contributing items to a large open-source item bank, from which an endless number of individualized tests might be generated, to be administered on-line at any suitable moment and fully integrated in the educational process. In cooperation with a monitoring system into which test outcomes and other information about student achievement would be fed, this might be a way to achieve a higher pedagogical goal: continuous education, not interrupted by large-scale tests. Students would still be tested and assessed, though, but hardly notice it.

It was the time that expectations of using computers to facilitate a smoother administration of more valid student assessments were rising high. Simple gaming technology had entered the classroom as a means of instruction, often in combination with (self)-assessments. But providers of large-scale and high-stakes assessments still continued to rely on paper-based techniques to administer their tests or examinations.

In 2003, for The IAEA conference in Manchester together with my colleague Gerben van Lent, I presented a paper that also assumed a strong role for using computer technology in the exams of the future. *'Computers should finally enable us to deliver national exams on demand, creating virtual real-life situations for candidates who will be allowed to demonstrate their real skills rather than their book learning. For students who are used to playing adventure games over the internet, it would only be normal if similar techniques would be used for the delivery of national exams. If we could gaze into a crystal ball and catch a glimpse of the exams of the future, we would not be surprised - in fact we would rather expect to see - students in the centres of secure virtual private networks connected to test banks, marking machines, data analysers and reporting engines. Results would be available in real time, of course…'* as the paper stated.

The paper concluded that at that time already many sophisticated applications had been developed for every key aspect of large scale testing, but that integration of such modules with on-line delivery tests still remained a problem. There were no adequate software programs commercially available for full computer-based delivery of exams, so institutions had to develop their own proprietary delivery platforms fit for serving large numbers of candidates at the same time while maintaining security. While the adequacy of the infrastructure outside testing centres (limited capacity of the cable network and servers) was often mentioned as an important source of uncertainty, a Dutch investigation of the feasibility of on-line large-scale testing indicated that, at that moment, the infrastructure within the schools (which would serve as testing centres) would not allow for a fully-fledged computer-based delivery of exams, over the web or from software installed on a centre-based server.

In 2011 after many years of experimenting, the Dutch Board of Examinations, the body that is responsible for the quality and secure administration of the national exams, commissioned the production of an integrated system for on-line testing to a commercial software provider. It was expected that by 2017 all national examinations would be fully computerized, meaning: secure administration of mainly linear tests on-line, using various item formats, also constructed response ones.

In Georgia the Ministry of Education and Science decided in September 2010 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. International experts were quite sceptical about the feasibility of a nation-wide rollout of such a logistically and technologically complex measurement instrument as a large scale CAT at such short notice.
In May 2011 44.000 students sat eight computer adaptive subject tests in one of the 1500 test centres established in Georgian schools. 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.

Why does it take so long to introduce ways of on-line assessments in large-scale high-stakes exams in countries with long-standing examination traditions, such as the UK, the Netherlands and even the US, and why could it be introduced in Georgia almost overnight? Is this another example of the dialectics of progress?

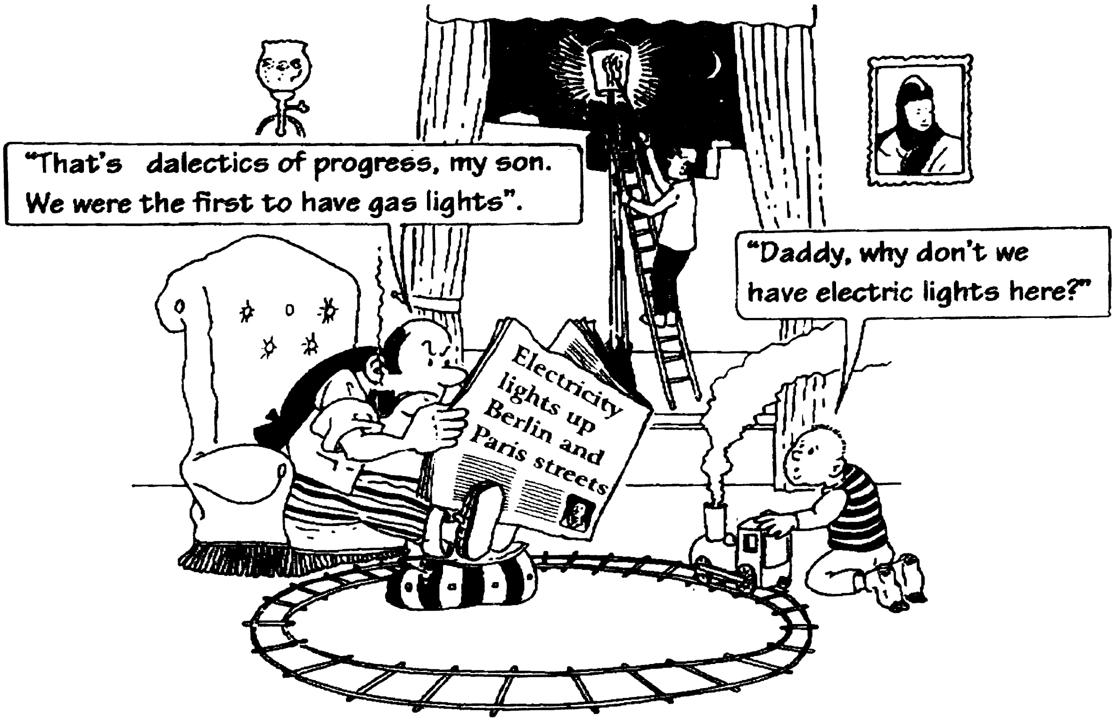

"That's dalectics of progress, my son. We were the first to have gas lights".
"Daddy, why don't we have electric lights here?"
Electricity lights up Berlin and Paris streets

Figure 1 The Dialectics of Progress

Some research commissioned by the English Ministry of Education in the first decade of this century aiming at finding out in what way on-line test taking and on-screen marking of student responses would influence student outcomes suggest a certain reluctance to step away from traditional modes. And it is true that in Georgia no year-long tradition of sophisticated paper&pencil exams with elaborate constructed-response questions and detailed marking schemes for human marking was in the way of introducing computerized machine-scorable tests.

The presentation will set out to identify the main success factors in the introduction of computer adaptive testing for large-scale high-stakes tests in Georgia and caveats for further use and development. The basis for this presentation is a background paper I did for the World Bank to support policy making for adopting CAT or CBT for national exams in other countries, and for which I interviewed stakeholders and actors in the process of introducing CAT in Georgia.

(http://siteresources.worldbank.org/INTREAD/Resources/Bakker_Introduction_to_CAT_Georgia_for_READ.pdf)

Why CAT? Political Context

In 2010 the MoES decided to replace the school-administered secondary school leaving exams – which since the introduction of the University Admission Exams quickly had lost their currency - by a national exam, to be administered by NAEC. This meant secure testing of 45.000 school leavers in their own school or a close-by facility. Gathering them in a few large centres for the time of testing – probably two weeks – would be impossible for various reasons: costs, stress and logistical complications. Computer-based testing immediately seemed the best option to avoid the costs and security risks of printing and moving massive amounts of papers. NAEC advised CAT because this way of testing makes more efficient use of an item bank and ability estimates would be more reliable than in linear testing. 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 definitely appealed to the Minister.

Planning the CAT; Some Figures

The following figures reflect the main investments in human resources, hardware and infrastructure that were done to make the CAT happen:

- Three psychometricians trained by CITO and US psychometricians
- Additional training for test developers
- 2300 proctors trained and certified by NAEC
- 200 regional IT school support staff trained
- Item banking software developed
- CAT algorithms developed
- Item banks for 8 subjects developed and calibrated
- Servers and routers purchased for national centre
- 1800 surveillance cameras bought by NAEC for test centres
- 11.000 computers purchased by MoES for use in the 1600 schools that were going to serve as a testing centre
- 1600 testing centres to be connected to the internet (570 glass fibre, 1100 wireless connections)
- Twelve major regional information meetings held by NAEC; brochures and web-based practice tests prepared, Q&A on NAEC Facebook page, mock tests for all students (45.000) in all 1600 testing centres

Testing Centres

Connectivity

At the start of planning the CAT, one of the main questions was how to connect the 1500 testing centres to the main server, allowing for sufficient speed and bandwidth to have many thousands of students taking the test at one time without the system going down or connections getting lost. And what to do if this would happen, nation-wide or locally.

The existing Virtual Private School Network was used to connect schools to the servers on which the CAT application was running. The physical networks supporting this VPN 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. 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 centre.

MAGTI has two options. For smaller centres, 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. The CDMA signal is transmitted over the existing MAGTI network. In principle, the common GSM signal 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 Wi-Fi signals are connected to the fibre-optic cables of the Delta-network

For testing centres 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 centre, with the tower and antenna in sight of each other. Distances between towers and antennae usually are three to five kilometres, but up to 25 to 30 kilometres is possible. The point-to-point wireless technology uses the IEEE802.11 standard (same as Wi-Fi). According to MAGTI, protocols used make it very unlikely that the signal may be intercepted and decoded by non-authorized third persons.

All data of students logged in to the system were continuously buffered at the central server. If the system would go down or internet connections were lost students should be able to resume as soon as the system would be up again or connections restored without answers they had given before the breakdown getting lost. Fortunately a nation-wide breakdown did not happen, and local ones were very few.

Security

The main security measure is the presence of well-trained and motivated external proctors in each testing centre. Also, a number of measures were taken to prevent access to outside sources during the testing and item leakage.

Once a school had proved to be fit to serve as a CAT centre, 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. It was also checked whether all internet access outside the testing room could be switched off during the testing.

The first time CAT was administered (2011), brand-new netbooks were used in the testing centres. 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[1], 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 centres.

During the testing all ports on the school router[2] 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 centres 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.

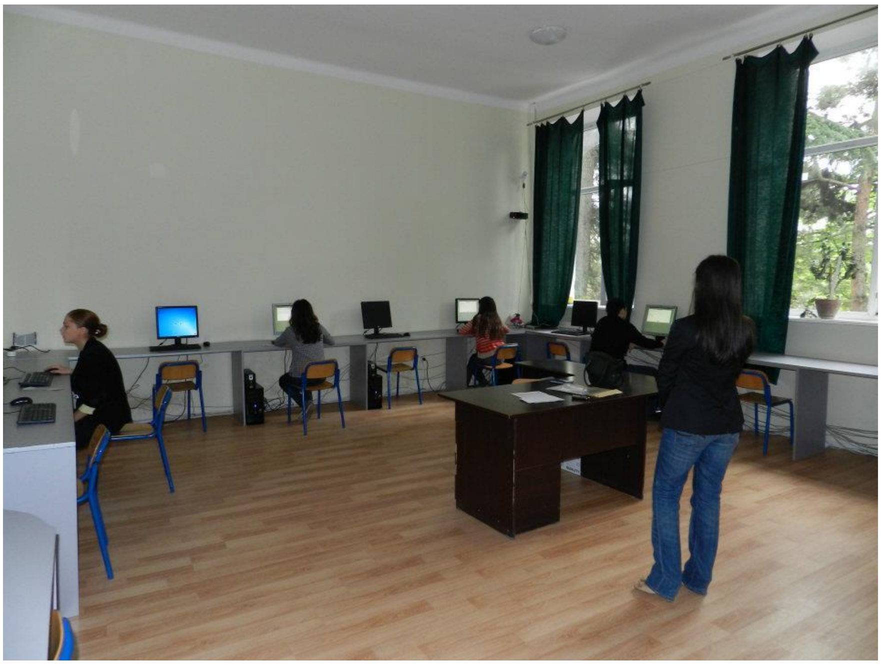


Figure 2 Testing room in 2013

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 2). 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

[1] The server used a database of IP addresses of all computers in the testing centres. 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.

[2] 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.

beyond expectations. In practice, this approach has made optimal use of the available facilities and proctoring capacity, at the same time guaranteeing maximal security. 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.

Administering the CAT

Two months before the administration student registration takes place (2011: 47.000; 2012: 45.000). This is done online by the school principals. Students receive a message indicating where and when they are expected at a testing centre for sitting a test.

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.

On the testing day proctors log in and subsequently log in the students. Several measures apply to prevent impersonation and unethical or even illegal conduct by proctors. Students receive their scores immediately after the test.

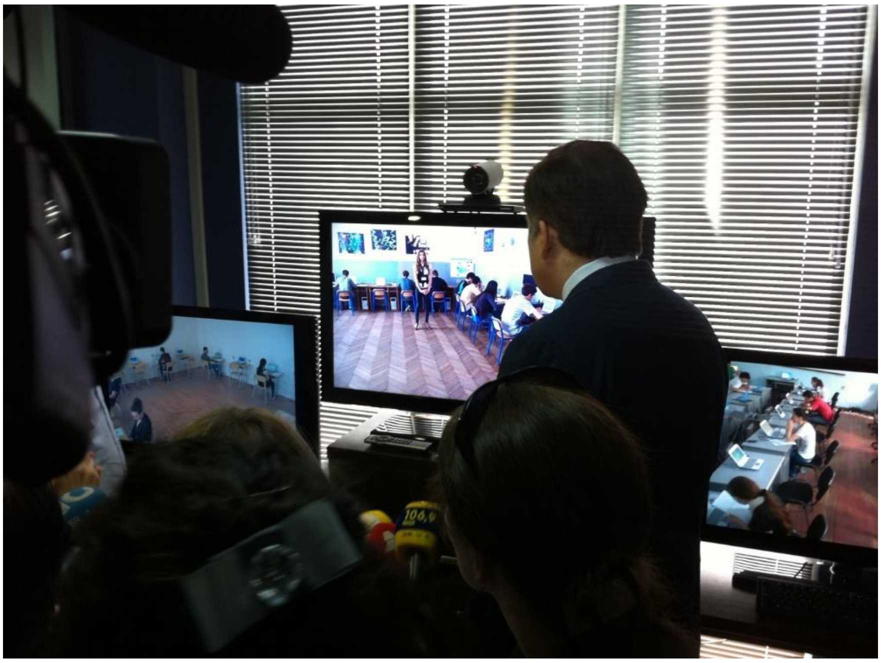


Figure 3 . President Saakashvili visiting the Educational Scientific Infrastructure Development Agency, where in 2011 the servers hosting the computer adaptive graduation exams were located. Video connections between a few testing centres had been set up, enabling the President to watch the process and have a chat with students before the testing started.

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 had to be cancelled as a consequence of an unfortunate incident that happened the year before. 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.

(see page **Error! Bookmark not defined.** for details). In the 2013/2014 school year, the science tests were administered in October 2013. The remaining tests were administered in May 2014.

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. Main cost items are the following:

- Computers for testing centres; in the first year 21 million GEL (9,4 million Euro) were spent on buying the netbooks ('bukis') that were first used in the testing centres and then given to all grade 1 students. In 2012, personal computers were used. The Ministry had invested a large sum for equipping the majority of Georgian schools with pc's, which also could be used in testing centres;
- Surveillance cameras;
- Item writers
- Test administration costs (registration, test centre management, NAEC office costs, transportation, accommodation and subsistence); and
- Proctors (the largest continuous cost item).

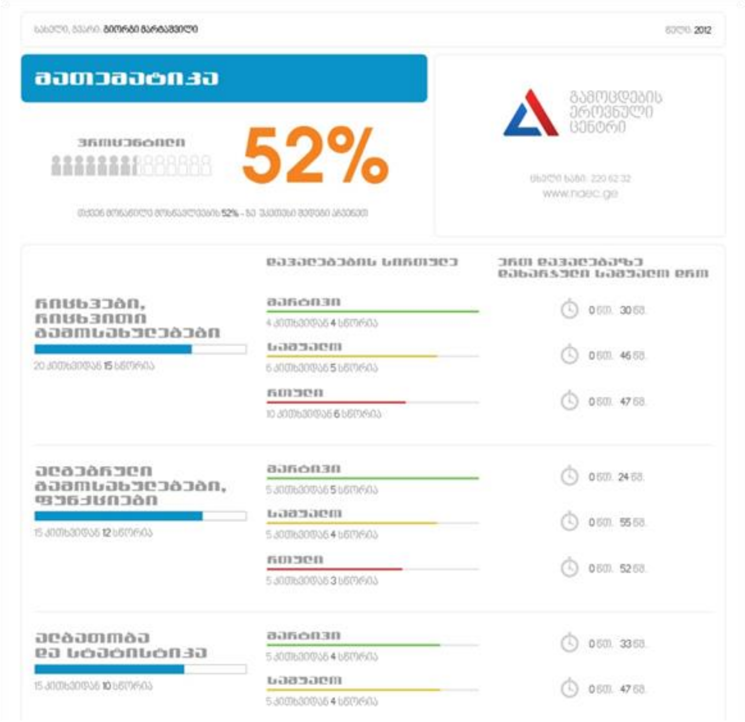
NAEC estimates the costs for item writing, test administration, proctor fees and providing transportation, accommodation and subsistence for proctors and their own staff for one campaign (8 subjects) at 4,28 million GEL (1,92 million Euro).

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.



სახელი, გვარი: გიორგი გარგაშვილი — წელი: 2012

მათემატიკა

პროცენტილი 52%

[illegible] 52% - ზე [illegible]

გამოცდების ეროვნული ცენტრი

ცხელი ხაზი: 220 62 32
www.naec.ge

	დავალებების სირთულე	ერთ დავალებაზე დახარჯული საშუალო დრო
რიცხვები, რიცხვითი გამოსახულებები 20 კითხვიდან 15 სწორია	მარტივი 4 კითხვიდან 4 სწორია	0 წთ. 30 წმ.
	საშუალო 6 კითხვიდან 5 სწორია	0 წთ. 46 წმ.
	რთული 10 კითხვიდან 6 სწორია	0 წთ. 47 წმ.
ალგებრული გამოსახულებები, ფუნქციები 15 კითხვიდან 12 სწორია	მარტივი 5 კითხვიდან 5 სწორია	0 წთ. 24 წმ.
	საშუალო 5 კითხვიდან 4 სწორია	0 წთ. 55 წმ.
	რთული 5 კითხვიდან 3 სწორია	0 წთ. 52 წმ.
გეომეტრია და სტატისტიკა 15 კითხვიდან 10 სწორია	მარტივი 5 კითხვიდან 4 სწორია	0 წთ. 33 წმ.
	საშუალო 5 კითხვიდან 4 სწორია	0 წთ. 47 წმ.

Web page showing results for a school with overall achievement of the school as a percentile score (52%, also indicated as a relative position within a group), a breakdown over subject content categories (blue bars to the left, achievement at different levels of difficulty (green bar: percentage of easy items correct, yellow bar: percentage of medium items correct, red bar: percentage of difficult items correct. Also time spent on the various items is displayed (column to the right)

Figure 4. 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 be 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 of5.5 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, 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 10[3]. 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,

[3] In practice only the 5-10 part of the scale was used. All scores below 5 were reported as 5

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.

Success factors and Caveats

Looking back to the first three years of using Computer Adaptive Testing in Georgia for the school leaving exams, a number of factors may be identified that explain why this has become a success, in spite of the fact that a system had to be developed from scratch in an environment with only limited experience with computer-based testing.

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. D**oubts 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.