The use of the Cito Monitoring and Evaluation System for school self-improvement in The Netherlands

35th IAEA Annual Conference
Brisbane, Australia, September 13~18, 2009

Marleen van der Lubbe
Cito, Arnhem, The Netherlands

1. Introduction

Monitoring the level of an educational system nationwide forms a serious challenge. Usually it is conceived that such a monitoring system should be designed and carried out by a central authority to guarantee validity of content and objectivity. Without doubt this kind of argument is convincing, but it is Cito's point of view that such a system may be complemented by a system of self-evaluation, reflecting the trust that educational authorities may have in the will and capacity of self correcting activities of the schools, given that they are informed in an objective and reliable way about the educational performance and progress of their pupils.

In The Netherlands Cito, Institute for Educational Measurement, has put in place a monitoring and evaluation system for primary education in the nineteen eighties. Although its primary purpose was to provide a unified system that enabled the 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.

It should be emphasized that the Cito Monitoring and Evaluation System is as non-directive as possible: use of the system is on a completely voluntary basis, as well as the number of modules (subjects) that a school wishes to use. The information collected by the school is only returned (voluntarily by the schools) to Cito in a highly anonymized way for the purpose of statistical analyses; in all other respects the information is owned by the school and the use they make of it is their full responsibility. Cito as a provider of the system can only judge on its usefulness by the qualitative feedback on the system and by its growing popularity.

In this paper an overview of the Cito Monitoring and Evaluation System as a pupil centered system and as a system for school self-evaluation will be given. In Section 2 and 3 the content and psychometric basis of the Cito Monitoring and Evaluation System for primary education is briefly discussed. Section 4 contains a sample of the many reports on pupil and group level that are available for the users of the system. In Section 5 specific attention is given to the reports at school level and to the use of these reports for school self-improvement in the Netherlands.

2. The content of the Cito Monitoring and Evaluation System for primary education

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[1] 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. Most tests are also available as computer-based tests. Some of the computer-based tests (all the test for grades 1 and 2) are adaptive.

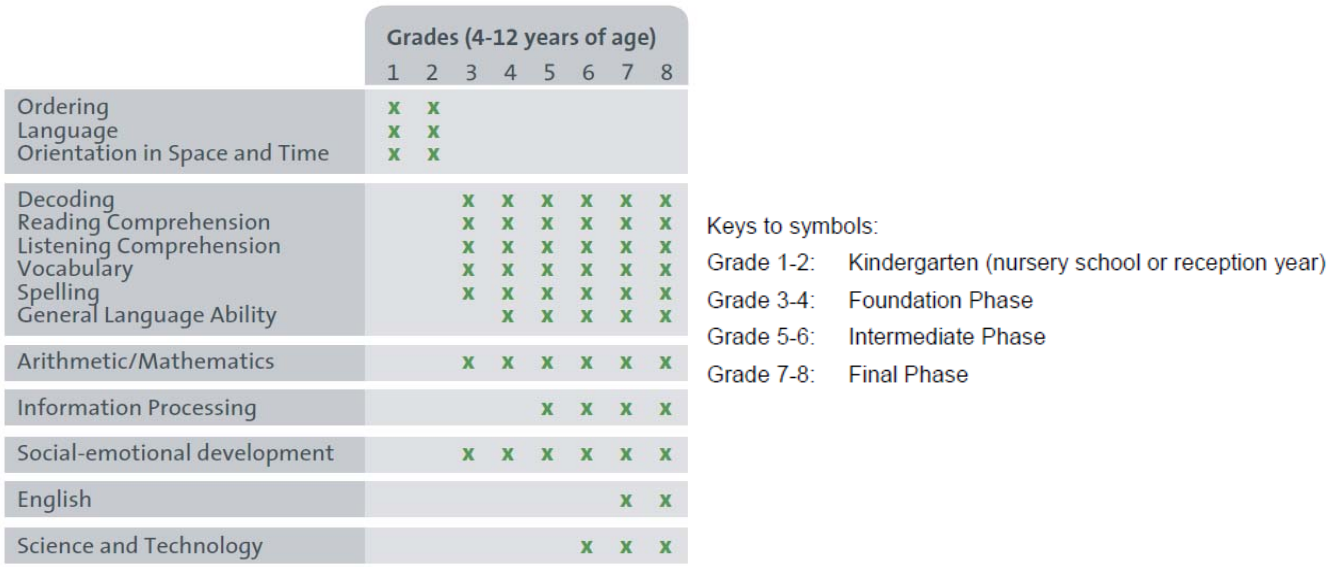


Figure 1. Tests in the Cito Monitoring and Evaluation System for primary education

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

[1] Primary education in the Netherlands comprises eight grades, the first two coinciding with what in most countries is kindergarten education. Grades are indicated with the term 'group'; pupils in 'group 5' are comparable to grade 3 pupils in most educational systems.

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.

3. 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 the spread of abilities in (sub) populations and it provides a framework for interpreting test results. In the IRT model used in the Cito 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 the 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.

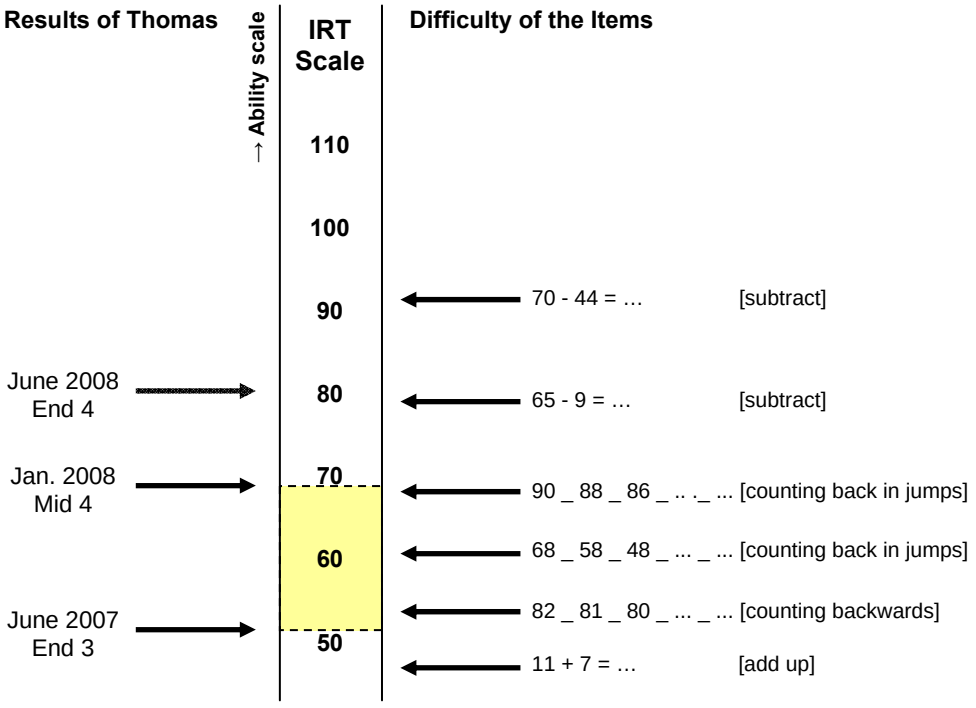
Particularly the fact that both pupil abilities and item characteristics can be put on the same scale and can be related to each other is of great advantage to the Cito monitoring and evaluation system:

- The results on tests that differ according to difficulty, contents and number of items can be compared. In other words: John's 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.

In Figure 2, a graphical display is given where the growth of a pupil can be related directly to the content of the test. The middle band in the figure represents the scale, in this example the scale for Mathematics. To the left the arrows indicate the positions of a single pupil, Thomas, measured at three consecutive points of time (June ’07, Jan. ’08, June ’08) halfway the school year and at the end of the school year. The right hand part displays some items and their location on the scale. From the relative position of item points and pupil’s points one can obtain a description of the kind of items the pupil mastered or did not master at each measurement. For example, the item ’11 + 7 =’ could be answered correctly with a higher than 50% probability at measurement point 1 (E3), while there was a lower than 50% probability of a correct response when the pupil was required to count back in units starting from 82. At the third point of time (E4), the probability of a correct response for the item ’65 – 9 =’ is over 50% while the more difficult item ’70 – 44 =’ is not yet mastered at that point of time. 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.

Figure 2: Part of the scale for Mathematics



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.

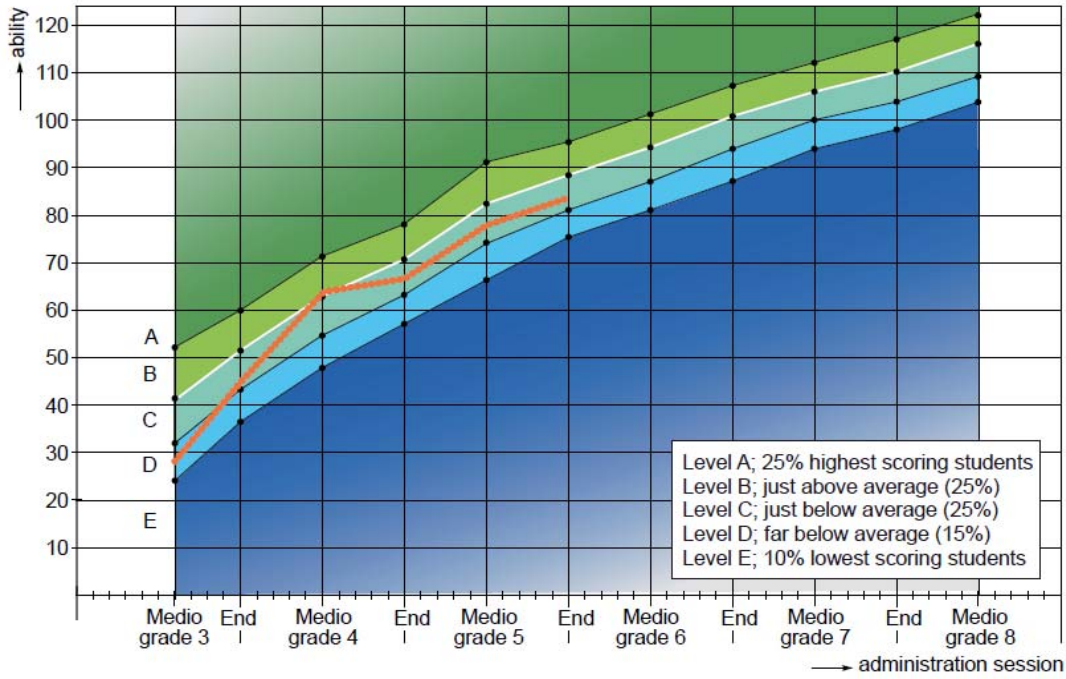
The index for comprehensive reading in the Cito monitoring and evaluation system is an example of a report that allows for different kinds of interpretations. On this scale the difficulty of reading texts and the reading ability of the pupil are presented. The raw test score of the pupil is transformed 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 pupil that correspond to his reading ability level at different moments in time. A similar index has been developed for decoding.

4. Reporting

When a school uses the computer-based version of the tests of the Cito monitoring and evaluation system, the computer of course automatically processes the test results and reports immediately after completing the test. The paper-based version of the tests can be processed and recorded manually or with the help of a computer program that has been developed to take over a number of the teacher's routine activities. After the test session the test results can be fed into the computer in three ways. The quickest way is to directly type in the pupils' test scores. However, it is also possible to click on the item that the pupil answered incorrectly. Both ways presuppose that the teacher has marked the test himself. In many cases a third way of feeding data into the computer is possible: directly feeding the answers given. For every item the pupil's answer is fed into the computer, after which the computer scores the test. Per pupil the most desirable way of processing data can be chosen. After the data have been fed into the computer, the computer calibrates the test- and ability scores and determines the level indication that goes with them. Then the computer can produce the various reports, such as a pupil report, a group report, an answer survey, an error analysis etc.

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



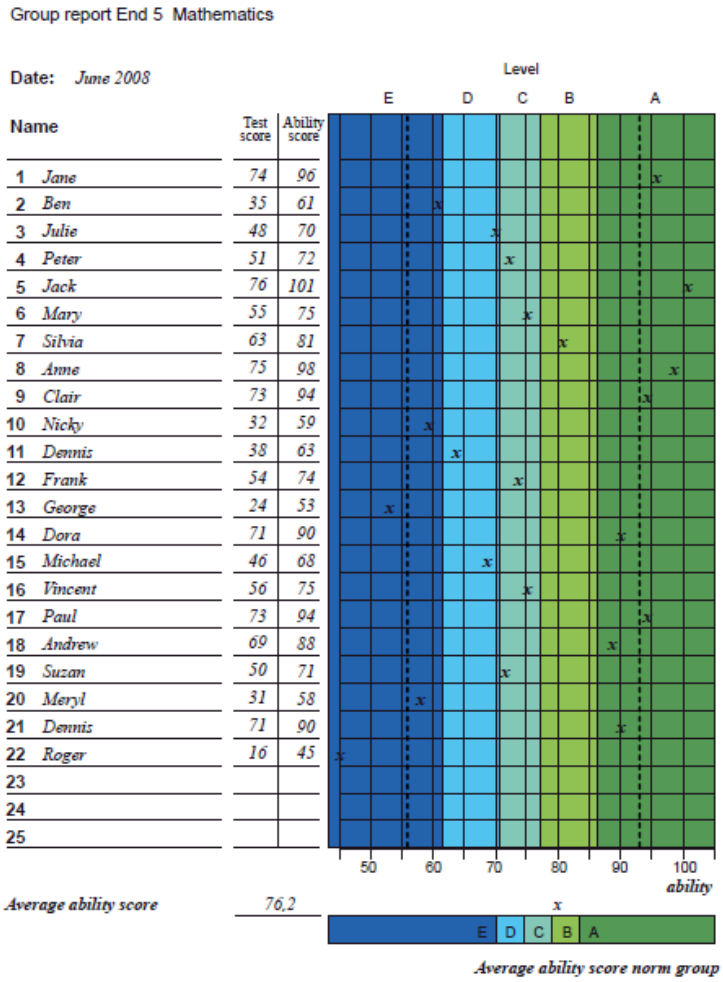
The pupil's report not only shows the growth of the ability but also the relative position of the pupil among their peers. The data collected from the various subpopulations in a national survey are used as a frame of reference.
In the graph four curves have been drawn that correspond to percentiles 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.

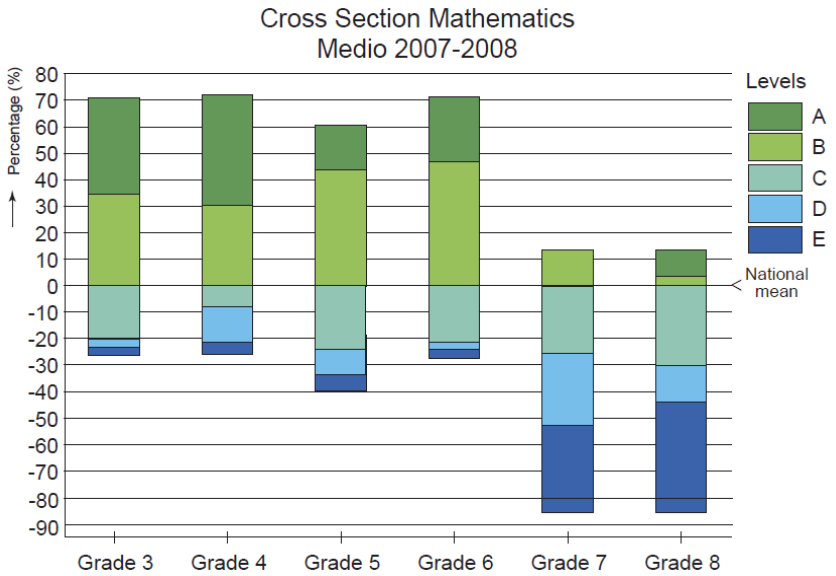
Figure 4. Example of a group report



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.

Figure 5: Example of a cross section for Arithmetic/Mathematics



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 5 for an example.

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 6: Trend analysis of cohorts for Arithmetic/Mathematics

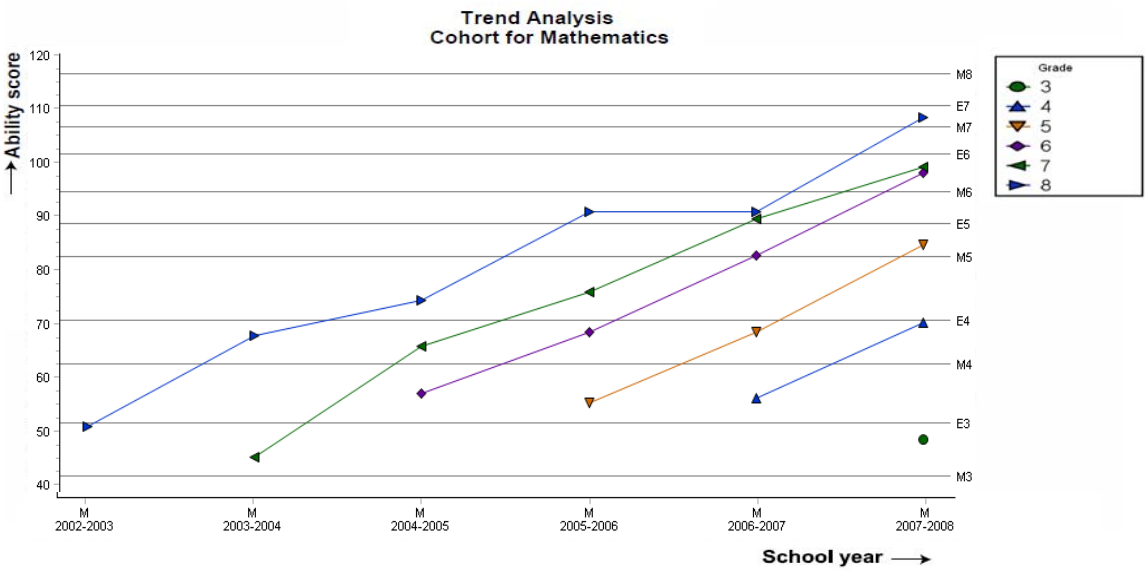
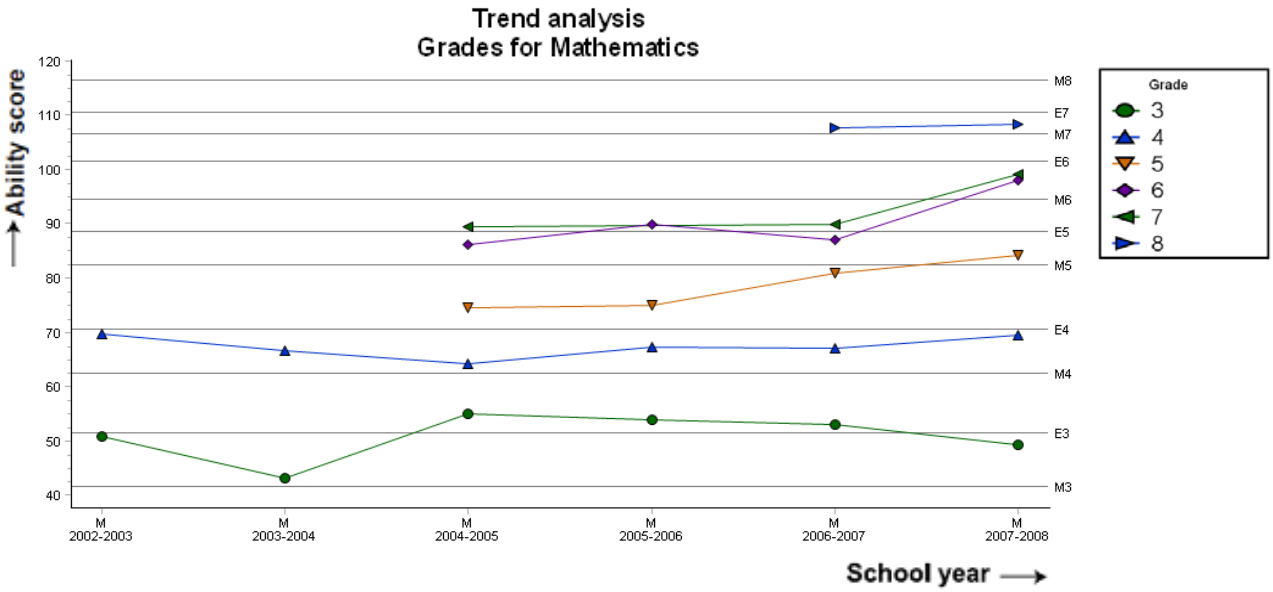


Figure 6 shows the results of several cohorts of pupils (same group of pupils) over the years compared to the national mean in the different grades. In this example only the results on the tests taken halfway the school year are displayed. The level of the national mean is displayed as the set of irregular 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 7: Trend analysis of grades for Arithmetic/Mathematics



In figure 7 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.

Although we do not know to which extent schools actually use the school reports of the Cito monitoring and evaluation system for the purpose of school self-evaluation, the fact that most of the schools (approximately 95%) for primary education in the Netherlands use at least one of the tests of the system indicates that at least these schools have the necessary information at their disposal to do so. Furthermore, the interest in the courses that Cito offers to help schools interpret and use the various school reports of the Cito monitoring and evaluation system, has significantly increased during the last couple of years. On the other hand we often observe that schools that followed the course experience difficulties putting into practice what they learned during the course. In general they complain about the lack of time they need to really carry out the activities involved in self-evaluation (e.g. studying the reports, discussing about the results and so on). Recent research (Vanhoof, Van Petegem & De Maeyer, 2009) reveals that a positive attitude towards self-evaluation is a precondition which favours successful school self-evaluation. In the Netherlands a study (Blok et al., 2005) shows that head teachers in the Netherlands regard self-evaluation as a useful and instructive undertaking, although they admit that it takes up a lot of time. Schools thus believe in the potential power and value of self-evaluations, but the accompanying process causes many to hesitate when it comes to actually carrying out a self-evaluation. The important question is how to change this.