Computerized adaptive testing in the Monitoring and Evaluation System for primary education in The Netherlands

36th IAEA Annual Conference
Bangkok, Thailand, August 22 – 27, 2010
Topic: Theory and it application

Henk Moelands
Cito, P.O. Box 1034, 6801 MG Arnhem, The Netherlands
T: +31 26 352 1562, F: +31 26 352 1135, E: henk.moelands@cito.nl

1 Introduction

The Dutch educational decentralization policy, aimed at given schools more freedom and making the more responsible for the quality of their education, has resulted in a growing interest in procedures on the pupil and school level. Schools are held more responsible for the quality of the education they provide and therefore have to carry out an active policy of quality control, on both levels.

In the Netherlands, the National Institute for Educational Measurement (Cito) has developed the monitoring and evaluation system for primary education. This system consists of a coherent set of nationally standardized tests for longitudinal assessment of pupil achievement throughout education including a system for manual or automated registration of pupil progress. The system for primary education contains not just tests for measuring sub-skills of language (including decoding and reading comprehension) and mathematics, but also tests for social-emotional development and study skills such as using study texts and schemes, tables and graphs. The results of the successive assessments are converted to the same fixed scale (this possibility is offered by a measuring technique based on item response theory) so that the progress of pupils can be monitored over a number of years. Though the primary purpose was to provide a unified system that enabled schools to follow the position and progress of individual pupils in a number of subjects, the system gradually evolved to serve a dual purpose: apart from providing schools and teachers with detailed information on individual pupils, it also gives information on higher levels of aggregation, such as the grade, the school or even the regional clusters of schools.
A next development of the monitoring and evaluation system is computerized adaptive testing. Since recent years the system contains several computer-based and computer adaptive tests for grade 1 and 2.

In this paper an overview of the Cito Monitoring and Evaluation System will be given. First the content of the system and the measuring technique will be discussed, next, attention will be paid to computerized adaptive testing in general and finally, an example of a computerized adaptive test for grade 1 and 2 will be schown.

2 Benefits of Cito Monitoring and Evaluation System

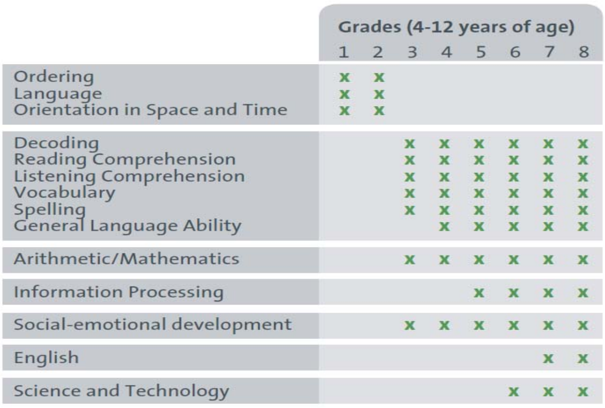
For instructional guidance it is important to have a good insight into the learning progress and a good knowledge of the strong and weak sides of a pupil. For day-to-day monitoring of a pupil's progress the classroom teacher can use common assessment techniques such as observations, teacher-made tests, performance assessments and portfolio assessment. Although classroom assessment is important, it has its limitations because it is subjective and person-related. Different teachers could assess the same performance differently. It is also possible that the assessment of the teacher is not consistent over time and over learners. This is very likely to happen in those cases where there are no clear objective standards. If teachers do not have the same framework of reference for assessing and recording achievements, then it is very difficult to realize continuous assessment and monitor a learner from year to year and from school to school.

Another problem is the rate of progress: how does the teacher know that a learner has made sufficient progress within a certain period of time? Or: how does the teacher know that sufficient assessment criteria have been met? Or: is the pace with which the teacher move through the teaching and learning process adequate? There is an almost inevitable danger according to international research that the pace slows down in systems that rely only on school-based assessment.

Cito's Monitoring and Evaluation System helps teachers in primary education obtain reliable data about the progress in the pupils' learning processes systematically. The system complements the knowledge that the teacher has of the learners on the basis of day-to-day progress assessment. As there is also continuity in the recording of results, all the teachers in the school have the same frame of reference. This is of great importance for early identification of any problems. Moreover, the Monitoring and Evaluation System allows the user to combine a norm-referenced and a domain-referenced (or content-referenced) interpretation of results that are gathered nationwide. The latter makes the system also very informative for curriculum developers and policy makers.

3 A short outline of the Monitoring and Evaluation System

The monitoring and evaluation system developed by Cito consists of a coherent set of nationally standardized tests for longitudinal assessment of a pupil's achievement throughout primary education as well as a system for manual or automated registration of pupil progress. The system contains tests for measuring subject skills of Language (including decoding and reading comprehension), Arithmetic and the social and emotional development of pupils. An overview of the various tests in the system is given in figure 1.



	Grades (4-12 years of age)							
	1	2	3	4	5	6	7	8
Ordering	x	x						
Language	x	x						
Orientation in Space and Time	x	x						
Decoding			x	x	x	x	x	x
Reading Comprehension			x	x	x	x	x	x
Listening Comprehension			x	x	x	x	x	x
Vocabulary			x	x	x	x	x	x
Spelling			x	x	x	x	x	x
General Language Ability				x	x	x	x	x
Arithmetic/Mathematics			x	x	x	x	x	x
Information Processing					x	x	x	x
Social-emotional development			x	x	x	x	x	x
English							x	x
Science and Technology						x	x	x

Keys to symbols:

Grade 1-2:	Kindergarten (nursery school or reception year)
Grade 3-4:	Foundation Phase
Grade 5-6:	Intermediate Phase
Grade 7-8:	Final Phase

Figure 1. Tests in the Cito Monitoring and Evaluation System for primary education

During the primary school period tests are usually taken once or twice a year. The results of the successive assessments are converted into a fixed scale for each subject in which a pupil’s progress over a number of years is monitored. The continuity in the collection of data is of great importance for early identification of any problems. In this way the Cito Monitoring and Evaluation System complements the impression that the teacher has of the pupils on the basis of day-to-day progress assessment. Moreover, the nationally standardized tests of the system make it possible to widen one's view beyond the classroom or the school. Thus the results of the pupils can be compared nationally with those of other children.

Working with the system does not merely involve testing and the registration of test results. It is an Educational System that allows teachers to make decisions about the progress of the learning process on the basis of the data collected. Should the data indicate that the pupil is not performing well, the problems will then have to be analyzed and, where needed, appropriate remedial actions will have to be taken. Therefore the system has been set up as a procedure that calls for a systematic, cyclic approach.

In the systematic approach three stages can be distinguished:

1. Identification
   This implies all the activities that have to do with recording the pupil's achievements and interpreting the results (testing, marking of the tests, registration and preliminary interpretations).
2. Analysis
   Should the results of the test show that the pupil’s development is not up to standard or that it even stagnates, then it is desirable to collect additional data. Firstly to verify the signal and secondly to pinpoint specific problems or gaps. The system offers the teacher the equipment to carry out this analysis.
3. Actions
   On the basis of the information of the former steps a specific plan of remedial actions can be set up, carried out and evaluated. Wherever useful and possible, exercises and directions for use are provided for teachers.

To sum up, the Monitoring and Evaluation System comprises the following elements:

- a coherent set of tests for Language/Reading, Arithmetic, World Orientation (including information processing), Social-Emotional Development, Science and Technology, and English;
- a recording system, based on a measuring technique with the help of which scores are comparable on the same fixed scale in the course of time;
- means and procedures to detect the nature of the learning problems;
- didactic directions for specific help.

4 Item Response Theory as a measuring technique

It is desirable for a system that is aimed at monitoring pupils' achievements over a number of years that the various tests of a subject matter measure the same abilities and that the results can be put on the same fixed scale. Only then it can be determined to what extent a pupil has made progress compared with a previous measurement. This possibility is offered by a measuring technique based on item response theory (IRT). IRT presents a general framework for constructing measuring instruments, validating measurements, estimating item and test characteristics, estimating individuals' abilities and spread of abilities in (sub) populations and it provides a framework for interpreting test results. In the IRT model used in the Monitoring and Evaluation System the chance that an item can be solved is specified as a function of a latent one-dimensional pupil ability and one or more item characteristics (e.g. difficulty). The difficulty of the items and the latent ability can be represented on a same scale. If the model fits, the scale that measures the ability is calibrated with the help of the estimated item characteristics. This is done with the help of OPLM, a computer program developed by Cito based on a One Parameter Logistic Model.

Particularly the fact that both pupil abilities and item characteristics can be put on the same scale and can be related to each other is of great advantage to a Monitoring and Evaluating system:

- The results on tests that differ according to difficulty, contents and number of items can be compared. In other words: Thomas' results on the math tests of mid grade 4 can be depicted on the same scale as the results he obtained six months before on the math test of end grade 3, so that the degree of progress can be determined. Furthermore, the position that the pupil takes on the scale can be compared to that of other pupils nationally.
- On the basis of the position on the scale a general conclusion can be drawn about the degree of mastery of a particular subject matter.

Figure 2 gives an example of a scale consisting of several types of math items and the ability estimate of a pupil (Thomas) on the basis of the test results on Arithmetic tests that have been taken at six months' intervals (June ’09, Jan ’10, June ’10) half way the school year and at the end of the school year. Thomas’s position on the scale (June ’09, end Grade 3) indicates that he has mastered the type of items that is below his ability level (11 + 7), but that the items that are above his ability level are still too difficult for him. Items that are on the same level are partially mastered. Six months later his ability has increased (Jan ’10, medio Grade 4). Now he has mastered items that were too difficult at an earlier moment (counting backwards).

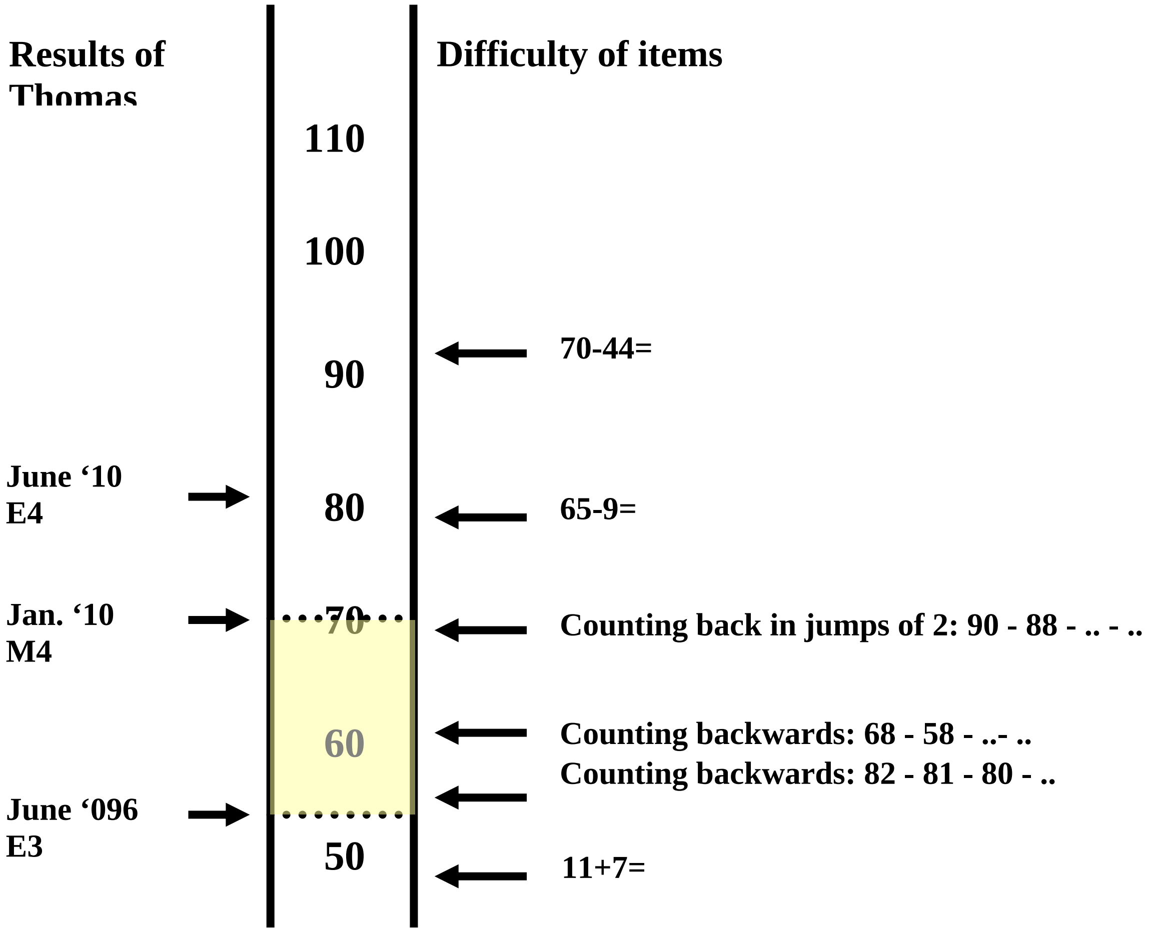

Results of Thomas
Difficulty of items
110
100
90
70-44=
June '10
E4
80
65-9=
Jan. '10
M4
70
Counting back in jumps of 2: 90 - 88 - .. - ..
60
Counting backwards: 68 - 58 - ..- ..
Counting backwards: 82 - 81 - 80 - ..
June '096
E3
50
11+7=

Figure 2: Part of the scale for Arithmetic

If, at the same time, for every measuring moment the spread of a (national) reference group is indicated on the scale, the relative position of the pupil compared to his 'peers' can be determined.

So we see that this technique allows three kinds of interpretations of the results:

- Self-referenced
  The degree of progress can be determined in relation to an earlier moment in time. After each measurement the raw score of a test is converted into a number on the ability scale, after which the difference compared to the previous scale score can be read just like measuring a child's length.
- Norm-referenced
  The position that the pupil takes on the scale can be compared to that of other pupils nationally.
- Domain- or content-referenced
  On the basis of the position on the scale a general conclusion can be drawn about the degree of mastery of a particular subject matter.

The ability-profile used for Arithmetic in the Monitoring and Evaluation System is an example of a report that allows for norm-referenced and domain-referenced interpretations. The index for comprehensive reading is another example of a multi-interpretable scale. On this scale the difficulty of reading texts and the reading ability of the learner are presented on the same scale. The raw test score of the learner is transformed to a reading-index, a number on the scale. The difficulty of all kinds of reading texts can also be expressed in a number on the same scale. In this way it is possible to select texts for a learner that correspond to his/her reading ability level. A similar index has been developed for decoding.

5 Reporting

After each measurement the raw score of a test is converted into a number on the ability scale, after which the difference compared to the previous scale score can be read just like one measures a child's length. Figure 3 is an example of the *pupil report*, a graph in which the pupil's progress is visible throughout the years. The horizontal axis represents time, while the vertical axis is the scale that represents the ability. The orange line summarizes the test performances of this pupil for six time points, from mid grade 3 until the end of grade 5.

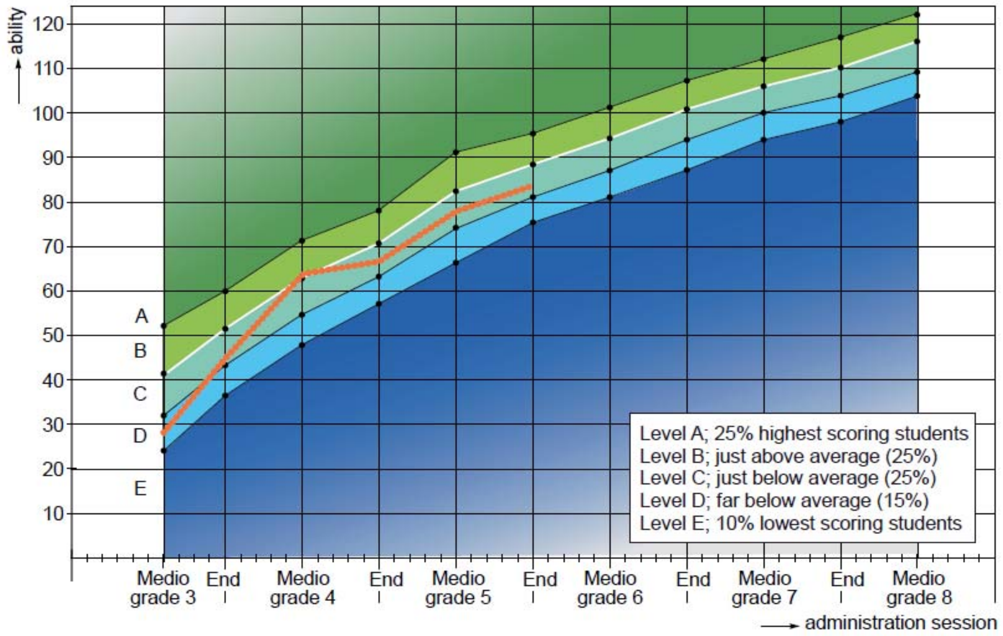


Figure 3. Example of a pupil report

The pupil's report not only shows the growth of the ability but also the relative position of the pupil among their peers. The data collected from the various subpopulations in a national survey are used as a frame of reference.

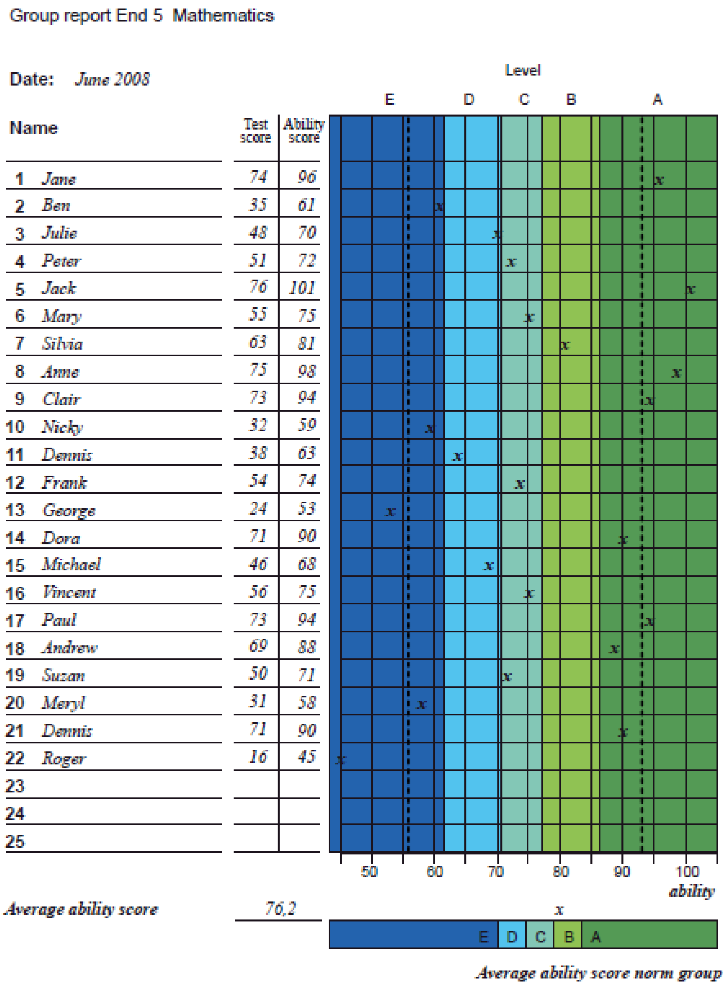
In the graph four curves have been drawn that correspond to percentiles 10, 25 and 75 and the population mean. On the basis of these data five levels can be distinguished:

Level A (dark green coloured area): 25% highest scoring pupils

Level B (light green coloured area):	just above average
Level C (turquoise coloured area):	just below average
Level D (light blue coloured area):	far below average
Level E (dark blue coloured area):	10% lowest scoring pupils

The orange line shows that the pupil started out far below average (as a level D pupil) and performs below average (as a level C pupil) for all successive time points although there is a relative improvement at mid grade 4 (see M4 where the mean is reached). From the end of grade 4 on this pupil is making the progress one could expect from a level C pupil.

Figure 4 is an example of a group report which graphically shows the results of all the pupils from one grade. At a glance a teacher can conclude which of the pupils' scores are below or above average when compared to the results of other pupils nationwide. Next to the ability scores of the individual pupils, the average ability score of the group as a whole is also included in the group report. The data collected from the various groups in the national survey are used as a frame of reference to compare the relative position of this specific group to other groups.



Group report End 5 Mathematics

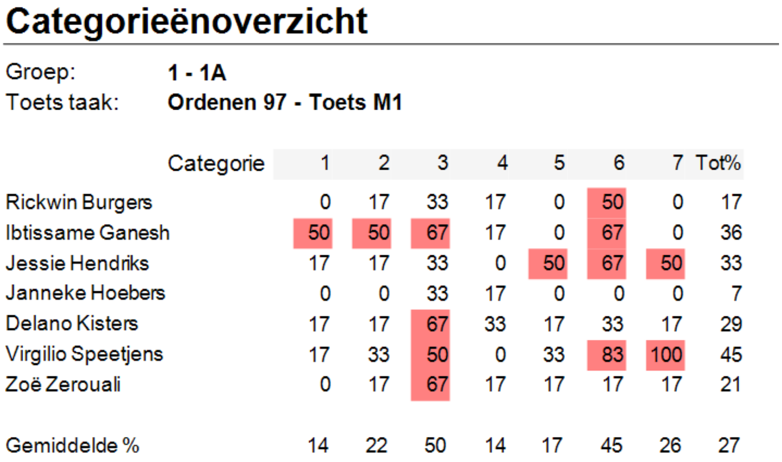
Date: *June 2008*

Name	Test score	Ability score
1 *Jane*	*74*	*96*
2 *Ben*	*35*	*61*
3 *Julie*	*48*	*70*
4 *Peter*	*51*	*72*
5 *Jack*	*76*	*101*
6 *Mary*	*55*	*75*
7 *Silvia*	*63*	*81*
8 *Anne*	*75*	*98*
9 *Clair*	*73*	*94*
10 *Nicky*	*32*	*59*
11 *Dennis*	*38*	*63*
12 *Frank*	*54*	*74*
13 *George*	*24*	*53*
14 *Dora*	*71*	*90*
15 *Michael*	*46*	*68*
16 *Vincent*	*56*	*75*
17 *Paul*	*73*	*94*
18 *Andrew*	*69*	*88*
19 *Suzan*	*50*	*71*
20 *Meryl*	*31*	*58*
21 *Dennis*	*71*	*90*
22 *Roger*	*16*	*45*
23		
24		
25		

Average ability score *76,2*

Figure 4. Example of a group report

Although the tests of the Monitoring and Evaluation System can be processed and recorded manually, a computer program has been developed to take over a number of the teacher's routine activities. This program is especially useful in the identification stage and in part of the analysis stage. The computer program enables teachers to make additional report. In Figure 5 an example is given of a category analysis for ordering. The red square in the figure marks the risk scores in a particularly category.



Categorieënoverzicht

Groep: **1 - 1A**
Toets taak: **Ordenen 97 - Toets M1**

Categorie	1	2	3	4	5	6	7	Tot%
Rickwin Burgers	0	17	33	17	0	50	0	17
Ibtissame Ganesh	50	50	67	17	0	67	0	36
Jessie Hendriks	17	17	33	0	50	67	50	33
Janneke Hoebers	0	0	33	17	0	0	0	7
Delano Kisters	17	17	67	33	17	33	17	29
Virgilio Speetjens	17	33	50	0	33	83	100	45
Zoë Zerouali	0	17	67	17	17	17	17	21
Gemiddelde %	14	22	50	14	17	45	26	27

Figure 5: Example of a category analysis for ordering

6. School self-evaluation

When the Cito Monitoring and Evaluation System has been implemented in the school for a couple of years in several grades, the data gathered can also be used for school self-evaluation purposes. It is possible to fill in some reports manually, but more advanced reports can be made with a separate module of the computer program specially designed for this function. The module allows the construction of cross-section reports and trend analysis for various subjects.

A ***cross section*** shows the distribution of pupils of the different grades over the 5 levels (A to E) at a certain moment in time. See Figure 6 for an example.

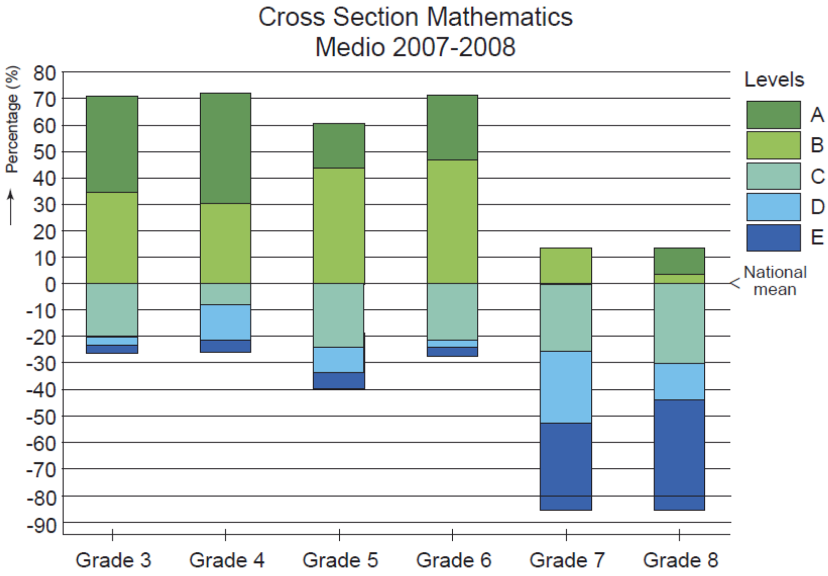


Figure 6: Example of a cross section for Arithmetic/Mathematics

The 0%-line shows the national mean. Above this line the percentage of pupils in the different grades with a level A or B are depicted. In the national reference group about 50% have an A or B-level. The other 50% have a C, D or E-level. The results of grades 7 and 8 are eye-catching. In the case of grade 7 only 15% of the pupils score above the national mean and there are no A-level pupils. Approximately the same percentage of the pupils of grade 8 score above the national mean (although there are pupils with an A-level!), while 85% of the pupils score below the 0% line (the national mean). Compared to the results of the other grades in this school, these results are remarkable.
Of course the system cannot find the reason for these remarkable results, but it points to a possible problematic area and it is up to the school to find a reasonable explanation for such a phenomenon. In the example given, the reason might be that the groups of pupils are exceptionally weak or it might be that something is going wrong systematically in grades 7 and 8. If the former explanation is correct, the performance of the same groups of pupils – a cohort – should show below average performance over several years. If the latter explanation is correct, different cohorts within the same school should show below average performance in grades 7 and 8.

To gather more information which makes it able to confirm or reject these hypotheses, the program allows two kinds of ***trend analysis***: cohort based trends and grade based trends.

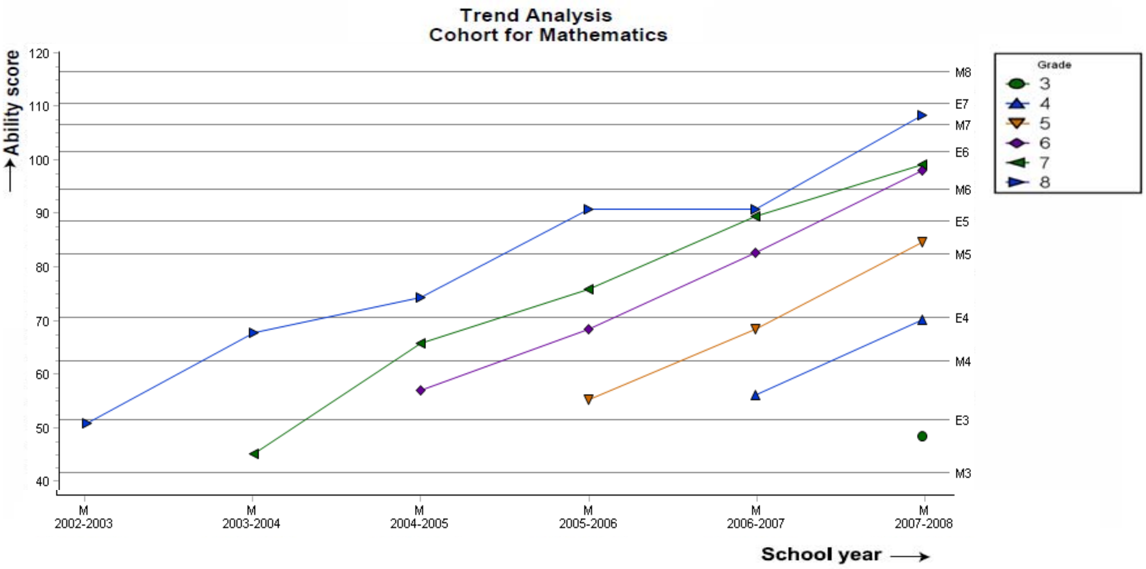


Figure 7: Trend analysis of cohorts for Arithmetic/Mathematics

Figure 7 shows the results of several cohorts of pupils (same group of pupils) over the years compared to the national mean in the different grades. In this example only the results on the tests taken halfway the school year are displayed. The level of the national mean is displayed as the set of irregular grid lines.
If we look at the results of the pupils from grade 8 in year 2007-2008 (the blue line), we see that they score (far) below average almost all the years compared to the national mean. The results of the tests these pupils took halfway the school year when they were in the school years 2002-2003 and 2003-2004 (at the start of the blue line) were above average, respectively above the M3-line and above the M4-line. This is also the case for grade 7 (green line). The cohort of grade 7 in year 2007-2008 started in their grade 3 (Mid 2003-2004) and grade 4 (Mid 2004-2005) above average, but score below average from Mid 2005-2006 on, respectively below the M5-line, the M6-line and below the M7-line.
The above formulated explanation - that the pupils of the groups 7 and 8 are exceptionally weak – can now be rejected. After all, the pupils in grades 7 and 8 started out above average in grades 3 and 4. Both cohorts started to perform below average from grade 5 on. If we look at the results from the pupils from grade 6 and grade 5 in year 2007-2008 (respectively the purple and the orange line), we see that they score on or above the average all the years compared to the national mean. But we can also see that they started out better in their grades 3 and 4 than they score nowadays. It looks as if the results decrease as the pupils move on to grade 5 and further. Something might be going wrong in the education from grade 5 on. If this assumption is right then different cohorts within the same school should show below average performance from grade 5 on. To see if this really is the case we can look at the grade based trend analysis. This trend analysis shows the results of different learner groups in a certain grade. Figure 7 shows an example of this kind of trend analysis.

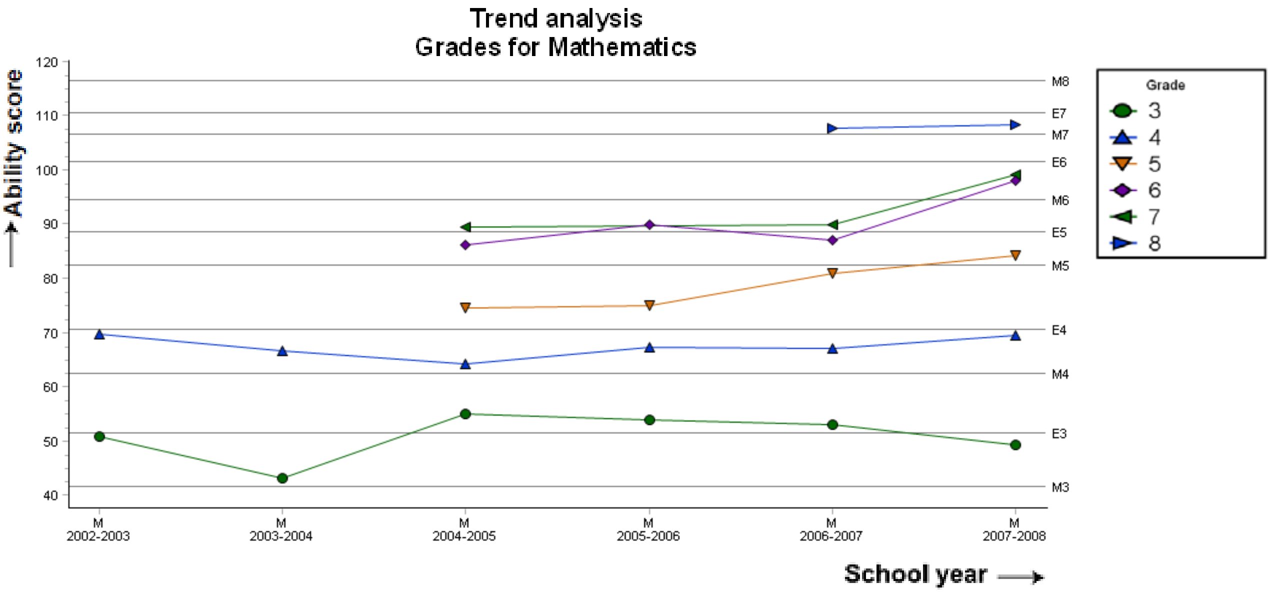


Figure 8: Trend analysis of grades for Arithmetic/Mathematics

In figure 8 we can see that although the average results vary, the average results for grades 3 and 4 are above the national mean throughout the years (respectively above the M3-line and above the M4-line). However in grades 5, 6, 7 and 8 the results are (far) below average almost all the years compared to the national mean. We can thus confirm the assumption that different cohorts within the same school perform below average from grade 5 on. Only in school year 2007-2008 the results in grades 5 and 6 are above the national mean. In this school year the results in grades 7 and 8 also show a (slight) increase. The question is what has changed in the education in mathematics in this school year and more importantly how can this school continue their efforts in such a way that a long term improvement is made in their education and subsequent also in their results in mathematics.

In the case of the above example we now know that something in the education in mathematics in this school is systematically going wrong from grade 5 on, but we also see that the results in the most recent school year show an increase. On the basis of the reports we don't know the explanation for this phenomenon. Changes or major deviations of the ability scores between the school years per grade can be caused by many factors, such as:

- a change in composition of the pupil population
- a major incident with a high impact on the pupils
- a long illness of the teacher
- the replacement of the teacher
- new textbooks or learning materials
- a change in the amount of teaching time spent on a specific subject
- additional counseling or learning projects

It is up to the school to find a reasonable explanation. In the opinion of Cito this is something that concerns the whole team in the school; all team members have to be involved in the discussion about the findings but, of course, the head teacher has the responsibility to initiate such a discussion.

7. Computer-adaptive testing[1]

Using the computer in administering test has great advantages. Not only reduces these tests the amount of test-administering time, the teacher no longer needs to give extensive instructions to the pupils, does not have to mark the tests and does not have to fill in reports. The computer processes the test results and reports immediately after completing the test.
Since 2003 the Monitoring and Evaluation System also contains computer-based tests. Some of these tests are even adaptive. Apart form the advantages mentioned earlier, the latter also means that better information is gained in less testing time and that pupils are no longer bored or frustrated. However, using computer adaptive testing has not become common practice yet. A start has been made with two tests, one referring to dyslexia and one to phonological awareness. In this chapter brief information will be given about the principles of computer adaptive testing. Moreover, an example of a CAT will be given.

Introduction

In computerized adaptive tests (CATs) the construction and administration of the test is computerized and individualized. For every test taker a different test is constructed by selecting items from an item bank tailored to the ability of the test taker as demonstrated by the responses given thus far. So, in principle, each test taker is administered a different test whose composition is optimized for the person. The main motive for computerized adaptive testing is efficient measurement. It has been shown that CATs need less items to measure the ability of the test taker with the same precision.

An important prerequisite in a CAT during testing is the direct accessibility of a complete (IRT calibrated) item bank and also that computational procedures for ability estimation and item selection must be carried out in real time. In general there are (still) main limitations in the permissible responses of the candidates, due to the fact that these is limited. Practical applications of scoring polytomous items are limited responses need automatic scoring. For this reason adaptive testing with polytomously scored items. This is why items in computerized tests often are of a choice of matching type format.

The availability of an IRT calibrated item bank is a necessary condition for CATs. In IRT a relation is specified between the non-observable ability θ which is measured

[1] The text is this paragraph is largely based on Eggen, Th.J.H.M (2004). *Contribution to Theory and practice of Computer Adaptive Testing*. Enschede

and the observable score on the measurement instrument. The properties of IRT make it excellently suited for application in CATs because:

- with any subset of the items in a calibrated item therefore not necessary to administer the same items to pupils in order to get comparable estimates of the ability;
- the difficulties of the items are expressed on the bank it is possible to estimate the ability on the same scale. It is same scale as the ability of the pupils. It is therefore possible to adapt a test to a level of a pupil;
- the information in an item is a function of the ability, by which the information function can serve as a basis for tailored item selection.

The testing algorithm

CATs are governed by a testing algorithm. The algorithm is a set of rules which determine the way CATs are started, continued and terminated. In Figure 9 a schematic representation of a CAT algorithm is given.

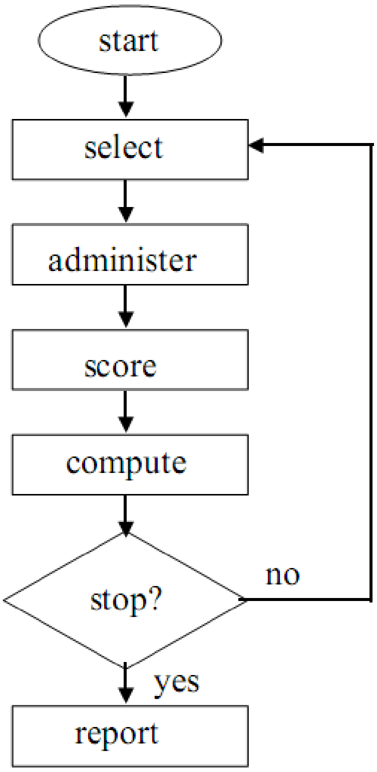


Figure 9. Schematic representation of an adaptive test

As data on a pupil’s ability are not always available, a CAT starts with presenting a randomly selected item from (a subset of) the item bank to the pupil. If there is any information on the ability of the pupil before testing, this information could be used in the first selection. After every administered item an item selection procedure is carried out. From the item bank an item is chosen that is in accordance with the answers given by the pupil so far: the test is adapted or tailored to the ability of the tested pupil.

Besides the ability estimate, an estimation of its standard error is determined. The standard error expresses the accuracy with which the ability of the pupil is known. This standard error is normally used as a criterion for stopping the testing. If the criterion has not yet been met, a new item is selected, otherwise the result of the test is reported.
In CAT practical considerations play an important role. The main examples of these constrains are with respect to the content of the test and to item exposure. With content constraints a tester wants to employ a desirable content specification for the test, establishing that certain components of the ability that are to be assessed occur in a given proportion. With exposure constraints a twofold problem with unrestricted item selection can be solved. The first is overexposure: some items are selected so frequently that confidentiality of the items is rapidly and directly compromised. The second is underexposure: there are items in the bank which are so seldomly used that one could wonder how the expense of constructing them can be justified.

Below the principle of a CAT will be illustrated. The examples are taken from a computer adaptive test for phonological awareness. Phonological awareness refers to the sensitivity to the sound structure of a spoken language. In the examples given three response alternatives are presented, both auditorily and visually in the form of pictures in order to reduce memory demands. A target word is then presented auditorily. The pupil task is to select the correct picture.

In figure 10 an example of an item is given.

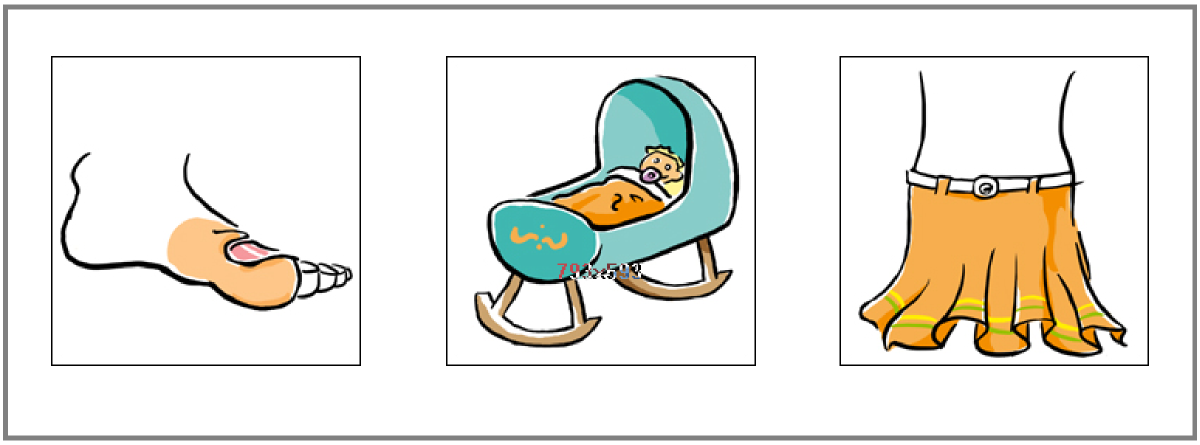
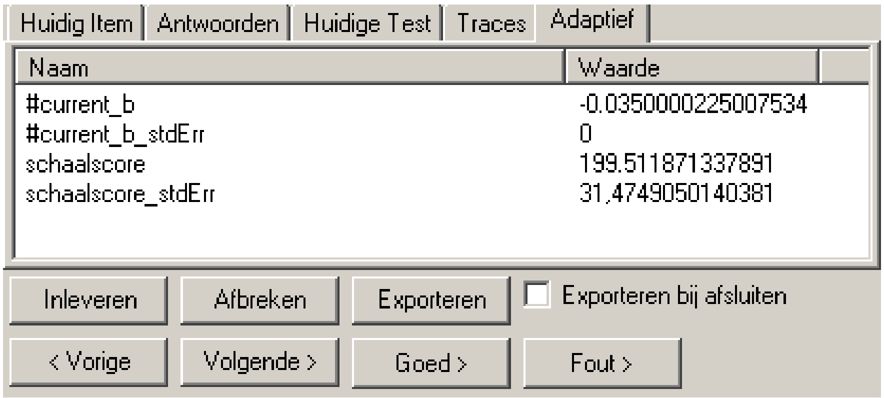


Figure 10. Example of an item of the test phonological awareness

Based on the answers of the pupil on the items presented so far, the estimated ability of the pupil is calculated. Figure 11 indicates the estimated ability of the pupil if the pupil has answered the item correct.



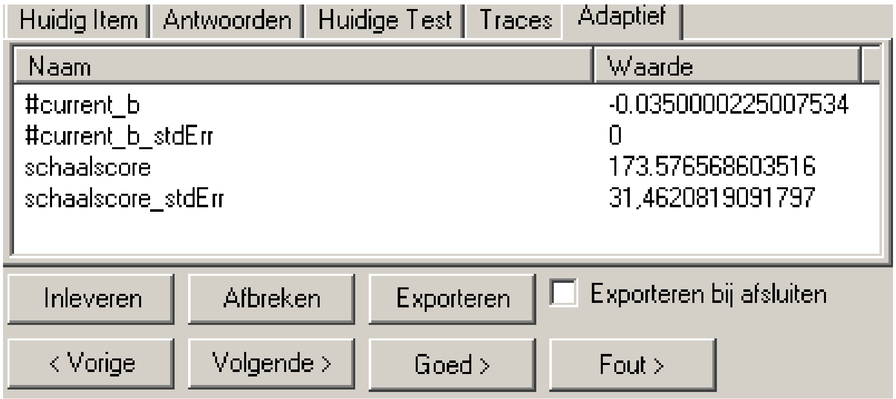
Naam	Waarde
#current_b	-0.0350000225007534
#current_b_stdErr	0
schaalscore	199.511871337891
schaalscore_stdErr	31,4749050140381

Figure 11. estimated of the pupil ability based on the (correct) answer on the item shown if figure 11.

As can seen from figure 11 the estimated ability ('schaalscore') of the pupil so far is 199.51.

If the pupil answers the item *correct* a next (more difficult item) will be selected from (a subset) of the item bank. Based on the presented items so far a new indication of the estimated ability will be given.

If the pupil answers the item presented in Figure 10 *incorrect* his estimated ability will be lower compared to the ability presented in Figure 11. Figure 12 represents the estimated ability when answering the item in Figure 10 incorrect.

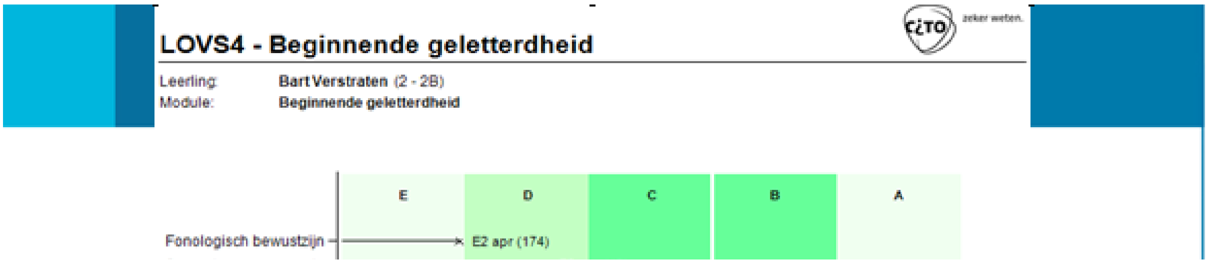


Naam	Waarde
#current_b	-0.0350000225007534
#current_b_stdErr	0
schaalscore	173.576568603516
schaalscore_stdErr	31,4620819091797

Figure 12. estimated of the pupil based on the (incorrect) answer on the item shown if figure 10.

From Figure 12 it can be seen that the estimated ability decrease to 173,58. By answering the item correct the estimated ability was 199.51 (see Figure 11).

After the test is taken a report is given as presented in Figure 13.



LOVS4 - Beginnende geletterdheid

Leerling: Bart Verstraten (2 - 2B)
Module: Beginnende geletterdheid

Figure 13. Example of a report phonological awareness

Figure 13 indicates the level of ability. The levels A – E are comparable to the level illustrated in Figure 5. The classification A – E is based on the distribution of the scale scores in the population. From this report it can be concluded that this pupil (‘leerling’) Bart is lacking behind (he belongs to the 10% lowest scoring pupils). To avoid reading problem as much as possible, extra attention is needed.

Unfortunately at this moment computer adaptive testing in the Netherlands is not yet common practice. The computer infrastructure in primary schools is still not of a satisfactory standard. Most schools do not have enough computers to organize computer-adaptive testing for all of their pupils. That is the reason why it is sometimes easier for schools to use paper-based tests for whole class testing.