Bergh, Huub H. van den / Thije, Jan D. ten (2005) Assessment of language competencies
The development of system and pupil assessment instruments in the Netherlands, in: K. Ehlich et al. (eds) Anforderungen an Verfahren der regelmässigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Sprachförderung von Kindern mit und ohne Migrantionshintergrund. Bonn: Bundesministerium für Bildung und Forschung, 217-241.

Assessment of language competencies

The development of system and pupil assessment instruments in the Netherlands[1]

Huub van den Bergh and Jan D. ten Thije (Utrecht)

1. Introduction

Recent European comparative studies on language competencies (e.g. PISA) have caused a passionate social and political discussion in many European countries on the possibility and necessity of developing national assessment instruments. Disappointing results in several federal states in Germany have aroused vehement discussions. In order to supply the discussion with more research-based arguments, Prof. Dr. Konrad Ehlich (Institut für Deutsch als Fremdsprache / Transnationale Germanistik of the Ludwig-Maximilians-University Munich) initiated a pilot study on this topic.

The German project aims at an evaluation of available German language assessment instruments and analyses the conditions for the realization of a periodical national assessment of language competencies. One of the central focuses of interest is the question ofhow national assessment allows assessing the individual development of native/national and migrant children. The German project also includes expert reports

1. We want to thank K. de Glopper (University of Groningen), H. Kuhlemeier (Cito), and Dr. M. Zwarts (National school inspection), and H. Breevelt for their valuable comments on a previous version of this text.

on relevant developments in countries that have a longer tradition in this field of interest, e.g. Sweden, Australia and the Netherlands. The present report characterizes the developments in language assessment in the Netherlands.

In the Netherlands some twenty years ago a first feasibility study on national assessment was executed. This study brought at light that seven percent of the pupils at the end of primary education were 'functional illiterates' (Wesdorp / Van den Bergh / Bos / Hoeksma / Oostdam / Scheerens / Triesscheijn 1986). This result then triggered a passionate discussion on the perceived level of achievement. Questions concerning the aim, types of tests, and applied norms were only seldom put forward. Since then there have been a national assessment of language competencies every five years, i.e. in 1989, 1994, and 1999.

In order to characterize the development and actual functioning of the national assessment nowadays, we discuss this type of language assessment in relation to the other important systemic and individual assessment instruments that traditionally are carried out in the Netherlands. In fact, during their compulsory school attendance (4 until 16/18 years old), Dutch pupils and students are confronted with four types of assessment. These instruments are applied at different stages of the pupil's school carrier and differ in scope and purpose:

- During primary education individual pupils take (standardized) tests within the framework of a *pupil-monitoring system.*
- In grade six and in grade eight in primary education randomly selected pupils are tested once every five years in the framework of the *national assessment of educational progress*.
- At the end of primary education (grade 8) all pupils (age ± 11 years) take a standardized *final test* (often the so-called Cito Primary Education Final Test) that determines in great part their choice for secondary education.
- Finally, pupils take tests in all school subject matters at the end of secondary school. These *exams* consist of two parts: a school exam and a centrally (nationally) developed exam with standardized tests.

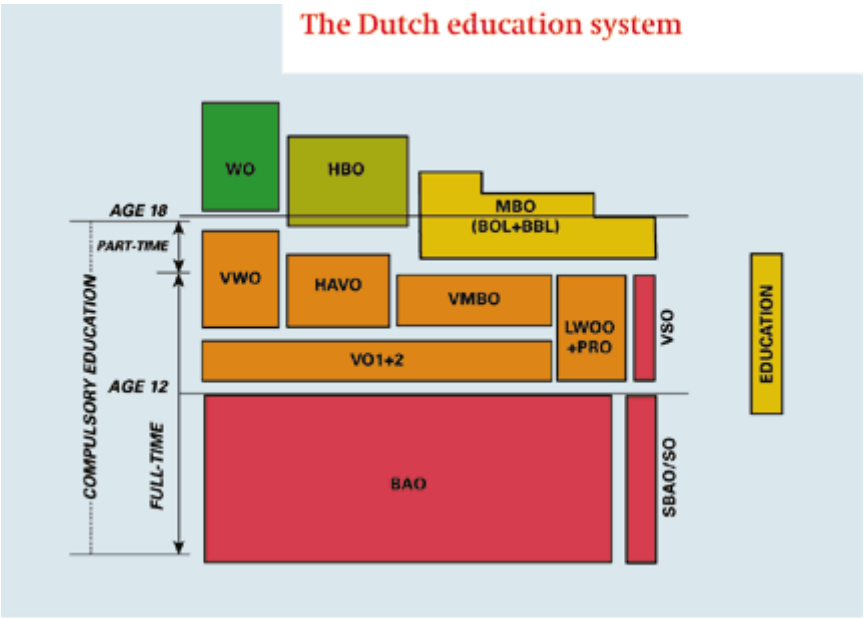
For a good understanding of the Dutch conditions for the realisation and management of the different types of assessment we will supply the German reader with some preliminary information about the Netherlands educational system and politics and give some background information on migrant children. At the end of the report we will raise questions and discuss linguistic, didactic and educational aspects of the Dutch system of assessment instruments.

2. Relevant characteristics of the Dutch educational system

2.1 Dutch educational structure

A striking difference with the German educational system is that Dutch children enter primary school at the age of five and finish their compulsory school career when they are eighteen years old. During the last two years education is part-time compulsory. Figure 1 represents the Dutch system. The size of each block represents the number of pupils attending that type of education.

Figure 1 The Dutch educational system (source: Minister of Education, 2004)



After the primary school (BAO) (i.e. 'Grundschule') the system opens up to various types of secondary education: the pre-university education (VWO, i.e. 'Gymnasium' or 'Fachoberschule'), next to senior general (HAVO, i.e. 'Fachoberschule'[2]) and prevocational secondary (VMBO, i.e. 'Hauptschule' or 'Realschule') education. After secondary education, pupils move on to senior secondary vocational education (MBO i.e. 'Berufsfachschule') or higher education (HO). MBO is divided into a vocational training programme (BOL) and a block/day release programme (BBL). Higher education encompasses higher professional (HBO i.e. 'Fachhochschule') and university (WO – 'Universität') education. Apart from these mainstream types of education, children with special needs attend learning support departments (LWOO), special primary (SBAO and SO) and secondary schools (VSO) or practical training programme (PRO).

In a recent comparison of different European educational systems the Dutch Social Planning Institute (SCP 2000) provides us with background information. Dutch and German educational systems have in common that pupils have to choose at a relatively young age between the different kinds of secondary school. In most other countries secondary school types are longer integrated and pupils choose not before the age of 15 or 16. Shared classes during the first one or two years of secondary school sometimes provide some relief from the need of early selection (see Figure 1: VO1+2). However in the Netherlands the gap between pre-university (VWO) and senior general (HAVO) on the one hand and primary and senior vocational education (VMBO) on the other becomes more and more deep and unbridgeable. This difference is comparable with distinction between 'Gymnasium' and 'Mittelschule' in many federal states of Germany. The Netherlands tends to have one of the most selective educational systems in Europe.

SCP (2000) reports that the Dutch educational level and participation were high for long time compared to other European countries, but have decreased to the European

[2] The names of the Dutch and German school system partly differ and partly corresponds.

average in the last years. School hours for Dutch pupils are still circa 10% higher than for children in most other European classes. Dutch expense on education is relatively low and has even decreased in the last years. Currently a debate is going on whether the Netherlands loses touch with international standards with respect to the reproduction of a modern knowledge society.

2.2 Freedom of education and changes in educational policy

The developments within the Dutch educational system cannot be understood without a short note on the constitutional freedom of education that underlies the national system since 1917. Due to this freedom of education Dutch schools have a large autonomy compared to schools in other European countries. The Dutch constitution not only allows private groups to establish schools, it also grants them a great - though not absolute - liberty to determine the content and form of education. In the course of time the national state developed various organizational instruments in order to control and determine indirectly the changes in school programs.

Koole / ten Thije (1994) describe in this respect the emergence in the sixties of a so-called constructive educational policy that for the first time recognized the state as initiator and stimulator of educational innovation and state: "*Social needs in society, such as the changing needs of the labour market due to the automation of the labour process and the subsequent need for and influx of unschooled immigrant workers urged this policy change.*" Under this policy, many local, regional and national advisory institutes were relocated in an overall structure offering service and support to schools and enabling to the government to conduct a coordinated policy aimed at a systematic innovation of both content and structure (Doornbos, 1986: 267). National advisory centres were installed and assigned with specific tasks with respect to curriculum development (SLO), educational research, school counselling and, finally, test and assessment development (Cito). The development of national assessment instruments in the eighties can be considered as a direct result of this incentive policy.

However, over the last ten years, successive governments gradually replaced this incentive policy with an input/output policy. This means that the national state does no longer initiate so many innovation projects itself, but instead provides schools with financial means so that they can hire support from national or local advisory institutes. This policy shift can be considered as the next stage in the societal transformation of the constitutional freedom of education. The national education policy restricts itself to so-called ‘core business’ and the Minister of Education concentrates on an input/output comparison. Within this framework national assessment of school achievement assumed new importance.

2.3 Multilingualism and migrant pupils in school

As migrants were brought to Northern Europe in the sixties of last century the Dutch migrant pupils were treated according to the idea of 'integration with maintenance of minority identity'. The aim of this 'two track' policy was to integrate immigrant pupils in the Dutch school and society, and at the same time to prepare them for their return to their country of origin. Immigrant pupils received extra lessons in Dutch, as well as in their mother tongue.

In the eighties it became clear, that migration would be a permanent condition and the Netherlands had become, in fact, a multicultural society. Consequently, among other things educational policy changed. Teaching of Dutch as a second language was more accentuated and the special attention for migrant pupils was integrated in the special attention for children from working class families. Since that time the funding of primary schools is based on a scale rate from 1.0 via 1.25 to 1.9 corresponding to the number of pupils with indigenous or foreign family background. In the weighing of the children the place of birth of one or both parents is considered as well as their educational level and professional career. This scale is the decisive factor in the allocation of supplementary means to schools for extra teachers or schooling materials.

Nowadays, teaching of minority languages is forbidden during school time and teaching Dutch is strongly emphasized in all sectors of education. In some secondary

schools Spanish, Turkish, Moroccan and other minority languages can be chosen as a third foreign language. However, this opportunity is rarely available in the vocational sectors of education, where migrant children are strongly represented. Following the process of European unification, in primary and secondary schools there have been tendencies towards bilingual education. Consequently, mainly English as language for instruction determines bilingual education policy nowadays in the Netherlands. Recently, language policy for primary school also considers French and German.

The current state of affairs concerning the educational career of migrant children is represented in the OESO report 'Education at a glance 2003' (Minister of Education 2003). The proportion of pupils with non-Dutch backgrounds in primary education was 15.3 % in 2001/02. In relative terms, minority children are represented most strongly in special education: 18.9 %. Over the past five years, the number of secondary school pupils from non-Dutch speaking groups has increased. The increase in the number of ethnic minority pupils is primarily reflected in special education (i.e. LWOO). On average, almost 10 % of the secondary school population can be said to come from ethnic minorities. In LWOO, this is well over 33 %. The ethnic minorities are still under-represented in HAVO and VWO with only 3.5%.

Following this short overview of Dutch educational system and policy and the special attention for migrant children in the Netherlands we will focus on the different assessment instruments available for language competencies.

3. Instruments for educational assessment

During their school career pupils' achievements are measured at various occasions. Besides the traditional teacher-made tests, pupils also take more standardized tests. These tests do not only allow to assess the pupils' progress, but also make it possible to compare schools and to gain insight into the changes in pupil skills over time. In many schools for primary education pupils take (standardized) tests of a pupil-monitoring system. The main purpose of such a pupil-monitoring system is to follow the

development of pupils' skills during primary education, in order to find out as early as possible which pupils are lagging behind in their development.

As we have seen before, at the end of primary education (grade 8) pupils (age ± 11 years) have to choose a type of secondary education. This choice is based on both the advice of the primary schools' principle and the scores achieved on a standardized test. In principle any scholastic aptitude test can be used for this purpose, as long as an independent agency takes responsibility. However, over sixty percent of the primary schools use the so-called Cito Primary Education Final Test. Although both types of advice (principle's advice and test score) weigh equally, the final advice becomes more and more dominated by the test scores. In some cities certain secondary school types only accept pupils with certain scores on the Cito Final Test. Hence, the importance of Cito test scores increases.

During primary education pupils can take tests for the national assessment of educational progress. These tests can be taken at two occasions, in grade six and in grade eight. Contrary to both systems of quality control discussed above, the national assessment is based on samples and not on populations. For the national assessment only a sample of schools need to co-operate, and within these schools only a sample of pupils take the tests. This procedure of the national assessment has advantages as well as disadvantages, which will be discussed below.

The last type of assessment traditionally takes place at the end of secondary education. For each subject matter each pupil takes two exams: a school exam and a central exam. Teachers of a school develop the school exams and define norms for them. The school exams are taken at intervals throughout the final school year. Central exams are centrally developed, norms are centrally provided, and test-taking conditions are more or less standardized. The central exam consists of one test per subject matter, taken at the end of the final school year. The arithmetic mean of grades for both types of exam is the final mark. Based on all marks a pass-fail decision for pupils is made.

3.1 Central exams

As said before, the exams in secondary education consist traditionally of two parts: a school-related and a central part. Both parts of the exam differ with respect to the educational objectives tested. Educational objectives that are difficult to assess are tested through the school exams. For instance, speaking skills in foreign languages, or understanding of literature are tested through school exams. Traditional goals of language education (reading and writing skills) are tested in the central exams. Dependent on the language tested and on the type of school reading skills are assessed by means of a summarizing assignment (pre-university education) or by means of a traditional multiple-choice test.

Secondary school exams from different schools differ with respect to both content and moment of measurement. In general, pupils take school exams about three or four times (per subject matter) during the final year. The conditions in which pupils from different schools take school exams differ between schools. Some schools apply very strict conditions, similar to those for the central exams, while in other schools conditions are much more relaxed. Here, pupils can take the exam in their own classroom, with only their own teacher supervising. Furthermore, different school exams may differ with respect to the norms applied to pupil work. Sometimes the norms are written down in advance, sometimes norms are drawn up only afterwards, and sometimes no norms at all are available (Van den Bergh / Rohde / Zwarts, 2003). All in all, it does not make much sense to compare the scores on school exams at different schools, nor is a comparison of different years very sensible, as the content of the exams as well as the norms can change from year to year.

In this respect, central exams are, in principle, much more stable: content, norms as well as test-taking conditions are more stable over the years. However, the central exams differ per type of school. It is not feasible to make one central exam for pupils in different types of education; the variance in skills of these pupils is much to large; one exam for all would result in an exam that is too easy for the pre-university pupils, and much too difficult for those in the lower vocational tracks. Hence, schools cannot be

compared on the scores of the central exams, unless they provide the same type of education. A comparison of scores is therefore, by necessity limited to school types.

This leaves the question of how comparison over time per type of school is done. First of all, please note that Cito is not allowed to pre-test central exam items. The demands concerning secrecy of central exam items are so tight that it is not allowed to pre-test items. Hence, the psychometric properties of the exams can only be assessed in retrospection. Therefore, despite Cito's efforts to construct exams as well as they can (without pre-testing), the difficulty of exams may vary over years. That's why the Commission for Exams in Secondary Education (CEVO) can change the norms (i.e. conversion of raw scores into grades) afterwards. Although Cito tries to maintain a kind of norm-reference procedure, by means of an IRT-modelling[3], this by no means guarantees that exams in different years measure the same skills, nor that pupils with equal abilities score the same grades in different years.

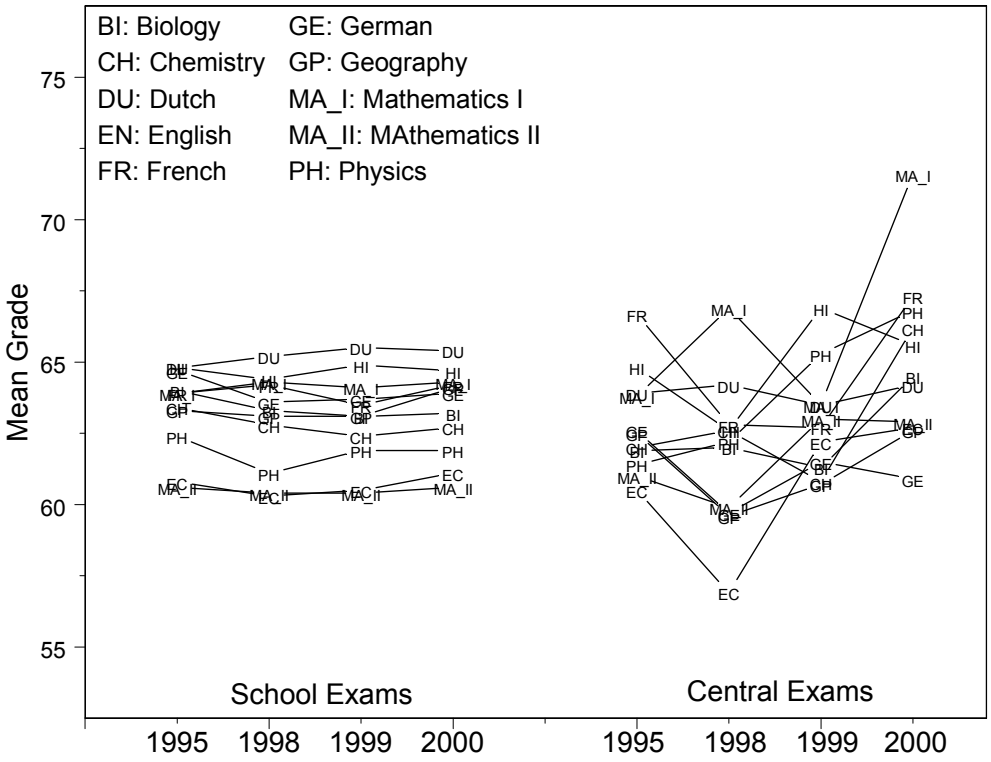
Cito maintains that grades from the central exams can be compared over years, but recent studies cast some doubt on this point (e.g. Van den Bergh et al., 2003). In an analysis of the scores of five years of school exams and central exams, the (mean) scores on school exams proved to be much more stable over different years than the scores on the central exams (see Figure 2).

As can be seen in Figure 2 the school exams show, contrary to the expectations discussed above, only very modest differences between different years: the lines for the different subject matters are almost horizontal. Apparently teachers adjust their norms in order to reach about the same mean grade each year.

[3] IRT-modelling is a statistical technique, which allows for a population-independent estimation of the difficulty of items in a test. Once the difficulty of the items is estimated, one can estimate the ability of the individuals who have taken part of the items [of: one can estimate the results of individuals on separate items]. Hence, skill can be expressed in terms of the probability of answering an item correctly; high-skilled pupils have a relatively high probability of answering an item correctly, while low skilled pupils have a lower probability of answering this item correctly. However, this does not necessarily mean that all high-skilled pupils have answered this item correctly (and v.v.). IRT-models have stringent assumptions. One of these assumptions is the assumption of local independence; there is only one source responsible for the differences in scores. Hence, it is assumed that the items are uni-dimensional in a psychological sense. If one would assume that both the school and the pupil contribute to differences in scores, this assumption would not be met by an IRT modelled design. This is one of the reasons why the number of pupils per school is kept low.

The outcome of the central exams is much more variable over the years. Take for instance French (FR): scores were high in 1995 and 2000, but only mediocre in the years in between. The other subject matters show similar variations. It is extremely unlikely that these differences in grades between years can be attributed to changes in skill in the population. Furthermore, the variability of scores on the central exams does not only appear in a graphical representation of mean scores, but also in a statistical analysis in which the total variance in scores was split up between pupils, schools, year-groups and exams.

Figure 2 Mean grades (on a scale from 1 (extreme low) to 10 (superb)) for School Exams (left) and Central Exams (right) for HAVO in four different years.



For most subject matters, the variance between exams proved to be substantial. It was concluded that the current construction rules for central exams do not warrant a comparison over years (Van den Bergh et al, 2003). Possibly, the differences between exams are not only attributable to differences in difficulty. There is also debate on whether the same cognitive skills are tested in each exam (De Glopper / Van Schooten, 2002). As, for instance, in different language exams different texts are used, there will be different types of questions asked. While one text may be more suitable for asking questions on relations between paragraphs (or sentences), another text may be more suitable for questions that require inferences. Possible small differences between texts and/or questions result in the measurement of different skills.

All in all it seems safe to conclude that central exams only warrant a comparison of schools of the same type in a specific year. Questions with respect to changes over years cannot be answered unequivocally. Much more stringent assumptions on the skills measured as well as on the difficulty of different exams are needed to allow a comparison of scores over years.

3.2. The Cito Primary Education Final Test

The Cito Final Test, taken at the end of primary education, is the successor of the Amsterdam-school test that was developed in the sixties in Amsterdam. The main purpose of the Amsterdam-school test was to achieve a more adequate advisory instrument to facilitate the choice for a secondary school, which was in those days primarily based on the advice of the primary schools' principal.

Each year during three mornings in February about 80% of the pupils in the last (eights) grade of primary education take the Cito Final Test. It consists of four parts: language, arithmetic, information-processing skills and world orientation (this last part is optional). The number of items per subject is announced in advance, and more or less the same each year. That is: 100 items for language (spelling: 20; writing: 30; reading 30; and vocabulary: 20), 60 items for arithmetic (numbers and calculations: 25; percentages and fractions: 25; measuring, time and money: 15), 40 items for

information-processing skills (use of study texts: 10; the use information sources like the yellow pages or an encyclopaedia: 10; reading and understanding of schemes, tables and graphs: 10; and map reading: 10), and 60 items on world orientation (geography, history and biology with 20 items each). The test consists completely of multiple-choice items, and pupils have to use a specialized answering form. This makes it possible to have all answers (of about 170 000 pupils) corrected centrally and reported back to the schools within a couple of weeks.

The results are reported (per pupil) in terms of number of items correct per subject (see Table 1), as well as in terms of a standardized score. This score is expressed on a scale ranging from 501 to 550. The school is provided with a school score, which is a mean of all pupil's scores.

Table 1. An example of an imaginary pupil report of the Cito Final Test

Results	Language	Arithmetic	Study skills	World orientation	Total	**Standard score**
Number of items	100	60	40	60	260	
Number of items answered correctly	73	48	23	52	195	**536**
Percentage score	56	52	42	57	53	

The imaginary pupil in Table 3 has answered 73 of the 100 language items correctly and 56% of the pupils have the same or a lower score. The scores on all items are transformed to a standard score, which equals 536 for this pupil. Only 2% of the pupils who choose lower vocational training have a higher standardized score; of the pupils choosing higher general educational 54% have a lower or the same score and 46 % have a higher score; of the pupils going to pre-university training, only 1% has the same or a lower score. Hence, this pupil will be advised to pursue his school career in higher general education.

The choice for a secondary school is in principle a choice of the parents (the above mentioned constitutional freedom of education). However, in some cities, the local education authorities have made admission to certain secondary school types dependent on certain scores on the Cito Final Test. For instance, the city of Amsterdam requires a score of 545 (or higher) for pre-university education (with classical languages; Gymnasium). Hence, in Amsterdam the Cito test is de facto not optional at all. Pupils need to take this test in order to qualify for secondary education. Therefore, nowadays it is mandatory for all children in Amsterdam to take this test, and the results are published. A problem is that some schools used to withhold the results of poorer learners (as this would have a negative influence on the schools' statistics).

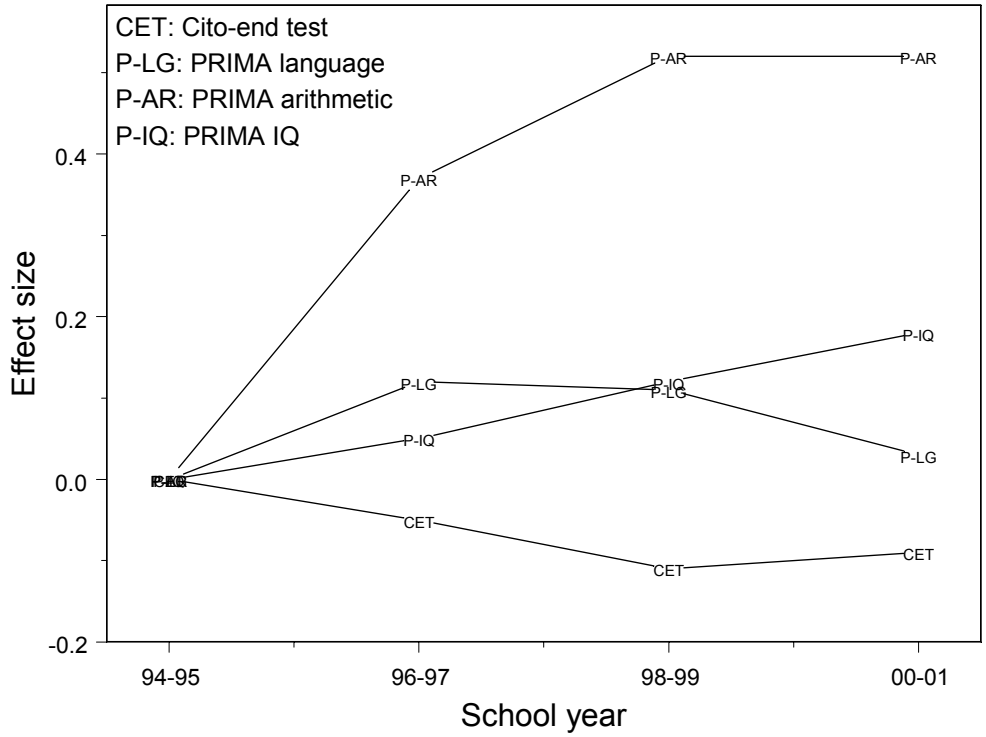
In practice this test is now a school placement test, although the difference between school placement and school aptitude is rather thin. Indirectly the score on the Cito Final Test also has bearing on aspects such as intelligence, concentration, learning speed, and perseverance. The predictive validity of the Cito Final Test has proven to be rather large. Even three years later, in the eleventh grade, correlations between school results and Cito Final Test scores are found. However, the predictive validity of the Cito Final test is not as well for migrant children (Uiterwijk, 1994). The predictions for these children are less precise.

Cito Final Test scores are an important output indicator. One of the reasons why the Cito Final Test is so important is that the scores in this test are submitted to an equivalation procedure that links the scores in one year to those in other years. Hence, these scores can be used to assess differences in achievement of output.

Unlike the central exams in secondary education, the Cito Final Test is not an exam. Therefore, in constructing the test Cito is not limited by rules concerning the pretesting of items. Hence, this test can meet the more stringent psychometric assumptions necessary to allow comparison of scores between different years. Besides, a classification matrix is used to construct the items, and, what is more, it is tested whether the items behave in the expected manner. This procedure facilitates the comparison of schools.

The Cito Final Test has, however, several drawbacks. Firstly, in practice only a part of the pupils take this particular test. Schools are obliged to subject their pupils to a to take a scholastic placement test, but they are free to choose which one. Hence, it is not unlikely that the choice for taking the Cito Final Test is related to aspects of educational practice, and/or educational values. Some schools, for instance, state that the skills of their pupils cannot be assessed by means of multiple-choice questions, and therefore reject the Cito Final test. Secondly, the Cito Final Test is a rather narrow test. With, for instance, only 100 items for language, one cannot test the complete domain of language skills. Important language skills like (real) writing, speaking or listening are not included. And even though reading is included, it can be measured only in a rather narrow way; it is not possible to generalize from the reading scores on the Cito Final Test to a diversity of texts and or situations.

Figure 3. Effect sizes for differences in scores between years for Cito-end test, as well as for three indicators of achievements taken form the PRIMA-cohort study.



In recent publications, it was suggested that there is a decrease in scores on the Cito Final Test. Careful statistical analysis showed, however, that the differences between years are only very small (Roeleveld, 2002) and in most instances not significant (Webbink, 2002). This holds even when the social economic status of parents, or the ethnic background of the pupils is taken into account. It was also shown that the Cito Final Test is rather limited with respect to the aspects measured. The test does not allow to generalize to the whole domain of language, arithmetic or reading. Other indicators, like PRIMA-cohort studies (a study, which focuses on the achievements of different cohorts of pupils), show an increase in pupil achievements for

arithmetic. Illustrative in this respect is Figure 3, in which the effect sizes for Cito Final Test scores, as well as three indicators for the PRIMA-cohort studies are presented According to Figure 3 (Roeleveld, 2002) language skills (P-LG) change marginally over years, whereas the maths skills (P-AR) definitely improve in the studied period. This conclusion differs from a conclusion based on the Cito Final Test scores, which show an extremely small decrease, or no change at all over years. However, we simply do not know whether the PRIMA-study gives an accurate description of changes in skills, or whether the Cito Final Test scores are more precise. In neither study the domain of, for instance language tasks, was sampled. That means that we do not know if, and to what extent the scores on both tests can be generalized to the domain to which the scores belong. Hence, one can only hope that the sample of items provides a correct mirror of the domain of possible items. Only in *national assessment* the domain of skills, texts, and tasks is sampled in such a way that one can make generalizations to the domain of reading skills (see below).

A separate point that we would like to mention in this respect concerns so-called backwash effects. Backwash effects refer to the possibility that schools teach their pupils only things that are tested in a test, or train their pupils with tests of previous years. In the Cito Final Test writing for instance is not really measured; children do not have to write. It has been suggested that this might have a deteriorating effect on education. It was feared that if writing were not tested, teachers would consider writing as less important and would spend less time on writing lessons. The general conclusion is that such backwash effects are hardly noticeable; teachers are not influenced as much as feared by the Cito Final Test (Wesdorp, 1979). Recently the time spent in classrooms on elements not tested in the Cito Final Test was investigated (e.g. Cito, 2002). No differences between years in time allocation to different subjects could be shown. Hence, it seems unlikely that refraining from measurement of skills has had a large influence on educational practice.

The second point of backwash effects concerns training pupils with previous versions of the test. Cito discourages this, but it undoubtedly happens. In many schools

and even at home children make previous versions of the Cito-end test, not only to become familiar with the type of items, but also as a means to increase their scores.

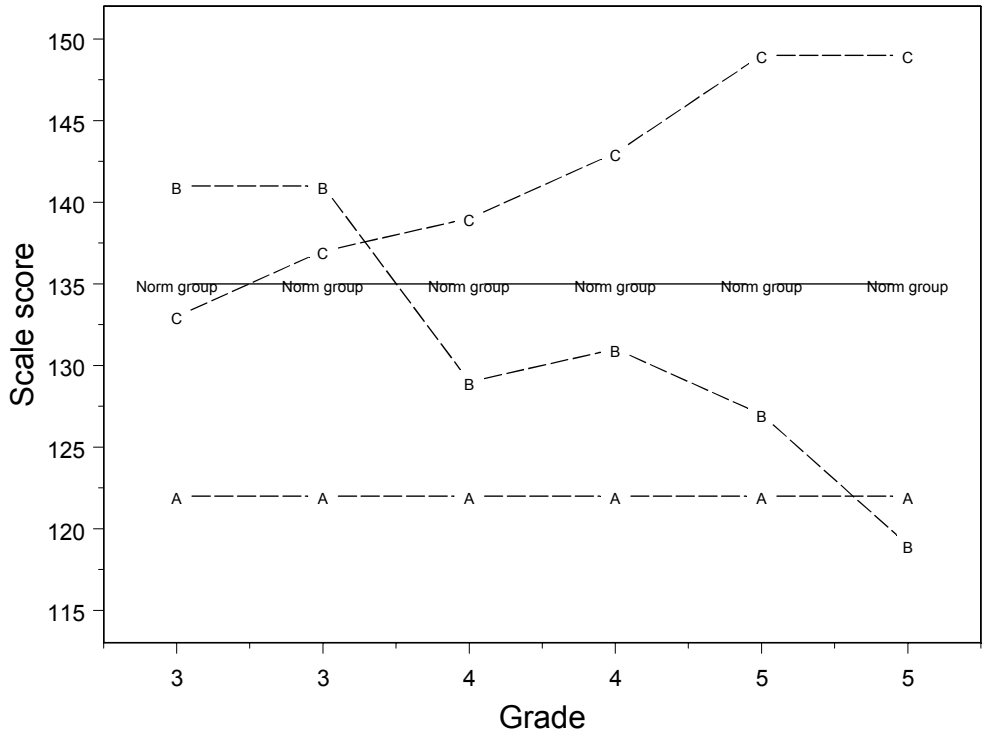
The Cito Final Test is by far the most influential test in basic education. This holds for individual pupils, who are advised in their choice of secondary schools on the basis of their Cito Final Test scores. But also for the schools themselves the Cito-end test scores are a relevant measure: schools are obliged to publish the mean score in their annual reports. And schools with high scores are considered to be good schools, whereas schools with low scores are considered to be weak. The dutch inspectorate is using the Cito test to evaluate the output of schools. Currently there is even a discussion going about the possibility of making the funding of schools (at least partially) dependent on their achievements. Schools that are doing well will get more funding than schools that are doing less well. As the achievement of a school is at least partially dependent on the quality of the pupils, efforts are made to introduce an achievement test in kindergarten. In essence, the difference between both tests will determine (partially) the funding of a school. Luckily however, this is still a plan in consideration, and not implemented as yet, but politicians appear to be in favour of this plan.

3.3 Pupil monitoring systems

There is a diversity of pupil monitoring systems. With pupil monitoring systems one can get reliable information on the learning progress of individual pupils. At more or less fixed occasions pupils take standardized (and sometimes) normed tests for the basic skills. The results are processed with specialized computer programs in order to get a progress report for each individual pupil as well as for each class. For instance, a pupil has taken tests at six occasions. Based on his scores one can identify whether he is doing well with respect to his previous test scores, whether he is doing well with respect to the other pupils in his class, and, in the case of a normed test, to other pupils in the population. Figure 4 shows how this works. It represents fictional scores of three pupils

on reading tests taken twice a year in third, second and fifth grade. The scores are compared to those of a reference group.

Figure 4 An example of the scores in a pupil monitoring system.



Pupil A in Figure 4 is a rather poor reader, during all tree years he scores about 13 scale points below the mean. Hence, his development in reading is at the same pace as that of the mean pupil. However, he is just a poor reader. The score of pupil B continuously decreases. His development of reading skills slows down. Hence, the teacher has to pay special attention to this pupil. Please note that the previous score(s) of a pupil provide the basis for interpretation. We can also compare the scores of this pupil with the scores in the reference group. It shows that this pupil was well above average at the first occasion, but is below average in the fourth grade. This is again an indication that this pupil ought to receive additional attention from the teacher. The third pupil in this figure

(C) starts below average, but the development of his reading skills is above average. This pupil is doing very well.

Pupil monitoring systems provide extremely useful information. Firstly, the results are utterly relevant for teachers. Not only does such a system help teachers to identify pupils who are doing worse than before (or who are at risk), but also will it enable them to evaluate their teaching. If too many pupils fail on a sub-subject, repetition of that subject might be useful. Secondly, the data of monitoring systems can be extremely useful for evaluation of the educational system in general. Unfortunately however, the data of monitoring systems are hardly gathered systematically. They are used at schools, and in classrooms, but they are not gathered for research purposes. Moreover, the results from different monitoring systems are hardly comparable. This holds both for the tests and for the way data are presented and/or stored. Due to the lack of standardization of different monitoring systems it is doubtful whether it is feasible to piece together a nationwide picture of the level of education based on the information gathered from different monitoring systems.

3.4 National Assessment

The main objective of national assessment is twofold. Firstly, it is aimed at assessing the achievement level at a certain moment in education. Secondly the observed achievements are compared to achievements in previous years in order to assess whether in general pupils perform better or worse than before on a certain subject matter. In this respect national assessment can be compared to a clinical thermometer that shows the state of education. This is the only way to lift the discussion on the achievement level of pupils above twaddle at the club. As everybody knows, everything used to be better, including educational achievements. We believe that complaints about the level of education are of all times and cannot, and should not be a basis for educational policy or changes in education. Already in 1892, a professor who rated the essays of candidates for pre-university education said '*these essays are below any criticism*'.

In national assessment the quality of education is measured in terms of output. However, this is not the only thinkable definition. Quality of education could also be defined otherwise, for instance in terms of learning processes. Are pupils challenged in the classroom? Do they get challenging, but feasible, tasks? Are they getting the right feedback, in a right way? Do they feel safe in the classroom? Obviously, quality of learning processes is something entirely different than quality of output. Although aspects of the learning process are inventoried in national assessment, learning processes are not the main objective of the assessment.

A third possible definition of quality could be stated in terms of availability of learning and development opportunities. Such a definition emphasizes the available qualification structure. National assessment does not imply aspects related to such a definition of quality. In short, of all possible types of quality assessment, national assessment only focuses on achievements and some aspects of the realized curriculum.

The first national assessment in the Netherlands was carried out in 1984, and was presented as a feasibility study on language skills assessment (Wesdorp et.al., 1986). The main question was whether it was feasible to let relatively large numbers of children take language tests and to rate the observed achievements afterwards.

In order to assess, for instance, the level of reading skills traditionally pupils are subjected to a reading test consisting of a text and questions about that text. However, two reading tests do not necessarily measure exactly the same skills. It is a well-known fact that the text on which the questions bear exerts a large influence on the (sub)skills measured (e.g. Van den Bergh, 1990). For instance, if readers have knowledge of the subject matter of a text, the test measures different sub skills than when the subject is relatively new to the readers. Or in other words, in general the correlation between the scores on different reading tests is not equal unity. This finding appears to hold for reading as well as for writing, listening and speaking (Kuhlemeier / Van den Bergh, 1998). Seemingly arbitrary decisions made in the course of developing measuring instruments for educational objectives turn out to be extremely important. In order to make a precise and reliable estimate of someone's writing skills, this person needs to

take up to twenty or more writing tasks (see, Van den Bergh / de Glopper / Schoonen, 1987). The random variance between language tasks, whether it concerns reading, writing, speaking or listening has proven to be large (although this is by far the best documented for writing). Hence, in order to reach conclusions and make judgments about pupils' skills at a certain moment in education one cannot just have pupils to take one reading, writing, speaking or listening test. A test consisting of only one assignment does not allow generalizing to the complete domain of the test. So, from the universe of (possible) tests a sample of tests has to be drawn and presented to pupils. Generally speaking, this means that pupils take a large number of different tests for each skill. Or to put it differently, the variance of tests has to be incorporated into the design of the assessment study as a random factor (see Clark, 1973). In the aforementioned feasibility study, not less then 17 reading tests were developed (with six to sixteen items per test).

Contrary to school exams or qualification tests, national assessment is not aimed at evaluating individual pupils' achievements. Hence, in national assessments samples can be used; samples of schools and samples of pupils. As we don't need a precise estimate of for instance an individual pupil's reading ability, there is no need to have all pupils in the sample to take all reading tests. A matrix design can be used to allocate tests to pupils, providing of course that the scores on individual tasks can be linked to one another (by means of some statistical model). The scores on all tasks can then be expressed on the same scale, and conclusions that go beyond the individual test results can be drawn. Hence, national assessment in the Netherlands draws heavily on IRT-modelling, in which the ability of the pupils can be estimated from the difficulty of the items. But items measuring different skills can never be expressed on the same dimension. Therefore, many different dimensions, or elements of a skill, have to be distinguished. For instance, for reading (at the end of primary education) a distinction is made between technical reading, reading of reporting texts, reading of contemplating texts, reading of directive texts, reading of argumentative texts, reading of fictional texts, using reference books, interpreting tables and figures, and map reading. For speaking and writing a distinction is made according to language acts. This results in seven text function types (for both writing and speaking), such as giving information,

describing, asking information, persuade, etc. For each text function type three different tasks are developed. In Table 2 some examples of items of the national assessment are presented.

Table 2. Two examples of language items form national assessment (source: Cito, 2002).

Reading figures and tables

Distances in kilometres

	Pelo	Puki	Suka	Tremp
Alomi	30	100	50	90
Septono	70	20	315	10
Manuk	54	210	38	20
Kura	165	85	340	310
Leksa	40	90	115	50

What is the distance between Kura and Suka?

_____ kilometres

Asking information (writing).

Read this first!

Imagine …

You regularly eat SMUCO potato chips. On a package you read the following:

ACTION	Dixys radio shops are found everywhere in the Netherlands. It is also possible to have the headphone set sent to your house, but this will cost EUR 3,-. You can pay this additional charge by
Incredible but true: with the new SMUCO potato	

chips you can get a free stereo headphone set that fits your walkman as well as your stereo. This is what you'll have to do: Each SMUCO bag has one SMUCO saving point. Gather three points and send these in a franked envelope to SMUCO. You will receive a voucher. You can trade this voucher at Dixys radio shop for a marvelous free headphone set (FOR FREE!)	means of a cheque. Send the envelop with the saving points to: SMUCO saving action P.O. Box 3333 1200 AD Hilversum Mention clearly your name, address and zip code and indicate whether you want a voucher or the headphone set sent to your house
Task You have three saving points and you want a voucher for the headphone set. You also want to know where you can find a Dixys radio store. Write a letter to SMUCO After that write the envelope.	

Speaking: evaluating

The pupil is asked to react to an advertisement and to send in a story for a radio program.

Imagine… In the newspaper you read the following ad:

RADIO VERY YOUNG
The youngest radio of the Netherlands.

A new summer a new program!
Between 7 and 8 in the evening

REALLY HAPPENED, REALLY TRUE.

Send in an audio cassette with your story about

NASTY PEOPLE

And you stand a chance it be on the air soon!

Send your cassette to:
RADIO VERY YOUNG
P.O. BOX 12345
1014 RP Jonge Tonge

You find this ad rather appealing. Telling a story on nasty people. You certainly can recall some really annoying event. You must have experienced that somebody was being really nasty. Think of things like:
Somebody treated you unfairly
Somebody let you down
Somebody teased you
Somebody gossiped about you
Somebody made a fool out of you
Somebody laughed at you
Somebody didn't keep his promises
Somebody didn't let you join
Somebody jumped the queue
Somebody ….

The aim of national assessment is to give insight into the achievement level pupils have reached at a certain age in a diversity of subject matters. If the aim of national assessment is taken seriously it has some far-reaching implications concerning the instruments, test-taking situations, scoring as well as the design of assessments.

It is of course unfeasible for pupils to take all 21 writing and all 21 reading tasks. This would take an enormous amount of educational time and would diminish the willingness of schools to participate in national assessment. What is more, there is no need for all pupils of a school to take all tasks. The aim of national assessment is not to compare schools, but to get information about the pupils. Hence, a sample of pupils will have to take a sample of tasks. In general about three pupils of a school will take the same tests.

This statement contains at least one pitfall. A sample of pupils means essentially a sample of schools in which a sample of pupils is drawn. Hence, a two stage sampling procedure is in operation; in the first stage a sample of schools is drawn, and in the second stage pupils within schools are sampled. Such two-stage samples are by their nature less precise than single-stage samples (with the same number of elements), because the achievements of two randomly selected pupils from one school are likely to be more alike than the achievements of two randomly selected pupils. In effect a sample of one pupil per school would be the smallest and most precise sample. However, the costs would be far to high, as for each pupil a different school needs to be visited. Therefore, the same test is generally given to (about) three pupils per school. In practice, however, more pupils per school are tested, but with different tests. For instance, one pupil of a school can take three writing tests, while another pupil of the same school takes three reading tests, and yet another pupil of the same school takes three speaking tests.

A crucial point in sample studies is of course the precision of the mean estimate. An estimate has to be rather precise if the aim is to evaluate the status quo as well as to identify changes in achievements between different assessments. In general a choice is made for a precision of 95% (Wesdorp et al., 1986). So, all other things being equal, the true mean score will not deviate much from the estimated mean score. This holds, however, for the population as a whole, but estimates of mean scores for several relevant subpopulations need to be precise as well. Hence, for each subpopulation the sample must be precise enough, but it will always be less precise than the overall sample.

Except distinctions between boys and girls also distinctions are made according pupil weighting (see above). Pupil weighting is an important factor in the funding of schools. Based on pupil weighting the population of schools are divided in three strata. The first stratum contains schools with primarily children whose parents finished secondary education. The second stratum includes schools with primarily Dutch working class pupils and few non-native pupils. The third stratum consists of schools with many Dutch working class pupils and non-native pupils.

For each stratum a sample is drawn, in order to arrive at a total precision of 95% and a somewhat lower precision per sample (see for instance, Cito 2002). This seems to be appealing figures, but please note that schools participate on a voluntary basis in national assessment. Possibly, the schools that are not willing to co-operate with national assessment differ in certain aspects from the ones that are co-operating. This means that there might be a bias in the sample due to nonresponse. In Table 4 bias due to nonresponse in the feasibility study of Wesdorp et al. (1986) is quantified.

Table 4. Possible bias due to nonresponse in the estimate of the population mean on one reading assignment (X = 4.17).

Score of nonrespondents	Percentage of respondents with score	Bias in estimate of X
0	2.8	2.00
1	4.5	1.52
2	8.1	1.04
3	12.3	0.56
4	21.6	0.08
5	31.9	-0.40
6	18.9	-0.88

If, for instance, the score of nonrespondents on a task is 0, then the bias in the estimate of the population mean is 2 score points. Such a large difference due to nonresponse, however, is very unlikely, as only 2.8% of the respondents did get this score. If, on the other hand all nonrespondents would get the maximum score on this task, the estimated mean would be 0.88 below the true population mean. Even if the mean difference between respondents and nonrespondents would be only one item answered correctly, the bias in the estimated mean would be considerable. In that case the bias due to nonresponse would be about six times higher than the standard error. Purpose of this

illustration is to show that the problem of nonresponse may be considerable, especially if one takes into consideration that the response percentages are below 50% and in the lowest stratum (with working class and non-native children) only about 30%.

The results of national assessments in the Netherlands are reported per scale. For instance, on the test of reading argumentative texts, that consists of 18 items, the 10% weakest pupils have slim chance (smaller than 50%) to answer any item correctly. The average pupil has answered six items correctly, six items moderately, and failed on twelve items. The percentile-90 pupils have answered eleven items correctly, seven items moderately and six unsatisfactorily (Cito, p.61). Next to such general descriptions, the achievements are broken down by gender, language spoken at home, age and pupil weighting. In general girls score slightly higher than boys (only on contemplative texts boys outperform girls). Moreover, the scores of pupils who speak standard Dutch are slightly higher than those of pupils who speak dialect of Dutch or those who speak more languages at home. Pupils who only speak a foreign language (usually their parents' mother tongue) at home have the lowest language scores. Pupil weighting proves to be an important indicator of language achievements. The higher the pupils' weight, the lower their achievements.

Over the years the national assessments revealed that most language achievements did not change; between 1993 and 1998; pupils in 1998 can read and write as well as those in 1993. The same holds for supportive skills like grammar, word knowledge, spelling, etc. Only on listening to reporting texts and fictional texts the achievements appear to decrease gradually between 1988 and 1998.

The achievements of pupils can be summarized per scale. Important however are value judgments on the achievements; what does it mean that 50% has answered six items correctly? Is such an achievement satisfactory, or does this represent a minimal achievement? Hence, it is essential to define norms in order to value achievements. Cito uses a method of norm setting, in which informed raters decide how many items pupils need to answer correctly in order to achieve a minimum, sufficient or advanced achievement. This procedure can be used for all subject areas. However, this method does not always produce relevant results. For instance, politicians who establish norms

were asked how many spelling errors they thought pupils make. Without fail they underestimated pupil achievements. Sometimes, other methods for standard setting might be preferable. For some writing tasks, for instance, it might be better to use the core-item method. This method consists of identifying the most important items in a task - items that a pupil must absolutely answer correctly. If a pupil fails to mention a return address in the SMUCO potato chips task (see Table 2), the communication can never succeed. Hence, one could argue that this information is more essential than other elements (compare, Kuhlemeier / Van den Bergh, 1990).

Pupil achievements depend, at least partially, on education. Therefore, national assessment also gathers information on aspects like methods applied and amount of time spent on different subject matters. In this way also information can be provided on which methods are used, and which trends get apparent over time. Product oriented methods are nowadays by far more popular than strategic or eclectic methods. On average almost 5 hours per week are spent on language activities in the last three years of primary education, of which 2:18 hours are spent on reading.

Although the achievements are many times broken down (e.g. by method), we cannot interpret these results as input-output analyses. The number of pupils per school is much too low to allow for precise estimates of achievements at school level. Neither the design of national assessment in the Netherlands, nor the item-response modelling allows for such type of analyses. The item-response method is in principle developed and should be used be used primarily for an efficient estimate of pupil achievements.

4. Discussion

After the description of a variety of assessment systems we will now discuss some main issues regarding the purposes and outcomes of the national assessment compared to the other instruments.

4.1 The language concept behind language assessment

From a linguistic point of view it is important to mention which domains of language competencies are tested in the national assessment instruments. Criticism has focussed on the concept of language testing. Traditionally language tests draw heavily on the distinction between receptive and productive skills, and are restricted to reading and listening (Van Berkel, 2002). As the skills are tested separately, only a fragmented picture of language competencies is likely to arise.

In national assessment the traditional language domains are divided up into sub domains. In reading a distinction is made between reading fiction and non-fiction, and with respect to listening the subtypes argumentation, reportage, opinion and fiction are distinguished. Next to these traditional domains the following so called 'supportive skills' are tested: (1) meaning of words, (2) the semantic relationships categorizations, generalizations, contradictions and the coherence within a semantic field, (3) the morphological formation of plural, gender, composites, diminutives, and degrees of comparison, (4) syllables, (5) function words, (6) the syntactic construction of compound out of simple sentences, the conversion of assertions into questions, and the construction of correct sentences by changing word order, (7) orthography, and (8) interpunction and, finally, (9) the alphabet. These linguistic elements of Dutch are tested separately. The degree to which knowledge of these characteristics interrelates with language use capacity is not considered.

This fragmented language concept is defended with reference to the necessity of constructing a valid test, that allows for comparison of language skills over several years as well as for a comparison of individuals. Only by dividing language

competencies up into these small, isolated sub skills, a comparison of these skill over years is guaranteed. Interestingly, the final report (Berkel et al., 2002) contains a plea to base the national assessment on an interactive language concept, which integrates the fragmented language subskills (Berends 2002, 12). This plea has not triggered any reactions. Instead, Cito uses, so-called 'curricular units' (units which have a specific place in schools' curricula), as basic elements in national assessment. One could wonder however, if the proposed way of measurement of Berkel is compatible with the aims of national assessment, which is comparison of scores over years and comparison of scores between sub-populations. Consequently statistical and psychometric assumptions will have to be met in order to allow for such comparisons. On the other hand the present method of matrix sampling (in which each pupil takes only a small part of the tests) does not allow for exhaustive analysis of the relations between scores on different tests. If one would like to analysis the relation between language skills, one would have to use a less efficient test-taking design (e.g. Kuhlemeier / Van den Bergh, 1998).

Compared to the Cito Final Test the national assessment test covers a relatively broad language domain by testing each sub domain and supportive skills by means of altogether 166 different tests, whereas the Cito Final test has only 100 language items. This is all the more striking since the Cito Final Test is far more influential with respect to the determination of the school career of individual pupils as well as with respect to the social relevance and regular publicity. Every year Cito Final Test attracts big media attention in the public debate on education.

4.2 National assessment, individual language development and quality control

From a didactic and curricular point of view an interesting question is how different assessment instruments allow to assess the language development of individual pupils. First of all, it should be noted that school exams and the Cito Final Test are derived from the Dutch qualification structure with its norms and purposes, and do not aim at

documenting the progress of individual pupils. National assessment tracks changes at the system level. So, individual pupils are of less importance for this type of evaluation.

Progress of individual pupils becomes only visible through regular tests within a framework of a pupil-monitoring system that periodically relates individual test scores to a standard norm. Although only few sub skills in linguistic sub domains are tested, the system has a general diagnostic function, since it warns teachers when pupils do not achieve the expected progress. Teachers then need to interpret the test result based on their own understanding of the pupil's language achievements and decide whether extra support is necessary.

The pupil-monitoring system starts from an import presupposition that is not self-evident in the Dutch educational structure, namely that all schools offer the same (language) curriculum. Only if a school offers the language curriculum that corresponds to the pupil-monitoring system its tests make sense. Under circumstances it may not be clear whether the individual pupil fails because he is delayed in his language development or simply because his school programme did not yet deal with the tested skills. Since the Netherlands cherish freedom of education, schools can set up their own curricula and they do so. In primary education, this implies that various pupil-monitoring systems function next to each other, but as there is more or less consensus on the curriculum (for the basic skills anyway) different monitor systems can function next to each other. In secondary education the language curricula are so diverse, that attempts to develop a pupil-monitoring system have failed.

In this connection we have to mention an institution with a long tradition within the Dutch educational system, that we have not mentioned before, namely the school inspection. This institute's task is to control the minimal educational quality on national and local level. The inspection has the right and duty to interfere in school curriculum and didactics when negative achievements occur, for instance when parents complain. All schools have to co-operate with the school inspection and comply with their advices. School boards are responsible when schools do not implement the inspector's directives. For example, the city of Amsterdam - in its position as school board for the Amsterdam public schools - was recently obliged to pay extra lessons in a case where parents

complained about the poor quality of language education in a school that had repeatedly ignored the directives of the local school inspection. The national school inspection reports periodically to the minister of education on important issues in all educational sectors and issues advices on the execution and development of the assessments instruments under discussion in this report. The actual impact of school inspection's advice and report is not always very apparent. Sometimes, however, the institute initiates a political dispute on the quality of education.

4.3 Language assessment and migrant children

An important question in this study is how suitable various types of language assessment are for the assessment of achievements of migrant children. As we stated before, despite extra efforts the position of these children in Dutch education is problematic. Generally, they perform rather poorly on various tests and they are over-represented in special primary and secondary education as well as in the vocational sectors of education. A tendency towards ethnic segregation in schools becomes evident and more and more accepted (Karsten et al, 2002). The results on language assessment tests have an important impact on the school career of migrant children, but can not completely account for their problematic situation: other factors such as school choice by parents and preference for more culturally homogenous schools are also important. The proverbial Dutch cultural tolerance looses its mythical value and the presence of everyday and institutional racism appears (Koole / ten Thije, 1994). Concerning the suitability of language assessment for migrant children there are three important issues to consider.

First of all, language tests for migrant children suffer from a cultural bias. Many studies have shown that item bias is an issue in the Cito Final Test (e.g. Kok, 1988; Uiterwijk, 1994). However, the effects of bias on the test scores are at best minimal; the overall test score is hardly influenced by bias against culture and/or gender. Hence, although the effect exists in the Cito Final Test, it is an effect on only a very small part of the items, nevertheless even for these biased items it is small but significant

(Uiterwijk, 1994). For effects of item bias on scores in central exams, or pupil monitoring systems, no data are available as yet.

The next issue concerns the question of how language assessment affects the opportunities for migrant children. Above we noted that the pupil-monitoring system helps to identify children who are at risk. This refers to migrant children as well. As for the Cito Final Test, it prevents migrant children to choose for an educational sector that does not correspond to their capacities. In a sense the Cito Final Test tempers parents' (too) high expectations. As the Cito Final test appeared to be less precise for migrant children, teachers have 'over-advised' them. Afterwards pupils could not satisfy school standards and abandoned school before they actually achieved their formal qualification (Van Veen / Berdowski, 2000; Vedder / Kloprogge, 2001). Illustrative in this respect is the Amsterdam situation. Since several years the Cito Final Test is obligatory for all pupils in Amsterdam, and the outcome of the test delimits the possibility of secondary school choice. As a consequence the school choice of migrant children among others has changed slightly and a tendency can be observed that less migrant children abandon school before their qualification (Minister of Education, 1999).

The accuracy of the Cito Final Test for low scoring children has been a concern of the Cito as well. Therefore, they have started last year with a special version of the Cito Final Test in which differences between low scoring children is optimized. Hence, standard error of the estimate is reduced considerably and advice for types of secondary education can be made more secure.

Finally, we want to stress the fact that we have only discussed mean scores, and reported (some) changes in mean scores over years. For instance, it was concluded that migrant pupils have on average lower language scores. However, we did not focus on differences between sub populations of migrant children; as these differences tend to be large as well the generalization to one group is hardly warranted. Neither did we discuss the differences between pupils with sub-populations. These tend to be large as well, especially within the population of migrant children.

4.4 The Netherlands: European testing champion?

What is the overall picture of our survey on the various assessment instruments? Central exams, Cito Final Test, pupil-monitoring systems measure pupil achievements at certain points during the educational career of pupils. Each of these pupil achievement-measuring systems has its own strengths and weaknesses. For instance, school exams are primarily developed to make pass/fail decisions. They can be used to compare schools from the same type in one year. However, schools from a different type cannot be compared, as their pupils take different exams. Nor do the exams allow for a comparison of different years' pupil achievements. However, the main drawback from using exam scores as indicator of the level of education is that exams are relatively small measures of achievement; one cannot generalize to the domain the test comes from.

Since national assessment has been institutionalised 15 year ago (economic) interest is involved, and therefore, the system needs to legitimise itself constantly. Currently little is known about the influence of its outcomes on educational policies on a national and local level. The first feasibility test in 1985 caused a vehement social discussion on ‘functional illiteracy’. The next three studies documented a relatively stable level with respect to reading and listening competencies. However, they also revealed the structural problems of migrant children. As long as the national assessment studies are not used to put these issues on the political agenda, they legitimise the current language and educational policy.

References

Berends, R. (2002). In Nederland missen we een goed concept voor taalonderwijs. In: S. van Berkel et al. (red.). *Balans van het taalonderwijs halverwege de basisschool 3. Uitkomsten van de derde peiling 1999.* Arnhem: Cito, 12.

Bergh, H. van den (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement*, *14*, 1-12.

Bergh, H. van den / Kuhlemeier, H. (1990). De haalbaarheid van eindtermen voor de basisvorming. *Pedagogische Studiën*, *67*, 1-15.

Bergh, H. van den, Glopper, K. de / Schoonen, R. (1987). Directe metingen van schrijfvaardigheid: Validiteit en taakeffecten. In, F.H. van Eemeren / R. Grootendorst (Red.), *Taalbeheersing in ontwikkeling*. Dordrecht: Foris Publications.

Bergh, H. van den, Rohde, E. / Zwarts, M. (2003). Is het ene examen het andere? Over de stabiliteit van schoolonderzoek en centraal examen. *Pedagogische Studiën*, *80*, 176-191.

Berkel, S. van, Schoot, F. van der, Engelen, R. / Maris, G. (red.) (2002). *Balans van het taalonderwijs halverwege de basisschool 3. Uitkomsten van de derde peiling 1999.* Arnhem: Cito.

Cito (2002). *Balans van het taalonderwijs aan het einde van de basisschool 3*. Arnhem: Cito.

Clark, H.H. (1973). The language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335-359.

Doornbos, K. (1986). De verzorgingsstructuur van het onderwijs. In: J.A. Kemenade et al. (red.) *Onderwijs: Bestel en beleid I: Onderwijs in hoofdlijnen*. Groningen: Wolters, 244-270.

Glopper, K. de / Schooten E. van (2002). Dalende leerlingprestaties op de centraal schriftelijke examens Duits, Engels en Frans in mavo, havo en vwo? *Pedagogische Studiën*, *79*, 5-17.

Karsten, S., Roeleveld, J., Ledoux, G., Felix, C. / Elshof, D. (2002). Schoolkeuze en etnische segregatie in het basisonderwijs. *Pedagogische Studiën* 79/5, 359-376.

Kok, F. (1988). *Vraagpartijdigheid*. Amsterdam: Universiteit van Amsterdam.

Koole, T. / Thije, Jan D. ten (1994). *The Construction of Intercultural Discourse. Team discussions of educational advisers* (Utrecht: diss.) Amsterdam / Atlanta: RODOPI.

Kuhlemeier, H. / Bergh, H. van den (1998). Relationships between language skills and task effects. *Perceptual and Motor Skills*, *86*, 443-463.

Minister of Education (1999). *Plan van Aanpak Voortijdig Schoolverlaten*. Den Haag: SDU.

Minister of Education (2003). *Onderwijsprofiel van Nederland. Samenvatting van de belangrijkste beelden van 'Education at a Glance'. Het onderwijs indicatoren rapport van OESO.* Website: http://www.minocw.nl/brief2k/2003 /doc/44136g. PDF, 15 January 2004.

Minister of Education (2004). *Fact and figure*. Website: http://www.minocw.nl/english/ figures2003/008.html, 15 januari 2004.

Roeleveld, J. (2002). De kwaliteit van het basisonderwijs: dalen de Cito-scores? *Pedagogische Studiën*, *79*, 389-403.

Sociaal en Cultureel Rapport (2000). *Nederland in Europa*, Sociaal en Cultureel Planbureau, Den Haag.

Uiterwijk, H. (1994). *Eindtoets basisonderwijs: De bruikbaarheid van de eindtoets basisonderwijs voor allochtone leerlingen*. Arnhem: Cito.

Vedder, P. / Kloprogge, J. (2001). *Onderwijskansen op tafel: het bestrijden en voorkomen van onderwijsachterstand*. Den Haag: Management Landelijke Activiteiten Onderwijskansen PMPO.

Veen, D. van / Berdowski, Z. (2000). *Preventie van schoolverzuim en zorg voor risicoleerlingen*. Leuven: Garant.

Webbink, D. (2002). Moeten we ons zorgen maken over dalende scores op de Eindtoets basisonderwijs? *Pedagogische Studiën*, *79*, 184-191.

Wesdorp, H. (1979). *Studietoetsen en hun effect op het onderwijs*. Staatsuitgeverij: Den Haag.

Wesdorp, H., Bergh, H. van den, Bos, D.J., Hoeksma, J.B., Oostdam, R.J., Scheerens, J. / Triesscheijn, B. (1986). *De haalbaarheid van periodiek peilingsonderzoek*. Lisse: Swets & Zeitlinger.