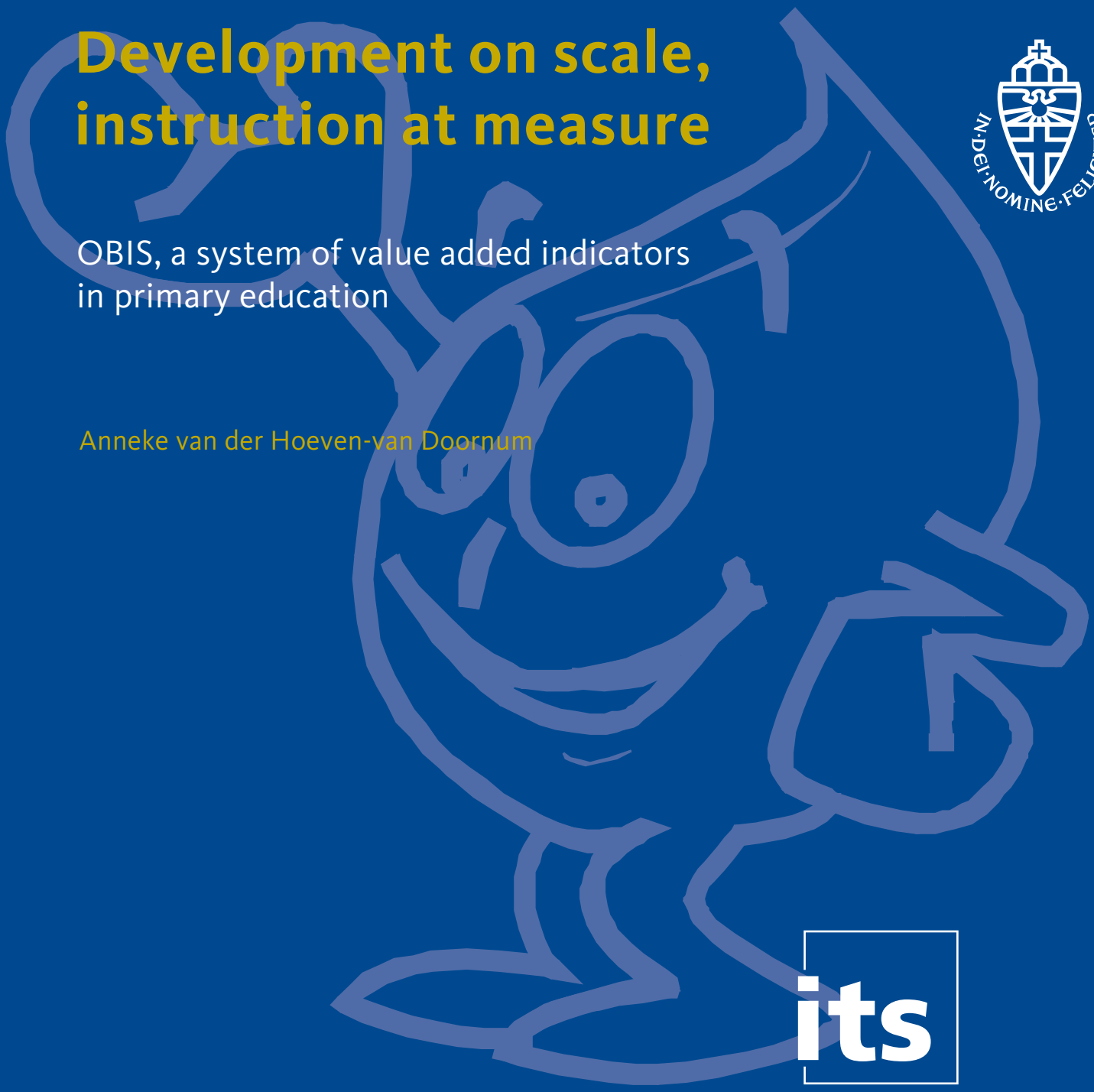
Het ITS maakt deel uit
van de Radboud
Universiteit Nijmegen

Development on scale, instruction at measure

OBIS, a system of value added indicators in primary education

Anneke van der Hoeven-van Doornum



its

DEVELOPMENT ON SCALE, INSTRUCTION AT MEASURE

Development on scale, instruction at measure

OBIS, a system of value added indicators in primary education

Anneke van der Hoeven-van Doornum

December 2005

ITS – Nijmegen

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK DEN HAAG

Van der Hoeven-van Doornum, Anneke

Development on scale, instruction at measure. OBIS, a system of value added indicators in primary education. Ontwikkeling op schaal, onderwijs op maat. OBIS een systeem voor het vaststellen van toegevoegde waarde in het basisonderwijs.

ISBN 90 – 5554 – 286 – 5
NUR 840

Contents

Acknowledgements

This report represents five years of research carried out in collaboration with the CEM Centre at Durham UK. In 1997 Peter Tymms of the CEM Centre kindly permitted me to translate and adapt the PIPS Baseline On Entry Assessment for use in The Netherlands. Therefore, first of all, I would like to express my sincere gratitude to Peter and the CEM Centre for the unconditional confidence and support I received right from the first e-mail exchanges starting Wednesday July, 2nd 1997. Obviously, the research borrowed heavily from the knowledge, the intellectual property, and publications produced by the researchers of the CEM Centre: Peter Tymms, Carol Taylor Fitz-Gibbon, Christine Merrell, Paul Jones, Brian Henderson, Richard Almond, Kate Bailey and all those other kind and helpful people.

Furthermore, I am grateful to the Netherlands' Organization for Scientific Research (NWO) for funding the project 'Development on scale, instruction at measure' under grant no. 411- 210-05. Without this support this study could not have been realized. I also thank the management of the Institute for Applied Social Sciences to allow the research project to be carried out. From colleagues and co-workers, I like to mention Nico van Kessel and Yvonne Meeuwsen for their various contributions to support the project.

Many others also have contributed in various ways- schools in the first place by participating in the project. I thank all the schools, the teachers and pupils to have participated in the project for all four years. In this regard, probably against the rule of anonymity, I would express my gratitude to both Beppy van Eck and Peter van Dijk for their genuine interest in the OBIS baseline assessment and their promotion of the wider use of the test.

Special gratitude and thanks are due for the insightful comments, unlimited and valuable support, continuous help and time I received from significant others.

Executive summary

This report deals with the OBIS[1], a test on the development and school performances of 4-6-years old children. OBIS was derived from the British PIPS[2] Baseline Assessment by translation and careful adaptation when necessary. The research with OBIS consists of two parts. The first part deals with questions concerning the construction of OBIS, and the reliability and validity aspects thereof. In the second part OBIS is applied to answer questions about:

- the impact of testing on pupils progress;
- the prediction of learning achievements using OBIS and background of pupils;
- the use of OBIS to identify pupils that need special education.

Part I Content and Structure of OBIS

Chapter 1 introduces the aim and background of the study, the theoretical framework and research questions. Methods and design are described, just as the baseline assessment related issues 'value-added' and 'monitoring progress'.

PIPS and OBIS are aimed to measure cognitive skills: literacy and numeracy of 4 to 6-years old children. In addition the child's attitude towards school is assessed. The test is very child-friendly and uses high-quality pictures attractive for children. It is available as an interactive computer program, and needs only about 20 minutes to be completed. As a rule, children are assessed individually preferentially by their own teacher. This appears to provide the teacher with valuable diagnostic information on the child.

Chapter 2 gives an extensive account of OBIS. High correlations were found between different reading and mathematics sections of OBIS and PIPS. OBIS fulfilled all requirements of reliability and validity. OBIS, and its extensions for later years (V-OBIS), were then used in a quasi-experimental longitudinal study with 450 children in 11 schools. Regular testing appeared to improve children's learning achievements and teachers' professionalism. OBIS scores were found to be a better predictor of future learning achievements then was the socio-economic and ethnic background of the child. The possible role of early testing, in particular using OBIS, in the allotment of finances to schools was discussed.

---

1 OBIS is the acronym for Onderbouwinformatiesysteem (Lower School Information System).

2 PIPS is the acronym for Performance Indicators in Primary Schools.

Differences between the educational systems in the United Kingdom and The Netherlands were discussed. In both countries there has been considerable debate about the type of assessment suitable for young children. At present, early testing is widespread and firmly established in the United Kingdom, while in The Netherlands it is in its infancy.

In the process of construction of OBIS from PIPS Baseline, the adaptation of the mathematics section was mainly limited to translation of the English items into Dutch. The correlation between the English and Dutch maths items was high (.90). The adaptation of the language section, requiring some more effort to adapt, yielded correlations between the difficulties of Dutch and English vocabulary and letter items of .80 and .90, respectively.

The aim of Chapter 3 was to describe the data collection and technical details of the instruments described in Chapter 2. It forms as such the transition from the descriptive to the explanatory part of this report. The chapter starts with a closer look at the sample of schools that formed the basis of the data. This is followed by details of the collected data. Next to this, the relationship between the various measures for reading, phonics and maths is set out, and comparisons between the groups broken down for date of intake, age, sex and socio-ethnic background (ses), are made.

Based on the idea of measuring what matters, PIPS and OBIS are constructed to be good predictors of later reading and mathematics achievements. Indeed, OBIS showed high correlations of .78 and .68 between baseline assessments and achievement scores one and two years later, respectively. The sections “Letter Identification” and “Digit Identification” appeared to be the best predictors. The reliability of the OBIS was found to be good to excellent as indicated by test-retest correlations of .92 for mathematics and .97 for reading. OBIS has been compared with existing Dutch instruments, mainly tests designed by CITO (Taal alle Kleuters and Ordenen). Correlations of OBIS with these tests were highly significant.

A continuation of OBIS, meant for grade 3 and named V-OBIS, was constructed on the basis of PIPS Year 1. The tests measure academic achievements, attitude towards school, and also developed ability, vocabulary and non-verbal ability. The latter are excellent predictors of academic attainment, and can be used to o determine if a child is making the expected progress. This so called ‘concurrent value added’ can be estimated for a child regardless of whether it was assessed earlier using OBIS.

Part II Studies using OBIS

The OBIS test was subsequently used in various studies, as described in the following chapters. In the first and main investigation using OBIS, data were collected when the children first entered school, and after 1 and 2 years in order to monitor their progress.

In Chapter 4, the question was investigated if assessments using OBIS can stimulate children's educational progress and teacher's professionalism. The study was carried out quasi experimentally in about 450 children in 11 schools at the start, of which 400 were still present after two years. The schools were randomly distributed among three treatment groups. Teachers in group 1 received, in addition to the OBIS scores, all relevant information and feedback on pupil and class level in order to stimulate their teaching. Teachers in group 2 also received the OBIS scores and feedback, but no detailed information about the individual target setting for the pupils. Group 3 only received feedback on test scores.

First of all, OBIS turned out to be a reliable instrument for baseline assessment. Generally, teachers' responses to the OBIS were very positive. The initial scores of children from ethnic minorities lagged considerably behind those of their peers. However, after one year of education in school their progress turned out to be higher than that of their higher SES peers. The findings supported the role of individual feedback to teachers on school achievements of their pupils. In this study the positive effect of feedback was most pronounced on reading abilities of pupils. The virtual absence of test effects between experimental groups 2 and 3 suggested the occurrence of a Hawthorne effect, e.g. throughout the study the experimental groups became more alike. In this regard, we are aware that the data were obtained under tentative experimental conditions, and it could be argued that there may be a discrepancy between the effects in experimental and the 'natural' situations in the school. We feel that the feedback effect merits further study.

Not surprisingly, the social and ethnic background (SES) appeared to be a good predictor of future learning achievements. However, using multilevel modelling, the effect of SES became insignificant as soon as better information on the competences and skills of the child was included into the model.

A following study (Chapter 5) described OBIS monitoring as the measure of choice to predict future learning achievements, rather than using a group characteristic i.e. the social and ethnic background of the child. This takes up an ongoing political debate in The Netherlands about the use of baseline assessment instead of social and ethnic background to allocate extra finances to schools. In connection, by measuring longitudinally, OBIS enables the estimation of value added by the school over the

years and serves as a monitoring system, the educational value of which has been described earlier in this report.

Data over three years indicate that the proportion of low-achieving Dutch working-class children remained constant, while that of the ethnic working class dropped from 56% to 44% of their total. Accordingly, the proportion of high-performing pupils in the ethnic working class increased from 14 to 23%. These results put stress on the use of socio-ethnic background as a sole measure of disadvantage. It seems that for valid and reliable decisions on pupil, school and state level, more appropriate to use learning achievements than social background.

Consequently (Chapter 6), OBIS was used to identify children at risk of future low school performance. These children should be identified in time so that they can be re-allocated to specialised schools or receive special attention in their own school. In recent years many children were referred to special education, and a new governmental guideline WSNS (Weer Samen Naar School / Going to School Together) is aimed to stop this trend. Analyses of cognitive scores over three years led to interesting observations and conclusions. Firstly, the group of children with the 16% lowest scores in a year had higher scores in the following or previous years. This indicates that the population of the lowest scoring group in a year is not constant, and children move in and out of this group. Secondly, the chance of being at risk, which is belonging to the group with the 16% lowest scores, was best predicted by the score in year 1. This latter score is measured 2 months after entry into school, and mainly represents preschool knowledge. These seemingly opposing results lead to the conclusion that the 'at risk' label of a child is not always a permanent one. The solution for the problem of risk identification lies in the estimation of the value added, the relative progress of the individual child compared with the mean progress. In this view, children with the same cognitive score in year 3 may have different risk status, depending on their scores in year 1, which shows the progress they have made.

The aim of the last chapter (Chapter 7) was to bring together the weight of evidence obtained so far that underpins the need for measuring progress starting from baseline. The central question of this conclusive chapter was how well baseline assessment – using OBIS – predicts learning progress and academic success? The second question was what classroom factors have to be taken into account for fair comparisons. The results of the longitudinal LISREL analyses show that when compared with SES and teachers' perceptions of their pupils' capacities, baseline assessment is the best predictor of future progress. Using multilevel analyses it was demonstrated that all pupils benefit from literature-based classroom practices. Cross-level interaction effects between teaching style and pupil characteristics could not be identified; differential

effects indicating that teachers adjust their instruction in favour of for instance disadvantaged pupils were not found either.

Conclusions and recommendations

Summing up, several conclusions can be drawn from this study and various recommendations can be made. Regarding the monitoring of progress it has been demonstrated that over a period of several years, baseline assessment is the best predictor of learning achievement. It is clearly demonstrated that in the course of time there is generally more progress for the more able children at entry.

Social background, teachers' perceptions and baseline assessment

During the last decade an ongoing debate on baseline assessment has taken place. The main riposte of opponents of measuring baseline at entry is that there is other information available to predict school success, such as the social background of the pupils or the teacher's perception of his or her cognitive capacities which are usually based on observations. Proponents however, argue that baseline assessment is the best approach for making fair comparisons of progress made by pupils, and of the value added by schools.

Overall, the present findings show that neither the social background nor the teacher's perception of the cognitive profile of the pupil is a better substitute for baseline assessment as a predictor of later achievement.

It also has become clear that the factor social background has several pathways of influencing learning progress. On one hand, the progress children of low social families make during their first year at school is considerable; on the other hand, over time the progress of children from a low social background is a sustaining factor hindering their progress in later years.

Our recommendation is that because of the stigmatising role of social background as a predictor of learning achievement, which leads to confusing 'messages' for teachers, information about the *actual* and *value-added* progress is exactly the information teachers need to know about their pupils (Fitz-Gibbon, 1996, 1997; Tymms, 1999).

Target setting

Based on our results, we conclude that there is no practice such as target setting in the lower years of primary school, either for individuals or for groups. This findings are in accordance with the conclusion of Van der Wel & Krooneman (2003) that schools are not familiar with the use of intermediate targets. The main conclusion of their study on the implementation of intermediate targets is that no systematic effect of target setting could be found. As we are convinced that target setting helps teachers to

adjust their curriculum for the benefit of their pupils, we recommend that target setting should be introduced on a large and systematic scale.

Teaching styles
As for the benefit of literature-based teaching styles for whatever pupils it concerns, we cannot emphasise enough that in the lower years of primary education the basis is laid for later reading skills – decoding and comprehensive reading. As it is known that good reading skills are essential for many other learning subjects and areas of life, we recommend that teachers are encouraged to provide all children with the strongest possible literacy foundations in the first years of their school career.

Children at risk: monitoring educational disadvantage versus progress
After an era of policy aimed at improving the educational opportunities of children from disadvantaged backgrounds, the question may be raised if we should still concentrate on assessment of disadvantage or that we should turn to monitoring of progress. From an economical perspective, policymakers may argue that there is no need to monitor educational progress or disadvantage of all pupils in the lower grades of primary school. However, progress can only be identified by comparison with the mean achievement. Information on learning progress over time is an indispensable instrument for early identification of stagnation or regression in learning. Thus, children at risk can only be identified by comparison with his peers over a prolonged period of time.

Chapter 1 Introduction

1.1 The contract

This is the final report of the research project 'Development on scale, instruction at measure' that was financially supported by the Netherlands' Organisation for Scientific Research. The project was registered under no. 411-210-05. According to the contract, the running time of the project was four years from1999 till 2003. The project is to be concluded by a final report and a presentation for the Programmaraad voor het Onderwijsonderzoek (PROO) committee.

This project concerned the educational impact of assessment at regular intervals on the development and school performances of young children. Over the past years, during which this research was conducted, many developments and changes regarding baseline assessment have taken place, internationally, as well as in The Netherlands itself. In The Netherlands, for example, successive governmental proposals to introduce baseline assessment in combination with a monitoring and accountability system met criticisms of different type. Without going into details, in the present situation, some politicians and professional organisations appear to be the major opponents of early testing.

Internationally, England has an advanced position; it is far ahead with legislation and implementation of testing and monitoring student achievement scores of young children. In the USA the introduction of the No Child Left Behind Act in 2002, has already had a substantial influence on educational practitioners and researchers. It elicited a lively debate on the questions whether testing helps to improve the quality of teaching, if it raises the level of student performances, and if it would be possible to give empirical proof of the progress made and the value-added.

This introduction describes the aim of the project and its background, followed by the theoretical framework and research questions, methods and design of the research, as is stated in the research proposal granted in 1999. Naturally, the recent developments have also had their effects on the concepts and approaches in this study. Subsequently, some related issues that are now widely used, as a result of the recent debate on assessment and monitoring, are briefly explained.

1.2 Aim and background

The aim of the project was to measure the effects of regular testing, and the use of explicit achievement goals on the cognitive development and school performance of young pupils (age 4-6). This may answer the question whether the level of development, at a certain time, is an indicator of future school performance, and to what extent the progress in development can be stimulated. Essentially, the research question was derived from the following educational model:

Assessment initial level -> diagnosis -> (steering) learning process -> assessment progress.

Freedom of education

To date, in Dutch primary schools there is no formal assessment of the learning capacity or potential of the young pupils starting school. Nationwide, the performances of pupils are only measured in Year 8, at the end of the 6-year curriculum, using a standard test indicating which type of school a child could attend after primary school. Clearly defined content standards of what is to be taught, and what kind of performance is expected, are lacking. And, the same applies to target setting. Controlling the curriculum, target-setting and testing of young children can only be introduced in the schools on a voluntary basis.

Various factors, historical and political as well, have worked as a hindrance to the introduction of monitoring and testing the progress of pupils in primary education. One of these is the relative freedom of education in The Netherlands. General attainment targets have been formulated indicating the basic minimum that schools are required by law to teach their pupils in each area of the curriculum. Schools have a considerable freedom to decide how they teach and attain these minimum requirements. They are also free to choose whether they award marks or indicate children's level of achievement in some other way, for example by describing their progress and results in words rather than in numbers. As a consequence, schools may have different priorities and topics which they consider as important for the pupils, and the measures of performance achievement and information for parents may vary considerably from school to school (Ministry of Education, Culture & Sciences, 1999). This freedom is guaranteed under Article 23 of the Constitution. It is only the Quality Act (1998) that prescribes that schools themselves are responsible for the quality of education they provide, and for pursuing a policy that ensures the improvement thereof (Hendriks, Doolaard & Bosker, 2002).

Yet, an increasing number of schools recognises the importance of accurately measuring, recording and assessing the progress of individual pupils, and the

comparison of the effectiveness of their school with other schools. In this context, there is a growing need for evidence-based education. In practice, teachers have a long tradition in a more intuitive type of evidence-based decision-making by observing and trying out what works. Still, for systematically evidence-based instruction valid and reliable tests are needed.

Performance indicators in primary schools

Probably as a result of the relative autonomy of schools, there is a lack of adequate instruments designed for baseline assessment in the context of classrooms.

Existing tests are usually meant for diagnostic purposes in case of special needs, and are not suitable for measuring skills and progress by teachers themselves. For our study we selected the PIPS Reception Assessment[1] or PIPS Baseline (BLA) developed by Peter Tymms and his colleagues in Durham (UK). PIPS was translated into Dutch and adapted for use in The Netherlands. PIPS stands for Performance Indicators in Primary Schools. The main reasons for using this particular test were:

- Its content and construction, including practical implications: the PIPS can be done in a short time, and is of a child-friendly character.
- There were methodological considerations and criteria: validity, reliability, correlation and prediction were determined to be very high for PIPS.
- Applications in education: schools are provided with feedback, presenting high quality information, that helps to diagnose problems.

The PIPS Reception Assessment test is constructed to assess cognitive development of young children, that is to say, their literacy and numeracy skills. In former days these skills were called early reading and early mathematics skills. The Reception Assessment provides a baseline for monitoring the progress of development and learning during the first three years at school. By comparing the scores on attainment levels collected at three different time points, for example – between starting school at age 4 and three years later – the progress that children make during this period can be measured.

1 The PIPS project is one of a suite of school monitoring systems developed by the Curriculum Evaluation and Management (CEM) Centre which is based at the University of Durham in the UK.

1.3 Theoretical framework and research questions

Research suggests that by giving teachers feedback, pupils' performances can improve. This is one of the aims of all monitoring systems – good feedback helps the teacher to work better. But we do not know how big this effect. To address this question, we set out to compare a situation in which no feedback is provided to the teacher, with a situation where feedback is provided to the teacher in combination with individual target setting. To optimize the learning process teachers need:

- understanding of subject-matter and didactic knowledge;
- explicit targets for individual pupils and on group level;
- feedback of the results, that is knowing their pupils' progress.

Development and instruction

Vygotsky's concept of the zone of proximal development (ZPD) offers a theoretical framework for studying relations between cognitive development and educational intervention. As defined by Vygotski (1978), the child's zone of proximal development is "the distance between the actual developmental level as determined by independent problem solving, and the level as determined under adult guidance or in collaboration with more capable peers."

According to Schneuwly (1994), the concept of the zone of proximal development is Vygotsky's response to a difficult problem, namely, the tension between the internal self-propulsing mechanisms of development and the impact of external cultural forces, as embodied in instruction. ZPD is a relational concept that describes the social interactions that allow progress and it is not measurable in terms of quantifiable information, for example the size of its upper limits. Stated otherwise, ZPD makes the interactions in the teaching-learning process from which a child can profit, understandable. The concept of ZPD also links assessment and instruction as a tool for regulating learning and fostering development, especially in school settings. Vygotsky's concept on cognitive development is still a major basis for research and didactic programs as well as on instruction and target setting. For example, SpelenderWijs, an instructional preschool program aimed to stimulate language development of toddlers in a disadvantaged situation, is based on Vygotski's theory (De Jong-Heeringa, 2003).

Teacher expectations

In addition, we know that teachers' expectations, basically setting of the aspiration levels for a child, influence performance and progress of the child. In a previous, longitudinal study it was shown that in the Years 7 and 8 teachers' expectations influence learning performances (Van der Hoeven-van Doornum, 1990; Van der Hoeven-van Doornum, Voeten & Jungbluth, 1993). The aspiration levels that

teachers define for their pupils were mainly based on perceived learning performance and social background of the pupils. Teachers' perceptions of the social background and the learning ability of their pupils, as well as anticipated achievement levels, were shown to be associated with the socio-economic background of the pupils. We found substantial indirect effects of social background via teacher expectations, on achievements and the advise for a particular type of secondary education. The sizes of the effects were 20% and 40% respectively of the total effects. It is likely that a similar mechanism of setting implicit educational goals, applies to younger children as well.

Research questions

To answer the main research question on the influence of assessment on learning progress, the following specific questions were investigated:

- How well does the OBIS, the Dutch version of PIPS-BLA, predict learning performances of Dutch 4-year old children?
- What are the effects of test results (diagnosis) and explicit achievement goals on the teacher's actions, and on the progress of the pupils?
- What effects have the implicit educational goals of the teachers (teachers' expectations), which arise from personal observations and expectations, on the progress and performances of pupils?

1.4 Methods and design

The investigation took four years. The first year was used for instrument development, in particular the translation and conversion of the PIPS Baseline Assessment (BLA) into Dutch. The Dutch version of the PIPS is called OBIS, the acronym for Onderbouw Informatiesysteem (Lower School Information System).

In the second and later years a quasi-experimental investigation was carried out with approximately 500 children in 26 groups coming from 11 schools. Measurements at the pupil level were the OBIS test scores, target-setting, and implicit goals set by the teacher. Measurements at the teacher level aimed to record the way teachers stimulate their pupils in the acquisition of knowledge and skills during play and learning activities. The schools were randomly distributed among 2 experimental groups (E1 and E2) and one control group (C3). The factorial design is shown in Scheme 1.1.

Scheme 1.1 – Design of the study with two experimental groups and a control group

Year 1				Year 2				Year 3	Condition
T1	ET1	Ex1	LP1	T2	ET2	Ex2	LP2	T3	
+	+	+	+	+	+	+	+	+	E1
+	(+)	(+)	+	+	(+)	(+)	+	+	E2
+	(+)	(+)	(+)	+	(+)	(+)	(+)	+	C3

(+) no feedback on OBIS test scores given to the teachers

Measurements at the pupil level were:
- T1-T2, test scores OBIS, Year 1-2;
- T3, test scores reading and mathematics, Year 3;
- ET1 and ET2, explicit targets, Year 1-2;
- Ex1 and Ex2, teacher's expectations per pupil, Year 1-2;
- LP1 and LP2, teaching processes, Year1-2, questionnaire, logbook or observations in classroom Year 1-2, to evaluate the teachers' performances to activate their pupils in the acquisition of knowledge and skills.

Measurements at class level were:
- Questionnaire for school staff, comparable with those used in earlier school effectiveness studies.

Feedback
The experimental group E1 got all available information and feedback on pupil level and class level that stimulates teaching and learning processes, including:
- test scores on the OBIS;
- expectations, compared with expectations for children with similar OBIS scores;
- individual targets;
- support of the teaching and learning process.

The experimental group E2 got information on the individual test-scores and information on target-setting on class level.

The control group C3 received only feedback on test scores. The test scores collected in the first year provide information about what children know and can do, when they start school. It is a fixed point from which the progress of the experimental and control groups can be assessed. The 'value-added' approach enables us to assess the progress of children in various educational settings.

1.5 Related issues

In the context of baseline assessment, the issues 'value-added' and 'monitoring progress' have become of increasing interest in educational research, policies, and school practice as well. The definitions and descriptions of these issues here summarised, draw heavily upon of Fitz-Gibbon (1997), Tymms (1999a, 1999b), Visscher & Coe (2002), Linn (2000), Raudenbush & Willms, (1995).

1.5.1 Value-added

School mean achievement scores or school mean progress

There are many ways in which schools and teachers add value to the lives of their pupils. In this research we are particularly interested in the progress made by pupils in learning performance in addition to their development. The assessment of student performances and the measurement of school effectiveness are neither simple nor straightforward (Linn, 2000; Stevens, 2000). It is important to be clear that evaluation of school performance differs depending on how data are modelled and how analytical methods are used. In the context of school quality, there are many misconceptions about value-added.

One issue of substantial importance is whether school mean achievement or school mean progress scores are examined. The problem is how to separate properly the effects of schooling from the intake characteristics of the pupils attending the school. Usually pupils are not randomly assigned to schools. The socio-economic background of the pupils is a well-known malefactor of the process that 'sorts' students into schools. It causes an unequal distribution of student characteristics that makes schools more or less advantaged when school effectiveness is at stake. Particularly, if school mean achievement scores are used as an indicator of school performance, then schools with disadvantaged intakes are at risk of unfavourable evaluation if there is no proper account for student background. Schools with low mean scores are not always poor performing schools. On the contrary, the use of students' progress scores to evaluate school performance enables schools that would otherwise be labelled as low performing, to demonstrate positive effects on student achievement.

Value-added as relative progress

Fitz-Gibbon (1997) and Tymms (1999a) define a value-added score for a pupil as relative progress. It is a measure of the progress made by that pupil relative to the progress made by other pupils. Starting with a baseline, measures of relative progress can be derived by computing a trend line showing how pupils have achieved, for example, over a period of one year. For each pupil a score can be calculated that

predicts the most likely score for pupils with the same starting point. The difference between the predicted and the actual score of a pupil is the value-added score for that pupil: the measure of that pupils' progress relative to that of similar others. Describing value-added as relative progress, Fitz-Gibbon (1997, par. 1.3) stated: "The statistical term is 'residual' denoting that [the variation] which is left over after the prior achievement has been taken into account. If a pupil makes more than average progress, the 'value-added' is positive, with higher achievement than predicted. If a pupil makes less than average progress, the value-added score will be negative, with lower achievement than predicted. Average progress is indicated by there being no difference between the statistically predicted and the actual score.

Value-added on school level is the average across all pupils in a school of the difference between their individual test results and their expected results, based on regression analysis. A school with an average value-added score of zero has kept up with other schools; on average progress is appropriate progress." With a typical correlation of 0.7 between prior achievement and a later test, the value-added score is the fairest indicator for academic success, so far. It is a simple and clear measure that can be regularly monitored.

Longitudinal designs

Another matter of interest is that, if the effects of teachers and schools are to be validly estimated, the use of longitudinal designs and analyses is needed (Goldstein, 1991; Bryk & Raudenbush, 1988; Linn & Haug, 2002). School performance cannot be estimated without bias when student test scores are aggregated at a single point in ime, nor with precision when successive cohorts of students are used. Estimates of year-to-year gains are affected in large part by sampling variation, measurement error and unique factors that are not associated with school practice. The only fair way to compare schools is on the basis of how much progress pupils make during their time at school. Longitudinal, multilevel analyses are highly recommended for research purposes.

1.5.2 Monitoring systems

There are many complex issues involved in the development and implementation of accountability systems that are not acknowledged or considered in public and political debates (Zvoch & Stevens, 2004). The most important distinction has to be made between official accountability systems and professional monitoring systems. An official accountability system is used with the express purpose of holding schools and/or teachers to account. On the other hand, a professional monitoring system is appropriate in order to generate data for the professionals to run a better system.

An example of an accountability system may be seen in England where a National Curriculum was introduced so that it would be clear what the teachers were supposed to do and then pupils are tested at the end of what were called Key Stages so that their progress in the curriculum can be judged. The data from statutory assessments given at the end of primary school and the end of secondary school are released to newspapers and school league tables appear on a yearly basis. The results of these inspections can have major consequences for the life of a school and the professional career of school-heads and teachers (see for example Fitz-Gibbon, 1998; Shaw, Newton, Aitkin & Darnell, 2003).

Professional monitoring systems are designed to provide information to schools to help themselves. In other words – to see which pupil is falling behind, which pupil is progressing rapidly, which class has difficulty with maths and so on. A very important element of this is the notion of value-added, that is to say, the progress of one pupil compared with other similar pupils in other schools. By looking at pupils with similar starting points, fair comparisons are made and like is compared with like. The technique employed is regression analysis and the value-added data are residuals from regression analysis, as described above.
A modern, professional monitoring system enables teachers to spot children who are falling behind and watch over those who have special needs or consistently extreme levels of attainment. Once this becomes integrated with professional knowledge a powerful pointer for educational action has been created (Tymms & Wilde, 2003). According to progress that is made, the teacher is far more important, for example, than the home background. The two key variables in predicting a child's academic standing at the end of one year at school are the child's starting point and who the teacher is (Tymms, Merrell, & Henderson, 2000).

Although the accountability and professional models can be presented as extremes with opposite purposes, they are, in fact, at two ends of a continuum and one would expect the data in accountability systems to be used to help pupils in some circumstances, and it is appreciated that the kind of data developed may be used to monitor teachers as well. It is nevertheless useful to make the distinction, and also to think about what happens if problems occur in the school. In an accountability system the school or the teacher may wish to hide the problem and disguise what is going on. On the other hand, if the school takes part in a professional monitoring system, for which they have paid to join, the aim is to detect problems in an early phase These are very different mindsets that have very different implications for the way the systems run, and also for the quality of the data that are developed as time goes on.

Well-constructed and well-used monitoring systems are needed, although the evidence for the positive impact of such systems has yet to be established. Nevertheless,

there is little doubt that well-constructed, and well-used monitoring systems can be of enormous benefit to the educational provision of children (Visscher & Coe, 2002).

1.6 References

Bryk, A.S., & Raudenbush, S.W. (1988). Towards a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. In R.D. Bock (Ed.), *Multilevel analysis of educational data* (pp.159-204). San Diego: CA: Academic Press.

Fitz-Gibbon, C. (1997). *The Value-Added National Project: Feasibility Studies for a National System of Value Added Indicators (Final report).* London: School Curriculum and Assessment Authority.

Fitz-Gibbon, C. (1998). Ofsted: time to go? *Managing Schools Today 7*(6): 22-25.

Goldstein, H. (1991). Better ways to compare schools? *Journal of Educational Statistics, 16 (2),* 89-92.

Hendriks, M.A., Doolaard, S., & Bosker, R.J. (2002). Using School Effectiveness as a Knowledge Base for Self-evaluation in Dutch Schools: the ZEBO-project. In: *School Improvement through Performance Feedback*. Eds. A.J.Visscher & R.Coe. Lisse: Swets & Zeitlinger.

Jong-Heeringa, J.L. de (2003). Het programma SpelenderWijs: Implementatie en analyse in de gemeente Huizen. GION: Groningen.

Linn, R.L. (2000). Assessments and accountability. *Educational Researcher, 29*, 4-16.

Linn, R.l., & Haug, C. (2002). Stability of schoolbuilding accountability scores and gains. *Educational Evaluation and Policy Analysis, 24 (1)*, 29-36.

Ministerie van Onderwijs, Cultuur & Wetenschappen, 1999. *Variëteit en waarborg: voorstellen in de ontwikkeling van het toezicht op het onderwijs*. [Diversity and a guarantee. Proposals for the development of the supervision of education.] Zoetermeer; Ministerie van Onderwijs, Cultuur en Wetenschappen.

No Child Left Behind Act of 2001 (2002). Publ. L. No. 107-110.

Raudenbush, S.W., & Willms, J.D. (1995). The estimation of school effects. Journal of Educational and Behavioral Statisctics, *20*, 307-335.

Schneuwly, B. (1994). Contradiction and development: Vygostky and Paedology. European *Journal of Psychology of Education, 9* (4), 281-291.

Shaw, I., Newton, D. P., Aitkin, M., & Darnell, R. (2003). Do OFSTED Inspections of Secondary Schools Make a Difference to GCSE Results? *British Educational Research Journal, 29* (1): 63-75.

Stevens, J.J. (2000). *Educational accountability systems: Issues and recommendations for New Mexico* (Technical Report). New Mexico State Department of Education.

Tymms, P. (1999a). *Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-added Indicators*. London: David Fulton Publishers.

Tymms, P. (1999b). Baseline assessment, value added and the prediction of reading. *Journal of Research in Reading, 22* (1), 27-36.

Tymms, P., Merrell, C., & Henderson, B. (2000). Baseline assessment and Progress during the First Three Years at school. *Educational Research and Evaluation, 6* (2), 105-129.

Tymms, P., & Wylde, M. (2003). *Baseline assessment and monitoring in primary schools*. Paper presented at the Symposium Connectable Processes in Elementary and Primary section. Bamberg, DL, April 2003.

Van der Hoeven-van Doornum, A.A. (1990). *Effecten van leerlingbeelden en streefniveaus op schoolloopbanen*. [Effects of teacher expectations and aspiration levels on school careers; Dissertation with a summary in English]. Nijmegen: ITS.

Van der Hoeven-van Doornum, A.A., Voeten, J.M.J., & Jungbluth, P. (1993). The influence of parents' expectations on school careers. *Tijdschrift voor Onderwijsresearch, 18* (6), 369-379.

Visscher, A. J., & R. Coe, Eds. (2002). *School Improvement Through Performance Feedback.* Contexts of Learning: Classrooms, Schools and Society. Lisse/Abingdon/Exton PA/Tokyo, Swetz & Zeitlinger.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Blackwell.

Zvoch, K., & Stevens, J.J. (2003). A multilevel longitudinal analysis of middle school math and language achievement. *Education Policy Analysis Archives, 11* (20). Retrieved from http"//epaa.asu.edu/epaa/v11n20.

Chapter 2 Methods

The aim of this chapter is to describe the instruments used in the study. The first year of the study was used to convert the PIPS BLA into the OBIS. In section 2.1 the content and structure of the OBIS will be described, with particular attention paid to validity and reliability of the test. Section 2.2 focuses on the features of the V-OBIS (i.e. Vervolg Onderbouwinformatiesysteem, the continuation of OBIS in Year 3), the instrument measuring later academic achievement in grade 3. Subsequently, section 2.3 deals with the other instruments, mainly questionnaires for the teachers. This chapter concentrates on the content of the instruments while the next Chapter 3 addresses the empirical features of the instruments.

2.1 OBIS

2.1.1 Validity and reliability

How well does the OBIS predict the learning performances of Dutch 4-year old children?

In view of the research questions we needed a test that would not only assess the actual cognitive developmental level, but whose content and measures can also be related to target setting, learning instruction and future performances. The conversion of the PIPS into the OBIS started with the inspection of the content and pictures of the test, while taking into account differences in language, culture, developmental level and typical socio-ethnic situations, as precisely as we could. It is obviously important to avoid disparities in the test, and consequently putting children with low socio-economic status, minority groups, and boys or girls at a disadvantage. The aim of testing is to search for differences between well performing and less weakly performing pupils. If the vocabulary scores of girls appear to be higher than boys', this may be because the girls performed better, but it is also possible that the higher scores were the result of an artefact, that is the way the skill was measured. For instance, the pictures presented to the pupils may have been more appealing to girls than to boys (Uiterwijk, 1994). The culture fairness of the baseline assessment was one of our major concerns. This inspection of the content and pictures prompted us with the following questions and considerations:

- What lies behind the assessment content and structure?
- Are there essential differences between the English and the Dutch educational context of the assessment?
- To what extent is the assessment rooted in a particular cultural context?

Usually, the quality of tests is judged by the theoretical basis, reliability, validity, norms, and practical issues. The conversion of the PIPS BLA into the OBIS constituted the first phase of the validation process, and a prerequisite for answering the first research question how well OBIS predicts the learning performances of Dutch 4-year old children.

Validity and reliability are considered as the main components of instrument development. Test validity focuses on what a test measures, and how well this is done. Reliability is commonly defined as the accuracy with which a test measures whatever it does measure (International Encyclopedia of Statistics, 1978). General acknowledged norms for the validity and reliability of tests are described in Standards for Educational and Psychological Testing (AERA, APA & NCME, 1985). In this investigation we used classical elaborations of those standards by Cronbach (1990), Messick, (1989) and Linn, Baker & Dunbar (1991). Validating is an iterative process, and therefore the order of the various components of it, namely content, construct and criterion validity, may not be strictly chronological; and the components are also strongly interrelated:

Content: Examining the indicators, e.g. items or questions in comparison to the content domain; in particular the relevance and representativeness, the quality, meaning and selection of the items.

Construct: Examining which processes are underlying the items, questions, and tasks. What are the cognitive domains: knowing facts and procedures; using concepts, problem solving or reasoning?

Criterion: Examining the instrument for practical and economical criteria. Correlating it with other, similar tests regarding external structure, similarities and differences in time, grouping and circumstances, procedures. Assessing also costs and efficiency of the instrument, social consequences and fairness.

2.1.2 Content and construct

What lies behind the assessed content and structure?

For understandable reasons, the validity of the content and structure of the OBIS correspond with those of the original PIPS test. Therefore we concentrate here on the features of the PIPS. The PIPS-test measures the cognitive level resulting from the developmental level and earlier learning. The assessment gives information about what children know and can do when they start school. When a child is tested for the first time at the age of four, this represents mainly pre-school learning.

The content of the PIPS-test matches the educational program in the first two years of school education. The On-Entry Baseline Assessment is not restricted to any curriculum objectives in particular, and it allows progress to be assessed in many different contexts.
The literacy section of the test contains writing, vocabulary, ideas about reading, letter identification and word identification. Children know a lot before they begin formal reading, and these skills are considered to be good predictors of later reading achievement (Whitehurst & Lonigan, 2001). The test starts with asking the child to write his or her own name. Even a child that scribbles has an idea of language. Picture vocabulary is generally considered as the kernel of language and language acquisition (Clark, 1993). The phonic section gives an indication of phonological awareness. It is based on repeats and rhymes. Judging by a vast amount of research, phonological awareness appears to be a good predictor of early literacy and later reading performances (Wagner & Torgesen, 1987; Bus & IJzendoorn, 1999; Aarnoutse, 2001; Schneider, 2001). Children with reading problems often have difficulty to recognise phonemes in words.

There is extensive literature upon which to draw particularly for progress in reading, but to a lesser extent for progress in mathematics (Tymms & Merell, 2004). The mathematic section consists of ideas about maths, counting, sums and digit identification. This section assesses early numeracy: the use of concepts such as more or less, greater or smaller.
The origin of the content is not theoretical in any formal sense; all of these measures have been documented in research literature on one hand as valid and reliable indicators of underlying developmental processes, and on the other as being good predictors of later reading (Blatchford, Burke, Farquhar, Plewis, Tizard, 1987; Bryant, Maclean, Bradley and Crossland, 1990; Riley, 1994; Stuart, 1995; Tymms and Williams, 1996) and mathematics achievements (Shaeffer, Eggleston & Scott, 1974; Saxe 1979; Vacc, Vacc & Fogelman, 1987; Bryant et all, 1990; Kilgallon and Mueller, 1996). The PIPS-test is based on the idea of ‘measuring what matters’, it is designed to be used in a broad range of cultures and languages. The test is curriculum oriented to such an extent that there is a correspondence between test and curriculum and objectives that has to be met. The test has been developed in conjunction with teachers over a period of several years. The content is annually reviewed for validity and reliability. All its items strongly correlate with future literacy and numeracy.

Figure 2.1 – OBIS sections: handwriting, vocabulary, ideas of reading, repeats and rhyming

Considering construct validity, the test is based on a number of tasks of increasing difficulty and complexity representing different levels of literacy and numeracy. For example, there are three pictures to assess a child's vocabulary, each with progressively more difficult items until it becomes too difficult for almost all children at the start of school with items such as saxophone and microscope (Tymms & Merrell, 2004). It should be clear that the PIPS is not an intelligence-test but an assessment that refers to future learning. For instance, it does not measure logical reasoning or spatial intelligence. In the past it was shown that traditional IQ-tests do not predict future learning achievements of young children very well. In particular, they tend to underestimate the cognitive abilities of children of disadvantaged socio-economic and socio-cultural backgrounds (Tellegen & Laros, 1993).

2.1.3 Criterion

What evidence is there to support the claims made for the assessment?

The data gathered with the PIPS Baseline Assessment are converted into standard scores. The scores indicate a child's present level of attainment in two ways; depending on how many times a child is tested:

- Concurrent value added, when a child is tested for the first time. This is the child's 'context' score; it is a measure of developed ability. In effect it gives a snapshot of where the child is at a certain moment in time. The concurrent value-added tells if a child is doing better or worse compared with the average child. The concurrent value added is the same score as the so-called 'norm-referenced' score that places the individual with respect to a group mean.
- Prior value-added, when a child is tested twice or more times over a period of several years. The attainment on the prior test(s) is used as a predictor of the present scores. This score provides a measure of the child's progress in time. Each pupil is compared with pupils with similar scores on the pre-test.

In terms of claims, the correlation between the PIPS BLA assessed at the age of four and the reading and mathematics scores of seven-year-olds is around .70. This means that value-added measures are possible for schools, while individual predictions are expressed as chances of gaining a certain level in the next years. The chances are calculated for each child using the PIPS score as a predictor. The overall chances for all of the children of the same year group in the school are also calculated. The chances are calculated using a statistical model based on data collected in previous years (Tymms, 1999, Tymms, 2000). The sections 'Letter Identification' and 'Digit Identification' are the best predictors. As it is, both concurrent and prior value-added measures provide little information about how children got to a certain level.

Data of three years in a row show that there is a considerable similarity between the PIPS correlations and OBIS correlations. The PIPS On-Entry Baseline correlates strongly with later achievement and the OBIS does so as well.

Table 2.1 – Correlations between baseline assessment and later achievement scores

Start of Reception	–	End of Reception (20,000 children)	r= 0.78
Start of Reception	–	PIPS reading and maths, year 1, age 7	r= 0.70
Start of Reception	–	End year 2 (End Key Stage 1), age 8	r= 0.60
OBIS 2000, Year 1	–	OBIS 2001, Year 2 (450 children)	r= 0.78
OBIS 2000, Year 1	–	V-OBIS reading and maths, Year 3, age 7	r= 0.68

Norms are checked on annual basis because, due to various causes such as changes in pre-school education, a mean rise in the average score requires regular updating of norms (Jones, 2003). Similarly, for western countries a mean rise in intelligence score of 2 to 3 points per decade can be expected (Lynn & Hampson, 1986).

Other criteria

Other important reasons to select the PIPS for our research were its high reliability and the short length of the test. The crucial measure of reliability in a baseline assessment must be the extent to which independent assessors arrive at the same result. The reliability of the PIPS assessments is determined each year. The most recent figures, given below, show the test-retest reliability (correlation coefficient).

Table 2.2 – Test-retest reliability for PIPS and OBIS

	Maths	Reading	Total
On-Entry Baseline (text)	0.88	0.92	0.95
On-Entry Baseline (CD)	0.91	0.93	0.93
On-Entry Baseline Follow up (text)	0.89	0.93	0.94
On-Entry Baseline Follow up (CD)	0.90	0.80	0.85
OBIS (CD)	0.92	0.97	0.98

In 2003 the test-retest reliability of the OBIS was measured on a random sample of children that was firstly assessed by their own teacher and approximately four weeks later, by a researcher. Thirty pupils in six different groups were re-assessed using the CD version of the OBIS. The correlations between the test and retest were as follows: maths 0.92, reading 0.97 and total 0.98. Based on these results both, the PIPS and OBIS may be considered to be reliable instruments for baseline assessment and prediction of later achievement.

Procedures

The test uses a combination of objective assessment and teacher ratings to provide information about children as they first enter school. A textbook or a multimedia computer program is used to assess the early reading, early mathematics, phonological awareness and short-term memory of children as soon as they have settled into school. This is carried out by an adult working with each child on an individual basis. The test takes about 20 minutes for each child. Using scoring rules, the test is adaptive so that children are not presented with questions that are beyond their ability. Teachers rate the personal, social and emotional development of pupils from observations over the first few weeks of their starting school. Clear instructions are given for the assessment of children whose first language is not English.

Usually the assessment is repeated about a year later in Year 2, giving value-added over the first year in primary school. Using the CD, the assessment moves on from the point the child reached at the assessment, so that questions are not repeated unnecessarily. Once again, the assessment takes around 20 minutes per child. Children's attitude to school and their behaviour are also assessed. To assess attitude, children are asked to rate how they feel about a number of different statements, related to school life, on a three point scale of faces ranging from happy to sad. Teachers assess the behaviour of pupils using a scale incorporated into the computer program.

Feedback

Schools receive two different sets of feedback, one following the Baseline Assessment in Year 1 and one a year later following the second assessment administered in Year 2. The feedback is intended to affirm teachers' professional judgements over their pupils. All feedback includes raw and standardised scores. These scores are based on a larger sample, characterised by a normal distribution. Because of this, standardised scores can be compared over time and between groups and individuals with confidence. Baseline feedback contains coloured bar charts, result tables and box-and-whisker plots. An example of the result table for the start of Year 1 is given in Fig. 2.2.

Figure 2.2 – Results table Year 1. Raw and standardised scores

	Toetsscores				Standaardscores			
Naam	rekenen	taal	nazeggen	**totaal**	rekenen	taal	nazeggen	**totaal**
Liesje	38	38	17	93	58	56	59	58
Ina	24	35	17	76	45	53	59	50
Danny	27	32	16	75	48	49	52	49
Toon	37	36	17	90	57	54	59	57
Imkei	29	32	15	76	50	49	48	50
Rosa	32	32	15	79	53	49	48	51
Michael	28	37	9	74	50	55	onder 40	49
Otto	46	34	17	97	66	52	59	60

The Year 2 feedback contains results in tables, as well as in line graphs and scatter grams. Line graphs enable teachers to see how pupils have progressed from their first assessment to their second one. Line graphs are also important because they contain that year's sample average against which pupils can be compared. Scatter grams with a line of best fit and 95% confidence intervals allow teachers to see whether pupils have progressed as expected given their performance on the Baseline Assessment. See Fig. 2.3 for an example of the results table in Year 2.

Figure 2.3 – Results table Year 2. Raw and standardised scores, value added and attitudes

	rekenscores			**taalscores**			**totaalscores**				**toegevoegde waarde**		**houding**
Naam	4-5	5-6 jaar		4-5	5-6 jaar		4-5 jaar		5-6 jaar		**rekenen**	**taal**	
	ruw	ruw	std.	ruw	ruw	std.	ruw	std.	ruw	std.			
Maple	17	16	37	4	13	29	22	35	33	29	gemiddeld	– –	··☺
Amine	15	5	28	16	25	41	38	43	33	29	– –	gemiddeld	··☺
Mirna	7	7	29	3	12	27	14	30	35	30	–	–	··☺
Erna	6	16	37	5	13	29	14	30	38	32	gemiddeld	–	··☺
Adam	10	18	39	9	27	43	25	37	56	40	gemiddeld	gemiddeld	·
Ali	11	18	39	10	24	40	28	39	58	41	gemiddeld	gemiddeld	··☺
Rashed	10	23	45	11	28	44	23	36	60	42	+	gemiddeld	··☺
Misja	23	22	44	25	29	45	55	53	61	42	–	–	··☺
Sabrina	17	23	45	10	28	44	32	40	62	43	gemiddeld	gemiddeld	··☺
Sil	6	19	41	6	32	49	17	33	65	45	gemiddeld	++	··☺

Practical considerations

The manner in which the assessment was constructed has already been outlined but there were other basic principles accounted for: that testing is something that children enjoy doing, that teachers see it as valuable and that it should involve as little work and time as possible. Another important feature of the test is that it can be conducted by teachers themselves as a natural part of the daily routine in the school life of young children. The experience of teachers in the UK is that the test situation doesn't distress the children. Teachers often comment that spending 20 minutes with each child helps to build a good relationship and that it is not just the child's reaction to assessment items that matters but the way in which they respond that gives valuable information (Tymms & Merrell, 2004).The Dutch teachers report similar experiences with the OBIS as is shown here below in the comment from a new user.

Figure 2.4 – Experiences of an user of OBIS

All children were very enthusiastic.
To have the teacher all for you is obviously very special.
It was great fun to take the test. In the beginning it took some more time, but after having tested 5 children I had gained skill, and could do it quicker.
The children liked 'Meaning of words' very much. I wasn't surprised by their reactions. I already had a fair image of the possibilities/knowledge of words/language of the children.
Rhyming appeared to be very difficult for some of the children; they didn't understand what to do.
During the week that followed the test, I have been doing this activity (rhyming) also with the youngest children. Before I did the rhyming games only with the eldest children in my group.
Knowledge of the letters was very exciting for me. At the introductory meeting on the test I was rather sceptical of this subject. My experience with the test made me change my opinion. Children reacted very differently. One started reading aloud all names of classmates from the name book. Another picked out the initials of his own name and those of his brothers and sisters. One reacted by saying that these were "adults letters that I don't read. I can only read picture letters." Very few children really knew how to name the letters.
Doing the topic numbers, I found out that one child recognised and could name all numbers up to 25. That didn't occur to me before. Although I knew he was skilful, this was new for me.
The test had clearly additional value for me. I'd like you to know that our whole team of teachers was enthusiastic about the quality of the test materials. The test book looks smart and inviting.

2.1.4 The educational system in the United Kingdom and in The Netherlands

Levels and age groups in the UK

Children start school at the age of four in England, and the years of compulsory schooling are from 5 to 16 years. Under the National Curriculum, compulsory schooling has been divided into four Key Stages. Key stage 1 contains year 1 and year 2 of Primary Education. Most children aged four/five to eleven are taught in mixed-ability classes with children of the same age, with one teacher in charge of the class. A teacher normally stays with a class for one school year, and at the end of each school year, pupils normally progress to the next class. However, many primary schools, particularly small schools, have one or more mixed-age classes, in which case some pupils might stay in the same class for more than one year.

Scheme 2.1 – Overview of grouping in primary education in the UK and The Netherlands

Age	Up to 5	4 – 5	5 – 6	6 – 7	7 – 11
UK	Nursery / play group	Reception	Year 1	Year 2	Year 3 to 6
Age	Up to 4	4 – 5	5 – 6	6 – 7	7 – 12
NL	Nursery / play group	Year 1	Year 2	Year 3	Year 4 to 8

Dutch educational system: age and grouping

In the Dutch system primary education begins at age 4 as well, and it is also compulsory from the age of five. Primary education in The Netherlands is intended for all children aged four to approximately twelve years old. In most primary schools the pupils are grouped by age. There are eight year groups or grades[1] in all and each 4-year old child begins in Year 1 and, in most cases, goes up to a higher class each year until they reach the top class. Years 1 to 4 (4 to 8-year-olds) are known jointly as the juniors and Years 5 to 8 (9 to 12-year-olds) as the seniors. Alternatively, the school may be divided into junior, middle and senior sections (years 1 to 3, 4 to 6 and 7 and 8 respectively).

In both countries, the UK and The Netherlands, there has been considerable debate about the kind of assessment appropriate for young children. Some argue that observation is the way forward; while others point out that the use of objective data is essential. The two countries also include large numbers of children starting school whose first language is not the dominant language of the host country.

1 Grade: a class organized for the work of a particular year of a school course.

2.1.5 On-entry baseline assessment across cultures

Cultural influences are likely to be greater in some parts of the assessment than in others. The development in early mathematics will be fairly consistent across different cultures (Tymms & Merrell, 2004). Although there is evidence for a universal starting point of newborn infants in arithmetic (Wynn, 1992), it is possible that different cultures induce different developmental pathways as the child grows up.

The adaptation of the mathematics section consisted of simple direct translations of the English items into Dutch. Using a Rasch model (see e.g. Bond & Fox, 2001) the difficulty levels of the items were estimated. The correlation between the English and the Dutch maths items was .90. Fig. 2.5 depicts the difficulty estimates of the math items for English and Dutch children.

Although the Dutch translation followed the English text as close as possible in terms of nature and difficulty of the questions, it seems clear that the differences between the two tests are more pronounced for language items than for the mathematics section. Therefore also lower correlations might be expected for the reading section. Still, the correlations between the difficulties of Dutch and English vocabulary and letter identification were .80 and .90, respectively. Figures 2.6 and 2.7 show the difficulty estimates of the vocabulary items and the section letter identification and words for English and Dutch children.

Figure 2.5 – Difficulty estimates of maths items for English and Dutch children

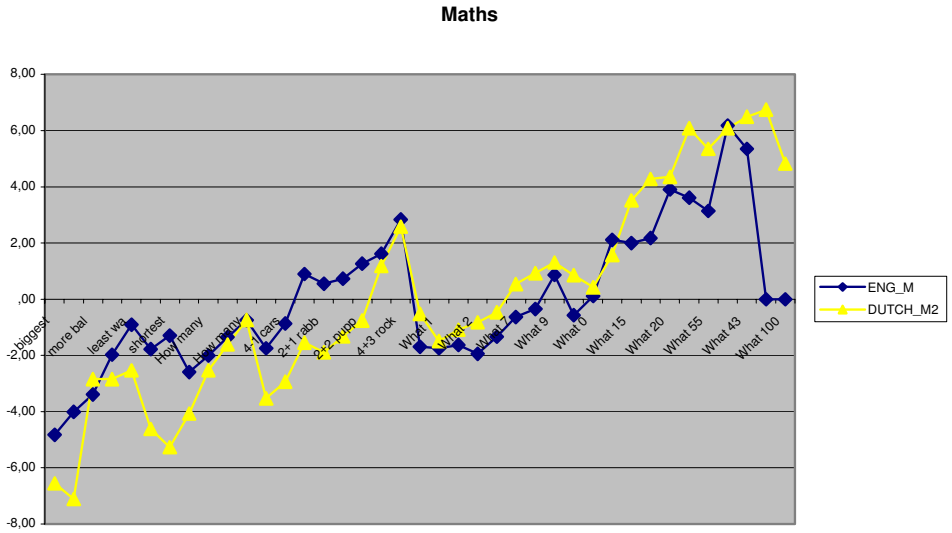


Figure 2.6 – Difficulty estimates of vocabulary items for English and Dutch children

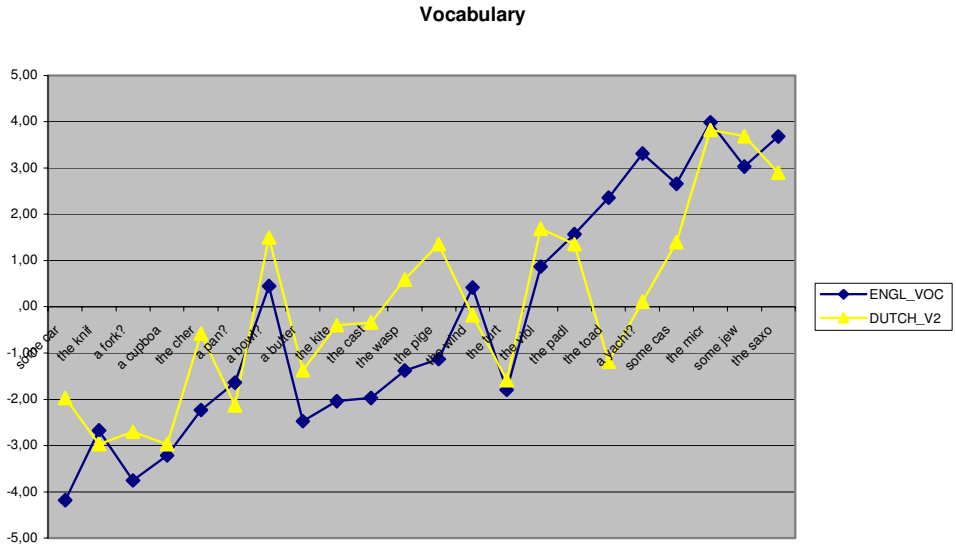
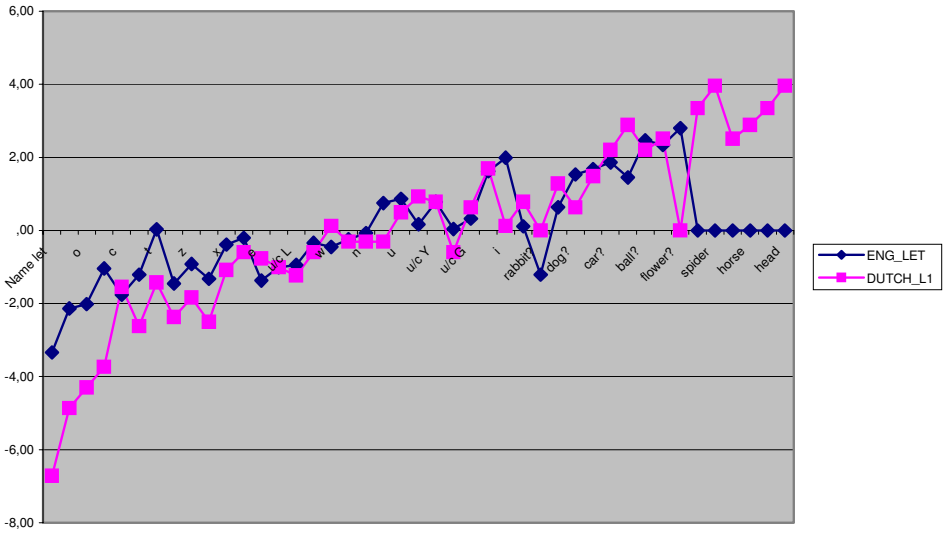


Figure 2.7 – Difficulty estimates of letter identification and words for English and Dutch pupils



Apart from the comparison of the English and Dutch results obtained with the same instrument, there were also qualitative and quantitative comparisons made with other Dutch assessments for young children. In the adaptation phase the content and procedures of several Dutch instruments were studied. Mooij and Smeets (1997) investigated a battery of instruments on intake characteristics of pupils in primary education varying from questionnaires for parents, observation and registration by the teacher to assessments of cognitive and social skills. The latter were mainly instruments made by CITO, the Dutch company for testing and measurement (Central Institute for Test Development). Mooij and Smeets recommended the assessment of early reading 'Taal voor Kleuters' (Van Kuyk, 1996) and of early mathematic competence 'Ordenen' (Van Kuyk, 1997); both instruments are part of the CITO Pupil Monitoring System.

The quality of both tests was rated by COTAN, the Committee on Test Affairs (Evers, Van Vliet-Mulder, Resing, Starren, van Alphen de Veer & Van Boxtel, 2002). The rating system results in grades for seven criteria: Theoretical basis and the soundness of the test development procedure (TB), Quality of the testing materials (TM), Comprehensiveness of the manual (MA), Norms (MO), Reliability (RL), Construct validity (COV), and Criterion validity (CRV). The COTAN rating for both tests is depicted in Scheme 2.2.

Scheme 2.2 – Overview of COTAN ratings for CITO tests for early maths and early reading

	Year	TB	TM	MA	NO	RL	COV	CRV
Ordenen (Order)	1997	●●	●●	●●	●●	●●	●○	○○
Taal voor Kleuters (TAK)	1996	●●	●●	●●	●●	●●	●○	○○

Rating: ○○ unsatisfactory / not applicable; ●○ sufficient; ●● good

For construct validity, both tests 'Ordenen' (Order) and Taal voor Kleuters' (TAK) as well as another CITO test named 'Begrippentoets' (Be) measuring Ideas and Concepts, were applied in order to investigate the correlation with the OBIS. Although the COTAN rating of the Begrippentoets is unknown and the test is no longer for sale since 2000, the test is still widely used. The correlations between OBIS assessed in Year 1 and Year 2 and both instruments of the CITO Pupil Monitoring System – Taal voor Kleuters and Ordenen (Order) – measured in Year 2 vary from adequate to good. The correlation between OBIS assessed in Year 1 and the Begrippentoets (Be) measuring Ideas and Concepts appeared to be inadequate (<.60).

Table 2.3 – Correlations between OBIS and CITO tests for early maths and early reading

	Obis Year 1			Obis Year 2			
	Maths	Reading	Total	Maths	Reading	Total	N
TAK	.40**	.72**	.69**	.67**	.70**	.74**	53
Order	.63*	.71**	.68*	.75**	.61*	.78**	13
Be	.36**	.58**	.54**	.72**	.54**	.65**	59

** Correlation is significant at the 0.01 level (1-tailed).
* Correlation is significant at the 0.05 level (1-tailed).

After careful inspection and comparing with several Dutch tests, including tests for young children with Dutch as the second language, we didn't find topics or items that had to be adapted or removed and only a few required some alteration. For the most items a proper translation was sufficient. It seems clear that in the section 'Rhyming' alterations were necessary to account for differences between the languages.

2.2 V-OBIS

2.2.1 Comparison and prediction

This section focuses on the instrument which measures the criterion variable in our study: the academic attainment in Year 3. The instrument we have used is the V-OBIS (i.e. Vervolg Onderbouwinformatiesysteem), which corresponds with the PIPS Year 1. The test measures academic achievement at the end of Year 3 being the first year of formal reading and maths instruction. It generates information on academic attainment, developed ability and attitude (self-concept). This information is used to calculate measures of relative progress or value added. Value added measures enable fair comparisons to be made because they take into account important factors from outside the school. In the event a child has been previously assessed using OBIS the relative progress over time can be calculated. This shows which children are progressing and which are failing to thrive, in relation to their starting point. We call this 'prior value-added'.
The main research question of this study is the extent to which it is possible to predict the reading and maths of 7-years olds from measures at the age of 4. It is well known that there is a limit to prediction, usually with a maximum correlation of 0.70. The correlations between the standardised V-OBIS test scores show that the cognitive scores in reading and mathematics in Year 3 are well predicted by the total OBIS scores in Year 2 and Year 1 as well. The latter score was assessed about two months after the start of the school year and it represents mainly preschool learning.

Table 2.4 – Predictors: correlations between OBIS, SES, and V-OBIS in Year 3

	Start 1 total	Start 2 total	SES
Maths 3	.57	.56	-.30
Reading 3	.57	.56	-.25
Total 3	.68	.64	-.33

It should be noted that baseline assessment at the start of Kindergarten is a substantially better predictor of future cognitive attainment than the socio-ethnic background of the pupils. Combination of both predictors explains 38% of the variance in the total score of cognitive attainment in Year 3. This is in accordance with the results of the PRIMA-Cohort study where 23% of the variance in language and maths scores in Year 4 was explained by socio-ethnic background and assessment in Year 2.

The V-OBIS also contains a 'context' section. This includes measures of vocabulary and non-verbal ability that are combined to provide a measure of a child's developed ability. This is an excellent predictor of academic attainment and can therefore be used to determine if the child is making the expected progress. This 'concurrent value-added' can be generated for any child regardless whether she or he has been assessed before using OBIS.

Table 2.5 – Correlations between context scores and cognitive scores over three years

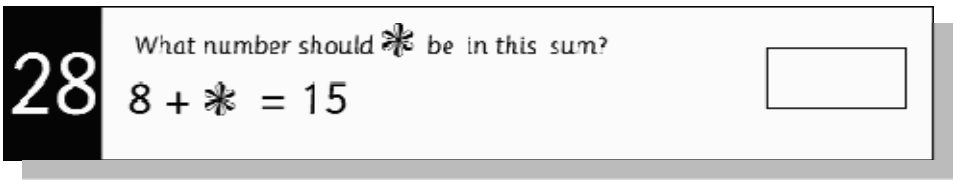
	Start Year 1	Start Year 2	End Year 3
Vocabulary 3	.56	.50	.41
Non-verbal Reasoning 3	.46	.39	.49
Context 3	.59	.52	.53
SES	-.41	-.25	-.27

The correlations between context indicators in Year 3, socio-ethnic background and cognitive attainment over a period of 3 years are reasonably high. The combination of cognitive scores assessed in Year 1 and 2, the context indicators of vocabulary and nonverbal reasoning, and the pupils' socio-ethnic background accounts for 45% (r=.68) of the variance in the academic achievement at the end of Year 3.

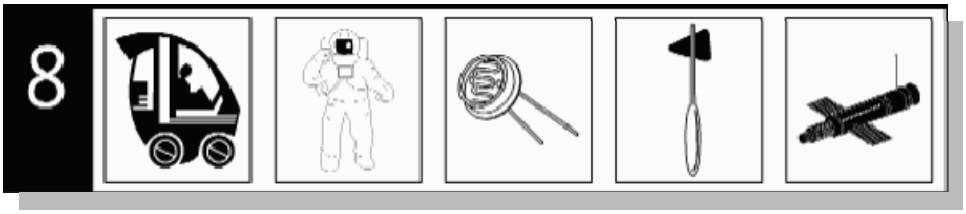
2.2.2 Content and structure

The assessment consists of three sections, each taking about half an hour to complete. The first two sections assess mathematics and reading. The final section is used to collect contextual and attitudinal data. Examples of a math item, a vocabulary and a non verbal item are given below.

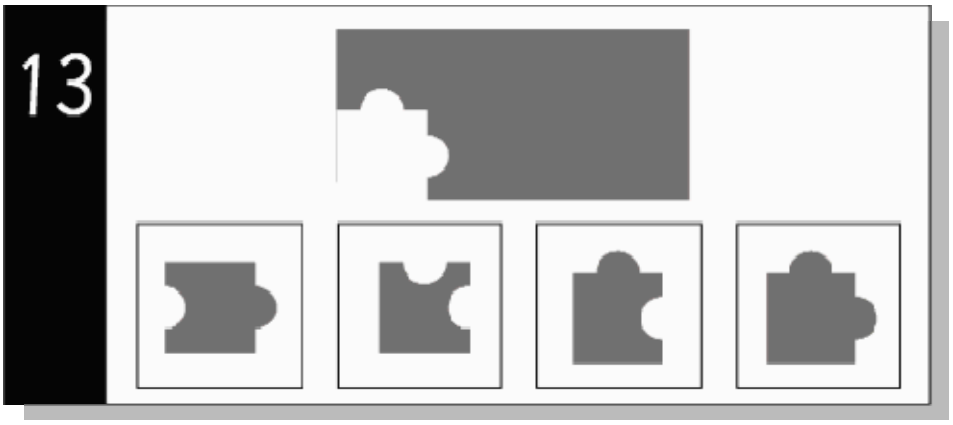
Figure 2.8 – V-OBIS: examples of a math item, a vocabulary and a nonverbal item



Can you put a tick through the astronaut?



Put a tick through the piece which is missing from the puzzle.



For the adaptation and translation of the English Year 1 assessment into the V-OBIS, the content and quality of the test was judged in roughly the same way as it was done before in processing the PIPS Baseline Assessment into the OBIS.

2.2.3 Curriculum and targets

The UK National Curriculum

Inspection of the content of the test started with a comparison between the guidelines of the UK National Curriculum for early literacy and numeracy (Department for Education and Skills) and the Curriculum Authority (http://www.qca.org.uk/)[2] with various Dutch publications on curriculum development and testing coming from the

2 The Qualifications and Curriculum Authority (QCA) is a non-departmental public body, sponsored by the Department for Education and Skills (DfES). QCA maintains and develops the national curriculum and associated assessments, tests and examinations; and accredits and monitors qualifications in colleges and at work.

Freudenthal Institute, the Expertisecentrum Nederlands, the CITO and other publications on assessment (Visser, 1997; Van Luit, Van de Rijt & Pennings, 1998). The UK National Curriculum contains detailed descriptions of the content and targets to be attained. On the Standard Site of the Department for Education and Skills the framework for teaching from Reception (YR) class to the end of primary school (YR6) is published.

Dutch objectives and targets

The Dutch government does not prescribe attainment targets. "The content of teaching and the teaching methods to be used are not prescribed. However, attainment targets have been formulated and schools are expected to organise their teaching in such a way that pupils are able to achieve the attainment targets by the end of primary school." Moreover, the Dutch Ministry of Education, Culture and Science commissioned the Freudenthal Institute and the Expertisecentrum Nederlands to develop intermediate targets and teaching guidelines for arithmetic and Dutch language, respectively (Min. OCW, 2001: http://www.eurydice.org/Eurybase/)[3]. Despite these different governmental policies, comparisons between the English national curriculum and the Dutch situation can be made, using various sources of information on e.g. instruction, intended and implemented curriculum.

2.2.4 Reading

Literacy in the UK: Key Stage 1

According to the national guidelines for Key Stage 1 there should be a strong and systematic emphasis on the teaching of phonics and other word level skills. Pupils should be taught to:

- discriminate between separate sounds in words;
- learn the letters and letter combinations most commonly used to spell those sounds;
- read words by pronouncing and blending their separate parts;
- write words by combining the spelling patterns of their sounds.

The phonics, spelling and vocabulary work to be covered in Years R to 2 (Key Stage 1) are summarized in the following objectives for teaching:

---

3 Eurybase is the EURYDUCE database providing detailed information on each Education System covered by the network. For each country, the information is available in English and in the country's national language(s). EURYBASE has been designed by the EURYDICE European Unit. The EURYDICE National Units prepare, annually update and, where necessary, translate into English, the information on their national education system. The Education System in the Netherlands (2001/2002).

Phonological awareness, phonics and spelling

To understand and be able to rhyme through:

- recognising, exploring and working with rhyming patterns, e.g. learning nursery rhymes;
- extending these patterns by analogy, generating new and invented words in speech and spelling.

Knowledge of grapheme/phoneme correspondences through:

- hearing and identifying initial sounds in words;
- reading letter(s) that represent(s) the sound(s): a-z, ch, sh, th;
- writing each letter in response to each sound: a-z, ch, sh, th;
- identifying and writing initial and dominant phonemes in spoken words;
- identifying and writing initial and final phonemes in consonant-vowel-consonant (CVC) words, e.g. fit, mat, pan.

Alphabetic and phonic knowledge through:

- pronouncing and naming each letter of the alphabet in lower and upper case;
- writing letters in response to letter names;
- understanding alphabetical order through alphabet books, rhymes, and songs.

To link sound and spelling patterns by:

- using knowledge of rhyme to identify families of rhyming CVC words, e.g. hop, top, mop; fat, mat, pat, etc.;
- discriminating 'onsets' from 'rimes' in speech and spelling, e.g. 'tip', 'sip', 'skip', 'flip', 'chip';
- identifying alliteration in known and new and invented words.

Word recognition, graphic knowledge and spelling

- to read on sight a range of familiar words, e.g. children's names, captions, labels, and words from favourite books;
- to read on sight the 45 high frequency words to be taught by the end of YR located in the Resource Area;
- to read on sight the words from texts of appropriate difficulty; to recognise the critical features of words, e.g. shape, length, and common spelling patterns.

Vocabulary extension

- new words from their reading and shared experiences;
- to make collections of personal interest or significant words and words linked to particular topics.

Handwriting

- to use a comfortable and efficient pencil grip;
- to produce a controlled line which supports letter formation;
- to write letters using the correct sequence of movements.

Reading in The Netherlands

As mentioned above, there are several sources to get an overview of the Dutch teaching-learning trajectory for reading in the lower years of Dutch primary schools. Commissioned by the government the CITO periodically assesses the educational content and results of primary education (PPON). In autumn 1999 the third survey for the year groups 3, 4 and 5 of primary education was carried out. It turned out that halfway through primary education the reading skills of the pupils were not at the intended level (Van Berkel, Van der Schoot, Engelen & Maris, 2002). According to Sijtstra (1998) reading instruction in the nineties lacked a systematic approach. The Expertisecentrum Nederlands developed and published the teaching line 'Early literacy' with intermediate targets and teaching strategies on literacy aspects for the year groups 1 to 4: concepts of print, ideas of reading, functions of reading and writing, linguistic concepts, alphabet knowledge and principles; functional reading and writing; technical reading and writing (decoding and word recognition); reading and writing comprehension. From the point of view of interactive reading instruction the teaching line describes the literacy development of young children as a cyclic process with objectives accumulating in difficulty level.

As far as spelling is concerned, there is a strong parallel between the content of English spelling objectives and the AVI (Analyse van Individualiseringsvormen) categorisation system that is widely used in the Dutch schools to determine the degree to which the students' decoding skill has been automated (Visser, 1997, p. 181). The AVI test includes nine levels for categorizing decoding skills of students. The AVI test is composed of texts with increasing difficulty, levels 1 through 4 vary with respect to the degree of difficulty from one-syllable words such as *'pan'* to three- or more-syllable words as *'gelukkig'*. The distribution of the AVI levels across the elementary school period shows that the largest increase in reading occurs in the first two years of reading instruction in Year 3 and 4.

On AVI-level 1 text characteristics are short sentences, one sentence per line; and occasionally composite sentences over two lines long; also capital letters may be used. Words are monosyllabic ones of various consonant-vowel combinations (c-v; v-c; c-v-c): *ga, ik, mak*.
AVI level 2 includes words with two or three initial or final consonants such as *houdt, bang, zink; stoel, pats, stelt; straal, kamt*. There are also dissyllabic or composite words as *voetbal, mamma* and diminutives as *huisje*. Characteristics of AVI level 3 are sentences over two lines long and an extension of words including all types of monosyllabic and dissyllabic words; and three- and four-syllabic words without spelling difficulties. In March of Year 4, the average AVI level is above 4, which means that all types of three- and four-syllabic words occur and also borrowings such as *portemonnee, diskette.*

2.2.5 Maths

Numeracy in the UK: Year 1

According to the Department for Education and Skills (DfES) and the National Numeracy Strategy (NNS) the key teaching objectives for year 1 comprise numbers and the number system, calculations, solving problems, measures, and shapes.

- Count reliably at least 20 objects.
- Count on and back in ones from any small number, and in tens from and back to zero.
- Read and write numerals from 0 to at least 20; understand and use the vocabulary of comparing and ordering numbers.
- Within the range between 0 to 30, say the number that is 1 or 10 more or less than any given number.
- Understand the operation of addition, and of subtraction (as 'take away' or 'difference'), and use the related vocabulary.
- Know by heart all the pairs of numbers with a total of 10.
- Use mental strategies to solve simple problems using counting, addition, subtraction, doubling and halving, explaining methods and reasoning orally.
- Compare two pair of lengths, masses or capacities by direct comparison.
- Suggest suitable standard or uniform non-standard units and measurement equipment to estimate, and then measure a length, mass or capacity.
- Use everyday language to describe features of familiar 3-D and 2-D shapes.

Maths in The Netherlands

In The Netherlands intermediate targets and teaching strategies on early numeracy are developed and published by the TAL-team of the Freudenthal Institute (1999). In 1997 the development of a learning-teaching trajectory for whole-number calculation in the lower years of primary school for the school year 3 was commenced with teaching objectives being: counting and arithmetic operations to 20 with emphasis on memorising and automating of addition and subtracting, splitting. Three levels are distinguished: arithmetic operations using counting materials; structuring operations using appropriate models and formal operations with numbers as mental objects not using any counting materials.

Counting includes saying the numbers starting from any number under 20, back and forwards. A distinction is made between using numbers in real life context (context numbers), for instance referring to age, days, time and hours; and structuring and positioning numbers on line and grouping models. Computations are adding and subtracting in the context of everyday life as well as in a mathematical context.

The next comparison was made with the content of CITO Periodical Survey assessed halfway through primary education which distinguishes the following domains:

Numbers & Operations and Measurement (Noteboom, Van der Schoot, Janssen & Veldhuijzen, 1998). Numbers & Operations consists of five subjects: basic operations (+ – x); counting and order; structuring of numbers; operations (+ – x); operations: applications. The measurement domain is on the subjects measuring, time and money.
At the start of year 5 almost all pupils master the domains consisting of basic operations, counting and order (20% of the items); the results for the sections of Numbers and of Measurement vary from unsatisfactory to moderate. The results on Measurement vary from satisfactory for Measuring (such as length, weight) and Time to unsatisfactory for Money.

TIMMS 2003

Another comparison was made by inspecting the content of the mathematics section of the V-OBIS on the basis of the framework of the international TIMMS study (2003). The mathematics assessment framework for TIMMS 2003 is framed by two organizing dimensions, a content dimension and a cognitive dimension. Each dimension has several domains. The content domains define the specific mathematic subject matter and the cognitive domains define the sets of behaviours expected of students as they engage the mathematics content.

The five mathematics content domains are: number, algebra, measurement, geometry, and data. Each content domain has several topic areas, i.e. number is further categorized by whole numbers, fractions and decimals, integers, ratio, proportion, and percent. Each topic area is presented as a list of objectives covered in a majority of participating countries, at either Year 4 or Year 8. There are four distinct mathematics cognitive domains: knowing facts and procedures; using concepts, solving routine problems, and reasoning.
The idea is that in order to respond correctly to the test items, students will need to be familiar with the mathematics content of the items, and just as important, the items have to be designed in such a way that they elicit the use of particular cognitive skills. Furthermore, the percentages of items devoted to each content and cognitive domain for the TIMMS (fourth grade assessments) are given. The consistencies and differences in content and domain structure between the TIMMS assessment and the V-OBIS are given in Scheme 2.3.

Scheme 2.3 – Percentages content and cognitive domain items, TIMMS (grade 4) and V-OBIS (Year 3)

Content domain	TIMMS Grade 4	VOBIS Year 3	Cognitive domain	TIMMS Grade 4	V-OBIS Year 3
Number	40%	52%	KFP*	20%	33%
Algebra	15%	11%	Using concepts	20%	33%
Measurement	20%	19%	Solving problems**	40%	33%
Geometry	15%	3%	Reasoning	20%	0%
Data		10%	15%		

*Knowing, facts, procedures; **including non routine and real life problems

Procedures

The V-OBIS assessment in reading and mathematics are group tests lasting 30 minutes each. If necessary, the assessments may be administered individually as well. Both sections are internally consistent with Cronbach's Alpha 0.94 and 0.87 respectively. Vocabulary and jigsaws have a Cronbach's Alpha of 0.80 and 0.83 respectively.

2.3 Student, classroom and school profile

For many years school effectiveness studies have sought to describe which other factors than cognitive ability, previous achievement and SES explain differences in actual learning achievement or progress. A variety of context indicators mediating school success have been presented as relevant. In particular, maintaining high expectations and high standards for student achievement – on pupil level as well as on class and school level – belongs to the characteristics of effective teachers. In this section the instruments used to measure teacher expectations, targets, classroom practices and school characteristics will be described.

2.3.1 Teacher expectations: cognitive and social profiles

The teachers of Year 1 and Year 2 were asked to answer for each pupil a series of questions on cognitive and social aspects of the child. Part of the items was derived from the Dutch PRIMA (Primary Education) cohort study (Jungbluth, Roede & Roeleveld, 200I) and also new items were developed. The topics were behaviour and attitude, social background, educational details, development and targets. The teachers had to indicate the extent to which a particular statement applied to the pupil according to a five-point scale ranging from 1 (certainly not) to 5 (strongly). The

collected information concerned developmental characteristics of the pupil, in particular the evaluation of the cognitive and social competences according to the teacher.

Next to this, information was collected to assess the educational level of the pupils in Year 3 according to the teachers in Year 2 and 3. Teachers were asked to indicate to what cognitive category the particular pupil belonged: the head, middle or tail group of his or her year. This information is linked to the teachers' information on that year's implemented curriculum. This approach is developed by Jungbluth (2003) and goes back to previous research where it was shown that over a longer period of time the combination of teachers' expectations and individual cognitive goals influences learning achievement (Van der Hoeven-van Doornum, 1990; Van der Hoeven-van Doornum, Voeten & Jungbluth, 1993).

2.3.2 School profile and classroom practices

Information on school background was collected at class level and completed with information on background characteristics coming from CFI (Centrale Financiële Instellingen) of the Ministry of Education and the school reports from the Inspectorate of Education. The inspectorate conducts a periodical assessment of the quality of educational institutions. The schools' own evaluation is an important input for the external evaluation by the Inspectorate.

The questionnaire 'Class and teacher profile Year 1' (Klas- en leerkrachtprofiel groep 1) contained the topics: personal teacher information; development and curriculum; Dutch as an additional language; pedagogical climate; teaching strategies; ICT; retraining; contacts with parents. The content of the section development and curriculum was mainly based on the learning-teaching strategies and intermediate targets for early literacy and early numeracy developed by the Expertisecentrum Nederlands (Verhoeven & Aarnoutse, 1999) and Freudenthal Institute (Treffers, Van den Heuvel-Panhuyzen, & Buys, 1999).

A second teacher questionnaire entitled 'The transfer from Year 2 to 3' ('De overgang van groep 2 naar groep 3) was developed. The purpose of this questionnaire was to collect information on differences between the teachers in implemented curriculum; topics regarded development in reading and maths in relation to targets and differences in tempo. The teachers were asked to indicate which pupils master those targets at the transfer from Year 2 to 3. Next to this, the teachers were asked to indicate which corresponding teaching and learning activities were emphasized in Years

2 and 3, and also whether this was done with all the children or just with the distinct head, middle or tail groups.

2.4 Concluding remarks

Chapter 2 gives an extensive account of the instruments used throughout this study. The instrument, called OBIS, is the Dutch adaptation of the PIPS Baseline Assessment, developed by the CEM Centre of the Durham University (UK). The test aims to measure cognitive skills: literacy and mathematics of 4 to 6-year old children. In addition the child's attitude towards school is assessed. The test is very child-friendly and uses high-quality pictures attractive for children. The test is very compact, and needs only about 20 minutes to take. As a rule, children are assessed individually, preferentially by their own teacher. This appears to provide the teacher with valuable diagnostic information on the child.

PIPS, and OBIS are constructed and based on the idea of measuring what matters, that is, to be good predictors of later reading and mathematics achievements. The sections "Letter Identification" and "Digit Identification" are the best predictors. Overall, the validity and the reliability of the OBIS were found to be from good to excellent. The test-retest correlations were .92 for maths and .97 for reading.

Differences between the educational systems in the U.K. and The Netherlands were discussed. In both countries there has been considerable debate about the type of assessment suitable for young children. At present, early testing is widespread and firmly established in the U.K., while in The Netherlands it is in its infancy.

The adaptation of the mathematics section was limited to translation of the English items into Dutch. The correlation between the English and Dutch maths items was high (.90). The adaptation of the language section, requiring some more effort to adapt, yielded correlations between the difficulties of Dutch and English vocabulary and letter items of .80 and .90, respectively.

The main difficulties for international studies lie in differences in language, script, school starting ages, consensus on what to cover, and also cultural differences. OBIS has been compared with existing Dutch instruments, mainly tests designed by CITO (Taal alle kleuters and Ordenen). Correlations of OBIS with these tests were highly significant and varied from adequate to good.

In summary, it can be concluded that the OBIS is a reliable basis on which to measure pupil progress and relative pupil progress (value-added) throughout the first year at school.

This is in accordance with the conclusion of The Education Council[4] that the OBIS may be considered to be a reliable instrument for assessment of young children.
The continuation of OBIS meant for Year 3 is called V-OBIS. It corresponds with PIPS Year 1. The test measures academic achievements and attitude. V-OBIS is an excellent predictor of academic attainment, and can be used to determine if a child is making the expected progress. This so-called 'concurrent value added' can be generated for a child regardless of whether it was assessed earlier using OBIS.

2.5 References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.

Baker, E.L., & Linn, R.L. (1993). The technical merit of performance assessments. *The CRESST Line: Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing.* Los Angels:UCLA.

Berkel, S. van, Schoot, F. van der, Engelen, R., & Maris, G. (2002). *Balans van het taalonderwijs halverwege de basisschool 3. Uitkomsten van de derde peiling in 1999.* Arnhem: Cito. (PPON-reeks nr. 20). (Periodical Survey: third assessment halfway primary education 1997, grade 3, 4, 5, details for start of grade 5).

Blatchford, P., Burke, J., Farquhar, C., Plewis, U., & Tizard, B. (1987). Associations between pre-school reading and related skills and later achievement. *British Educational Research Journal*, *13*, 15-23.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental Measurement in the Human Sciences.* Mahwah, NJ, Lawrence Erlbaum Associates.

Bryant, P.E., Maclean, M., Bradley, L.L., & Crossland, J. (1990). Rhyme and alliteration, phoneme detection, and learning to read. *Developmental Psychology, 26,* 429-438.

Bus, A., & van IJzendoorn, M. (1999) Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology, 72,* 665-669.

Cito (1993). *Begrippentoets.* Arnhem: Cito.

Clark, E.V. (1993). *The lexicon in acquisition.* Cambridge: Cambridge University Press.

4 The Education Council is an independent governmental advisory body which advises the Minister, Parliament and local authorities. The Council provides advice, both solicited and unsolicited, to the Minister of Education, Sciences and Cultural Affairs and the Minister of Agriculture, Nature Management and Food Quality. Moreover, the Council may be asked for advice by the Dutch upper and lower chambers of parliament.

Commissie Indicatiestelling Onderwijsachterstanden (1996). *Zo onvoorspelbaar als het leven zelf. Advies*. Zoetermeer: Ministerie van Onderwijs, Cultuur en Wetenschappen.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 317). Hilssdale, NJ: Erlbaum.

Cronbach, L.J. (1989). Construct validity after thirty years. In R.E. Linn (Ed.), *Intelligence: measurement, theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L.J. (1990). Essentials of psychological testing (5th ed.). New York: Harper & Row.

Dave Bartram, D. (2002). EFPA Review model for the description and evaluation of psychological tests: Notes for reviewers. http://www.4tp.nl/products/efpa 3.2/notes.doc

Dunn, G. (1989). *Design and Analysis of Reliability Studies.* New York: Oxford University Press.

Dutch Eurydice Unit (2002). *Information Dossier on the Dutch Education System 2001*. Zoetermeer: Ministerie van Onderwijs, Cultuur en Wetenschappen.

Eiting, M.H. & Blok, H. (1992). Een vergelijking van twee schatters voor de intraklasse correlatie coëfficient: een simulatie in de context van het onderzoek naar de effectiviteit van scholen. *Tijdschrift voor Onderwijsresearch, 17, 6,* 339-342.

Elsäcker, W. van, & Verhoeven, L. T. W. (1998). Motiverend lees- en schrijfonderwijs op de basisschool. *MOER, 5*, 195-203.

Evers, A., Vliet-Mulder, J.C. van, Resing, W.C.M., Starren, J.C.M.G, Alphen de Veer, R.J. van, Boxtel, H. van (2002). COTAN. *Testboek voor het onderwijs*. Lochem: Boom.

Groot, A.D. de & van Naerssen, R.F. (1977). *Studietoetsen: construeren, afnemen, analyseren. Deel II*. Den Haag: Mouton Publishers.

Haertel, E.H. (1991). Validity, vaudeville, and values. *American Psychologist, 36 (10)*, 1021-1034.

Hatch, T., & Gardner, H. (1990). If Binet had looked beyond the classroom: The assessment of multiple intelligences. *International Journal of Educational Research, 14,* 455-473.

International Encyclopedia of Statistics (1978). W.H. Kruskal & J.M. Tanur (Eds.) London / New York: The Free Press, Collier Macmillan Publishers.

Jones, P. (2003). PIPS TES Vocabulary news item. http://www.pipsproject.org/pipstes.asp

Jong, P.F. de, & Leij, A. van der (1999). Specific contributions of phonological awareness to early reading acquisition. Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology, 91,* 450-476.

Jungbluth, P. (2003). *De ongelijke basisschool*. Nijmegen: ITS.

Jungbluth, P., Roede, E., & Roeleveld, J. (2001). *Validering van het PRIMA leerlingprofiel.* Amsterdam: SCO-Kohnstamm Instituut.

Kilgallon, M.K., & Mueller, R.J. (1996). *A Longitudinal Study of the Predictive Validity of a Kindergarten Screening Battery.* ERIC Document Reproduction Service No ED276756.

Kjeldsen, A.-C., Niemi, P., & Oloffson, A. (2003). Training phonological awareness in kindergarten level children: consistency is more important than quantity. *Learning and Instruction, 13,* 349-365.

Kuyk, J.J. van (1996). *Leerlingvolgsysteem. Ruimte en Tijd. Handleiding*. Arnhem: Cito.

Kuyk, J.J. van (1996). *Leerlingvolgsysteem. Taal voor Kleuters. Handleiding*. Arnhem: Cito.

Kuyk, J.J. van (1997). *Leerlingvolgsysteem. Ordenen. Handleiding*. Arnhem: Cito.

Leseman, P.P.M. (1989). *Structurele en pedagogische determinanten van schoolloopbanen.* Rotterdam: Rotterdamse Schooladviesdienst. Project Onderwijs en Sociaal Milieu.

Lindberg, I., Frost, J., & Petersen, O. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23,* 263-284.

Linn, R., Baker, E.L., & Dunbar, S.B. (1991). Complex, performancebased assessment: Expectations and validation criteria. *Educational Researcher, 20*, (8), 15-21.

Linn, R.L. (1994). Performance Assessment: Policy Promises and Technical Measurement Standards. *Educational Researcher, 23* (9), 4-14.

Luit, J.R.H.van, Rijt, B.A.M. van de, & Pennings, A.H. (1998) *Utrechtse Getalbegrip Toets.* Doetinchem: Graviant.

Mehrens, W.A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11,* 1, 3-9.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2), 13-23.

Mooij, T., & Smeets, E. (1997). *Beginkenmerken van leerlingen in de basisschool.* [Intake characteristics of pupils in primary education] Nijmegen: ITS; Ubbergen: Tandem Felix.

Mosteller, F., Elashoff, J.D., & Elashoff, R.M. (1978). Errors. In W.H. Kruskal & J. M. Tanur (Eds.), *International Encyclopedia of Statistics* (p. 210). London / New York: Collier Macmillian Publishers.

Noteboom, A., Schoot, F. van der, Janssen, J., & Veldhuijzen, N. (2000). *Balans van het rekenonderwijs halverwege de basisschool 3. Uitkomsten van de derde peiling in 1997.* Arnhem: Cito. (PPON-reeks nr. 15). Periodical Survey: third assessment halfway primary education 1997, grade 3, 4, 5, details for start of grade 5).

Onderwijsraad (2003). *Wat scholen toevoegen.* Onderwijsraad: Den Haag.
Riley, J.L. (1994). *Why being able to identify and label letters of the alphabet at school entry is valuable.* Paper presented at the British Education Research Association Conference, Oxford, 1-21.
Roeleveld, J. (2003). *Herkomstkenmerken en begintoets. Secundaire analyses op het PRIMA-Cohortonderzoek.* Ongepubliceerd onderzoek.
Saxe, G.B. (1979). Developmental relations between notational counting and number conservation. *Child Development, 50*, 180-187.
Schaerlakens, A.M., Kohnstamm, D., & Lejaegere (1999). *Streeflijst woordenschat voor zesjarigen.* Derde herziene versie. Lisse: Swets.
Shaeffer, B., Eggleston, H., & and Scott, J.L. (1974). Number development in young children. *Cognitive Psychology, 6,* 357-379.
Shavelson, R.J. (1991). Generalizability theory and military performance measurements: I. Individual performance. In A.K. Wigdor & B.F. Green (Eds.), *Performance assessment for the workplace (*Vol. II Technical Issues). Washington D.C.: National Academy Press.
Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405-450.
Sijtstra, J. (1998). *Raamplan, deel I, Taalonderwijs op de basisschool. Een stand van zaken.* Expertisecentrum Nederlands. Nijmegen: ITS.
Stuart, M. (1995). Prediction and qualitative assessment of 5- and 6-year-old children's reading: a longitudinal study. *British Journal of Educational Psychology, 65,* 287-296.
TAL-team (1999). *Jonge kinderen leren rekenen.* Eindredactie: A. Treffers, M. van den Heuvel-Panhuizen & K. Buijs. Groningen: Wolters-Noordhoff.
Tellegen, P., & Laros, J. (1993). The Construction of a Nonverbal Test of Intelligence: the revision of the Snijders-Oomen tests. *European Journal of Psychological Assessment, 9, (*2), 147-157.
Thorndike, R.L. (1971). *Educational measurement* (second edition). Washington D.C. American Council on Education.
Treffers, A., Van den Heuvel-Panhuyzen, M., & Buys, K. (1999). *Jonge kinderen leren rekenen. Tussendoelen Annex leerlijnen. Hele Getallen Onderbouw Basisschool*. Groningen: Wolters-Noordhoff.
Tymms, P., & Merrell, Ch. (2004). On-Entry Baseline assessment across Cultures. In: A. Anning, J. Cullen & M. Fleer (Eds.) *Early Childhood Education. Society and Culture.* London: Sage.
Tymms, P. (1999). Baseline assessment, value-added and the prediction of reading. *Journal of Research in Reading*, *22*, 27-36.
Tymms, P.B., & Williams, D. (1996). *Baseline assessment and Value-added: A report to the School Curriculum and Assessment Authority.* London: The School Curriculum and Assessment Authority.

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

Vacc, N.A., Vacc, N. N., & Fogelman, M.S. (1987). Pre-school screening: Using the DIAL as a predictor of first grade performance. *Journal of School Psychology*, *25*, 45-51.

Van der Hoeven-van Doornum, A.A. (1990). *Effecten van leerlingbeelden en streefniveaus op schoolloopbanen*. [Effects of teacher expectations and aspiration levels on school careers; Dissertation with a summary in English]. Nijmegen: ITS.

Van der Hoeven-van Doornum, A.A., Voeten, J.M.J., & Jungbluth, P. (1993). The influence of parents' expectations on school careers. *Tijdschrift voor Onderwijsresearch, 18* (6), 369-379.

Verhoeven, L., & Aarnoutse, A. (1999). *Tussendoelen beginnende geletterdheid. Een leerlijn voor groep 1 tot en met 3.* Nijmegen: Expertisecentrum Nederlands.

Verhoeven, L., & Vermeer, A. (1986). *Taaltoets allochtone kinderen. Diagnostische toets voor de mondelinge vaardigheid bij allochtone kinderen van 5-9 jaar. Handleiding*. Tilburg: Zwijsen.

Verhoeven, L., & Vermeer, A. (2001). *Taaltoets Alle Kinderen. Diagnostische toets voor de mondelinge vaardigheid Nederlands bij kinderen van groep 1 tot en met 4.* Handleiding. Arnhem: Cito.

Verhoeven, L. (1995). *Leerlingvolgsysteem. Drie-minuten-toets. Technische Leesvaardigheid. Groep 3 tot en met 8. Handleiding.* Arnhem: Cito.

Verhoeven, L. (1996). *Leerlingvolgsysteem. Woordenschattoets groep 3 en 4. Handleiding.* Arnhem: Cito.

Verhoeven, L., & Aarnoutse, C. (1998). Naar interactief taalonderwijs. *School en Begeleiding, 15*, 1, pp. 5-7.

Visser, J.J.C.M. (1997). Lezen op AVI-niveau. *De schatting, normering en validering van de nieuwe AVI-toetskaarten*. 's Hertogenbosch: KPC-Groep.

Wagner, R., & Torgesen, J. (1987). The nature of phonological processes and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101,* 192-212.

Whitehurst, G.J., & Lonigan, C.J. (2001). Emergent Literacy: Development from prereaders to readers. In S.B. Neumann & D.K. Dickinson (Eds.), *Handbook of Early Literacy Research.* New York /London: The Guilford Press.

Wynn, K. (1992). Addition and subtraction in human infants. *Nature 358,* 749-751.

http://www.4tp.nl/products/efpa 3.2/notes.doc
http://www.eurydice.org/Eurybase/
http://www.pipsproject.org/pipstes.asp
http://www.qca.org.uk/ages3-14/
http://www.standards.dfes.gov.uk/numeracy/teaching_resources/year1/

Chapter 3 Data Collection and Results

The aim of this chapter is to describe the data collection and technical details of the instruments described in Chapter 2. The chapter starts with a closer look at the sample of schools that formed the basis of the data. This is followed by details of the collected data. Next to this, the relationship between the various measures is set out, and comparisons between the groups are made.

3.1 Sample of schools

Eleven primary schools participated in the study, and at the start 458 children joined in, of which 402 were still present after two years. Data were collected when pupils first entered school at the age of 4 or 5, and then one year later. It is important to note that even though the data used for particular analyses overlap they are not always identical. On specific variables some data may be missing due to transition from one year to the next or due to mobility of pupils. In relation to transition, it is the entry system that pupils start school on their fourth birthday which means that some pupils spend about two years in kindergarten and others almost three years. When necessary, it will be mentioned with which data the analyses deal.

The representativeness of the OBIS sample was checked against the data of the total Dutch population of schools for primary education (N=7019). The data on the population were obtained from the Ministry of Education (Cfi, 2002). The Tables 3.1 to 3.3 give the details of the analysis of four characteristics associated with representativeness: denomination and urbanisation, number of pupils, percentages of educationally disadvantaged pupils (SES) per school.

Table 3.1 – Comparison of the distributions of denomination among the Dutch population of schools and the sample

Denomination	Population		Sample	
	N	%	N	%
Public	2325	33	5	46
Roman-Catholic	2122	30	3	27
Protestant	2102	30	3	27
Else	470	7		
Chi-square 1.57, df 4, p= 0.86				

Table 3.2 – Comparison of the distributions on urbanisation of the Dutch population schools and the sample

Urbansiation	Population		Sample	
	N	%	N	%
Very densely populated area	795	11		
Densely populated area	1415	20	3	27
Moderately populated area	1405	20	2	18
Intermediate area	1852	26	4	36
Thinly populated area	1552	22	2	18
Chi-square 2,15, df 4, p= 0.73				

Table 3.3 – Differences between the mean number of pupils and percentage disadvantaged pupils per school for the Dutch population and the sample[1]

		N	mean	sd	t-value	df	p (2-tailed)
N pupils	Population	7008	221	124	-.74	7016	0.46
	Sample	11	249	101			
1.00	Population	7008	.73	.24	.62	7017	0.53
1.00	Sample	11	.69	.32			
1.25	Population	7008	.15	.12	.16	7017	0.87
1.25	Sample	11	.14	.12			
1.90	Population	7008	.12	.21	-.79	7017	0.43
1.90	Sample	11	.17	.29			

Categories of socio-ethnic status:
1.00 children without disadvantage (middle and higher education and occupation);
1.25 children from a Dutch cultural background whose parents have a low level of education and low-skilled occupations;
1.90 children from a non-Dutch cultural background whose parents have a low level of education and low-skilled occupations.

It appeared that the characteristics of the population of schools were evenly distributed among the schools participating in the study. There were no significant differences found between the values of the total population of schools and those of the sample.

It is important to note that the data collection was not affected by loss of participating schools. All schools remained in the study for 4 years. Keeping in mind that in the biennial Dutch PRIMA (Primary Education) cohort study a decrease of about 30 % of the schools occurs (Driessen, Van Langen & Vierke, 2004), this is a remarkable result.

Classes and teachers
Within the 11 schools a total of 26 classes were participating in the study. The number of classes within the participating schools varied from one class to six classes per school. Since more than half of the classes (54 %) had at least two part-time teachers; during the first two years of the study in total 53 teachers were involved in the study. More than half of the classes were taught by the same (part-time) teacher in both years of Kindergarten. About 58 % of the teachers received their professional education more than 25 years ago.

3.1.1 Overview of the data collection

The longitudinal data were collected over a period of 4 years and were generated through various respondents. The baseline assessments and cognitive assessments were administered from the pupils themselves, while the questionnaires on the pupils' profiles were filled out by the teacher for each individual child. The teachers also filled out questionnaires on classroom practices.

Scheme 3.1 – Overview of type of time of collection; type of assessment, respondents and instruments

Year	Instrument	Time of collection	Type of assessment	Respondent
2000-01	OBIS	Autumn & Spring	Baseline assessment	Pupil
	Questionnaire	Autumn & Spring	Pupil profile	Teacher
	Questionnaire	Autumn	Classroom practices	Teacher
2001-02	OBIS	Autumn & Spring	Baseline assessment	Pupil
	Questionnaire	Summer	Classroom practices	Teacher
	Questionnaire	Autumn & Spring	Pupil profile	Teacher
2002-03	V-OBIS	Summer	Cognitive Assessment	Pupil
	Questionnaire	Summer	Pupil profile	Teacher
2003-04	V-OBIS	Summer	Cognitive Assessment	Pupil

3.2 OBIS

3.2.1 Sample of pupils

Scheme 3.2 gives a summary of the numbers of pupils who participated in this study. In total, 488 children participated in the study; from 314 (69%) of the 458 children starting in the first year of data collection 2000-01, the scores on the OBIS and V-OBIS are still available. Summer 2004, the last assessments were done at the end of the school year 2003-04. This is due to the above mentioned entry intake system, most pupils of the autumn intake were transferred to group 3 in 2002; and most pupils of the spring intake who spent more than two years in kindergarten, in 2003. Therefore the V-OBIS scores were collected in two different school years, e.g. 2002-03 and 2003-04. 78 % of the pupils of the autumn intake at entry (Year 1) were after two years (Year 3a) still present in the study[1].

1 The biennial Dutch PRIMA (Primary Education) cohort study knows a pupil mobility varying from about 48% in the lower groups to 39% in higher groups (Driessen, Van Langen & Vierke, 2004).

Scheme 3.2 – Overview of numbers of pupils per measurement moment

Year	Autumn	Spring	Summer	Year	new	left	longitudinal
Year 1	344	114	-	458	-	-	458
Year 2	318	100	-	418	16	56	402
Year 3a	-	-	269	269	16	165	253
Year 3b	-	-	61	61	-	-	61

Year 3a is the Autumn intake; Year 3b is the Spring Intake

3.2.2 Distributions of the scores on the subscales and total scores

The OBIS provides an overall score as well as scores in Early Reading, Early Maths and Phonics, see also Section 2.1.2. The Early Reading score is formed by totalling the scores in the sections on Writing, Picture Vocabulary, Ideas of Reading, Letters and Words. The Early Maths score is calculated from Ideas about Maths, Counting, Sums and Numbers. The Phonics section gives an indication of phonological awareness. It is based on repeating Words and Rhyming Words. For both assessments administered in year 1 and 2 of Kindergarten the descriptive statistics of the reading, maths, phonics and total score are given in Table 3.3.

Table 3.4 – Descriptive statistics for the sections of the OBIS (raw scores)

	mean	sd	min – max
Year 1			
Reading	23.59	10.03	1 – 70
Phonics	10.49	4.58	1 – 17
Maths	17.32	7.99	1 – 38
Total	51.30	19.75	7 – 123
Year 2			
Reading	36.57	18.43	8 – 156
Phonics	14.54	3.13	2 – 17
Maths	29.57	9.72	3 – 53
Total	80.68	27.04	16 – 221

In order to make comparisons between children and classes these scores are standardised on a representative sample of pupils doing the same assessment. The average pupil has a standardised score of 50. The standard deviation is 10, which means that approximately two thirds of the children in the sample will fall between 40 and 60.

On the basis of this teachers can compare the scores of children in their own group with the average. If the child is situated above 60, they belong to the top 16% of the sample. If their score is below 40, they belong to the bottom 16% of the sample. Scores above 70 or below 30 are exceptional, only 2-3% will have scores as high as 70 and a similar proportion as low as 30 (Using PIPS, CEM Centre, 2004).

3.2.3 Baseline assessments across cultures: distributions of OBIS and PIPS

One of the main research questions is how well Dutch pupils are doing on the adapted version of the PIPS Baseline assessment? Therefore, in this section the distributions are presented on the basis of the responses to scales within the OBIS and PIPS assessed at the entry of primary education. The distributions of OBIS and PIPS are given in percentages (CEM Centre: Technical report. Text version, 1999; CEM Centre: Technical report. Text version, 2001). The average age of the Dutch pupils was 4 years and 6 months; the average age of the English sample was 4 years and 6 months as well (CEM Centre: Using the PIPS Baseline Assessment 1999/00, 1999).

1 Name writing
In this section the pupils simply had to write their name and the teacher rated the work on a 0-5 scale using descriptions provided by OBIS. Around 12 % of the English children and 25 % of the Dutch children were not able to write their own name at all.

2 Picture Vocabulary
About 9 percent of the Dutch pupils score 3 or less on Picture Vocabulary and about 66 % of the pupils scored between 16-22 points, while the English percentages in 2000 are 2.5 % and 51.5 % respectively (UK-00 %).

Picture Vocabulary	Dutch-00 %	UK-99 %	UK-00 %
0-3	8.5	1.7	2.5
4-15	25.1	43.2	46.1
16-22	66.3	55.2	51.5

3 Ideas about reading
In the year 2000 the distributions of both samples are quite similar; 75 percent of the English pupils did not pass the Scoring Rule and likewise, 77 percent of the Dutch pupils have a score below 5 on a scale with a maximum of 10.

IAR	Dutch-00 %	UK-99 %	UK-00 %
0-1	2.9	4.0	3.4
2	16.6	12.7	9.8
3	31.7	15.6	36.4
4	25.6	13.5	25.1

4 Repeating words

Although teachers reported that repeating nonsense words was new and unfamiliar for most pupils, the Dutch results on this section are surprisingly good. The differences between the Dutch and English results are evident and probably caused by major differences in phonological characteristics between the two languages.

Repeats	Dutch-00 %	UK-00 %
0	1.8	3.4
1	2.5	3.4
2	3.1	6.3
3	4.3	6.7
4	5.2	9.4

5 Rhyming

The distributions of the English and Dutch scores on Rhyming are fairly similar, on the understanding that the Dutch pupils accomplish slightly better than the English pupils.

Rhyming	Dutch-00 %	UK-99 %	UK-00 %
0	26.5	28.1	26.5
1	12.4	15.6	15.9
2	9.2	8.4	9.1
3	2.9	6.0	5.5
4	5.2	6.6	5.3

6 Letter Identification

The English pupils perform better than the Dutch pupils on the section Letter Identification. Over a third of the Dutch pupils could not identify a single letter, including the first letter of their name; and almost 48 % only identified 1 to 4 letters correctly.

Letter identification	Dutch-00 %	UK-90 %	UK-00 %
0	37.5	26.8	21.8
1	23.8	16.2	13.0
2-4	24.5	18.0	21.0

7 Words

Although most children can't read a single word at the age of four, the English pupils accomplished better than the Dutch pupils on the section Words.

Words	Dutch-00 %	UK-99 %	UK-00 %
0	95.1	84.5	86.4
1-4	4.9	15.5	13.6

8 Ideas of Maths

With regard to Ideas of maths the Dutch pupils did surprisingly well, 95 % scored 4 points or higher, of whom 55 % scored all items correctly.

IAM	Dutch-00 %	UK-99 %	UK-00 %
0	1.1	.7	1.1
1-5	23.2	46.0	33.5
6	20.4	34.2	20.8
7	55.3	19.0	44.6

9 Counting

The pattern of the distributions of the English and Dutch scores on the section Counting is quite similar, but in favour of the English pupils.

Counting	Dutch-00 %	UK-99 %	UK-00 %
0	11.5	9.3	9.4
1-3	44.3	42.8	39.5
4	44.3	47.9	51.1

10 Digit identification

The English pupils scored higher on digit identification than the Dutch pupils. Approximately 60 percent of the Dutch pupils have a score of 4 or less on identifying the numbers from 1 to 10 (first digits), against 34 % of the English pupils.

Digit identification	Dutch-00 %	UK-99 %	UK-00 %
0	34.4	17.7	15.7
1-3	20.9	14.7	14.5
4	4.5	4.1	3.6

11 Sums

The Dutch pupils score evidently higher on the section Sums. About 52 % of the Dutch pupils made 5-8 sums correctly in comparison with 41 % of the English pupils.

Sums	Dutch-00 %	UK-99 %	UK-00 %
0	13.5	11.5	12.3
1-3	23.8	37.2	27.1
4	11.0	10.6	10.1

Collectively, the distributions of the scores show that the English pupils performed best on the sections Name Writing (1), Letter Identification (6), Counting (9) and Digit Identification (10). The Dutch pupils performed better on the sections Picture Vocabulary (2), Repeats (4), Rhyming (5), Words (7), Ideas of Maths (8) and Sums (11). For the sections Ideas about Reading (3), Words (7), the patterns of the distributions of the English and Dutch scores appear to be quite similar.

Table 3.5 – Year 1. Comparison of correlations between standardised test scores in The Netherlands (N=458) and England (N=98571).

Dutch scores	Maths	Reading	Phonics
Reading	.71		
Phonics	.57	.64	
Total	.89	.92	.78
English scores	Maths	Reading	
Reading	.78		
Phonics	.48	.50	

Furthermore, Table 3.5 shows that the correlations between the subscales of the OBIS and the correlations between the PIPS BLA are not much different. This also supports the validity of the instrument.

Over the last years, the use of the PIPS Baseline in different countries all over the world has been extended rapidly. Tymms, Merrell & Jones (2003) presented the results of a study on the use of translated or adapted versions of the PIPS in different countries or cultural groups – Australia: preschool children aged 4, Year 1 aged 5, indigenous pupils; New Zealand, England: children aged 4, deaf children, children in disadvantaged areas; The Netherlands; and Scotland. Rasch scaling was used to estimate the relative difficulties of the items for each country or group. From the international and cultural comparison, it may be concluded that the use of OBIS as an instrument to assess the developmental levels and skills of pupils starting school is justified.

3.2.4 Comparisons of groups

In addition to the assessment results data on background characteristics were collected. Here follows a breakdown by four selected variables: intake, age, sex, socio-ethnic status of parents (SES), see Table 3.6 to 3.9.

Table 3.6 – Breakdown of OBIS standardised test scores by intake

Intake	Autumn 2000		Intake spring 2001		
	mean	sd	mean	sd	p
Maths	50.30	9.93	49.39	10.05	.38
Phonics	50.10	9.67	49.35	9.56	.45
Reading	49.70	10.08	50.70	9.80	.33
Total	49.96	10.05	50.12	9.90	.87

In The Netherlands nearly all children start school right after their fourth birthday; so they may start school at the same age but not on the same day. The OBIS assessments were administered during two fixed periods of time in autumn and spring. In autumn all the pupils were assessed, while in spring only those pupils were assessed who started school after the autumn intake. As a consequence, the pupils of the spring intake are younger than those of the autumn intake.

The average age of the pupils in year 1 varied from 3 years and 11 months to 5 year and 8 months, the average age was 4 years and 6 months. Usually, older pupils score higher than younger pupils (Table 3.7), therefore a t-test was used to compare the differences between the autumn and spring intake in age and OBIS scores. Although the average age of the autumn and spring intake pupils differed significantly (p=.00), respectively 55 months (4,7 years) and 52 months (4,4 years), there were no significant differences between the OBIS scores.

Table 3.7 – Year 1. Descriptive statistics of the OBIS scores by year of birth

	Maths		Reading		Phonics		Total	
	mean	sd	mean	sd	mean	sd	mean	sd
1995	54.70	9.1	53.28	10.0	53.08	8.6	54.46	9.1
1996	49.15	9.8	49.80	10.0	49.10	9.9	49.50	9.9
1997	47.52	10.5	49.60	9.9	47.10	9.8	48.50	9.5

Table 3.8 – Year 1. Descriptive statistics of the OBIS standardized test scores by sex

Year 1	Maths		Reading		Phonics		Total	
	mean	sd	mean	sd	mean	sd	mean	sd
Boys	49.7	10.2	49.0	10.2	49.1	9.7	49.4	10
Girls	49.9	9.8	51.3	9.8	49.5	10.1	50.7	9.7
p	.97		.03		.23		.23	

As it can be seen in Table 3.8, boys and girls hardly differed in the mathematics section, but at reading the girls were significantly better than the boys. For the phonics section and the total scores, the differences between the both sexes were not significant. One year later, in group 2, the differences between boys and girls were almost the same as in group 1 (data not shown).

Obviously, there is a strong relationship between SES category and performances on reading, maths, phonics or total scores on the OBIS test. Table 3.9 displays the means and standard deviations of the SES groups on the OBIS in year 1 and year 2. The starting point of children from ethnic minorities lagged considerably behind that of their peers (p=.00). Although their performances were still below average, it appeared that after one year of education their progress on the reading section was significantly higher compared with the progress of their Dutch peers.

Table 3.9 – Descriptive statistics of the OBIS standardized test scores by socio-ethnic status of parents (SES), indicating degree of disadvantage

Year 1	Maths		Reading		Phonics		Total	
SES	mean	sd	mean	sd	mean	sd	mean	sd
1.00	51.5	9.4	53.3	8.4	51.5	8.9	52.8	8.6
1.25	45.2	8.5	48.5	7.9	44.6	11.8	46.4	8.6
1.90	46.2	10.6	40.3	9.9	44.1	9.2	42.5	10.0
Year 2	Maths		Reading		Phonics		Total	
SES	mean	sd	mean	sd	mean	sd	mean	sd
1.00	51.9	9.4	51.1	9.3	52.0	7.8	51.8	9.2
1.25	45.4	9.6	48.1	9.6	45.8	9.8	46.3	9.9
1.90	45.7	9.7	47.0	11.8	43.1	8.7	45.5	10.8

Next to this, for both years the scores were also broken down for performances – high, middle and low. The proportions of pupils with high, middle or low reading performance scores within the various SES categories are displayed in Table 3.10.

Table 3.10 – Socio-ethnic status of parents (SES) and reading performance groups (high, middle and low) in year 1 and year 2, indicating changes in distributions of test scores

		SES	
Reading Year 1	1	1.25	1.90
Performance group			
High	41 %	33 %	5 %
Middle	36 %	33 %	24 %
Low	23 %	33 %	71 %
Reading Year 2	1	1.25	1.90
Performance group			
High	32 %	26 %	35 %
Middle	41 %	35 %	24 %
Low	27 %	39 %	41 %

The table shows that although in year 1 the percentage of high performances scores among SES group 1 is considerably higher than the incidence of high performance scores among SES group 1.90 (ethnic minorities), after one year there is a considerable group of 1.90 pupils within the group of high performing pupils. The same applies to the percentage of pupils from SES group 1 within the group of low performing pupils.

The SES-weighting score overlaps the performance categories only partially. The percentages of children in the SES group 1.90 with a high score for Reading[2] increased from 5 % to 35%. In Chapter 4 and Chapter 5 this phenomenon is discussed in greater details.

Here we take a closer look at the question on what subscales of the reading section the progress of the non-Dutch minority pupils significantly differs from the progress of the Dutch pupils. Therefore, the data on the subsections were broken down by

2 The percentages of overlap in the first and last assessment for Maths, Phonics and the Total score of the pupils of the ethnic minority group (SES-weighting score 1.90) were 21 % and 24 %, 14 % and 11 %, 10 % and 27 %, respectively.

three SES categories, 1.00, 1.25 and 1.90 respectively representing children without educational disadvantage, children from Dutch working-class families and pupils from ethnic minorities (Table 3.11).

Table 3.11 – Progress on reading subscales related to family background: performances on reading scales (raw scores) in year 1 and year 2, broken down by SES weighting score

SES	Year 1	mean	sd	Year 2	mean	sd	N
Vocabulary: In the Kitchen, maximum score is 7							
1.00		6.66	0.63		6.83	0.42	183
1.25		6.23	1.16		6.66	.68	44
1.90		3.67	2.02		5.59	1.09	61
Vocabulary: In the Country, maximum score is 10							
1.00		8.47	1.40		9.37	0.93	183
1.25		7.48	2.45		8.61	1.48	44
1.90		3.43	3.47		6.28	2.72	61
Ideas about Reading (IAR), maximum score is 5							
1.00		3.61	1.05		4.12	.98	183
1.25		3.57	1.30		4.34	.96	44
1.90		3.15	1.22		4.31	.79	61
Repeats, maximum score is 8							
1.00		6.88	1.65		7.50	.99	183
1.25		5.77	2.61		7.30	1.50	44
1.90		5.74	2.18		7.30	1.49	61
First Letters, maximum score is 13							
1.00		2.07	3.13		4.71	4.20	183
1.25		1.23	2.49		3.39	3.88	44
1.90		.89	1.91		4.56	4.03	61
Words, maximum score is 14							
1.00		.26	1.47		1.37	3.54	183
1.25		.02	.15		1.00	3.24	44
1.90		.05	.22		3.92	6.09	61

The progress of the allochtonous pupils applied to most of the reading subsections: vocabulary, phonics, letter identification and words. The progress on the latter subsection is remarkable and exceptional, because while checking the data it turned out that one particular school was responsible for this effect. Probably the teacher started to give reading instruction already in kindergarten.

3.3 V-OBIS

The V-OBIS (Vervolg onderbouwinformatiesysteem) assessment includes academic attainment on mathematics and reading; and also developed ability – vocabulary and non-verbal ability – which acts as a control to provide concurrent value-added measures. Attitudes and self-esteem measures relating to reading, mathematics and school were also collected. The descriptive statistics of the subscales are given in Table 3.12.

Table 3.12 – Descriptive statistics for the sections of the V-OBIS (raw scores)

Year 3	mean	sd	N items	Cronbach's α
Maths	16.97	5.20	27	.87
Reading	32.67	11.13	50	.94
Vocabulary	18.84	4.14	27	.80
Non-verbal Reasoning	21.38	4.62	30	.83
	mean	sd	min – max	
Attitude to Maths	2.50	.46	1 – 3	.69
Attitude to Reading	2.63	.53	1 – 3	.69
Attitude to School	2.70	.39	1 – 3	.74

Score 3 is a positive attitude ☺

Statements as 'I enjoy doing sums'; 'I am good at reading'; and 'I learn a lot at school' were read out to the pupils. More than half of the children responded positively when asked how they felt about maths (56%); reading (78%) and school (66%). Table 3.13 displays how positive attitudes towards maths and reading are linked with the scores on maths and reading, and shows that a positive attitude towards school does not correlate significantly with academic attainment.

Table 3.13 – Correlations between attainment scores and attitudes to maths, reading and school

	Attitude maths	Attitude reading	Attitude school
Attitude reading	.23		
Attitude school	.51	.08 (ns)	
Maths year 3	.18	.30	.07 (ns)
Reading year 3	.15	.46	.02 (ns)
Total 3 year 3	.19	.45	.02 (ns)

Doing well at school in maths and reading encourages positive attitudes towards learning, but the lack of academic success was not correlated with a positive or negative attitude towards going to school.

3.4 Pupil, classroom and school profiles

3.4.1 Pupil profile

The information on the pupils' profiles was collected by means of presenting a series of questions to the teachers. The data were collected separately and before the feedback of the OBIS assessment was returned to the schools. The questions the teachers were asked to answer regarded different kinds of expectations and perceptions of the child's social-emotional development; cognitive capacities and development; their home background and parental support. The summaries of Year 1 and Year 2 are given in Table 3.14 and Table 3.15 respectively.

Table 3.14 – Year 1. Descriptive statistics for the scales Pupil profile

	mean	sd	min – max	range[1]	N	Cronbach's α
School well-being	4.12	.49	2.17 – 5	1-5	6	.88
Self-confidence	3.53	.75	1.71 – 5	1-5	7	.85
Social behaviour	3.56	.75	1.0 – 5	1-5	3	.75
Attitude	3.27	.84	1.0 – 5	1-5	3	.87
Parental support	4.07	.85	1.0 – 5	1-5	3	.82
Cognitive capacities	3.42	.67	1.13 – 5	1-5	8	.88
Development	4.10	.58	2.43 – 5	1-5	7	.80
Test-effect	2.11	.53	1.0 –4.2	1-5	5	.93

1 Five point scale ranging from (1) 'no, certainly not' to (5) 'yes, strongly'. The meaning of score (5) is:

school well-being	the child feels happy at school
self-confidence	is self-confident
social behaviour	pleasant, behaves according to the rules
attitude	doesn't give up quickly, works steadily and accurately
parental support	stimulating home environment, ethnic parents speak Dutch at home.
cognitive capacities	bright child, belongs to the >better= pupils
development	mental and physical development according to age
test-effect	teacher changed expectations and teaching practices after assessment.

Test-effect

After completing the assessment it is possible to apply the information obtained from the OBIS immediately within the schools. Teachers may inspect the raw scores for important information they could act upon. In order to see more precisely if there was any effect of the assessment on the teachers, they were asked to indicate to what degree their perceptions or teaching practices were changed after the OBIS assessment was carried out.

The average score of 2 on the scale 'test-effect' indicated that in their opinion teachers did not change their judgements and teaching practices, despite the newly available information on academic attainment of the pupil. As it can be seen in Table 3.14 and Table 3.15, this was the case in both years.

Table 3.15 – Year 2. Descriptive statistics for the scales Pupil profile

	mean	sd	min – max	range	N	Cronbach's α
School well-being	4.15	.52	2.71 – 5	1-5	7	.88
Self-confidence	3.67	.62	1.86 – 5	1-5	7	.84
Social behaviour	3.57	.74	1.2 – 5	1-5	5	.83
Attitude	3.46	.82	1.0 – 5	1-5	3	.86
Parental support	4.03	.73	1.75 – 5	1-5	4	.82
Cognitive capacities	3.61	.70	1.13 – 5	1-5	8	.91
Development	4.42	.53	2.86 – 5	1-5	7	.87
Test-effect	2.07	.61	1.0 –4.3	1-5	3	.87

The correlations between the different kinds of teachers' perceptions are given in Table 3.16 and 3.17.

Table 3.16 – Year 1. Correlations between scales Pupil profile and SES

	sw	sc	sb	at	ps	cc	de	te
School well-being	1.00							
Self-confidence	.48	1.00						
Social behaviour	.31	-.16	1.00					
Attitude	.22	.14	.45	1.00				
Parental support	.21	.06	.17	.11	1.00			
Cognitive capacities	.32	.41	.23	.60	.37	1.00		
Development	.44	.37	.27	.45	.50	.63	1.00	
Test-effect	.04	.06	.08	.07	.37	.07	.21	1.00
SES	.02	.05	-.00	.03	-.74	-.17	-.25	-.31

Correlations < .10 are not significant (p=< .05)

In Year 2 the correlations between the 'triangle' school well-being, self-confidence and parental support have increased, which may indicate that there is a relationship between the teacher's perception of parental support and a 'happy' child at school.

Another remarkable increase can be seen in relation to the test-effect. The correlation between the effect of objective and comparable information with the more subjective judgments is almost absent in Year 1 and weak in Year 2. Nevertheless, the increase is remarkably illustrating that when the teacher has a more positive perception of a child it is less likely that the teacher will change his or her teaching attitude towards the child.

Table 3.17 – Year 2. Correlations between scales Pupil profile and SES

	sw	sc	sb	at	ps	cc	de	te
School well-being	1.00							
Self-confidence	.62	1.00						
Social behaviour	.27	-.14	1.00					
Attitude	.23	.19	.47	1.00				
Parental support	.40	.23	.12	.17	1.00			
Cognitive capacities	.39	.49	.20	.61	.37	1.00		
Development	.45	.35	.22	.41	.45	.59	1.00	
Test-effect	-.19	-.18	-.15	-.16	-.27	.19	.21	1.00
SES	.06	.09	-.09	.03	-.73	-.22	-.32	-.21

Correlations < .10 are not significant (p=< .05)

Table 3.18 and table 3.19 display the correlations between the OBIS scores – math, reading, phonics and the total scores – and the teachers' perceptions on the social-emotional development, cognitive capacities and home background of the children for both years. In addition to the teacher's judgments, the correlation between OBIS scores and the SES weighting scores is also given.

Table 3.18 – Year 1. Correlations between OBIS, Pupil profile and SES

	Maths	Reading	Phonics	Total
School well-being	.13	.12	.10	.13
Self-confidence	.16	.17	.17	.18
Social behaviour	.10	.09	.04	.10
Attitude	.29	.19	.24	.27
Parental support	.22	.47	.29	.39
Cognitive capacities	.49	.46	.41	.50
Development	.33	.44	.39	.43
Test-effect	.06	.18	.07	.11
SES	-.23	-.50	-.31	-.42

With some exceptions, the correlations between the OBIS scores and the perceptions in Year 1 and Year 2 are rather close. However, remarkable decreases (>.10) have arisen in the correlations between the teacher's perception of the personal development of the child and the cognitive development assessed with OBIS; while the correlation between the perception of cognitive capacities and the OBIS scores remained quite stable over time.

Table 3.19 – Year 2. Correlations between OBIS, Pupil profile and SES

	Maths	Reading	Phonics	Total
School well-being	.21	.19	.24	.23
Self-confidence	.26	.19	.25	.25
Social behaviour	.05	.05	.06	.05
Attitude	.26	.17	.23	.25
Parental support	.28	.26	.45	.33
Cognitive capacities	.55	.45	.46	.56
Development	.23	.14	.26	.22
Test-effect	.02	.01	-.10	.01
SES	-.25	-.16	-.39	-.25

The information given by the teachers provides a fairly reliable personal and social profile of the pupil. There is a vast amount of research on how teachers' judgments can affect achievement by the mediating role of self-fulfilling prophecies (see e.g. Van der Hoeven-van Doornum, 1990; 1993). But obviously, the mechanism of self-fulfilling prophecies can only be useful if there is certain stability in perceptions over time.

Table 3.20 – Year 1 and Year 2. Correlations between scales >Pupil profile

Year 1	sw1	sc1	sb1	at1	ps1	cc1	de1	te1
Year 2								
School well-being	.49	.38	.21	.19	.31	.25	.35	-.00
Self-confidence	.36	.62	-.04	.16	.21	.35	.36	-.01
Social behaviour	.06	-.23	.65	.29	.11	.08	.07	.05
Attitude	.06	.08	.32	.58	.17	.44	.33	.08
Parental support	.17	.05	.12	.06	.84	.29	.39	.14
Cognitive capacities	.25	.35	.22	.47	.33	.67	.54	.12
Development	.32	.35	.22	.41	.42	.52	.58	.09
Test-effect	-.00	-.06	-.03	-.06	-.27	-.18	-.13	-.15

Correlations < .10 are not significant (p=< .05)

Research outcomes pertaining to older pupils revealed that teachers' perceptions across grades are fairly stable (Van der Hoeven-van Doornum, 1990; Luyten, 1994). Nevertheless, it is not impossible to imagine that for young pupils, teachers may still adapt their perceptions according to their fast cognitive growth after starting school. Therefore, the correlations between the perceptions in both years are presented in Table 3.20.

Considering the correlations, the perceptions of teachers in Year 1 and Year 2 are rather similar. The correlations vary from .49 for school well-being to .84 for parental support, which is in the same order as the results found in the last two years of primary education.

Cognitive category

In addition to the cognitive and social profile of the child assessed in year 1 and 2, in year 3 the teachers were asked to indicate to which cognitive category the pupil belonged in comparison with peers in the same group: head (3), middle (2) or tail (1) group. This part of data collection only pertained to the pupils in year 3 (see also Scheme 3.1); therefore the data for 216 pupils were available. According to their teachers, 36% of the pupils belonged to the 'head' group, 44% to the 'middle' group

and 20% to the 'tail' group. There was no significant correlation between SES and the cognitive category (-.07).

Table 3.21 – Correlations between Pupil profile in Year 1 and 2, Cognitive category Year 3 and SES

Pupil profile Year 1	sw	sc	sb	at	ps	cc	de	te
Cognitive category, Y3	.02	.15	.05	.34	.17	.46	.27	.13
SES	.01	.05	-.00	.03	-.75	-.17	-.23	-.30
Pupil profile Year 2	sw	sc	sb	at	Ps	cc	de	Te
Cognitive category Y 3	.05	.16	.07	.42	.08	.58	.12	.03
SES	-.16	.09	-.09	-.08	-.73	-.22	-.32	-.21

Correlations < .10 are not significant (p=< .05)

3.4.2 Predicting V-OBIS

As indicated in section 3.2.2, the V-OBIS scores were collected in 2003 and 2004 as well. Even though the pupils of the Spring intake achieved better results than the pupils of the Autumn intake, the differences were not significant (maths: t-value .55, df 329, p .59; reading: t-value .94, df 329, p .35; total score: t-value .82, df 329, p .41). Therefore the scores of both groups were merged and standardized; Table 3.22 displays which aspects of the pupils' profiles in year 1 and 2 correlate significantly with V-OBIS. The correlation with the pupil's cognitive category – head, middle, tail – and social status are also given. Please note that self-confidence and social behaviour at the start of school are positively related with future achievement, but also that a year later this is no longer the case.

Table 3.22 – Year 3 (2003 and 2004). Correlations between V-OBIS, pupil profile, cognitive category and SES

	Y3 Maths	Y3 Reading	Y3 Total
Year 1			
Maths	.52	.45	.53
Reading	.44	.42	.47
Phonics	.38	.34	.39
Total score	.51	.47	.54
Self-confidence	.15	.14	.17
Social behaviour	.14	.10	.13
Attitude	.26	.22	.27
Parental support	.19	.13	.16
Cognitive capacities	.40	.41	.45
Development	.26	.26	.29
Year 2			
Maths	.59	.50	.58
Reading	.41	.43	.47
Phonics	.31	.29	.32
Total score	.54	.50	.56
Attitude	.25	.29	.30
Parental support	.24	.12	.17
Cognitive capacities	.39	.37	.42
Development	.19	.17	.20
Year 3			
Cognitive category	.54	.63	.67
SES	-.30	-.25	-.27

To see more precisely which aspects of the pupil profile contribute to the explanation of the V-OBIS attainment scores, regression analyses were performed between the V-OBIS as the dependent variable and the aspects of the pupil's profile and previous assessments as independent variables. Several analytic strategies were used: a) standard multiple regression (method enter) which gives us the unique contribution of the variables as if they entered the regression after all the other independent variables; b) the hierarchical method in which the order of entry of the variables is specified beforehand; c) the stepwise method which assigns the order of entry according to the highest initial correlation between the dependent and independent variables. For all three strategies the outcome was that previous learning assessment is the best predictor of maths, reading and total scores on the V-OBIS in year 3, sometimes followed by the perception of the pupil's cognitive capacities, self-confidence or parental sup-

port. Although the correlations between several aspects of pupil's profile and V-OBIS scores were significant, in a model with previous learning assessment they did not contribute to regression at all or they contributed very little. Apparently, if better information was available, teacher's perceptions did not improve the prediction of learning outcome. However, one exception to this was the contribution of the cognitive category (head, middle or tail). Added to the model, this perception of teachers exceeded the contribution of previous learning assessment. The total explained variance (adj R^2) by previous learning assessment and cognitive category according to the teacher varied from .40 for maths, .46 for reading and .54 for the total score.

To conclude this section on predicting the learning performances of pupils on the V-OBIS in year 3, we present correlations between V-OBIS and the corresponding CITO tests 'Begrippentoets' (Ideas and concepts) and 'Taal voor Kleuters' (Language Young Children). As the available data have been obtained from a small group of pupils (N=59), we also report the correlations of this particular sample which appeared to be of about the same magnitude as for the total sample of more than 300 pupils. Please note that the correlations of the OBIS and CITO instruments with the V-OBIS are very similar, indicating that the criterion validity for those instruments is quite acceptable.

Table 3.23 – Year 3. Correlations between V-OBIS and Cito-scores (N=59)

	Y3 maths	Y3 reading	Y3 total
Year 2			
Begrippentoets (Ideas and concepts)	.54	.47	.53
Taal Kleuters (Language Young Children)	.45	.57	.55
Year 2			
Maths	.57	.56	.59
Reading	.46	.53	.54
Total score	.52	.54	.56

3.4.3 Classroom Profile

In school effectiveness studies it is assumed that variation in instructional practices may influence the achievements of pupils. "Teachers of young children have a unique responsibility to promote children's literacy development. Recent research has demonstrated that access to books, opportunities to write, oral language play, instruction in letter-sound correspondences and story-book reading can significantly accelerate the literacy development of very young children at risk" (McGill-Franzen

and Goatley, 2001). Recent reviews of experimental research support the view that developing teachers' knowledge of various aspects of literacy development by direct instruction and interactive strategies is more effective than providing them with lots of materials and text books.

There were two waves of data collection for all the teachers participating in the study. Using questionnaires, data were collected at the start and at the end of the study.
Furthermore, at the start of the data collection the schools and the teachers were invited to participate in an additional data collection on teaching and learning processes. By means of classroom observations and the use of logbooks additional information on teaching practices should be collected. However, during the instructional meetings, it became quite clear that the teachers were not inclined to participate in those instruments for data collection at all. Their main objections included unnecessary disturbance of the classroom and the time consuming nature of keeping logbooks. As participation in the study was on an optional and voluntary basis, it was decided to refrain from classroom observations and logbooks and to concentrate on the questionnaires and meetings with the experimental schools.

Teachers profile in Year 1
Table 3.24 presents means for the variables in the first questionnaire 'Class and teacher profile year 1'. There were three sets of questions. The first set was about emerging and early literacy and numeracy. As described in Chapter 2, the questions were mainly based on the learning-teaching strategies and intermediate targets for early literacy and early numeracy developed by the Expertisecentrum Nederlands (Verhoeven & Aarnoutse, 1999) and the Freudenthal Institute (Treffers, Van den Heuvel-Panhuyzen, & Buys, 1999).

Five scales on the 'intermediate targets' for early literacy were distinguished, referring to the various aspects of conceptual knowledge and skills of the present curriculum. The teachers were asked to point out to which extent the statements applied to their situation or view along a five-point scale ranging from (1) certainly not to (5) strongly. A high score on the scale 'Ideas of reading 1' indicates a strong emphasis on language development in general, while the scale 'Ideas of reading 2' asks about the extent to which attention is paid to activities promoting rhyming, letter knowledge, links between letters and sound. The third scale 'Story schemas' asks about activities which facilitate the children's comprehension and sense of story structure; and the next scale 'Book orientation' asks for the focus on aspects such as punctuation and printing conventions. The fifth scale 'alphabetical and print principles' refers to activities supporting the understanding of the meaning of print, print awareness and letter knowledge and phonological processing abilities.

Table 3.24 – Year 1: Teacher profile and classroom practices

Intermediate targets	N teachers	N items	minimum	maximum	mean	sd	Cronbach's α
Ideas of reading 1	26	23	2.39	3.91	3.22	.46	.88
Ideas of reading 2	26	6	1.33	4.17	2.76	.78	.88
Story schemas	26	9	2.33	4.78	3.33	.63	.87
Book orientation	26	4	2.00	4.75	3.37	.82	.89
Alphabetical principles	26	4	2.25	4.50	3.56	.57	.77
Early numeracy	26	10	2.90	4.60	3.84	.51	.80
Development and learning	26	2	3.00	5.00	4.31	.65	.81
Monitoring development	26	4	1.00	4.50	2.62	.12	.72
Acquisition 2nd language	13	7	1.71	5.00	4.47	.92	.99
Classroom activities	26	4	2.25	3.00	2.92	.22	.84
Preparing activities	26	2	1.67	3.67	2.59	.66	.74
School organisation	26	9	2.22	3.78	3.20	.40	.77

In the same way, the scale 'Early numeracy' contains ten questions on counting and the understanding and use of ordinal numbers in different meaningful contexts:

- counting on and back in ones, count reliably up to 10 everyday objects;
- saying and using the number names in order in familiar contexts (rhymes, songs, stories, counting games and activities);
- reciting the number names in order, continuing the count forwards or backwards from a given number;
- recognising small numbers without counting;
- estimating a number in the range that can be counted reliably;
- comparing and ordering numbers, the use of language such as more or less, greater or smaller.

The second set of questions is at the heart of teaching practices in the early years of kindergarten; they deal with contrasting views on development and learning which influence the timing and function of early literacy and numeracy instruction. The first known as the Piagetian, is the view that development precedes learning in such a way that learning cannot take place unless appropriate prior cognitive development has occurred. The other, contrasting view, in which readiness plays a minor role, is the Vygotskian position stating that learning leads to development. In this view children are socialized into a set of social practices, beliefs and values, through guided, socially meaningful participation (Johnston & Rogers, 2001). Applying this view to cognitive learning means that by guided learning a higher level of development can be elicited. An example of an item is 'young children learn spontaneously'; and 'also for young children intentional instruction promotes learning'.

The items of the scale 'Monitoring development' ask the teachers about their preference for informal or formal assessment of development and learning: observation, keeping logbooks, using screening lists or assessment by testing. The next scale 'Acquisition of second language' contains items on the pedagogical and didactical activities teachers use to facilitate and optimize the language proficiency of second-language children.

The scales 'Preparing classroom activities' and 'school climate' both refer to a professional attitude towards the class and school organisation as well. Preparing classroom activities indicates the extent to which the daily educational program is prepared by the teacher in advance; the items vary from a very informal program to a precisely prepared program. 'School climate' gives an indication of professional skills such as team work, coaching, sharing knowledge and experiences, contacts with parents, retraining.

Teacher and class profile in Year 2

At the end of the study the questionnaire 'The transfer from year 2 to 3' was filled out by 16 teachers from 7 schools. As described in Chapter 2, differences between teachers in intended and implemented curriculum for reading and maths were investigated. By presenting the intermediate targets to them and asking to indicate which groups of pupils master the targets at the end of Year 2, differences in the implemented curriculum level were identified. By presenting similar targets and asking which aspects and skills were meant to be taught, the intended curriculum was identified.

The content of the questions was derived from the intermediate targets for early literacy and early numeracy developed by the Expertisecentrum Nederlands (Verhoeven & Aarnoutse, 1999) and the Freudenthal Institute (Treffers, Van den Heuvel-Panhuyzen, & Buys, 1999); while the alternative response categories indicating groups of pupils with a distinctive cognitive level compared with other (groups of) pupils, were derived from Jungbluth (2003). The response categories for the items on the implemented curriculum comprised a five-point scale ranging from (1) 'head group in Year 2' to (5) 'all pupils in Year 2: head, middle and tail group'; and the response categories for the items on the intended curriculum formed a four-point scale ranging from (1) 'all pupils in Year 2' to (4) Year 3.

Using factor and reliability analyses, scales for the intended and implemented curriculum were constructed. Although the *N-to-p ratio* where N is the number of participating teachers and p the number of observed variables, is rather critical[3], the results are

3 Guidelines are varying from 2:1 to 20:1.

satisfactory and do not seem to be affected by instability. The descriptive statistics of the scales are outlined in Table 3.25 and Table 3.26.

Table 3.25 – The transfer from Year 2 to 3. Descriptive statistics of scales for the implemented and intended reading curriculum

	N teachers	N items	minimum	maximum	mean	sd	Cronbach's α
Implemented curriculum			Head group (1)	All pupils (5)			
Ideas of Reading	16	10	1.38	4.88	3.52	1.00	.91
Alphabetical principles	16	4	1.00	4.75	2.77	1.31	.91
Book orientation	16	2	3.00	5.00	4.59	0.63	.91
Reading (start)	16	5	1.60	4.20	2.69	0.72	.67
Intended curriculum			All pupils (1)	Year 3 (4)			
Ideas of Reading	16	7	1.00	3.29	1.62	0.80	.91
Alphabetical principles	16	7	1.00	4.43	3.04	1.00	.96
Book orientation	16	2	1.00	4.00	1.75	1.05	1.00

The high mean scores on the scales 'Ideas of Reading' and 'Book orientation' of the implemented curriculum indicate that according to the teachers almost all pupils master these aspects of early literacy at the end of year 2; while only the pupils of the 'head' and 'middle' groups master alphabetical principles and already start to read.
The mean scores on the scales of the intended curriculum are in accordance with this, only the alphabetical principle is intended for the cognitive 'head' group, and 'ideas of reading' and 'book orientation' are aimed for almost all pupils in year 2 to master.

Table 3.26 – The transfer from Year 2 to 3. Descriptive statistics of scales for the implemented and intended maths curriculum

	N teachers	N items	minimum	maximum	mean	sd	Cronbach's α
Implemented curriculum			Head group (1)	All pupils (5)			
Counting up to 20	16	10	0.40	2.60	1.36	0.79	.93
Counting up to 10	16	4	3.25	5.00	4.38	0.61	.92
Use language: more/less	16	5	3.20	5.00	4.15	0.59	.74
Intended curriculum			All pupils (1)	Year 3 (4)			
Counting up to 20	16	8	1.00	4.50	2.82	1.39	.97
Counting up to 10	16	6	1.00	4.00	1.64	1.00	.94
Use language: more/less	16	5	1.00	1.80	1.41	0.42	.93

'Counting up to 10' and the 'use of language such as more or less, greater or smaller' are skills that most teachers intend to teach their pupils in year 2 and of which they say that most pupils master these skills; while counting up to 20 is generally not intended nor implemented in year 2.

In this questionnaire teachers were also asked how often they use direct instruction approaches to support early literacy and reading skills. On a scale of 7 items (Cronbach's $\alpha = .96$) ranging from (1) never to (4) very often, the mean score was 3 (often). On the other hand, it appeared that the use of computer programs to promote reading or maths development was not a regular classroom routine. On a three-point scale of 9 items (Cronbach's $\alpha = .95$) the mean was 2 (incidentally).

3.4.4 The experimental design and the effects of feedback

As described in Chapter 1, in this research project, the test scores collected in the first year function as a reference for the assessment of the progress in the experimental and control groups (see par. 1.4 Methods and design). According to this experimental set up of monitoring the progress of children in various educational settings, regular meetings with the teachers of the experimental groups were organised. The meetings were intended to help the teachers to become more reflective and intentional about their practice: targets, classroom practices and individual pupils were discussed.

Need for feedback
In the meanwhile, the teachers of the control group should not receive any further information on the feedback of the test results. However, during the first instruction meetings with the teachers it became already clear that withholding feedback was practically impossible and ethically not acceptable. Getting feedback on the test results even appeared to be a major reason for schools to participate in the study. So, there was no other option than to provide the control group with some feedback.

The nature of the instrument
Apart from the apparent need of schools for feedback, it was also the nature of the assessment that evoked the interest of the user. The instrument is designed to give immediate information for the teacher to act upon in the classroom. During and after completing the assessment teachers were already getting valuable information about what children know and can do when they started school. As a result, in the course of time schools and teachers grew more and more enthusiastic about the OBIS test.

The experimental design reconsidered

These factors have influenced the experimental design of the study. Whether or not as a result of the assessment done by the teachers themselves, some schools in the control group started to behave as the experimental group E2. They became interested in test scores and asked for more feedback; while at the same time one school assigned to the experimental group E2, had developed so many initiatives that it was hardly possible to make a distinction between this school and the schools in experimental group E1.

This Hawthorne effect caused the differences between the experimental groups to become smaller and experimental and control groups to be more alike. Practically, there was no other option than to admit that after one year of baseline assessment there were two conditions left instead of three as designed: one group E1 receiving all possible information and available feedback to stimulate teaching and learning processes; and the other group E2 just getting feedback on the test-scores. Group E1 received individual information on target-setting, while group 2 was only provided with limited general information.

Teachers and classroom practices

In year 1, several scales on the intermediate targets for early literacy and numeracy were identified as well as scales on the view of teachers on development and learning. What were the differences between teachers of the experimental and the control group in their classroom practices and educational perspectives? Comparing the means for both groups (Table 3.27) differences were found in the extent to which attention was paid to 'book orientation'; their view on monitoring development and also on classroom activities. Compared with the teachers of the control group, the teachers of the experimental group were significantly more focused on book orientation; monitoring development by regular assessment rather than observation; and a variety of classroom activities.

Table 3.27 – Teachers in year 1. Descriptive statistics for scales on classroom practices broken down for the experimental (1) and control group (2)

Classroom Practices	Book orientation		Monitoring development		Classroom activities		
	mean	sd	mean	sd	mean	sd	N
1	3.68	.86	3.04	1.15	3.00	.00	14
2	3.00	.61	2.13	.90	2.83	.31	12
p	.03		.04		.05		
E^2	.18		.17		.15		

About two years later, at the end of Year 2, significant differences were found between the experimental and control group with regard to the implemented reading curriculum (Table 3.28). The teachers of the experimental group scored significantly higher than those in the control group on 'ideas of reading' (IAR) and 'alphabetical principles', which means that almost all pupils master those intermediate targets. No differences were found with respect to book orientation and the intended reading curriculum, nor for the implemented or intended and numeracy curriculum.

Table – 3.28 Transit from Year 2 to Year 3. Descriptive statistics for scales on the implemented curriculum for early literacy broken down for the experimental (1) and control group (2)

Implemented curriculum	IAR		Alphabetic principles		Book orientation		Reading (start)		
	mean	sd	mean	sd	mean	sd	mean	sd	N
1	4.04	0.68	3.31	1.15	4.72	0.79	2.91	0.44	9
2	2.86	.99	2.07	1.23	4.42	.83	2.40	.53	7
p	.01		.06		.38		.17		
E^2	.36		.23		.06		.13		

Pupil achievements in the experimental and control group

How did the achievements of the pupils in the experimental and control group develop in the course of time? Table 3.29 displays the standardised differential scores indicating the progress made between year 1 and year 2 (Year 1-2), year 1 and year 3 (Year1-3), and year 2 and year 3 (Year 2-3) in case of both groups. By using the differential scores the change from baseline assessment is measured. With a mean differential score of 50, it can be seen that in general the experimental group achieved better than the control group.

Unlike the differences between years 1 and 2, the differences between years 2 and 3 are no longer significant. The progress in reading and total scores made by the experimental group should probably be accredited to the progress made in the first year that the pupils are at school.

Table 3.29 – Descriptive statistics for standardized differential scores of the OBIS broken down for the experimental (1) and control group (2)

	Maths score		Reading score		Total score		
Year 1-2	mean	sd	mean	sd	mean	sd	N
1	50.31	10.64	51.46	10.54	51.11	10.60	217
2	49.64	9.21	48.28	9.04	48.70	9.10	185
p	.51		.001		.02		
E^2	.00		.03		.02		
Year 1-3	mean	sd	mean	sd	mean	sd	N
1	51.05	10.53	50.69	10.90	50.89	10.82	186
2	48.38	8.90	48.93	8.34	48.63	8.44	121
p	.02		.13		.05		
E^2	.02		.00		.01		
Year 2-3	mean	sd	mean	sd	mean	sd	N
1	50.68	9.99	49.83	10.63	50.20	10.54	181
2	49.00	9.95	50.27	8.93	49.70	9.17	124
p	.15		.69		.67		
E^2	.00		.01		.00		

3.4.5 School profile

According to literature on school effectiveness, high achievement of pupils is associated with schools characterised by a strong educational leadership, high expectations, emphasis on basic skills, a safe and well-structured school organisation and above all frequent monitoring of learning achievement.

Information on school background was collected at class level and completed with the school reports from the Inspectorate of Education. The inspectorate conducts periodical assessments of the quality of educational institutions. The schools' own evaluation is an important input for the external evaluation by the Inspectorate. According to the Inspectorate (2004), in recent years the Dutch primary schools were doing fine with regards to the effective management of learning time (time on task), and a safe and well-structured pedagogical school climate, while the continuity in the educational program, pupil's learning strategies, and adaptive education were considered to be major weak points.

School performance indicators

Scheme 3.3 presents an overview of the Inspectorate's judgments of the schools in the study in the period 2000-2004. The judgments of the schools' achievement performances are based on aggregated data of the pupil performances in year 8 at the end of primary education provided by the schools themselves (Inspectie van het Onderwijs, 2004).

Scheme 3.3 – Overview of the Inspectorate's judgements of the schools in the study

School number	1	2	3	4	5	6	7	8	9	10	11
*Instrument ISP**	RST	RST	PKO	JO	JO	PKO	JO	JO	JST	PKO	RST
Year of assessment	2000	2001	2004	2003	2003	2003	2003	2003	2001	2004	2001
SES category	1	2	2	2	4	2	1	1	3	2	1
Achievements											
a. End of PE	S	S	S	S	S	S	S	S	I	S	S
b. Continuity	S	-	S	S	-	-	-	-	S	S	-
Teaching process											
a. Teachers/pupil	-	-	S	-	-	S	-	-	S	S	-
b. Safety	-	-	S	-	-	S	-	-	S	S	-
c. Motivating	-	-	S	-	-	S	-	-	S	S	-
Methods materials											
a. Modern	I	S	S	-	-	S	-	S		S	S
b. Adaptive	-	S	S	-	-	S	-	S	I	S	S
c. Coherence	I	S	S	-	-	I	-	I	I	S	S
Quality											
a. Lessons	S	S	S	-	-	S	I	S	S	S	S
b. Independent	S	S	S	-	-	S	I	S	S	S	S
c. Heterogeneity	S	S	S	-	-	I	I	I	S	S	I
d. Individual	S	S	S	-	-	I	I	S	S	S	I

* RST: Regulier Schooltoezicht (Regular Supervision)
JST: Jaarlijks Schooltoezicht (Yearly Supervision)
PKO: Periodiek Kwaliteitsonderzoek (Periodically Quality Assessment)
JO: Jaarlijks Onderzoek (Yearly Assessment)

** S is 'sufficient, adequate'; I is insufficient, inadequate'.

Compared with multilevel analyses this method is considered as relatively accurate (Veenstra, Bleker & Knuver, 2004). The Inspectorate collected the school data using different instruments: until 2003 via regular or annual supervision (RST and JST, 'Regulier Schooltoezicht' and 'Jaarlijks Schooltoezicht', respectively) and after that

via periodical quality assessment and annual assessment (PO and JO, 'Periodiek Kwaliteitsonderzoek' and 'Jaarlijks Onderzoek', respectively). Irrespective of other differences between the systems, the similarity in content is such that the comparison of data is relevant.

The disadvantage of the Inspectorate's performance indicator is that it is not a genuine value-added indicator calculated on on-entry baseline scores and a follow up test score (Janssens & Visscher, 2004). Ideally, the value-added indicator on school level should indicate the average of the difference between their individual tests results measured at different points of time. When value-added is measured on the pupil by pupil basis, the data can be accumulated to provide a single indicator for the school.
Scheme 3.4 shows differential scores calculated on pupils' on-entry baseline scores and their achievements scores one year later. It is an immediately understandable score: schools with an average value-added score of zero have kept up with other schools. According to Fitz-Gibbon (1997) "the average progress is the appropriate progress."

Whereas on pupil level the correlations between the OBIS total baseline scores in year 1 and the later OBIS test scores in year 2 and 3 and between the latter years are close (.78, .68 and .64< respectively), the correlations between the scores of pupils' *progress* for years 1-2, years 2-3 and years 1-3 show a considerable variation. The correlation between the distinct progress in year 1-2 and year 2-3 is -.32, which means that the pupils 'slow' on progress in the first 2 years of kindergarten caught up in year 3. Pupils' progress in years 1-2 correlates at .35 with the progress in the period of year 1 to year 3. On the other hand, the progress between years 2-3 and years 1-3 correlates at .76.

The value-added measures on school level (Scheme 3.4) show considerable year-to-year variation. According to Fitz-Gibbon (1997) this variation may simply be inherent to a complex system such as schooling. For example, children from disadvantaged families may gain up throughout their first year at school compared with other children and even make initially more progress than their more advantaged peers; while the progress of the latter children stays much the same over the same period of time.

Scheme 3.4 – Mean value-added score (OBIS) per school over 3 successive years

School number	1	2	3	4	5	6	7	8	9	10	11
SES score	1	1.11	1.13	1.11	1.83	1.11	1.05	1.04	1.57	1.11	1.03
Value-added school											
Year 1-2	-2.70	-2.50	-.95	4.87	3.52	-.75	-.95	-2.52	.26	-2.42	-.21
Year 2-3	7.01	-1.91	-3.09	-7.12	.63	-.5.46	-8.83	4.40	-3.42	-2.49	-1.41
Year 1-3	4.99	-5.24	-3.69	-3.86	5.07	-6.03	-9.48	2.86	-.299	-4.47	-3.31

Teachers are, above all, interested in the performance of each pupil and the progress made by the pupils in their class. The scores can help to detect lagging behind of pupils in good time. While the value added is the fairest indicator for academic success so far, it is tempting to use the value-added score as a key parameter for accountability or financing of schools. This would meet serious concerns. Fitz-Gibbon (1997, par. 3.4 and 4.8) outlined the implications clearly, out of which the most significant are that value added measure vary from subject to subject and from year to year. Compared with differences between pupils the differences between schools are relatively small and will also frequently be placed within a large margin of uncertainty.

3.5 Concluding remarks

This chapter addresses the data collection and results of the study and forms as such the transition from the descriptive to the explanatory part of this report. By setting out the relationships between various measures there is also a starting point created for further analyses. Namely, if the outcomes of conventional statistical technique are of no significance, there is little point in carrying out further analyses, testing hypotheses or complicated models. If relationships appeared to be acceptably strong there is a good reason for further exploration using linear structural models (longitudinal analysis) or hierarchal linear models (multilevel analysis). Relevant analyses of this type are given in Chapter 7.

In the course of the years the data collection of the study was not affected by loss of schools, while the differences in the number of pupils on distinct measure moments were caused by natural causes including mobility and transfer within the school itself. Based on the results described in this chapter, the data collection and the quality of the instruments may be called satisfactory. It is the design of the study that seems to have deteriorated over the years mainly due to processes initiated by exogenous factors.

Confrontation of the experimental design with real life in the schools

One of the questions in this study was whether feedback on Obis test results given to the teachers would positively influence the progress of their pupils. Until now, the study design included three groups of schools receiving increasing levels of feedback (from none to high) on Obis test results. The data subsequently collected in this chapter are in support of the hypothesis, as the overall scores on maths and reading are generally higher when feedback is given. Although the effects are rather small, there are reasons to believe that the effect size is underestimated.

In The Netherlands in the mid nineties it was still possible to collect data on pupil achievement in primary schools for research purposes while giving the schools hardly any feedback or insight in the results of the study. Particularly regarding baseline assessment meeting difficulties by having a control group not receiving feedback is not to be expected. It was a period of relative silence and rest around baseline assessment at the entry of primary education – the effect of which was probably reinforced by the negative advice of the Commissie Indicatiestelling Onderwijs-achterstanden (Committee Indicating Educational Disadvantage, 1996) on baseline assessment for young children.

In the year 2000, at the time the data collection in the schools started, the situation apparently had changed. There was a shift in the educational policy of the government characterised by a growing interest in educational standards and goal setting, and also rising debate on the positive and negative aspects of baseline assessment in general. Schools also took part in the debate, thus alerting teachers of the various aspects of early testing. In this situation it was no longer possible to withhold feedback on test results from the participating schools.

In conclusion, during the study major environmental factors eroded the design and affected the external validity of the study. These so-called 'reactive' effects limited the generalizability of the results as far as it regards the effects of the type and amount of feedback teachers received (Tuckman, 1994). This so-called Hawthorne effect caused that the differences between the experimental groups became smaller because experimental and control groups were more alike.

It is sobering to recognize that a straightforward experimental design does not survive the challenge with the real life conditions in the schools.

3.6 References

Batenburg, Th.A. van (2002). *Een instrument voor het observeren van leerkrachten bij Technisch lezen in groep 3 en een eerste indicatie over effecten van het Groninger Experiment Achterstandsbestrijding. GEA*. Groningen: GION.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ, Lawrence Erlbaum Associates.

CEM Centre (1999). *Using the PIPS Baseline Assessment 1999/00.* Durham: CEM Centre.

CEM Centre (1999). *Technical report. Text version.* Durham: CEM Centre.

CEM Centre (2004). *Using the PIPS Baseline Assessment 2004/05.* Durham: CEM Centre.

Driessen, G., Van Langen A., & Vierke, H. (2004). *Basisonderwijs: veldwerkverslag, leerlinggegevens en oudervragenlijst. Basisrapportage PRIMA-cohortonderzoek*. ITS: Nijmegen.

Fitz-Gibbon, C. (1997). *The Value-Added National Project: Feasibility Studies for a National System of Value Added Indicators (Final report).* London: School Curriculum and Assessment Authority.

Inspectie van het Onderwijs. (2004). *Onderwijsverslag 2002/2003*. Utrecht: Inspectie van het Onderwijs.

Janssens, F., & Visscher, A. (2004). Naar een kwaliteitskaart voor het primair onderwijs. [Towards a School Report for Dutch primary schools.] *Pedagogische Studiën, 81,* 371-383.

Johnston, P., & Rogers, R. (2001). Early Literacy Development: "The case for informed Assessement". In S.B. Neumann & D.K. Dickinson (Eds.), *Handbook of Early Literacy Research*. New York /London: The Guilford Press.

Jöreskog, K.G., & Sörbom, D. (1983). Lisrel, analysis of linear structural relationshipps by the model of maximum likelihood. Version 6.

Jungbluth, P. (2003). *De ongelijke basisschool*. Nijmegen: ITS.

Luyten, H. (1994). *School effects: stability and malleability*. Dissertation. Enschede: Copy Print.

McGill-Franzen, A., & Goatley, V. (2001). Title I and Special Education: Support for Children Who Struggle to Learn to Read. In S.B. Neumann & D.K. Dickinson (Eds.), *Handbook of Early Literacy Research*. New York /London: The Guilford Press.

Treffers, A., Van den Heuvel-Panhuyzen, M., & Buys, K. (1999). *Jonge kinderen leren rekenen. Tussendoelen Annex leerlijnen. Hele Getallen Onderbouw Basisschool*. Groningen: Wolters-Noordhoff.

Tuckman, B.W. (1994). *Conducting Educational Research*. Fort Worth: Harcourt Brace & Company.

Van der Hoeven-van Doornum, A.A. (1990). *Effecten van leerlingbeelden en streefniveaus op schoolloopbanen. [*Effects of teacher expectations and aspiration levels on school careers; Dissertation with a summary in English]. Nijmegen: ITS.

Van der Hoeven-van Doornum, A.A., Voeten, J.M.J., & Jungbluth, P. (1993). The influence of parents' expectations on school careers. *Tijdschrift voor Onderwijsresearch, 18* (6), 369-379.

Van der Wel, J.J., & Krooneman, P.J. (2003). *De implementatie van 'Tussendoelen beginnende geletterdheid'*. Regioplan, publicatienr. OA-257. Regioplan beleidsonderzoek: Amsterdam.

Veenstra, R., Bleker, L., & Knuver, A. (2004). Accuratesse en gebruik van schoolprestatie-indicatoren. [Accuracy and usefulness of two publications of school achievement.] *Pedagogische Studiën, 81,* 354-370.

Verhoeven, L., & Aarnoutse, A. (1999). *Tussendoelen beginnende geletterdheid. Een leerlijn voor groep 1 tot en met 3.* Nijmegen: Expertisecentrum Nederlands.

Chapter 4 What baseline assessment is doing for children's progress and teachers' professionalism

Paper presented at the ECER 2002 Lisbon

Summary

Eleven primary schools participated in the study, and at the start 450 children joined in, of which 400 were still present after two years. The aim of the study was to record the effects of regular testing on pupils' learning progress and teachers' professionalism. This was done by longitudinal assessment of the OBIS test scores and the individual targets set for each pupil by the teacher. Measurements at the teacher level included the way teachers stimulate their pupils' learning and other instructional characteristics. The data describe in particular the effects of testing on performances of pupils with low socio-economic status and minority groups whose educational opportunities lags considerably behind that of their peers. It appeared that after one year of education, their chances had improved. Although their performances were still below average, the relative progress of children of ethnic minorities was significantly higher than that of other Dutch pupils. There is also evidence that a feedback effect occurred, positively influencing the learning achievements, in particular in reading abilities. The feedback effect existed, despite of the fact that the experimental conditions were difficult to maintain so that groups became more alike in the course of the experiment (Hawthorne effect).

4.1 Introduction

Monitoring the learning achievements of pupils in primary schools may serve various purposes including improvement of the quality of education and, at present, providing academic standards for additional financial help for schools and students where it is needed most (Title I, 2002)[1]. With regard to the quality of education, it is generally accepted that regular testing can improve the achievements of pupils by influencing

1 In the USA the importance of assessment is strongly emphasized in the No Child Left Behind Act of 2001. The federal government expanded its role in public education with new legislation to motivate annual student performance testing, teacher improvement programs, and a plan to identify underperforming schools (Hiebert, Gallimore & Stigler, 2002).

their teachers. A good feedback of the pupils' test results may also help the teachers to perform better. Yet, little is known about the size of the feedback effects (Visscher & Coe, 2002).

The present project was aimed at measuring the effects of regular testing, and the use of explicit achievement goals on the cognitive development and school performance of young pupils (age 4-6). This may answer the question whether the level of development at a certain time is an indicator of future school performance, and to what extent the progress in development can be stimulated by regular testing. Essentially, the research question is derived from the educational model, depicted as follows:

Assessment initial level -> diagnosis -> (steering) learning process -> assessment of progress

At present, learning disadvantage, achievement assessment, and educational opportunities are related issues. Since the mid-eighties Dutch schools have been receiving extra financing for additional staff, based on the socio-economic and ethnic composition of the school population (Driessen, 2000). The underlying theory of this policy was that immigrant children and working-class children are considered to have comparable educational disadvantages mainly caused by the socio-economic position of their parents. The aim of this educational priority policy (EPP) was to try to eliminate or reduce these children's disadvantages. Despite of the fact that this policy has been sharply criticised as weak and vague (Mulder, 1996), and that since the beginning of the 1990s many studies indicated that the policy did not result in the desired effect, it is only now that politics and opinions are changing in The Netherlands. Probably likewise the No Child Left Behind Act in the USA (2001), it was proposed by the Dutch government (2002) to introduce an intake test for 4-year old children.

In his latest advice to the Minister of Education, August 2002, the Onderwijsraad (Educational Council of The Netherlands) recommends to distinguish between group related educational disadvantage and individual reading underachievement. Next to that, recommendations were made on the implementation of apparently successful educational programs such as Reading Recovery.

4.2 Theoretical framework of the project

It is Vygotsky's concept of the zone of proximal development (ZPD) that offers a theoretical framework for studying relations between development and educational intervention. ZPD is, according to Schneuwly (1994), Vygotsky's response to a difficult problem, namely, the tension between the internal self-propulsing mechanisms of

development and the impact of external cultural forces, as embodied in instruction. ZPD is a relational concept that describes the social interactions that allow progress and it is not measurable in terms of quantifiable information, for example the size of its upper limits. Stated otherwise, ZPD makes the interactions in the teaching-learning process from which a child can profit understandable. The concept of ZPD also links assessment and instruction as tools for regulating learning and fostering development, especially in school settings.
Research suggests that by giving teachers feedback, pupils' performances can improve (Coe, 2002). This is one of the aims of all monitoring systems – that good feedback helps the teacher work better. But we do not know how big this effect is when compared with no feedback at one hand and the combination of feedback and individual target setting at the other hand. To optimise the learning process teachers need the understanding of subject-matter and didactic knowledge; explicit targets for individual pupils and on group level; feedback of the results, that is pupils' progress. The following research question was here investigated: 'what does baseline assessment for children's progress and teachers' professionalism?

4.3 Methods

The investigation was carried out in relation to 4-6-years-old children in Dutch primary schools. The aim of this study was to investigate the educational impact of regular testing on the development and school performances of young children. The PIPS Reception Assessment-test was selected as an adequate intake test for use in this research project titled 'Development on scale, instruction at measure'. This baseline test was developed by Tymms and colleagues in Durham, the test was translated into Dutch and adapted where necessary.
The reasons to select the PIPS test were its balanced content and efficient construction and attractiveness for young children in general. For example, unlike many other tests, the PIPS can be taken in a short time by the teacher or any other competent adult. Furthermore, there were methodological considerations and criteria: validity, reliability, correlation and prediction are sorted out very well. The cognitive measures generated within the PIPS are reported as relative measures (norm-referenced). They are based on representative data and are expressed as T scores with a mean of 50 and a standard deviation of 10 (Tymms & Albone, 2002).

The investigation takes four years. The first year was used for instrument development and testing. The Dutch version of the PIPS is called OBIS which is the acronym for Onderbouw Informatiesysteem (lower school information system). In the second and later years the quasi-experimental investigation was, and is being carried out with about 300 children in 26 classes of 11 primary schools. Measurements at the pupil

level are test scores, target-setting, and implicit goals set by the teacher. Measurements at the teacher level aim to record the way teachers stimulate their pupils in acquisition of knowledge and skills in play and learning activities. The schools are randomly distributed among 2 experimental groups (E1 and E2) and one control group (C3). The factorial design is shown in Scheme 4.1 below.

Scheme 4.1 – Design of the study with two experimental groups and a control group

Year 1				Year 2				Year 3	Condition
T1	ET1	Ex1	LP1	T2	ET2	Ex2	LP2	T3	
+	+	+	+	+	+	+	+	+	E1
+	(+)	(+)	+	+	(+)	(+)	+	+	E2
+	(+)	(+)	(+)	+	(+)	(+)	(+)	+	C3

(+) no feedback on OBIS test scores given to the teachers

Measurements at the pupil level were:

- T1-T2, test scores OBIS, Year 1-2.
- T3, test scores reading and mathematics, Year 3.
- ET1 and ET2, explicit targets, Year 1-2.
- Ex1 and Ex2, teacher's expectations per pupil, Year 1-2.
- LP1 and LP2, teaching processes, Year1-2, questionnaire, logbook or observations in classroom Year 1-2, to evaluate the teachers' performances to activate their pupils in the acquisition of knowledge and skills.

Measurements at class level were:

- Questionnaire for school staff, comparable with those used in earlier school effectiveness study.

Feedback

The experimental group E1 got all available information and feedback on pupil level and class level that stimulates teaching and learning processes, including:

- test scores on the OBIS;
- expectations, compared with expectations for children with similar OBIS scores;
- individual targets;
- support of the teaching and learning process.

The experimental group E2 gets information on the individual test-scores and information on target-setting on class level. The control group C3 receives only feedback on test scores.

4.4 Results and discussion

In the process of translation from English into Dutch, one of our major concerns was if there would be any essential differences in culture between the English and the Dutch context. So we checked the test content as well as we could for differences in culture, developmental level and typical socio-ethnic situations. It is obviously important to avoid unbalances of the test, and putting children with low socio-economic status, minority groups, and boys or girls at a disadvantage. At careful inspection and comparison with several Dutch tests, including tests for young children with Dutch as an additional language, we didn't find topics or items that had to be adapted or deleted. In fact, only a few items required some adaptation. For most of the items a proper translation was sufficient. It seems clear, that in the section Rhyming alterations were necessary to account for differences between the languages.

Prior to the explanatory multivariate analyses, the descriptive statistics of the various sections of the OBIS were examined to identify the distribution characteristics and quality of the test. Comparison of the distributions of Dutch and English test scores in Year 1 showed that the English pupils performed best on the sections Name Writing, Letter Identification, Counting and Digit Identification. The Dutch pupils performed better on the sections Picture Vocabulary, Repeats, Rhyming, Words, Ideas of Maths and Sums. For the sections Ideas about Reading, Words, the patterns of the distributions of the English and Dutch scores appeared to be quite similar. At this moment there is no comparison available between the English and Dutch scores after one year of school. However, in Table 4.1 the correlations between the item difficulty parameters[2] of the English and Dutch scores indicate strong relations between the scores of children with different cultural backgrounds and age.

2 Difficulty measures were estimated using Rasch analysis. Rasch estimates are based on probability of success to probability of failure ratios (Bond & Fox, p.133). The measures are based on a probabilistic relation between any item's difficulty and any person's ability.

Table 4.1 – Correlations between the item difficulty parameters of the English and Dutch scores

Vocabulary	English	Dutch 4-5 year
Dutch 4-5 year	.74	
Dutch 5-6 year	.80	.98
Letters	English	Dutch 4-5 year
Dutch 4-5 year	.90	
Dutch 5-6 year	.76	.83
Mathematics	English	Dutch 4-5 year
Dutch 4-5 year	.90	
Dutch 5-6 year	.92	.98

The comparison of Table 4.2 and 4.3 shows increases in the pupils' raw scores for reading and mathematics, one year after their entering into primary school. It can be seen that the Cronbach's alpha scores were overall high, indicating a high reliability of the scales. In Year 1 the average age of the Dutch pupils (N=445) was 4 years and 6 month.

Table 4.2 – Descriptive statistics for the sections of the OBIS. Raw scores of 445 pupils that started (Autumn assessment) or just entered Year 1 (Spring assessment) of primary school

	mean	sd	min-max	N items	Cronbach's α
1 Name writing	1.80	1.50	0 – 5	1	-
2 Picture vocabulary	15.15	5.50	0 – 22	22	.93
3 Ideas about Reading	3.76	1.70	0 – 10	10	.72
4 Repeats	6.43	2.06	0 – 8	8	.85
5 Rhyming	4.01	3.57	0 – 9	9	.94
6 Letter identification	2.50	4.48	0 – 26	27	.95
7 Words	0.19	1.20	0 – 14	14	.94
8 Ideas about Maths	6.06	1.40	0 – 7	7	.80
9 Counting	2.59	1.47	0 – 4	4	.81
10 Digit identification	4.36	4.85	0 – 20	20	.93
11 Sums	4.25	2.53	0 – 8	8	.86

Table 4.3 – Descriptive statistics for the sections of the OBIS. Raw scores of 301 pupils taking part in the Autumn assessment one year later

	mean	sd	min-max	N items	Cronbach's α
1 Name writing	3.25	1.50	0 – 5	1	-
2 Picture vocabulary	17.96	3.79	0 – 22	22	.84
3 Ideas about Reading	5.35	2.27	0 – 10	10	.74
4 Repeats	7.43	1.19	0 – 8	8	.73
5 Rhyming	6.87	2.77	0 – 9	9	.90
6 Letter identification	6.67	6.93	0 – 27	27	.95
7 Words	2.08	4.49	0 – 14	14	.98
8 Ideas about Maths	6.54	.97	0 – 7	7	.66
9 Counting	3.59	.91	0 – 4	4	.70
10 Digit identification	9.98	5.73	0 – 22	22	.93
11 Sums	6.00	1.81	0 – 8	8	.75

At the start of their school education, children from families with a high socio-economic status score on average one standard deviation higher than their peers from minority groups (children from a non-Dutch cultural background whose parents have a low level of education and low-skilled occupations). The largest differences between the socio-economic groups were found in the scores in the reading section. Table 4.4 displays the differential scores for reading, maths, phonics and total scores between the assessments at the start (Year 1) and after one year (Year 2). Regarding the sections of maths and phonics, there were no significant differences found between the progress made by children from various SES backgrounds. A striking result is the progress after one year of education made by the children from SES-group 3. Their progress in reading is approximately one third higher than the progress of the Dutch children.

Table 4.4 – Descriptive statistics of the differential raw scores of the OBIS between Year 1 and Year 2, broken down by socio-ethnic status of parents (SES), indicating degree of disadvantage

SES	Maths		Reading		Phonics		Total		
	mean	sd	mean	sd	mean	sd	mean	sd	N
	11.50	6.9	12.79	8.2	3.53	3.9	27.83	16.98	292
1	11.96	5.7	11.50	14.4	3.33	3.6	26.79	16.62	184
2	11.24	6.2	10.20	8.5	4.04	4.3	25.49	13.38	45
3	10.32	7.9	18.42	12.18	3.79	4.8	32.54	19.53	63
P	.20		.00		.46		.04		
E2	.01		.05		.00		.02		

1. Children without disadvantage (parents with middle and higher education and occupation.
2. Children from a Dutch cultural background whose parents have a low level of education and low-skilled occupations.
3. Children from a non-Dutch cultural background whose parents have a low level of education and low-skilled occupations.

Additionally, multi-level modelling was used to examine the differences in pupils' scores due to individual and group characteristics. Multi-level is seen as appropriate to study the variation between pupils and classes or schools. Separate models for reading and mathematics were analysed. Starting their school careers at the age of four, young children already differ significantly in the level of cognitive development.

Table 4.5 – Differences between pupils and classes in reading and maths scores at the start of primary education and after 1 year. N=291

	Reading 1		Reading 2		Math 1		Math 2	
Variance components								
Intercept	50.40	(1.34)	50.10	(1.07)	50.54	(0.93)	49.96	(0.97)
Pupils	66.11	(5.72)	77.44	(6.70)	80.88	(6.00)	84.35	(7.30)
Class	37.03	(12.41)	20.32	(7.86)	13.48	(5.94)	15.29	(6.56)
Total	103.14		97.76		94.36		99.64	
% unexplained variance								
Pupil	64%		92%		86%		85%	
Class	36%		8%		14%		15%	
Deviance	2093.80		2125.13		2130.15		2143.70	
Df	3		3		3		3	

Table 4.5 shows the extent to which the variation in cognitive development can be ascribed to differences between pupils and classes. At the beginning of Year 1, the differences between classes in reading scores are surprisingly high; specially in comparison with the maths scores. This remarkable result occurs due to differences in social composition of the class. After one year of schooling the variation in reading scores on class level is much smaller, and even smaller than the variation on class level in maths scores that remained almost the same as the year before. Probably, these results indicate that education might have been compensating for differences in school readiness due to socio-ethnic background (SES).

If baseline assessment and feedback are meant to help teachers to improve their pupils' learning, it is important to identify which indicators provide good quality information. Hence, the influence and contribution of SES and prior learning was explored.
Table 4.6 shows the results for reading scores and Table 4.7 gives the results for mathematics. Compared with the null model in Table 4.5, SES appeared to explain 8% and 69% respectively of the variance in reading scores on pupil and class level, dropping back to 3% and 0% in Year 2.

In Year 1 and Year 2 hardly any variation in mathematic scores was found on the pupil's level, while the variation on the class level increased from 28% in Year 1 to 44% in Year 2.

Subsequently, prior learning achievement was added to the model in order to explain the math and reading scores in Year 2 (Table 4.6 and 4.7, model 2b). Now, the contribution of SES to the model was no longer significant, which applied to the reading scores and math scores, as well. After one year of schooling, approximately half or more of the variation in reading and math scores on both levels, pupil and class, could be explained by the developmental level at entering primary school.

Table 4.6 – Reading scores. Background characteristics of pupils and classes explaining differences between pupils and classes at the start of primary education and after 1 year. Unstandardised regression weights and standard deviations. N=291

	Reading 1		Reading 2a		Reading 2b		Reading 2c	
Pupil								
SES	-14.03	(1.82)	-6.87	(2.14)	2.18	(1.60)	0.92	(1.60)
OBIS Y1					0.74	(0.05)	0.72	(0.05)
Class								
Book orientation							2.73	(0.77)
Variance components								
Intercept	67.83	(2.4)	58.62	(2.86)	9.97	(3.59)	3.11	(3.75)
Pupils	60.93	(5.30)	74.78	(6.48)	40.02	(3.47)	39.70	(3.44)
Class	11.58	(4.93)	20.64	(7.87)	8.51	(3.48)	4.80	(2.39)
Total	72.41		95.42		48.53		43.50	
% unexplained variance								
Pupil	84 %		78%		82%		91%	
Class	16 %		22%		18%		9%	
% explained variance compared with null model								
Pupil	8%		3%		48%		49%	
Class	69%		0%		58%		76%	
Deviance	2042.80		2108.63		1922.80		1911.73	
Df	4		4		5		6	

To answer the question whether pupils benefit from a particular teaching approach, several indicators of teaching characteristics were added to the model. The indicators included various teaching practices to stimulate early reading and early mathematic competence skills. After controlling the contribution of SES and the prior level of cognitive development, it appeared that out of the various indicators for early reading teaching approach only a strong emphasis on books and stories[3] had a small, but significant effect on the reading scores. The explained variance of a 'book-and-

3 'Book orientation' refers to relations between cognitive development and educational instruction. The scale contains questions as "To what extent do you give attention to the following aspects of reading development: knowing that books are read from front to back; b) knowing that books are read top down; c) lines are read from left to right; and d) stories generally have a structure".

stories-oriented' approach was 18%. We could not find any effect of regular use of such teaching methods on the mathematics scores in Year 2.

Table 4.7 – Mathematics scores. Background characteristics of pupils and classes explaining differences between pupils and classes at the start of primary education and after 1 year. Unstandardised regression weights and standard deviations. N=291

	Maths 1		Math 2a		Maths 2b	
Pupil						
SES	-6.22	(1.97)	-6.55	(1.97)	-2.78	(1.44)
OBIS Y1					0.79	(0.05)
Variance components						
Intercept	58.23	(2.57)	58.02	(2.56)	6.37	(3.43)
Pupils	79.84	(6.92)	84.32	(7.30)	40.63	(3.52)
Class	9.66	(4.79)	8.56	(4.60)	3.89	(2.15)
Total	89.50		92.88		44.52	
% unexplained variance						
Pupil	89%		91%		91%	
Class	11%		9%		9%	
% explained variance compared with null model						
Pupil	1%		0%		52%	
Class	28%		44%		75%	
Deviance	2114.35		2127.84		1915.68	
Df	4		4		5	

The last research question so far was whether the beneficial influence of 'book orientation' could be the result of the type of feedback given to the teachers. Concerning the experimental design of the study, it was noted that in the course of time, schools and teachers grew more and more enthusiastic about the OBIS test. Whether or not as a result of the assessment done by the teachers themselves, the control group started to behave as the experimental group E2. They became interested in test scores and asked for more feedback; while at the same time one school assigned to the experimental group E2, had developed so many initiatives that it was hardly possible to make a distinction between this school and the schools in experimental group E1.

Practically, there was no other choice than to admit that after one year of baseline assessment there were two conditions left instead of three as designed: one group E1 receiving all possible information and available feedback to stimulate teaching and learning processes; and the other group E2 just getting feedback on the test-scores. Next to that, group E1 received individual information on target-setting, while group 2 was only provided with limited general information. Analysis of variance between the two types of feedback showed a significant difference between both groups on the scale 'book orientation ($F = 5.137$; $df=1$; $p = 0.03$), indicating that the teachers in both groups vary in the extent to which they put emphasis on this aspect of early reading.

4.5 Conclusion

It is generally accepted that improving classroom teaching is the best way to increase students' learning. It is also obvious that in order to improve classroom teaching in a steady, lasting way, the teaching profession needs the best available knowledge and instruments that the educational researchers can provide. The present study was focused on what baseline assessment does for pupils' progress and teachers' professionalism. The results here presented deal with the following aspects of learning and teaching: baseline assessment, progress, performance indicators, feedback and teaching practices.

We have used the OBIS test, the Dutch version of the PIPS, to assess the level of cognitive development of four-year-old children at the beginning of their school career. So far, the data of two consecutive years show that the OBIS may be considered as a reliable instrument for baseline assessment. This is in line with the positive and enthusiastic responses from the teachers participating in the study. At present, additional data collection is carried out to investigate concurrent and predictive validity of the instrument. Just like PIPS, OBIS aims to monitor pupil progress in specific curriculum areas using prior attainment as the starting point (Tymms, 2002). From the perspective that every child deserves a fair chance of a successful school career, the starting point of children of ethnic minorities lags considerably behind that of their peers. It appeared that after one year of education, their chances have been improved. Although their performances were still below average, compared with their Dutch peers without or with less educational disadvantage, the progress of children of ethnic minorities was significantly higher.

The results of multilevel modelling indicated that if there is no other information available, the socio-economic background of a child is the best predictor of later learning performances. However, the meaning of SES as a performance indicator

becomes insignificant when better information on the competences and skills of the child is added to the model. Furthermore, it appeared that a book-and-stories-oriented environment has a favourable influence on early reading competence and skills of young children. Since the teachers participating in the study were assigned to groups receiving different types of feedback, we also compared their educational targets and practices. The teachers who were provided with individual help and feedback demonstrated a significantly stronger approach orientated on books and stories. Although this finding supports the hypothesis that individual feedback stimulates the pupils' learning, this conclusion is still premature. Since the school environment is dynamic and constantly changing, we also should take into account the possibility that experimental designs are not appropriate for longitudinal studies.

The present findings give considerable confidence in the quality of the OBIS as an instrument to predict progress. The baseline test provides the teacher with immediate information on the school readiness of a child compared with peers as well as a baseline for measuring the progress. Conversely, it is almost common knowledge, SES might be a good predictor, but when used as a 'performance indicator' it also has a sustaining, and self-fulfilling effect on educational opportunities via teacher's expectations (Van der Hoeven-van Doornum, 1990).

What can we expect of the teachers' responses to straightforward, reliable information about their pupils on entry to school and their relative progress in the first year? Do they adapt their practices? According to Coe (2002, p. 3) "The evidence about feedback effects is mixed, complex, and not well understood." In the present study we have obtained evidence that a feedback effect exists, positively influencing the learning achievements, in particular reading abilities. This may be so under the experimental conditions in our study, but would it also apply to the natural conditions in the classroom? In this regard, it has been argued by Taylor Fitz-Gibbon (2002) that there may be a significant discrepancy between the effects in experimental and natural situations. We are aware that for any alternative approach to win acceptance, it is to prove why it is a better choice that the existing system.

Acknowledgement

I would like to express my gratitude to the Netherlands' Organization for Scientific Research (NWO) for funding by grant 411- 210-05 the project 'Development on scale, instruction at measure' on which this paper is based.

4.6 References

Andrews, S. P., & Slate, J. R. (2001, May 17). Prekindergarten programs: a review of the literature. *Current Issues in Education* [On-line], 4 (5).

Bond, T., & Fox, C.M. (2001). *Applying The Rasch Model. Fundamental Measurement in the Human Sciences.* New Jersey: Lawrence Erlbaum Ass.

CEM (2001). *Technical Report 2001*. Durham: Curriculum Evaluation and Management Centre, University of Durham.

Coe, R. (2002). Evidence on the Role and Impact of Performance Feedback in Schools. In A.J. Visscher and R. Coe (Eds.), *School Improvement through Performance Feedback.* Lisse: Swets & Zeitlinger.

Driessen, G. (2000). The Limits of Educational Policy and Practices? The case of ethnic minorities in The Netherlands. *Comparative Education, 36,* 1 ,55-72.

Hiebert, J., Gallimore, R., & Stigler, J.W. (2002). A knowledge base for the Teaching Profession: What would it look like and how can we get one. *Educational Researcher, 31* (5), 3-15.

Mulder, L. (1996). *Meer voorrang, minder achterstand. Het Onderwijsvoorrangsbeleid getoetst.* [More priority, less disadvantaged? The Educational Priority Policy put to the test]. Nijmegen: ITS.

Shanan, T., & Barr, R. (1995). "Reading Recovery: An Independent Evaluation of the Effects of an Early Instructional Intervention for At-risk Learners". In: *Reading Research Quarterly. 30,*4, 958-996.

Schneuwly, B. (1994). Contradiction and development: Vygotsky and Paedology. European *Journal of Psychology of Education, 9* (4), 281-291.

Taylor Fitz-Gibbon, C. (2002). A Typology of Indicators. In A.J. Visscher and R. Coe (Eds.), *School Improvement through Performance Feedback.* Lisse: Swets & Zeitlinger.

Title I. Resources for Reform: New Hope for America's Most Disadvantaged Public Schools (2002). *A Congressional Analysis of the Impact of President Bush's No Child left behind reforms*, by U.S. Rep. John Boehner (R-OH), Chairman House Committee on Education & the Workforce and U.S. Sen. Judd Gregg (R-NH), Ranking Member Senate Committee on Health, Education, Labor, & Pensions.

Tymms, P. (1999). Baseline assessment, value added and the prediction of reading. *Journal of Research in Reading, 22* (1), 27-36.

Tymms, P., Merrell, C., & Henderson, B. (2000). Baseline assessment and Progress during the First Three Years at school. *Educational Research and Evaluation, 6* (2)*,* 105-129.

Tymms, P., & Albone, S. (2002). Performance Indicators in Primary Schools. In A.J. Visscher and R. Coe (Eds.), *School Improvement through Performance Feedback.* Lisse: Swets & Zeitlinger.

Van der Hoeven-van Doornum, A.A. (1990). *Effecten van leerlingbeelden en streefniveaus op schoolloopbanen*. [Effects of teacher expectations and aspiration levels on school careers; Dissertation with a summary in English]. Nijmegen: ITS.

Van der Hoeven-van Doornum, A.A., Voeten, J.M.J., & Jungbluth, P. (1993). The influence of parents' expectations on school careers. *Tijdschrift voor Onderwijsresearch, 18* (6), 369-379.

Visscher, A.J., & Coe, R. (2002). *School Improvement through Performance Feedback.* Lisse: Swets & Zeitlinger.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Blackwell.

Chapter 5 The prediction of learning progress in young children using social-ethnic background or baseline assessment as predictors

Anneke van der Hoeven-van Doornum[1] and Peter Tymms[2]
[1] ITS, University of Nijmegen, Postbox 9048, 6500 KJ Nijmegen, The Netherlands.
[2] CEM Centre, Mountjoy Research Centre 4, University of Durham, Stockton Road, Durham, DH1 3UZ, United Kingdom.

Paper presented at ECER, Hamburg, September 2003.

Summary

This contribution is on the use of the OBIS (Onderbouw-informatiesysteem) baseline assessment, in particular for purposes reaching beyond the school level. The OBIS is the Dutch version of the English PIPS (Performance Indicators in Primary Schools) Baseline Assessment. Although the baseline assessment was constructed in the first place to act as a basis for value added measures, it has stimulated other uses which may be very valuable.

In The Netherlands there is an ongoing debate about the use of baseline assessment to determine the amount of extra support a school needs when children show low test scores due to developmental problems or language deficits; or secondly using it as a baseline to assess the value added by the school over a longer period of years.

The aim of the Dutch longitudinal study was to record pupils' learning progress and teachers' activities. The data presented here describe the performances of pupils on the OBIS related to their social and ethnic background. The results showed that during the first three years of primary school, the performance scores within a SES group are not constant. Quite substantial changes have taken place, specially in the lower class of SES groups, this means that being a low performer is not per se a typical and inflexible group quality. Furthermore, after three years of primary school the mean scores of Dutch working class and the non-Dutch working class are nearly the same, while the distribution of extra money is not. For a fair distributions of money a recalibration of the approximately 20-year old weighting scores is a precondition.

Finally, the discussion will concentrate on the question whether or not baseline assessment should be used or not as a means for provision of extra resources for schools and policy decisions.

5.1 Introduction

Children start school at the age of four in England and The Netherlands. The English school system comprises seven years at primary level including one Reception class (Kindergarten) and six years of formal instruction; while the Dutch system has eight years on primary level, out of which there are two years of Kindergarten (years 1 and 2) and six years of formal instruction starting in year 3. Although the compulsory age of school entry in The Netherlands is being five years old, practically all children enter school on their fourth birthday in year 1 of Kindergarten. The two countries also include large numbers of children starting school whose first language is not the dominant language of the host country.

5.1.1 UK and The NETHERLANDS

PIPS, Performance Indicators in Primary Schools
The Performance Indicators in Primary Schools project (PIPS) deals with the progress and attitude of children in primary schools (Tymms 1999a; Tymms and Albone 2002). It is designed to provide information for teachers about their pupils, and information for headmasters at elementary schools about the running of their school. Pupils are assessed in the schools, the data are processed by the CEM Centre and these processed data are fed back to the school. At the heart of the system there are baseline assessments which assess the point from which the children start. In this way attainment in the curriculum can be put into context (Tymms & Wylde, 2003).

The *PIPS On-Entry baseline* assesses early reading, early mathematics and phonological awareness in an adaptive test that takes between 15 and 20 minutes per child to administer. It includes sections on name writing, vocabulary, concepts about print, phonological awareness, letter identification, reading, concepts about mathematics, simple sums, simple counting, digit identification, formal sums, and short-term memory (see for example Tymms 1999b). Personal, social and emotional development is also assessed from teachers' observations. Standardised scores for reading, mathematics and phonological awareness are fed back to schools giving an indication of children's skills compared with a nationally representative sample. Age is not taken into account in the standardised feedback. Children are standardised against other pupils of the same intake.

OBIS, Onderbouw-informatiesysteem
The OBIS (Onderbouw-informatiesysteem) baseline assessment was first introduced into primary schools as part of a NWO (The Netherlands Organization for Scientific Research) project in 2000. The OBIS is the Dutch version of the PIPS Baseline Assessment. Just like PIPS, the OBIS baseline aims to monitor pupil progress in specific curriculum areas using prior attainment as the starting point.

5.2 Monitoring systems

It is generally accepted that regular testing can improve the achievements of pupils by influencing their teachers. Specially, a good feedback of the pupils' test results may help the teachers to perform better (Van der Hoeven-van Doornum, 2002). Therefore well-constructed and well-used monitoring systems are needed, although the evidence for the positive impact of such systems has yet to be established. Nevertheless, there is little doubt that well-constructed, and well-used monitoring systems can be of enormous benefit to the educational provision for children (Visscher & Coe, 2002).

Usually, a distinction is made between official accountability systems and professional monitoring systems. In both systems value added is the key issue.
In a professional system it mainly comes down to measuring the progress of one pupil compared to other similar pupils in other schools. In both systems, the technique employed is regression analysis and the value-added data are residuals from regression analysis.

Assessment as a part of a professional monitoring system is a means to diagnose and measure if there is adequate progress of pupils over a longer period of time, for example a year. Comparisons of the progress of (similar) pupils in different schools can also be made. The intended use of monitoring progress of pupils on different levels is to provide valid and reliable information for clear and fair judgments on the pupils and schools in question. Based on such information early interventions can be carried out for those children who are falling behind, whatever the cause may be, their home background or other factors associated with a lack of learning progress. If regular testing is integrated in the school system, sustained support can be given.

Monitoring progress
How well pupils are doing compared with similar pupils is defined as value added. Value added is not simply the difference between an intake test and an output test, it is the average across all pupils in a school of the difference between their individual test results and their expected results. Correlations and regression analysis are used as a basis for value added calculations using the full dataset.

The PIPS On-Entry Baseline correlates strongly with later achievement. When the same assessment was administered to children 9 months later at the end of Reception class, the correlation between the two assessments for the academic year 2001/02 was 0.78 for a sample of 20,000 children (Merrell, 2002; Tymms & Wylde, 2003). The correlation between the PIPS On-Entry assessment and PIPS reading and maths assessments at age 7 was 0.7 and approximately 0.6 with the end of Key Stage 1 assessments, (Tymms, 1999b). The test/retest reliability of the PIPS is .95. This was established by selecting a sample of pupils across several schools after teachers had carried out the assessment and then it was repeated by a researcher.

Although the research in The Netherlands was carried out with a much smaller sample of 450 pupils in 26 classes at 11 schools, a considerable similarity between the PIPS correlations and OBIS correlations (2003) was found. The first assessments with OBIS were carried out in 2000-2001. When the same assessment was repeated one year later the correlation between the two tests was .77. Likewise, the correlation between the OBIS baseline assessment in year 1 and the total scores of a reading and maths test almost three years later at the age of seven, was .68 (Table 5.1) Just as the test/retest reliability of the PIPS, OBIS has a very satisfying test/retest reliability of .98 (reading .97, maths .92). Therefore, our conclusion is that the data of three consecutive years show that the OBIS may be considered as a reliable instrument for baseline assessment and prediction of later achievement.

Monitoring School Performance

The main difference between the two systems is in the purposes for which the value added data is used and the consequences involved for schools and teachers. The essence of an accountability system is using the value added of the school as a measure for the educational quality of schools. The express purpose is to hold schools and/or teachers accountable for the output results: exams, scores on (national) tests which are published in newspapers and on internet in e.g. league tables.

Table 5.1 – Correlations between standardised test scores over a period of 3 years

	Start Maths	Start Reading	Start Total	Year 3 Maths	Year 3 Reading
Start Reading	.71				
Start Total	.89	.92			
Year 3 Maths	.56	.51	.57		
Year 3 Reading	.50	.55	.57	.61	
Year 3 Total	.62	.63	.68	.78	.90

Monitoring in The Netherlands

In The Netherlands, the socio-ethnic status or educational level of the parents is used as a proxy measure to assess value added. The assessment of the educational quality of school is based on the mean school scores on the CITO Eindtoets (output test). Schools are assigned to groups ('families') of schools with a more or less equal school population according to their social background. SES is considered as an alternative for value added. The advantage of such proxy indicators for baseline assessment is that they are easy to collect and that schools are not 'bothered' with testing.

Nevertheless, the ongoing debate in The Netherlands on monitoring the educational quality of schools is recently extended with the question whether it is recommendable or not to use value added (test) scores for the allocation of extra financial support to schools with children in disadvantaged learning situations.

As learning disadvantage, achievement assessment, and educational opportunities are considered to be related issues, since the mid eighties Dutch schools receive extra financing for additional teaching staff, based on the social and ethnic composition of the school population. The underlying theory of this policy is that immigrant children and working-class children are considered to have comparable educational disadvantages mainly caused by the socio-economic position of their parents.

In the allocation of these resources various categories of disadvantaged children are distinguished by means of a weighting factor. Roughly speaking, this means that ethnic minority children count as 1.90; Dutch working-class children count as 1.25; other children 'simply' count as 1 (this is the non-disadvantaged category).

Despite the fact that since the beginning of the 1990's it has become clear that this system of money allocation is too rigid; it is only since last year that politics and opinions are changing. In 2002 it was decided by the Dutch government that an intake test for 4-year old children should be used instead. In advice to the Minister of Education, August 2002, the Onderwijsraad (Educational Council of The Netherlands) recommends to distinguish between group related educational disadvantage and individual reading underachievement. However, one year later, in August 2003, the advice of this Board is to assess children at the age of six and not at the age of four, mainly because of 'practical' reasons. At present no decision has been made yet, but it seems that the weight factor regulation will not survive in the end.

Research questions

The aim of the Dutch longitudinal study was to record pupils' learning progress and teachers' activities. Here we concentrate on the question whether background indicators or baseline assessment better predict future learning achievements, and which of

the two should be used as a means for provision of extra resources for schools. Using the data of three consecutive years we'll describe the performances of pupils on the OBIS related to their social and ethnic status.

- What is the relationship between different SES groups and test scores?
- Could we better use test scores as a means to allocate extra financial support to the schools instead of the weighting factor based on a membership of a socio-economic group?

5.3 Results

To answer the first research question concerning the extent to which the SES-weighting score used in The Netherlands is associated with OBIS performance scores at the start of primary school and almost three years later, the data were broken down by SES (Table 5.2). The categories are 1, 1.25 and 1.90. Obviously, a strong relationship was found between SES and the OBIS baseline test (eta =.43) and reading and mathematics performance three years later (eta=.40).

Table 5.2 – Mean test scores split up by socio-ethnic status of parents. Standardized scores, mean = 50, sd = 10

SES		Year 1 Baseline	Year 3 Primary education
1.0	High: middle class	53	53
1.25	Low: Dutch working class	46	47
1.90	Low non-Dutch working class	43	46

This shows that scores of different SES groups remain constant over the years. It suggests that nothing changes: 'if you have low scores at the start of your school career, you will keep them over the years'.

Performance scores and weighting scores

However, this is not the complete story. By looking in greater detail at the effects of each SES group separately we obtain a better view on the performances of the pupils within the SES groups. Furthermore, the data were broken down by performance group: the performance categories are high, middle and low.

Tables 5.3-5.5 depict for every SES group the changes in the distributions of test scores which are split up into three performance groups.

The results in Table 5.3 show that over a period of three years, the proportion of high scoring children from high and middle class families decreases and that the proportion of low scoring children increases. Next, the proportion of Dutch working class children with high performances also increases over the years, while the proportion of low scoring pupils in this group remains almost the same (Table 5.4). Subsequently, the proportion of high performing pupils from non-Dutch working class families increases during the years from 14% to 23%, and simultaneously the proportion of low scoring children drops from 56% into 44% (Table 5.5). This means that the performance scores within a SES group are certainly not constant. Quite substantial changes have taken place, especially in the lower class of SES groups.

Table 5.3 – SES (high and middle class) and changes in distribution of test scores

Test scores	Year 1 Kindergarten	Year 3 Primary education
High	46%	41%
Middle	37%	35%
Low	18%	25%

Table 5.4 – SES (Dutch working class) and changes in distribution of test scores

Test scores	Year 1 Kindergarten	Year 3 Primary education
High	19%	30%
Middle	32%	21%
Low	48%	49%

Table 5.5 – SES (ethnic working class) and changes in distribution of test scores

Test scores	Year 1 Kindergarten	Year 3 Primary education
High	14%	23%
Middle	30%	33%
Low	56%	44%

Although the fact that the percentage of high performances scores among SES group 1 is considerably higher than the incidence of high performance scores among SES group 1.90 (ethnic minorities), there is also a considerable group of 1.90 pupils within the group of high performing pupils. The same applies the other way round to the percentage of pupils from SES group 1 within the group of low performing pupils.

Compared with the mean test scores for the different SES groups, it appears that the SES-weighting score only partially overlaps the 'expected' performance categories.

5.4 Making valid and reliable decisions on pupil, school and state level

Based on mean test scores of distinct SES groups, it looks as if educational disadvantage of children of lower class families is a more or less permanent and constant characteristic. Taking a better view on the distribution of scores within the SES groups, it appeared that being a low performer is not per se a typical and inflexible group quality. The scores of different SES groups do not remain constant over a period of three years. This suggests that learning performances of pupils with a different social background do not have to remain constant.

Decisions on pupil and school level

The findings make it clear that regardless of family background, there is a need for sustained support for all children whose performances are lagging behind, but in particular for children who start in 'poor' educational circumstances. Secondly, baseline assessment and value added monitoring systems provide clear and valid information on which pupils need extra care and attention.

Decisions on state level

The second question to answer was whether we could better use test scores as a means to allocate extra financial support to the schools instead of the weighting factor based on a membership of a socio-economic group?

In The Netherlands, schools receive extra finances according to the distribution of their pupils over SES levels. In fact, finances are allotted according to the ratio: high: middle: low = 1.00 : 1.25 : 1.90. Thus, a pupil from the non-native lower class families represents 1.9 times as much money for the school than a Dutch pupil whose parents (at least one of them) have an education exceeding that of primary school. Disadvantage of the present system is the inflexibility of it, as illustrated by the fact that a bright non-Dutch child still brings in 1.9 times more money than strictly necessary.

Furthermore, after three years of primary school the mean scores of Dutch working class and the non-Dutch working class are nearly the same, while the distribution of extra money is not. For a fair distributions of money a recalibration of the approximately 20-year old weighting scores is a precondition.

5.5 Discussion

At present, the question was raised if finances could be allotted according to test scores, and whether any suitable tests existed. PIPS, and its Dutch version OBIS, were specially designed for monitoring progress, and they seem to fulfill the requirements to measure learning disadvantage (Educational Council of The Netherlands, 2003). Based on the data shown above, it seems clear that although the total proportion of low performing pupils in the study group remained fairly the same over a period of at least three years, for a substantial number of pupils learning disadvantage is not a constant feature. Expanding this fact on criteria for allotment of finances to schools would mean that not necessarily the whole budget must change, but rather the distribution of the finances among the schools.

Methodological considerations

If test scores were used as the major criterion for determining pupil progress and school accountability, several methodological considerations can be made. For example, important issues are vertical and horizontal scaling of tests, standards from year to year, reliability of school effectiveness measures and cohort size.

Regarding test content and vertical or horizontal scaling, test scores should be comparable from test to test and year to year. Scaling is a measurement technique that facilitates test score comparability. Regardless of changes in the test from year to year, the scores reported to the public should be on the same scale. The major assumption for equating is that the tests are assessing the same general content. In the case of horizontal scaling, this is not usually a problem since each (form of a) test should be designed to examine the same curriculum material. Vertical scaling has the same assumption of comparable content. This implies that the same dimensions are covered in each grade, as well. The equating of two tests in the horizontal scaling context is fairly easy using an item response theory (IRT) approach (e.g., Stocking & Lord, 1983). Since the nature of the items and the assessment process often changes over grades, vertical equating mixes or confounds content changes with method changes. This makes interpretation of results difficult and violates the assumption of comparable assessment across grades. Creating the vertical scale is also a difficult task from the technical point of view, even with the use of IRT models (Lissitz & Huynh, 2003).

Furthermore, with regard to cohort or group size, the reliability of the school effectiveness measure is related to the numbers of pupils and the proportion of variance accounted for by schools. A reliability of .90 together with an intra-school correlation of 0,11 – 0,20 and a turnover rate of circa 50% during primary school requires quite substantial cohort sizes (See the formula for reliability, Goldstein, 1987).

5.6 Concluding remarks

To sum up, this study has shown that group characteristics appeared to be bad predictors and not specific enough for the individual child. SES as a background indicator of learning progress is unfair and reinforces prejudices.
Considering the results, is there any inherent reason for decision-making on state level why baseline assessment could not be used as a means for provision of extra resources for schools and policy decisions? Can assessment data be used beyond pupil and school level, on district or state level? What are the concerns? Is there a risk that when teachers do the assessment themselves, some of them may carry out the assessment less accurate than others, or are there other reasons to compromise the data simply because of the fear of public accountability and of how national government might use the data? What are the benefits and drawbacks and what precautions can be taken? The experience is that teachers familiar with value added assessment show an increasing professional acceptance and interest in testing. Together with acknowledged reliability, validity and manageability, it is worthwhile to argue for baseline assessment beyond a mere intake test.

Acknowledgement I would like to express my gratitude to the Netherlands' Organization for Scientific Research (NWO) for funding by grant 411- 210-05 the project 'Development on scale, instruction at measure' on which this paper is based.

5.7 References

Educational Council of The Netherlands (2002). *Over leerlinggewichten en schoolgewichten.* Den Haag: Onderwijsraad.

Educational Council of The Netherlands (2003). *Wat scholen toevoegen.* Den Haag: Onderwijsraad.

Fitz-Gibbon, C. (1997) *The Value-added National Project* Durham: CEM Centre./ University of Durham.

Lissitz, R.W. & Huynh H., (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research and evaluation, 8* (10) Available online

Merrell, C. (2002). *Adapting to changing arrangements. Value added in early years*. Paper presented at the Symposium: One Assessment Many Cultures: the issues confronted in assessing children starting school in different countries. European Conference on Educational Research (ECER), Lisbon, September 2002.

Tymms, P., Merrell, C., & Henderson, B. (1997). The First Year at School: A quantitative investigation of the attainment and progress of pupils. *Educational Research and Evaluation, 3*(2), 101-118.

Tymms, P. B., Merrell, C., & Henderson, B. (1999). 'Baseline assessment and progress during the first three years at school.' *Educational Research and Evaluation*.

Tymms, P. (1999a). *Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-added Indicators*. London: David Fulton Publishers.

Tymms, P. B. (1999b). Baseline assessment, value-added and the prediction of reading. *Journal of Research in Reading*, *22* (1), 27-36.

Tymms, P. (2001). The development of a computer-adaptive assessment in the early years. *Educational and Child Psychology, 18* (3): 20-30.

Tymms, P., & Albone, S. (2002). Performance Indicators in Primary Schools in. *School Improvement Through Performance Feedback*. A. J. Visscher and R. Coe (Eds). Lisse/Abingdon/Exton PA/Tokyo, Swetz & Zeitlinger: 191-218.

Tymms, P., & Wylde, M. (2003). *Baseline assessment and monitoring in primary schools*. Paper presented at the Symposium Connectable Processes in Elementary and Primary section. Bamberg, DL, April 2003.

Van der Hoeven-van Doornum, A.A. van der (2002). *What Baseline Assessment is Doing For Children's Progress and Teachers' Professionalism.* Paper presented at the Symposium: One Assessment Many Cultures: the issues confronted in assessing children starting school in different countries. European Conference on Educational Research (ECER), Lisbon, September 2002.

Visscher, A. J., & R. Coe, Eds. (2002). *School Improvement Through Performance Feedback.* Contexts of Learning: Classrooms, Schools and Society. Lisse/Abingdon/Exton PA/Tokyo: Swetz & Zeitlinger.

Chapter 6 Cognitive and social profiles of children at risk

Predictors and outcomes of educational achievement in the lower years of primary school, from the perspective of WSNS (Going To School Together)

Paper presented at the Onderwijs Research Dagen 2004

Summary

In the Dutch school system, low performing pupils can be transferred to special primary education. It is felt that the outflow into this type of school is too high, and a policy called Going to School Together (WSNS) was introduced in order to keep children in the normal primary school. As a corollary, better methods are needed to identify the children 'at risk' who should still be referred to special education. In this study the question is raised whether children at risk can be identified at an early stage using cognitive and social characteristics, and the development thereof. Nearly 450 children with a mean age of four and a half year were assessed upon school entry, using OBIS, a broad baseline test. The test was repeated one year later, and reading and mathematics achievements were assessed again in year 3. The latter scores appeared to be well predicted by the total scores in Year 1, which are mainly determined by pre-school learning. The children at risk, with the 16% of the lowest scores, were not the same over the years, indicating that a single assessment may not be sufficient for identification of the 'at risk' status. The longitudinal assessments using OBIS allowed for estimation of value added and the relative progress of the child compared with the progress of pupils with the same scores one year earlier. The progress of the children at risk lags behind. This constitutes essential information for the teacher, providing the opportunity of intervention.

6.1 Introduction

This contribution is on predictors and outcomes of educational achievement in the lower years of primary education. Previous investigations have shown that the mean cognitive scores of pupils with weighting factor 1, 1.25 and 1.90 are fairly constant. Yet, in the first three years of primary education, within these groups significant shifts in the numbers of high and low scoring pupils occur (Van der Hoeven-van Doornum, 2003). Learning disadvantage appeared to be no constant characteristic.

Using longitudinal data collected with the OBIS, that is the acronym for Onderbouw-informatiesysteem (Lower School Information System), the progress of children identified at their entry of school as being at risk, is investigated.

Background of this investigation is the perspective of the WSNS which is the governmental policy to reduce the outflow of low performing pupils from the mainstream primary education (BAO) to schools for special primary education (SBAO). Previous studies have reported that very little is known about which children at risk benefit more from mainstream primary education or from special education (Van der Veen, 2003). This ignorance places those pupils and also their teachers at the disadvantage, as it appears from facts and conclusions reported by teachers and researchers (Smeets, e.a., 2003, pag. 104 and 105):

- 'Schools refer pupils to special education on child characteristics if no other solutions are available', according to a participant of a WSNS conference in 2003.
- The capacity of schools is under pressure, there is a need for better educational quality of primary schools: identifying, diagnosing and remedying children at risk.
- The classroom teacher is mainly responsible for extra care and instruction.
- Six out of ten children (61%) who are referred to special education have learning difficulties, 35% have both learning and behavioural difficulties; and 4% have only behavioural difficulties.

From the view of WSNS (Going to School Together) it is not only important to identify which children are at risk, but also to know if this is a permanent status or not. For example, is it known what educational progress children at risk have made after they have been identified as such? Previous studies have not reported the cognitive achievement of pupils at risk taking into account their attainment level when they first started school. Such information is crucial when trying to interpret later reading and mathematic levels of children. Without a baseline assessment at entry, it is not possible to assess the progress made at school. Otherwise stated, have the schools served the pupils badly or have they taught the children well? This is an essential question since schools should only be held responsible for the progress that their pupils make and not for the attainment levels, particularly in the case of children at risk (Tymms, Brien, Merrell, Collins & Jones, 2003). As a consequence, the progress of low scoring pupils and not the attainment level seems to be a better criterion to decide whether a child is better off in mainstream or in special education.

6.2 Research questions

The purpose of this study is to explore the school career of children with identical cognitive and social profiles at the entry of primary education. It will be investigated which cognitive and social characteristics at the start of primary education are predictors of the learning achievements of children in group 3. The findings are discussed in terms of identification and intervention of children at risk. The outcomes are intended to inform policy and practice about children at risk. The research questions are:

- Can children at risk at the beginning of group 1 be identified using their cognitive and social profile?
- What is the development of the cognitive and social profile of pupils at risk in the first years of primary education?

6.3 Methods

Nearly 450 children with a mean age of four and a half year were assessed upon entering school, using a broad baseline assessment, and then again one year later in Kindergarten. The reading and mathematics achievements of the same children were assessed one year later in Year 3. These data were used to look at the prediction of the academic achievement of children with varying degrees of cognitive and social profiles.

The assessments were done using the OBIS (Van der Hoeven-van Doornum, 2002). The OBIS is the Dutch version of the PIPS baseline (Performance Indicators in Primary Schools). The PIPS was first created by the CEM Centre (Durham, UK) by examining longitudinal studies that had assessed children at the ages of four-five and then reassessed their reading and mathematics at a later date. The best predictors were incorporated into the assessment. In year 3 cognitive attainments on reading and mathematics was assessed using the PIPS End of Year 1 assessment, selected for this purpose after examining and comparing the content with Dutch and international studies on curriculum in the first years of primary education.

Table 6.1 – Descriptives for the sections of the OBIS (raw scores). N=445, Year 1

	mean	sd	min – max	N items	Cronbach's α
1 Name writing	1.80	1.50	0 – 5	1	-
2 Picture vocabulary	15.15	5.50	0 – 22	22	.93
3 Ideas about Reading	3.76	1.70	0 – 10	10	.72
4 Repeats	6.43	2.06	0 – 8	8	.85
5 Rhyming	4.01	3.57	0 – 9	9	.94
6 Letter identification	2.50	4.48	0 – 26	27	.95
7 Words	0.19	1.20	0 – 14	14	.94
8 Ideas about Maths	6.06	1.40	0 – 7	7	.80
9 Counting	2.59	1.47	0 – 4	4	.81
10 Digit identification	4.36	4.85	0 – 20	20	.93
11 Sums	4.25	2.53	0 – 8	8	.86

Table 6.1 and 6.2 give the descriptive statistics for raw scores of the sections of the OBIS-test in year 1 and 2. Social profiles were measured using questionnaires for class-room teachers with items on non-cognitive skills of their pupils. Most of the items come from the PRIMA 4 cohort study 2000-01. Table 6.3 gives the descriptive statistics of those teacher assessments.

Table 6.2 – Descriptives for the sections of the OBIS (raw scores). N=418, Year 2

	mean	sd	min – max	N items	Cronbach's α
1 Name writing	3.08	1.53	0 – 5	1	-
2 Picture vocabulary	17.94	3.68	2 – 22	22	.84
3 Ideas about Reading	5.30	2.31	1 – 10	10	.74
4 Repeats	7.41	1.23	1 – 8	8	.73
5 Rhyming	6.12	2.56	0 – 9	9	.90
6 Letter identification	6.83	7.04	0 – 27	27	.95
7 Words	1.82	4.23	0 – 14	14	.98
8 Ideas about Maths	6.55	.96	0 – 7	7	.66
9 Counting	3.57	.93	0 – 4	4	.70
10 Digit identification	10.21	5.59	0 – 22	22	.93
11 Sums	6.02	1.77	0 – 8	8	.75

The criterion ´at risk´ can be approached in two ways: using a preset percentage (16%) of the children with the lowest scores on the test. Thus, children are labelled as at risk group (1) or not at risk (0).

The second approach is to use value added measures that give an insight into the progress that a pupil makes compared with similar pupils. In this way value added measures look at the relative progress of children. The standard way of doing this is to establish a measure, which is a good predictor of later achievement, and then to use regression analysis to compare the outcomes of pupils with the same initial scores on the predictor (Fitz-Gibbon, 1996; Fitz-Gibbon, 1997; Tymms, 1999). The procedure is used by the CEM Centre in a monitoring system designed to help schools to track the progress of their pupils (Tymms & Albone, 2002) and for educational research.

Table 6.3 – Social profile. Descriptives of non-cognitive skills (raw scores) Years 1 and 2. Attitude towards learning and school in Year 3. Eta^2 and p-value for at risk and non-risk groups

Year 1*	mean	sd	min – max*	N items	alpha	p	eta^2
School well-being	4.12	.49	2.17 – 5	6	.88	.15	.01
Self-confidence	3.53	.75	1.71 – 5	7	.85	.03	.01
Social behaviour	3.56	.75	1.0 – 5	3	.75	.27	.00
Attitude	3.27	.84	1.0 – 5	3	.87	.03	.02
Parental support	4.07	.85	1.0 – 5	3	.82	.00	.10
Cognitive capacities	3.42	.67	1.13 – 5	8	.88	.00	.05
Development	4.10	.58	2.43 – 5	7	.80	.00	.15
Year 2*	mean	sd	min – max*	N items	alpha	p	eta^2
School well-being	4.15	.52	2.71 – 5	7	.88	.00	.04
Self-confidence	3.67	.62	1.86 – 5	7	.84	.00	.04
Social behaviour	3.57	.74	1.2 – 5	5	.83	.10	.01
Attitude	3.46	.82	1.0 – 5	3	.86	.00	.05
Parental support	4.03	.73	1.75 – 5	4	.82	.00	.10
Cognitive capacities	3.61	.70	1.13 – 5	8	.91	.00	.18
Development	4.42	.53	2.86 – 5	7	.87	.00	.05

* Main source of items: PRIMA cohort 4, 2000-01. Meaning of high score (5):

School well-being	the child feels happy at school
Self-confidence	is self-confident
Social behaviour	pleasant, behaves according to the rules
Attitude	doesn't give up quickly, works steadily and accurately
Parental support	stimulating home environment, ethnic parents speak Dutch at home.
Cognitive capacities	bright child, belongs to the >better= pupils
Development	mental and physical development according to age

6.4 Results

Predictor and outcomes

Table 6.4 gives the correlations between standardised test scores over a period of 3 years. It shows that the total cognitive score (reading and mathematics) in year 3 is well predicted by the total score in Year 1 assessed about two months after the beginning of school year. It is noted that the latter score mainly represents preschool learning.

Table 6.4 – Correlations between standardised test scores over a period of 3 years

	Start Maths	Start Read	Start Total	Year 2 Maths	Year 2 Read	Year 2 Total	Year 3 Maths	Year 3 Reading
Start Reading	.71							
Start Total	.89	.92						
Year 2 Maths	.73	.61	.72					
Year 2 Reading	.63	.67	.70	.71				
Year 2 Total	.74	.71	.78	.90	.92			
Year 3 Maths	.56	.51	.57	.62	.55	.56		
Year 3 Reading	.50	.55	.57	.43	.50	.56	.61	
Year 3 Total	.62	.63	.68	.56	.55	.64	.78	.90

At risk or not at risk

Table 6.5 depicts the standardised scores of the ‘at risk’ groups over the first three school years. The ‘at risk’ group is represented by the 16 percent of the lowest scoring pupils. In each year the ‘at risk’ group was assessed according to this criterion. Furthermore, the scores in the following and previous years are given. It can be seen that the group with the lowest scores in one year has higher scores in the following and previous years. This indicates that the label ‘at risk’ is not a constant feature for a group of children and that shifts of pupils from the lower scoring groups to higher scoring groups and vice versa must occur.

Table 6.5 – Standardised maths, reading and total scores: mean (sd) of 'At risk' groups over the first three school year

	At risk at Start		At risk Year 2		At risk Year 3	
	mean	(sd)	mean	(sd)	mean	(sd)
Start Maths	37	(7)	41	(7)	43	(8)
Start Reading	35	(6)	40	(8)	40	(10)
Start Total	35	(8)	40	(6)	41	(8)
Year 2 Maths	40	(8)	37	(6)	42	(10)
Year 2 Reading	40	(10)	36	(6)	44	(10)
Year 2 Total	39	(8)	35	(4)	42	(9)
Year 3 Maths	41	(9)	41	(9)	36	(6)
Year 3 Reading	42	(10)	42	(10)	36	(5)
Year 3 Total	40	(9)	40	(9)	35	(4)

The group of children at risk in year 3 can be identified by cognitive and social predictors in earlier years. Table 6.6 shows the results of logistic regression. The values (Exp) are odds ratio's indicating the probability of being an 'at risk' pupil in year 3 predicted by previous risk status, achievement level and background indicators of the pupils.

It appeared that the cognitive level (p=.00) and the risk status (p=.02) in year 1, assessed about 8 weeks after starting school, are the best predictors of the risk status three years later. Given this model the other variables related to risk status have no significant contribution.

Table 6.6 – Logistic regression. Children at risk in Year 3, identified by cognitive and social predictors in earlier years. N=234

		B	p	Exp (B)
At risk Year 1		1.90	.02	6.67
At risk Year 2		0.03	.97	1.03
Total score OBIS at start		-.14	.00	0.87
Total score OBIS year 2		-.06	.20	0.94
Boy		0.68	.16	1.97
SES	1.25	-.36	.66	0.70
	1.90	.73	.38	2.07
Year 1				
	Cognitive capacities	-.08	.88	0.93
	Development	-.53	.35	0.59
Year 2				
	Parental support	-.38	.41	0.67
	Development	-.93	0.11	0.40
Constant		13.93		
Nagelkerke R2		.47		

Progress

Figure 6.1 shows the raw and standardised scores in year 2, and also the added value, that is the relative progress of a child compared with the progress of pupils with similar cognitive levels assessed one year earlier at the start of schooling. Value added was derived from the regression analysis (Fitz-Gibbon, 1996; Tymms 1999).

Figure 6.2 is a scatter gram that gives the standardised reading level of a child against its total cognitive level one year earlier. It shows what progress the individual child has made in relation to the score at the start of schooling. The pupils in the lower left corner of the diagram represent the 'at risk' group, their progress has lagged behind. This is essential information for teachers, giving the opportunity of intervention.

6.5 Discussion

The discussion addresses two issues: the flexibility of cognitive profiles of children at risk, and the interest and advantage of early identification of learning difficulties. Van der Veen (2003) concludes that being a child 'at risk' is not a permanent condition; she also raises the question of whether a criterion of belonging to the lowest scoring pupils is an adequate method to identify children at risk. According to Van der Veen

the interventions teachers use may even have a negative impact on the educational achievement of children at risk and enlarge their learning disadvantages. It is suggested that in classrooms with high percentages of 'at risk' pupils less attention is paid to early reading and maths.

According to Tymms and Merrell (2004), the effect of identifying and labelling children 'at risk' should be thoroughly investigated. There are arguments for both identification and non-identification. Identification can lead to effective communication between parents, schools and external agencies to provide appropriate resources and support for a child. Conversely, some argue that identification gives a child a 'label' that causes others to automatically perceive them in a negative way, which may be detrimental to their progress and development.

The first issue is the flexibility of cognitive profiles. Table 6.5 shows that shifts occur in mean scores of annually assessed 'at risk' groups of pupils with the 16 percent of the lowest scores. This indicates that the composition of the risk groups over the years is not constant, which complies with Van der Veen's observations. Next to this, the odds ratio's in Table 6. 6 show that the chance of being at risk in year 3 is almost seven times higher for pupils who were already identified as 'at risk' at the start of primary education than for the non-risk pupils. Those two independent observations may lead to opposing conclusions regarding the identification of risk pupils. However, it is clear that the 'at risk' label of a child is not always a permanent one.

More important than the attainment level per se is the relative progress, the value added. The whole purpose of a value added assessment is to compare the attainments of pupils with the same starting points. Figures 1 and 2 give insight whether the progress in development and learning is as expected or not. Measuring the value added provides information on progress on group level as well as on individual level. The on–entry or baseline assessment such as the OBIS opens up the possibility of providing teachers and schools with clear and fair information on the relative progress that pupils are making. The obtained information can be used directly for daily classroom instruction on early reading and maths.

It is our conclusion that early identification of learning difficulties applied in this way is in the interest and advantage of children at risk. This is in line with the perspective of the WSNS policy of going to school together.

Figure 6.1 – Results Table Year 2. Raw scores, standardised scores, value added grades and pupil attitudes

vervolg

OBIS 01-02

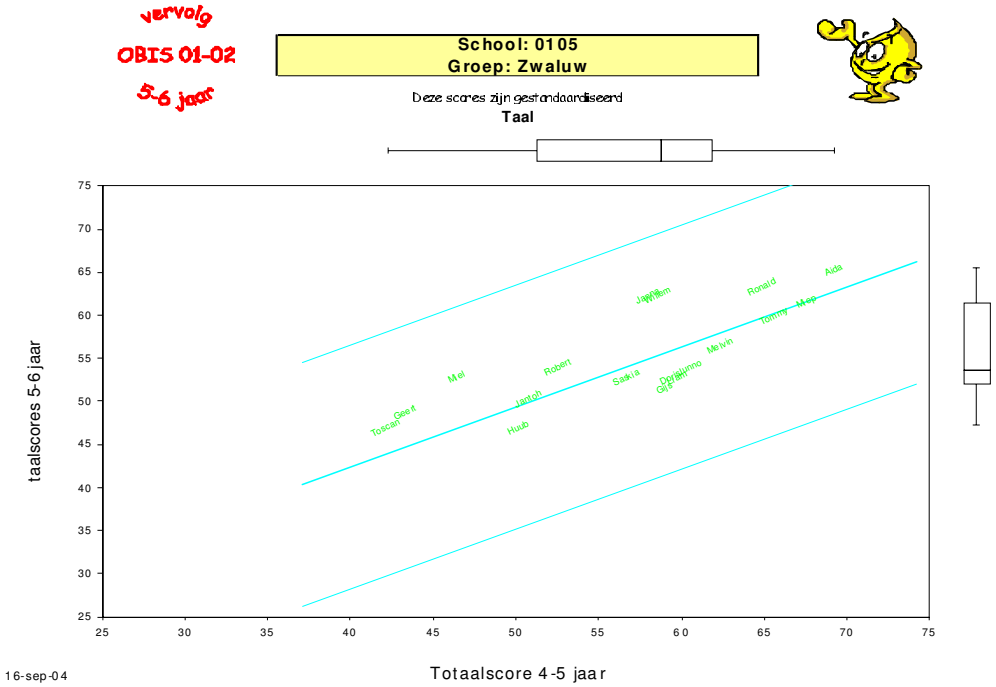
5-6 jaar

School: 0105
Groep: Zwaluw

Deze scores zijn gestandaardiseerd

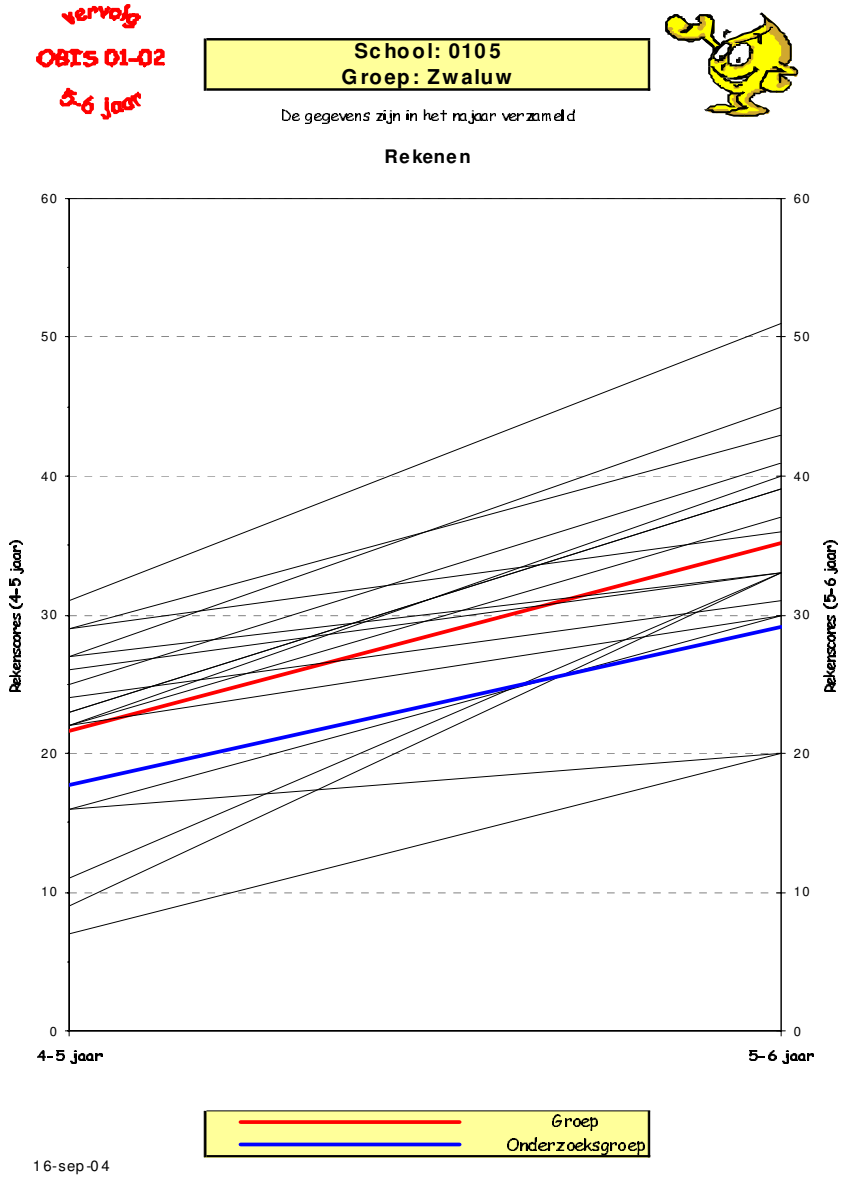
naam		rekenscores			taalscores			totaalscores				toegevoegde waarde		houding
		4-5 jaar	5-6 jaar		4-5 jaar	5-6 jaar		4-5 jaar		5-6 jaar		rekenen	taal	
		ruw	ruw	std.	ruw	ruw	std.	ruw	std.	ruw	std.			
Toscan		7	20	41	23	31	47	36	42	61	42	gemiddeld	gemiddeld	··☺
Huub		16	20	41	23	31	47	50	50	62	43	–	gemiddeld	··☺
Jantoh		16	30	52	24	33	50	51	51	80	52	gemiddeld	gemiddeld	··☺
Geert		9	33	54	21	32	49	38	43	81	52	++	gemiddeld	··☺
Gijs		24	31	52	26	34	52	67	59	82	53	gemiddeld	gemiddeld	··☺
Saskia		22	30	52	29	35	53	63	57	82	53	gemiddeld	gemiddeld	··☺
Doris		27	33	54	26	35	53	68	59	85	55	gemiddeld	gemiddeld	··☺
Miel		11	33	54	22	35	53	43	46	85	55	+	+	··☺
Robert		23	39	58	19	36	54	55	53	92	57	+	gemiddeld	··☺
Melvin		29	36	56	29	39	57	75	62	92	57	gemiddeld	gemiddeld	··☺
Junno		25	41	60	31	36	54	71	60	93	58	gemiddeld	gemiddeld	··☺
Liam		27	45	65	28	35	53	69	60	96	59	+	gemiddeld	··☺
Willem		23	39	58	28	53	62	66	59	109	62	gemiddeld	+	··☺
Janna		22	40	59	27	53	62	65	58	110	63	gemiddeld	+	··☺
Miep		29	43	62	42	51	62	87	68	111	63	gemiddeld	gemiddeld	··☺
Ronald		22	37	57	41	57	63	80	65	111	63	gemiddeld	gemiddeld	··☺
Tommy		31	51	75	34	46	60	81	66	114	64	++	gemiddeld	··☺
Aida		26	33	54	55	69	65	98	69	118	64	– –	gemiddeld	··☺

Figure 6.2 – Scatter plots. Reading score Year 2 (vertical) plotted against Total cognitive score Year 1 (horizontal)



The middle blue line is the best fit, the two outer blue lines enclose 95% of all pupils.
Pupils falling on or close to this line are making progress as expected.
Pupils above this line are making more progress than expected (positive value-added).
While pupils below this line are making less progress than expected (negative value-added).

Figure 6.3 – Line charts. Progress made by individual pupils and by the whole class; in comparison with the average research group



Each individual is represented by a thin black line, going from their start of primary education (year 1) raw score on the left to their raw score a year later on the right. The class average is displayed as a red line. The chart also includes a blue line representing the average.

Acknowledgement I would like to express my gratitude to the Netherlands' Organization for Scientific Research (NWO) for funding by grant 411- 210-05 the project 'Development on scale, instruction at measure' on which this paper is based.

6.6 References

Fitz-Gibbon, C. T. (1996). *Monitoring Education: Indicators, Quality and Effectiveness.* London:1 Cassell.

Fitz-Gibbon, C. T. (1997). *The Value-Added National Project: Feasibility Studies for a National System of Value Added Indicators* (Final report. London: School Curriculum and Assessment Authority.

Hoeven-van Doornum, A.A. van der (2002). *What Baseline Assessment is Doing For Children's Progress and Teachers' Professionalism.* Paper presented at the European Conference on Educational Research, Lisbon, September 2002.

Hoeven-van Doornum, A.A. van der (2003). *One Baseline Assessment and Many Uses: the issues confronted in using intake tests for different purposes.* Paper, presented at the European Conference on Educational Research, Hamburg, September 2003.

Onderwijsraad (2003). *Wat scholen toevoegen.* Den Haag: Onderwijsraad.

Smeets, E.F.L., Hoeven-van Doornum, A.A. van der, & Smit, F.C.G. (2003). *Wachtlijsten in de Samenwerkingsverbanden 'Weer Samen Naar School'*. Nijmegen: ITS.

Tymms, P. (1999). *Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-added Indicators*. London: David Fulton Publishers.

Tymms, P., & Albone, S. (2002). Performance Indicators in Primary Schools in. *School Improvement Through Performance Feedback*. A. J. Visscher and R. Coe.(eds) Lisse/Abingdon/Exton PA/Tokyo, Swetz & Zeitlinger: 191-218.

Tymms, P.B., Brien, D., Merrell, C., Collins, J., & Jones, P. (2003). Young deaf children and the prediction of reading and mathematics. *Early Childhood Research, 1* (2): 197-212.

Tymms, P.B., & Merrell, C. (2004). *Screening and Classroom intervention for Inattentive, Hyperactive and Impulsive Young Children – A longitudinal Study*. Paper presented at the AERA, 2004.

Chapter 7 Progress on scale

Summary

The previous chapters have been realised during the course of the study and rely on the data available at that time. The aim of this chapter is to bring together the weight of evidence that underpins the need for measuring progress starting from baseline obtained so far. The central question of this conclusive chapter is how well baseline assessment – using OBIS – predicts learning progress and academic success? The next question related to the central question is what other factors inside and outside the classroom have to be taken into account for fair comparisons. To unravel the complex amalgam of individual, pedagogical and educational factors that predict and explain cognitive progress, models are investigated on the individual pupil and on class level.

The results of the longitudinal LISREL analyses show that compared with SES and teachers' perceptions of their pupils capacities, baseline assessment is the best predictor of future progress. Using multilevel analyses it was demonstrated that all pupils benefit from literature-based classroom practices. Cross-level interaction effects between teaching style and pupil characteristics could not be identified, differential effects indicating that teachers adjust their instruction in favour of for instance disadvantaged pupils, were not found either.

7.1 Introduction

Individual characteristics, social background, teacher expectations, classroom practices and target setting are the main ingredients for successful school careers. From the perspective of learning as social interaction in a dynamic educational context, the aim of this study was to answer the question whether the level of cognitive development at the beginning of school is an indicator of future school performance, and to what extent the progress in development can be stimulated. At the start of primary school valid, reliable and objective comparable information on the cognitive capacities of pupils is generally not available. Instead, the socio-ethnic background of the pupil is the most straightforward and reliable information in stock to predict future cognitive progress.

However, it has been demonstrated that the use of socio-ethnic background may well lead to biased expectations of the social and cognitive opportunities for a successful school career (Van der Hoeven-Van Doornum, 1990; Van der Hoeven-van Doornum, Voeten, & Jungbluth, 1989; 1990a; 1990b).

Starting school, pupils with a high social background are privileged compared with other pupils with a less advantageous starting point, insomuch that they have better verbal skills, are acquainted with the prevalent culture and have access to available educational information (Bowles & Gintis, 1976). Next to this better starting point, teachers influence performance and progress of their pupils by setting higher or lower expectations and aspiration levels which are related to the social background of their pupils. Generally speaking, expectations of cognitive capacities may be fairly correct if they are based not only on the social background of pupils but also on information stemming from their achievement test scores. Since the nineties a substantial body of research on teachers' expectations has come available revealing that high expectations may even work out as a form of teacher effectiveness (Van der Hoeven-Van Doornum, 1993; Lee, Bryk, & Smith, 1993).

Because the various aspects of teacher expectations are strongly intertwined, for clarity's sake a distinction should be made between teachers' perceptions of the social and cognitive background of pupils, and the cognitive aspiration level teachers set for their pupils. Perceptions on the social and emotional development and cognitive capacities may form a personal profile, while aspiration levels may function as implicit educational targets.

Explicit target setting fits in the contemporary view on Vygotski's concept of the zone of proximal development (ZPD). Target setting in a dynamic pedagogical-didactical context should stimulate pupils in their development and elicit progress. Although formulating teaching-learning objectives is the major intent of curriculum development, this has hardly ever resulted in individual target setting for pupils. The curriculum in the first two years of primary school is mostly child-directed and development-oriented, which probably means that progress is mainly determined by the personal skills pupils already have at their disposal and less by intended instruction based on explicit educational goals.

7.2 Research questions

In this concluding chapter we are interested in longitudinal effects of baseline assessment at one hand and in separating out the effects of pupil and class characteristics on test scores over time at the other hand. The first aim is to seek the best predic-

tor of achievement on individual level over a period of several years; and the second one is to explore which class and teacher characteristics can explain the differences in achievement outcomes. The essence of these research questions is what teaching practices – including regular assessment – enable teachers to help more children do well. For the former aim rather extensive literature is available on which to draw upon for indicators of developmental processes and good predictors of later achievement; but with regard to the latter there is little insight in the efficiency of the teaching-learning trajectory in the lower years of primary schools and which practices help pupils do well in school.

Longitudinal hypothesis

In the previous Chapters 3 and 4 it was already shown that initial differences in achievement due to socio-ethnic background were partly caught up in the first year of school. Children with a disadvantaged background appeared to make a real developmental spurt, in particular with respect to their reading acquisition. Therefore, we hypothesize that over a period of three years baseline assessment will prove to be a better predictor of later achievement than the socio-ethnic background. Thus, baseline assessment is considered to measure pupils' own starting point reflecting the cognitive level as a result of preschool learning and implicitly controls for socio-economic status. With regard to the information on pupils' profiles – at least among teachers – it is generally assumed that happy, well-adapted children will do better at school than children with inappropriate behaviour. With regard to the information on pupils' profiles we'll test the extent to which teachers' expectations and judgments on cognitive capacities and social and behavioural aspects affect the cognitive progress measured by OBIS baseline assessment.

Multi-level hypothesis

As we already stressed in Chapter 3, teachers are not really familiar with explicit target setting. Notwithstanding that teachers in the experimental group experienced the meetings on target setting as very useful and informative; we expect little or no effects of the experimental condition on progress. It is more so because it is very possible that in the course of the study teachers in the control group have adapted their teaching practices also according to the feedback on the OBIS they received. We hypothesize that if variation in progress on group level can be explained this will be the impact of teaching practices in general that are basically related to the learning progress of the pupils, regardless of the experimental condition of the teachers.

7.3 Methods

Nearly 450 children with a mean age of four and a half year were assessed upon entering school, using a broad baseline assessment, and then again one year later in Kindergarten. The reading and mathematics achievements of the same children were again assessed one year later in grade 3. These data were used to look at the prediction of the academic achievement of children with varying degrees of cognitive and social profiles. To answer the first research question, the LISREL (Linear Structural Analysis, Jöreskog & Sörbom, 2001) procedure has been applied to the correlation matrix. For the second question of which class and teacher characteristics explain the differences in achievement outcomes, MLwin (2001) for multi-level analysis was used.

7.4 Results

Table 7.1 presents the descriptive statistics of the variables used in the longitudinal and multilevel models. Data were available for 295 pupils. Pupils with a missing value on one or more assessments were excluded from the analysis.

Table 7.1 – Descriptive statistics for the variables in the model for cognitive progress N=295, Year 1 to year 3.

	mean	sd	min – max
Year 1			
Maths	50.58	9.66	20.51 – 79.49
Reading	51.33	10.24	17.83 – 79.48
Total	51.19	9.76	20.51 – 79.49
Pupil profile			
School well-being	4.18	.49	2.17 – 5.00
Self-confidence	3.60	.63	1.71 – 5.00
Social behaviour	3.59	.73	1.00 – 5.00
Attitude	3.34	.80	1.00 – 5.00
Parental support	4.09	.84	1.33 – 5.00
Cognitive capacities	3.51	.62	1.88 – 5.00
Development	4.18	.56	2.43 – 5.00
Test-effect	1.88	.86	1.00 – 4.20
Year 2			
Maths	50.95	9.72	24.09 – 80.37
Reading	51.12	9.64	19.63 – 80.37
Total	51.10	9.58	23.11 – 80.37
Pupil profile			
School well-being	4.17	.50	2.71 – 5.00
Self-confidence	3.69	.58	1.86 – 5.00
Social behaviour	3.57		1.20 – 5.00
Attitude	3.47	.79	1.00 – 5.00
Parental support	4.03	.71	1.75 – 5.00
Cognitive capacities	3.62	.65	1.13 – 4.88
Development	4.41	.50	3.00 – 5.00
Test-effect	1.98	.74	1.00 – 4.33
Year 3			
Maths	49.91	10.00	24.36 – 76.63
Reading	50.11	10.12	22.04 – 75.63
Total	50.00	10.10	19.87 – 80.13
Pupil profile			
Vocabulary	49.36	10.03	22.04 – 80.13
Non-verbal reasoning	49.72	9.97	24.37 – 80.12
Cognitive category	1.41	1.18	0 – 3
Socio-ethnic status	2.46	.82	1 – 3
Age in months (End of Year 3)	101.62	5.50	85 – 112

Table 7.2 presents the correlations between the variables used in the analyses. As a consequence of variables recoding, all correlations are positive, except for the correlation between SES and age. Apparently children with a higher social background are younger. The descriptive statistics and correlations may differ slightly from those in previous chapters because of the different samples used in the analyses.

Table 7.2 – Correlations between variables over a period of 3 years. N=295

	Start Total score	Year 2 Maths	Year 2 Reading	Year 2 Total score	Year 3 Maths	Year 3 Reading	Year 3 Total score	SES
Year 2 maths	.72							
Year 2 reading	.68	.69						
Year 2 total	.78	.90	.91					
Year 3 maths	.52	.59	.40	.53				
Year 3 reading	.48	.49	.42	.49	.59			
Year 3 total	.54	.57	.46	.55	80	.95		
SES	.46	.28	.13	.25	.27	.17	.21	
Age	.25	.17	.17	.17	.01	.02	.02	-.21

7.4.1 Prediction of cognitive progress

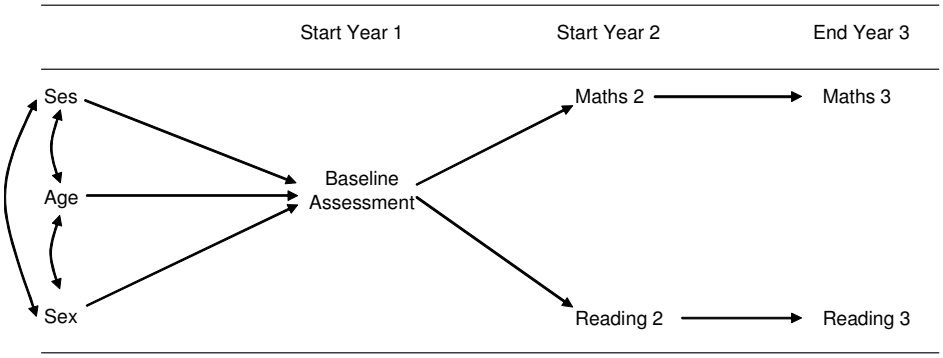
LISREL model specification

The starting point for the analyses is specification of a structural model that represents the causal hypotheses. On the left side there are the exogenous variables socio-ethnic background (SES), sex and age. The causes of these variables are unknown and thus not represented in structural models, the associations among exogenous variables are unanalysed and are assumed to covary (<–>). SES, sex and age are assumed to affect other variables in the model. Specific hypotheses about causal relations are represented by unidirectional effects (–>). If variables are concurrently measured and there is no plausible rationale for causality, the variables are also assumed to covary.

Figure 7.1 depicts the conceptual model for cognitive progress in the lower years of primary education with pupil characteristics at the start of Year 1: three exogenous variables – SES, sex and age – followed by five endogenous variables representing the levels of cognitive attainment assessed on distinct measure moments in three consecutive years. Each of the variables was assessed with a single observed measure. The endogenous variables can represent a predictor variable and a criterion in the same analysis as well. It is noted that the first assessment score is a baseline mainly

representing preschool learning, while the following scores are supposed to represent primarily learning at school.

Figure 7.1 – Initial model for the prediction of test scores over a period of 3 years

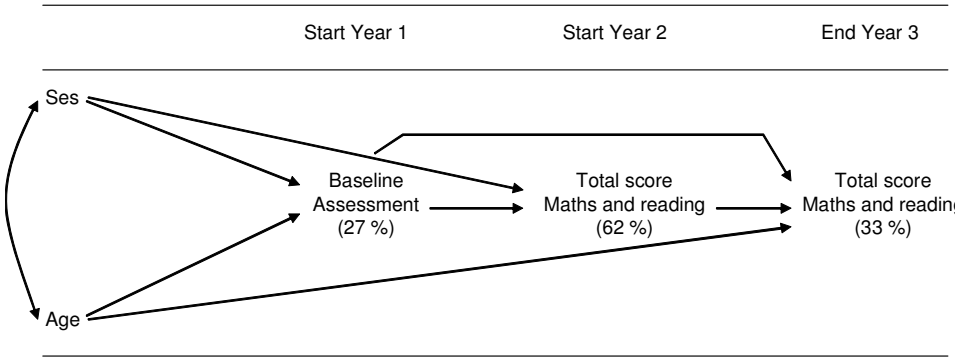


The approach taken was to test the model against the data and develop new models including the effects of teacher expectations. The cognitive level measured at entry of year 1 was, in addition to the exogenous variables, used as the determinant of teachers' expectations. In turn, teachers' expectations were seen as antecedent to achievement later on in time. Teachers' expectations were also supposed to affect the curriculum level aimed by the teacher for an individual pupil at the transition from year 2 – at the end of kindergarten – to year 3.

Cognitive progress

The correlation matrix was used to analyse the differences between variables. The initial model did not fit the data: $chi^2 = 82.62$, df =16, p = .00. With no shift of meaning of the conceptual model, the separate maths and reading assessment scores were replaced by one total score of maths and of reading on the OBIS. This model is depicted in Figure 7.2. The fit of this model is very satisfying: $chi^2 = 2.13$, df = 2, p =.35, AGF = .98. The OBIS baseline assessment at the start of year 1 predicts the total scores of the maths and reading tests almost three years later.

Figure 7.2 – Model 1 for the prediction of test scores over a period of 3 years



Chi^2 = 2.13, df = 2, p = .35, AGF = .98

Path coefficients are regression coefficients under control for correlations among other causal relations in the model. Table 7.3 shows the standardised direct and indirect effects of the model. The standardised solution is used for comparing the sizes of the variables coefficients that are expressed in different measuring scales[1]. Usually beta weights are less than their correlations, because they are corrected for intercorrelations among the predictors (Kline, 1998).

Table 7.3 – Model 1. Final model for the prediction of total scores on maths and reading in Year 3, over a period of 3 years. Standardised direct effects and total effects.

	Direct effects				Total effects			
	SES	Age	BLA	Year 2 Total score	SES	Age	BLA	Year 2 Total score
Y1 BLA	.48	.33			.48	.33		
Y2 Total score	-.12		.85		.28	.19	.85	
Y3 Total score	-	-.09	.27	.34	.23	.15	.56	.34

Chi^2 = 2.13, df = 2, p =.35, AGF = .98

1 The unstandardised regression coefficient (b) is affected by the measuring scale. A large b can be insignificant, while a small b can be significant.

Note that the direct effect of SES on the total score in Year 2 is -.12 and the total effect is .28. The indirect effect of .41 is the product of the direct effects of SES –> BLA and BLA –> total score in Year 2: (.48 * .85). The indirect effect is a mediator that transmits a portion of the effect of SES via the Baseline Assessment (Year 1) on the total score in Year 2. The direct effect suppresses[2] the intercorrelation among the predictors.

Thus, when both SES and BLA are predictors of the total score in Year 2, the total beta weight for SES is negative (-.12), which is the opposite sign of its correlation (.25). Furthermore, the beta weight for BLA (.85) is greater than its correlation with the total score (.78). Correcting both associations reveals that the relation of SES to the total score in Year 2 is actually negative, once BLA is controlled.

It is known that children with a low social background are more likely to get low test scores. However, these results suggest that when correcting for baseline assessment at entry, after one year of school low SES children achieve better once baseline assessment is controlled. Also, the relation of BLA to the total score in Year 2 appears even to be stronger once the effect of SES is held constant.

Regarding the negative effect of age on the total score in Year 3, the same phenomenon of suppression shows that the absolute value of the regression coefficient of age on the total score in Year 3 is greater than its correlation (.02) once other variables are controlled. This is the result of intercorrelations among SES, age and achievement scores: a) children from low social families are likely to get lower scores and b) older children are likely to get better scores. As we mentioned above in relation to the negative correlation between SES and age: the children with a higher social background are younger. The negative beta weight of age on the total score in Year 3 indicates that it took lower SES children with lower scores more time, for instance because of a prolonged Kindergarten year, to get in Year 3 than higher SES children with higher scores.

2 Suppression occurs either when the absolute value of a predictor's beta weight is greater than its Pearson correlation with the criterion [....] or when the two have different signs (Kline, 1998, p.39).

Table 7.4 – Correlations between variables over a period of 3 years. N=295

	BLA Total score	Year 2 Total score	Year 3 Total score
Year 1			
School well-being	.15	.18	.06
Self-confidence	.19	.24	.17
Social behaviour	.09	.16	.14
Attitude	.26	.26	.27
Parental support	.39	.28	.16
Cognitive capacities	.54	.51	.44
Development	.43	.37	.29
Test-effect	.14	.11	.04
Year 2			
School well-being	.16	.23	.01
Self-confidence	.24	.26	.07
Social behaviour	.02	.04	.08
Attitude	.24	.25	.30
Parental support	.42	..33	.18
Cognitive capacities	.54	.56	.42
Development	.30	.23	.20
Test-effect	-.01	.01	.06
Year 3			
Cognitive category	.42	.43	.61

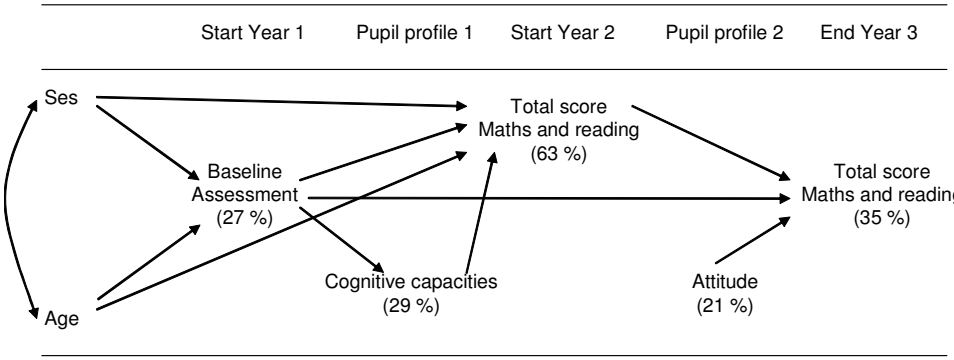
Pupils' profiles

Next, one by one the variables comprising the pupil's profiles were tested on their contribution to the model separately. Table 7.4 presents the correlations between assessment scores and the pupil characteristics. It appeared that in Year 1 self-confidence, attitude and cognitive capacities have a significant contribution to the prediction of the OBIS assessment scores in Year 2 or 3, or both. In Year 2 school well-being, self-confidence and attitude have a significant contribution to the prediction of the OBIS assessment scores in Year 3. Just as in Year 1, one might expect significant effects of 'cognitive capacities' and 'cognitive category' in Year 2, but notwithstanding the high correlations with later achievement scores, these teacher's judgment had no significant contribution to the model. In Table 7.4 the variables with significant effects are printed in bold.

Upon that, the six significant pupil characteristics were added to the model depicted in Figure 7.2 and were tested simultaneously. Yet, in the company of other characteristics of pupil's profile four of them, self-confidence and attitude in Year 1, and school well-being and self-confidence in Year 2 lost their significant contribution to the model. Together with the pupil's social background, age and assessment scores, the pupil's cognitive capacities assessed in Year 1 and the attitude in Year 2, re-

mained among the predictors of cognitive progress over a period of three years. The fit of the model is excellent: chi^2= 8.42, df = 8, p =.39, AGF = .97. The model is depicted in Figure 7.3; and the results are given in Table 7.5.

Figure 7.3 – Model 2 for the prediction of test scores over a period of 3 years



Chi^2 = 8.42, df = 8, p =.39, AGF = .97

Table 7.5 – Model 2. Final model for the prediction of Total scores on maths and reading in year 3 by on-entry assessment in Year 1, follow-up assessment in Year 2 and pupil's profile. Standardised direct effects and total effects.

	Direct effects					
	SES	Age	BLA on entry	Cognitive capacities	Year 2 Follow-up	Attitude
BLA on entry	.48	.33				
Cognitive capacities			.54			
Year 2 Follow-up	-.11	-.09	.79	.12		
Attitude		-.11		.45		
Year 3 Total score	-	-	.25		.32	.16
	Indirect effects					
	SES	Age	BLA on entry	Cognitive capacities	Year 2 Follow-up	Attitude
BLA on entry	-	-	-	-	-	-
Cognitive capacities	.26	.18	-	-	-	-
Year 2 Follow-up	.40	.28	.06	-	-	-
Attitude	.13	.08	.24	-	-	-
Year 3 Total score	.23	.14	.31	.11	-	-
	Total effects					
	SES	Age	BLA on entry	Cognitive capacities	Year 2 Follow-up	Attitude
BLA on entry	.48	.33	-	-	-	-
Cognitive capacities	.26	.18	.54	-	-	-
Year 2 Follow-up	.29	.20	.85	.12	-	-
Attitude	.13	-.03	.24	.45		-
Year 3 Total score	.23	-.14	.56	.11	.32	.16

Chi^2 = 8.42 df = 8, p = .39, AGF =.97

Answering research questions

It was hypothesized that over a period of three years baseline assessment is a better predictor of later achievement than the socio-ethnic background. This hypothesis was tested by omitting in Model 1 (Fig. 7.2) the direct effect of baseline assessment on the total score in Year 3. If the model without the relation BLA –> total score in Year 3 is not significantly worse than model l, then the hypothesis is not supported. The trimmed model without BLA –> total score in Year 3 did not fit the data: chi^2 = 13.97, df =3, p=.03, AGF=.91 and the difference with Model 1 was significant: $chi^2_{difference}$ = 11.84, df = 1, p < .001. The addition of the direct effect of BLA on the total score in Year 3 results in a significant reduction in the model's chi^2. The modification index for BLA –> total score in Year 3 is 11.93 with an expected change of

.27; the modification index for SES –> total score in Year 3 is .69 with an expected scores of -.04. In Table 7.3 it can be seen that if information on entry is available there is no direct effect from SES on the total score in year 3. Thus, baseline assessment measures pupils' own starting point reflecting the cognitive level as a result of preschool learning and implicitly controls for socio-economic status.

With regard to the information on pupils' profiles – at least among teachers – it is generally assumed that happy, well adapted children will do better at school than children with inappropriate behaviour. It appeared that 'school well-being' and 'self-confidence' had significant, but marginal contributions to Model 1 which disappeared when tested simultaneously with cognitive capacities in Year 1 and attitude in Year 2. This indicates that these two pupil characteristics have valid, but rather restricted contribution to the prediction of cognitive progress.

With regard to the two remaining aspects of the pupils' profiles the extent to which cognitive capacities and attitude contribute to the prediction of cognitive progress was tested; and secondly, it was tested if they can be considered as substitutes for baseline assessment. The difference between Models 1 and 2, that means without and with pupil characteristics, is $chi^2_{difference} = 6.29$, df = 6, $p > .30$. The nonsignificant value of the $chi^2_{difference}$ statistic suggests that the overall fits of the two models are comparable. This means that adding this information to the model did not improve the prediction of cognitive progress.

Yet, this does not necessarily implicate that teachers do not have a valid perception of the cognitive abilities of a pupil. What if there is no baseline information available: to what extent are cognitive capacities and attitude useful substitutes for baseline assessment? To answer this question, the direct effect of Baseline assessment Year 1 on the total score in Year 3 was deleted; while the effects of baseline assessment on the pupil's profile were kept in the model. All of this occurs because it can be argued that the teachers' perceptions are mainly based on a conscious or unconscious evaluation of the child's academic performance. The fit of this testing model was poor: $chi^2 = 18.64$, df = 9, p = .03, AGF .94. The difference between the models was highly significant: $chi^2_{difference} = 10.22$, df = 1 indicating that the teacher's perception of the cognitive capacities and attitude are not equivalent to baseline assessment as predictors of later achievement.

Comparing the magnitude of the effects in the models with and without the path BLA –> total score in Year 3, it was found that the small effects of the cognitive capacities and attitude in Model 2 (Table 7.5) hardly increased after constraining the path between BLA and total score in Year 3. The total effect of cognitive capacities –> total

score in Year 3 increased from .11 to .14 and the total effect of attitude –> Total score from .16 to .18.
Overall, the findings show that neither the social background nor the teacher's perception of the cognitive profile of the pupil is a better substitute for baseline assessment as a predictor of later achievement.

7.4.2 Cognitive progress: differences between pupils and classes

Multilevel model specification
Multi-level analysis was used to account simultaneously for variation on pupil and group level. This distinction is essential because there is much more variation between pupils in a group than between groups. Furthermore, pupils within a group tend to be more alike than pupils in different groups. This can be understood from the fact that if particular characteristics of the classroom context favour educational achievement, all pupils in that group are likely to benefit from it.

The first task in causal inference is to identify the units, treatments and potential outcomes. Longitudinal data are not strictly hierarchically nested since pupils do not remain together as a class over time; pupil's teachers are not the only ones who may change each year, their class mates are also changing. Since we are here dealing with three-year data comprising year 1 to year 3 in which period pupils may have changed classes and teachers, we have chosen to assign pupils to the initial 25 groups in which they started Year 1. This is not only done for simplicity of analyses but because the majority of pupils remained in the same group and in the first two years often with the same teachers too. The models were estimated with assessment scores as the dependent variable, the estimation sample on 295 pupils comprised those pupils with adjacent–year observations.

Measuring progress requires controlling initial level of achievement. This is most transparently done if the pre- and post-tests are on the same achievement scale ("vertically equated"). Because the value added method measures gain from a pupils' own starting point, it implicitly controls for socio-economic status and other background factors to the extent that their influence on the post-test is already reflected in the pre-test score. Test scores are also a function of both student- and class-level variables. The overall goal of the multilevel approach is to clarify the interplay between estimated teacher effects and repeated outcomes of students over time.

Class effects are characterized as random variables that contribute to tests scores. The class effects are contributions of teachers to student achievements. The implicit feature in the model is that the teacher has a constant effect on all pupils in the class.

According to Scheerens & Bosker (1997, p. 79) the school effectiveness research identifies school effects for about 10 percent or less.

Differences between pupils and classes

The basic model that serves as the starting point for answering the research questions is the empty model – null model – only containing the data on the response variable at two levels of observation – the individual and the group. The null models, intercepts only, for the total OBIS scores in Years 1, 2 and 3 show how the variances of the achievement scores are distributed over the individual and group level[3]. The intercept of each model simply represents the average OBIS score across all pupils, respectively 51.48, 51.24 and 49.72 in the Years 1, 2 and 3. The intra-class correlations vary over the years from .30 in Year 1 to .13 in Year 3. A high intra-class correlation indicates that groups are homogeneous and very different from each other as well (Kreft & De Leeuw, 1998, p. 3). In educational research values between 0.05 and 0.20 are common.

Comparison of the variance components of the null models for the different years in Tables 7.6 and 7.7, shows that at the start of school the differences between the groups are very high. This is mainly due to the social background of the pupils. In Chapter 4 (Tables 4.6 and 4.7) it was already shown that this applies in particular to reading in Year 1. In Year 1 the effect of SES on reading (t-value 7.71) is more than two times the effect of SES on math (t-value 3.15) indicating that early numeracy is much less associated with social background than early literacy.

Table 7.6 – Years 1 and 2. Differences between pupils and classes in OBIS scores.

Model 0	Total 1		Total 2	
Variance components				
Intercept	51.48	(1.07)	51.24	(0.89)
Pupils	72.27	(6.22)	79.87	(6.87)
Class	21.53	(8.00)	11.84	(5.43)
Total	93.70		91.71	
% variance				
Pupil	77 %		87 %	
Class	23 %		13 %	
Intra-class correlation	.30		.15	
Deviance	2136.45		2153.91	
Df	3		3	

3 Here we concentrate the analyses on the total OBIS scores of each year.

Note also the impact of schooling after one year (Tables 7.6 and 7.7): the null model in Year 2 shows that the intra-class correlation of .30 in Year 1 has decreased to 0.15 in Year 2 which indicates that the differences between the groups caused by social background have diminished considerably. In the Years 2 and 3, the ratios between the variance on pupil and group level, respectively .15 and .13 are rather close.

Table 7.7 – Year 3. Differences between pupils and classes in OBIS scores. N=295

Model 0	Reading 3		Maths 3		Total 3	
Variance components						
Intercept	49.85	(0.83)	49.65	(0.98)	49.72	(0.88)
Pupils	92.39	(7.94)	82.77	(7.12)	89.31	(7.67)
Class	8.60	(4.77)	16.25	(6.78)	11.18	(5.50)
Total	100.99		99.02		100.49	
% variance						
Pupil	91 %		84 %		89 %	
Class	9 %		16 %		11 %	
Intra-class correlation	.09		.20		.13	
Deviance	2190.36		2168.96		2184.37	
Df	3		3		3	

After the null model and before exploring which class and teacher characteristics can explain the differences in achievement outcomes, variables on pupil level were included in the model. To compare their unique contribution to the models, the variables were initially included in the models on a variable-by-variable basis. After that, the variables with a significant effect were tested simultaneously.

SES and achievement scores explaining differences on pupil and class level

Regarding the impact of SES on achievement scores, Tables 7.8 and 7.9 reveal the decrease in the course of the years of the contribution of SES to the explanation of differences in achievement scores. SES has an important contribution in Year 1 (t-value 6.83); it explains 71 % of the differences between the groups. In Year 2 the effect is about half the size (t-value 3.5), and in Year 3 it is comparatively low but still significant (t-value 2.62).

Table 7.8 – Years 1 and 2. SES and test scores explaining differences between pupils and classes in OBIS scores. Unstandardised regression weights (sd)

	Total year 1		Total Year 2	
Model	1 (Y1)	1 (Y2)	2 (Y2)	3 (Y2)
Pupil				
SES	5.26 (0.77)	2.94 (0.83)		1.38 (.58)
OBIS Y1			0.80 (0.04)	0.83 (0.04)
Variance components				
Intercept	38.54 (1.98)	44.04 (2.16)	10.11 (2.00)	11.93 (2.09)
Pupil	68.36 (5.87)	78.33 (6.73)	31.14 (2.68)	30.97 (2.66)
Class	6.29 (3.52)	7.61 (4.17)	5.09 (2.26)	4.01 (1.94)
Total	74.65	85.94	36.23	34.98
% variance				
Pupil	92 %	91 %	86 %	88 %
Class	8 %	9 %	14 %	11 %
% explained variance compared with null model				
Pupil	5 %	2 %	60 %	61 %
Class	71 %	36 %	57 %	66 %
Deviance	2101.35	2142.22	1877.64	1872.37
Df	4	4	4	6
Dev_{dif} with model 0	35.1	11.69	276.27	281.54
p	< .001	< .001	< .001	< .001
Dev_{dif} with model 2 (Y2)				5.27
p				.005

Model 1 (Y1)= SES;
Model 1 (Y2) = SES; Model 2 (Y2) = OBIS Y1; Model 3 (Y2) = SES and OBIS Y1.

Substituting in Year 2 the effect of SES by the effect of OBIS test score in Year 1, shows that an achievement score (t-value 20) is a much better predictor of later achievement than the social background of the pupil. The OBIS Year 1 score explains differences between pupils (60 %) and classes as well (57 %). In Year 3, inclusion of OBIS achievement scores and SES in the model (Table 7.9, model 4 Y3) led to an insignificant remaining effect of SES (t-value .92).

Table 7.9 – Year 3. SES and test scores explaining differences in OBIS scores between pupils and classes. Unstandardised regression weights (sd)

	Total year 3			
Model	1 (Y3)	2 (Y3)	3 (Y3)	4 (Y3)
Pupil				
SES	2.25 (0.86)			0.77 (0.83)
OBIS Y1		0.61 (0.05)		0.35 (0.08)
OBIS Y2			0.60 (0.05)	0.35 (0.08)
Variance components				
Intercept	44.26 (2.25)	18.29 (2.80)	18.78 (2.69	15.77 (3.06)
Pupil	89.06 (7.65)	59.85 (5.15)	59.23 (5.09)	55.59 (4.78)
Class	7.46 (4.38)	11.05 (4.70)	9.85 (4.33)	10.59 (4.47)
Total	96.52	70.90	69.08	66.18
% variance				
Pupil	92 %	84 %	86 %	84%
Class	8 %	16 %	14 %	16 %
% explained variance compared with null model				
Pupil	0 %	33 %	34 %	38 %
Class	33 %	1 %	12 %	5 %
Deviance	2.178.25	2.072.30	2067.50	2.051.06
Df	4	4	4	6
Dev_{dif} with model 0	6.12	112.07	116.87	133.31
p	< .02	< .001	< .001	< .001
Dev_{dif} with model 3 (Y3)				16.42
p				< .001

Model 1 (Y3)= SES; Model 2 (Y3) = OBIS Y1; Model 3 (Y3) = OBIS Y2;
Model 4 (Y3)= SES, OBIS Y1 and OBIS Y2.

As in the previous Years 1 and 2, Table 7.9 (model 1 Y3) shows that SES explains no variance on pupil level but only variance on class level, while the achievement scores explain variance on both levels. Even with an intermediate period of almost three years, the OBIS score Y1 – the baseline assessment at the age of 4 – appears to be an equally good predictor of learning achievement at the end of Year 3 as the OBIS score in Year 2.

After SES and achievement scores, the age and the other characteristics of the pupil's profile were included in the models. See Tables 7.10 and 7.11.

Year 1: Pupil's age and profile

The results for Year 1 are given in Table 7.10. Differences in age explain only differences in OBIS scores at the start of school, which means that older children perform better than younger children. In Year 1 we estimated a negative value for the intercept as a result of including age in the model. In later years, the age differences – yet to the disadvantage of older pupils – are no longer of significance.

Table 7.10 – Year 1. Pupil's age and profile explaining differences in OBIS scores between pupils and classes. Unstandardised regression weights (sd)

	Total Year 1		Total Year 2		
Model	2 (Y1)	3 (Y1)	4 (Y2)	5 (Y2)	
Pupil					
SES	5.61 (0.72)	4.16 (0.70)	1.60 (.59)	1.40 (.59)	
OBIS Y1			.86 (.04)	.76 (.05)	
Age	.62 (.10)	.47 (.09)	- .17 (0.08)	-. 13 (.08)	ns
Cognitive capacities		7.09 (.65)		2.42 (.60)	
Variance components					
Intercept	-25.89 (10.16)	- 31.30 (9.06)	27.89 (7.57)	20.16 (7.65)	
Pupil	60.01 (5.16)	41.25 (3.91)	30.30 (2.61)	28.62 (2.46)	
Class	5.33 (3.02)	10.05 (3.91)	4.38 (2.02)	4.44 (2.00)	
Total	65.34	51.30	34.68	32.06	
% variance					
Pupil	92 %	80 %	87 %	89 %	
Class	8 %	20 %	13 %	11 %	
% explained variance compared with					
Model	1 (Y1)	2 (Y1)	3 (Y2)	4 (Y2)	
Pupil	12 %	31 %	2 %	6 %	
Class	15 %	+ 47 %	+ 8 %	+ 1 %	
Deviance	2062.48	1967.33	1867.63	1851.87	
Df	5	6	6	7	
Dev_{dif} with model	1 (Y1)	2 (Y1)	3 (Y2)	4 (Y2)	
	38.86	95.15	4.74	15.76	
p	< .001	< .001	< .05	< .001	

Model 2 (Y1) = SES and Age; Model 3 (Y1) = SES, AGE and profile;
Model 4 (Y2) = SES, OBIS Y1 and Age; Model 5 (Y2) = SES, OBIS Y1, Age and profile.

Included in the model on a variable-by-variable basis several of the aspects of the pupils' profile perceived by the teacher – school well-being, self-confidence attitude, development and parental support and cognitive capacities – had a significant contribution to the explanation of differences in OBIS scores in Year 1. Tested together, all coefficients but one were no longer significant: only the teacher's perception of the pupil's cognitive capacities remained significant, see model 3 (Y1).

The effect of cognitive capacities with a t-value of 11 is almost twice the effect of SES (t-values 5.94). Inclusion of the teachers' perceptions of the cognitive capacities of their pupils explains 31 % of the variance on pupil level, but surprisingly the variance on class level has almost doubled. It seems strange and against intuition that a variable on pupil level increases the unexplained variance on class level, a meaningful interpretation is not available. Similar problems described by Snijders and Bosker (1999) and Hox (2002), are attributed to statistical difficulties. It appears that so far there is no satisfactory solution for this.

As described earlier, the baseline assessment was carried out at school entry, therefore the OBIS score in Year 1 mainly reflects preschool knowledge which can not be explained by differences in teaching style or other class characteristics. Hence, we conclude that model 3 (Y1) is the final model for Year 1. Compared with the null model (Table 7.6) 43 % of the variance on pupil level and (53 %) on class level is explained by the social background, age and the teacher's perception of the cognitive capacities of the pupil.

Year 2: Pupil's age and profile

The same approach of including the variables on a one-by-one basis was used for the analyses of Year 2. The anomaly of increasing class level variance caused by age and the perception of cognitive capacities did occur again, but on a smaller scale, probably because the contribution of age to the explanation of OBIS scores in Year 2 was insignificant yet, and secondly the inclusion of the baseline assessment OBIS in Year 1 improves the model considerably. The results for Year 2 are given in Table 7.10.

Regarding the effect of the several aspects of the pupils' profiles, similar results as in Year 1 were found for Year 2. If introduced as a stand-alone variable, the school well-being and the cognitive capacities of the pupils, and the test-effect as well, displayed significant coefficients, but when tested simultaneously only the contribution of cognitive capacities remained in the model 5 (Y2). The effect with a t-value of 4 is also rather large considering the presence of the previous OBIS score (Year 1) in the model.

So far we have tested the models with all pupil variables fixed, because no random effects for the variables on pupil level were found, otherwise. All in all, the level-one variables in Model 5 (Y2) explain about two thirds of the variance on pupil level and class level in model 0 (Table 7.6). The variable 'test-effect' referring to individual target setting for the pupil based on OBIS baseline test scores did not contribute to the explanation of differences in achievement scores. Apparently teachers did not change their expectations and teaching practices after the baseline assessment. So far model 5 (Y2) concludes the analyses using the level-one variables.

Year 3: Pupil's age and profile

In contrast with the Years 1 and 2, model 5 (Y3) in Table 7.11 shows that the social background and the age of the pupils have no significant effects in Year 3. In the more economical model 6 (Y3), the explained variance on pupil and class level is the outcome of previous achievement scores. Next to this, the characteristics of the pupil's social and cognitive profiles in the Years 2 and 3 were tested for their contribution to the model. Three of them, the score for non-verbal reasoning, the attitude and cognitive category of the pupils according to their teachers had significant coefficients. On pupil level these characteristics add another 16 percent to the explanation of the differences between the OBIS score at the end of Year 3, and also some 5 percent to those on class level.

Using again the parameter estimates of the null model (Table 7.7) as a yardstick, model 7 (Y3) explains almost half of the variance on pupil level, but unlike the results of the previous years, only 13 % of the differences between the groups.

Random slope for Cognitive category

Testing for random effects of pupil variables, a significant effect for 'cognitive category' was found. The coefficient for the pupil variable 'cognitive category' was allowed to be random, because we expected that the effect of 'cognitive category' might be different among groups. Table 7.11 – model 8 (Y3) – shows that by adding the random effect the intercept variance (class level) increased from 9.73 (4.01) to 27.04 (12.22). This is the result of multicollinearity, i.e. the high correlation between intercept and slope variances, as indicated by the negative covariance (- 11.58). Intercepts and slopes are negatively correlated as it can be seen in Table 7.11.

Because of the correlations between the random variances, they cannot be summed anymore, and the R^2 can no longer be calculated. Based on the difference in deviance between the models 7 (Y3) and 8 (Y3) it can be concluded that model 8 is a better fit, particularly because the significance of the fixed coefficient for 'cognitive category' has almost remained the same (respectively a t-value of 2.89 in model 7 and a t-value of 2.59 in model 8). The difference in deviance between models 7 (Y3) and 8 (Y3) is

highly significant: chi^2 11.64, df= 2, p= < .01. The random model 8 (Y3) is the starting model for the next set of analyses where variables on class level will be introduced to answer the question of which group characteristics are responsible for differences between classes.

Table 7.11 – Year 3. Pupil's age and profile explaining differences in OBIS scores between pupils and classes. Unstandardised regression weights (sd)

	Total year 3			
Model	5 (Y3)	6 (Y3)	7 (Y3)	8 (Y3)
Pupil				
SES	.99 (.83) ns			
OBIS Y1	.39 (.09)	.32 (.08)	.20 (.08)	.22 (.08)
OBIS Y2	.33 (.08)	.36 (.08)	30 (.07)	.30 (.07)
Age	-.16 (.11) ns			
Cognitive category			1.53 (.53)	1.76 (.68)
Non-verbal reasoning			.24 (.05)	.22 (.05)
Attitude			1.29 (.57)	1.19 (.57)
Variance components				
Intercept	31.69 (10.62)	14.69 (2.82)	5.64 (3.12)	4.82 (3.20)
Pupil	55.43 (4.78)	55.88 (4.80)	47.10 (4.05)	44.18 (3.90)
Class	9.63 (4.82)	10.24 (4.38)	9.73 (4.01)	-
Total	65.06	66.12	56.83	-
Random intercept				27.04 (12.22)
Random slope				5.58 (2.93)
Covariance slope-intercept				-11.58 (5.65)
% variance				
Pupil	85 %	85 %	83 %	-
Class	15 %	15 %	17 %	-
% explained variance compared with				
Model	4 (Y3)	4 (Y3)	6 (Y3)	-
Pupil	0 %	1 %	16 %	-
Class	9 %	3 %	5 %	-
Deviance	2048.67	2051.90	2003.54	1991.90
Df	7	5	8	10
Dev_{dif} with model	4 (Y3)	4 (Y3)	6 (Y3)	7 (Y3)
	2.39	.84	48.36	11.64
p	ns	ns	< .001	< .01

Model 5 (Y3) = SES; OBIS Y1, Y2, Age; Model 6 (Y3) = OBIS Y1, Y2; Model 7 (Y3) = OBIS Y1, Y2 and profile.

Group characteristics in Year 2 and Year 3

The individual variables explained part of the individual and part of the group variances as well. In this step group characteristics referring to teaching style are introduced into the models for Year 2 and Year 3. If group variables influence differences between pupils, we should find a reduction of variance either in intercepts or slopes on class level, or on both, because group level variables can only explain class level variance. In the next set of analyses depicted in Table 7.12, variables on class level will be introduced to answer the question which group characteristics are responsible for differences between classes.

Year 2: Teaching style

Before including the level-two variables in model 5 (Y2), the insignificant variable 'age' was removed. This happened because large and complex models may seem more realistic, but these models are often very instable: small changes in the model may result in large changes in the results due to e.g. collinearity (Kreft & De Leeuw, 1998).

In Year 2, two variables for teaching style – 'book orientation' and 'Ideas of Reading' – both referring to teaching style were found to contribute significantly to the model. Subsequently, the variable SES lost its significance to the model and was removed. The final model 7 for Year 2 is depicted in Table 7.12. The results show that with a t-value of 17.25, the OBIS total score assessed in Year 1 is the best predictor of the OBIS total score in Year 2, followed by the teacher's perception of the cognitive capacities of the pupil (t-value 4.74); book orientation (t-value 3.58) and ideas of reading (t-value 2.59) on class level. The latter variables indicate the emphasis and attention teachers spend on language results in higher achievement scores of their pupils. As we did not find any random effects or interactions, we may conclude that all pupils regardless of their background benefit from this teaching style.

Earlier we reported in Chapter 4 the significant contribution of 'book orientation' to the explanation of differences in reading scores. These results were based on the sample of pupils assessed in the autumn of Year 1 and Year 2 (see Scheme 3.2). Here, the comparable results are based on a partly different sample of pupils with longitudinal data from school entry to the end of Year 3. The results confirm the earlier findings. Although a significant difference was found between the experimental and control group in book orientation (Table 3.27), we cannot prove that the differences between pupils in learning achievement are in some way or another related to differences between the experimental and the control group. Therefore, model 7 (Y2) is considered as the final model for Year 2 explaining differences between pupils in learning achievement.

Table 7.12 – Years 2 and 3. Class characteristics explaining differences in OBIS scores between groups. Unstandardised regression weights (sd)

	Total Year 2 (N=295)		Total Year 3	
Model	6 (Y2)	7 (Y2)	8 (Y3) **N=295**	9 (Y3) **N= 178**
Pupil				
SES	1.22 (.57)			
OBIS Y1	.73 (.05)	.69 (.04)	.22 (.08)	.22 (.10)
OBIS Y2			.30 (.07)	.30 (.09)
Cognitive capacities	2.56 (.60)	2.75 (.58)		
Cognitive category			1.76 (.68)	3.32 (.99)
Non –verbal reasoning			.22 (.05)	.22 (.05)
Attitude			1.19 (.57)	.72 (.61) ns
Class				
Y 1-2 Book orientation		1.72 (.48)		
Y 1-2 Ideas of Reading 2		1.40 (.54)		
Y 2-3 IC Alphabetical				3.66 (1.21)
Variance components				
Intercept	7.63 (2.28)	- 4.25 (3.04)	4.82 (3.20)	-8.28 (5.36)
Pupil	29.07 (2.50)	29.31 (2.52)	44.18 (3.90)	32.27 (3.71)
Class	4.20 (1.93)	1.28 (1.93)	-	-
Total	33.27	30.59	-	-
Random intercept			27.04(12.22)	27.27(19.25)
Random slope			5.58 (2.93)	6.48 (4.20)
Covariance slope - intercept			-11.58 (5.65)	-13.14 (8.85)
% variance				
Pupil	87 %	96 %	-	-
Class	13 %	4 %	-	-
% explained variance compared with				
Model	5 (Y2)	6 (Y2)	-	-
Pupil	+ 2 %	+ 1 %	-	-
Class	+ 5 %	70 %	-	-
Deviance	1854.79	1843.84	1991.90	1139.39
Df	6	7	10	11
Dev_{dif} with model	5 (Y2)	6 (Y2)	7 (Y3)	-
	2.92	10.95	11.64	-
p	≤ .10	≤ .001	≤ .01	-

Model 6 (Y2) = SES OBIS Y 1 and profile; Model 7 (Y2) = OBIS Y1, profile and class characteristics; Model 8 (Y3) = OBIS Y1, Y2, profile and class characteristics.

Year 3: Teaching style
As described in paragraph 3.4.3, at the end of the study the questionnaire 'The transfer from year 2 to 3' was only filled out by the teachers of 16 groups from 7 schools. Using this questionnaire, differences between teachers in intended and implemented curriculum were investigated. The results are given in Table 7.12. Due to the missing information of 10 groups, these results are based on a sub-sample of pupils (N=178).
For comparison of the results of both samples, model 8 (Y3) that was already presented in Table 7.11, has been repeated in Table 7.12. Note that the coefficients for the OBIS scores and non-verbal reasoning are very close; the coefficients for the teachers' perceptions of 'cognitive category' and 'attitude' are not. In the sub-sample the perception of 'cognitive category' has a considerable larger contribution to the explanation of the differences between pupils in learning achievement, while the perception of the pupil's attitude lost its significance.

Model 9 (Y3) shows that teachers indicate that alphabetical principles are explicitly taught in Year 3, which leads to higher learning achievement of their pupils. Alphabetical principles comprise reading and writing skills such as: rhyming; being able to read and write words varying in difficulty (one to three syllables); reading and writing little stories and letters. Other aspects of teaching practices in Year 2 or the curriculum in Year 2 or 3 did not explain differences between pupils' scores.

The last step in this set of analyses was to examine whether differences between groups may be explained by cross-level interactions. Significant interactions will lead to a decrease in the variance of the random slopes. Hence, the interaction between the cognitive category and the teaching style of 'alphabetical principles' was tested; just like the interaction between the social background and 'alphabetical principles'. The latter was done because there is always a possibility that while the main effect is insignificant, the interaction is significant. Anyhow, both interactions appeared to be insignificant: neither the cognitive category nor the social background of the pupils interacted with the teaching style. Therefore, model 9 (Y3) is the final model for Year 3 explaining differences between pupils in learning achievement.

7.5 Discussion and Conclusions

The study
This chapter focused on the issues of prediction and promotion of learning progress in the lower years of primary education. The main purpose was to get a better insight which *educationally relevant factors* help teachers to teach their pupils as well as possible. It takes a longitudinal study to address these issues. Generally speaking, longitudinal designs suffer from limitations due to inevitable real life events. Yet, the

present study could be accomplished without major problems, by and large as it was originally planned.

Progress on scale and social background

During the last decade an ongoing debate on baseline assessment has taken place. Proponents argue that baseline assessment is the best approach for making fair comparisons of progress made by pupils, and of the value added by schools. Moreover, baseline assessment provides valid information for teachers to act upon in the classroom. On the other hand, the main riposte of opponents of measuring baseline at entry is that there is other information available to predict school success, for instance the social background of the pupils or the teacher's perception of his or hers cognitive capacities usually based on observations.

In the light of these controversies we examined theextent to which the different indicators, baseline progress, social background and teacher perceptions are associated with later cognitive progress. Using the longitudinal data available, it was clearly shown that young pupils make considerable progress during their first year at school. Although differences between the pupils at entry are in particular caused by their social background, baseline assessment appeared to be the best predictor of their progress. Once baseline was controlled for, after one year of schooling children from low SES families had made more progress than their more privileged peers from higher social families. Nevertheless, their *absolute* achievement gap remained still considerable.

Further, the results also show the extent to which the social background of pupils has an indirect impact on achievement over a long period of time. The standardised effect of SES on the OBIS score at the end of Year 3 is striking – amounting to half the effect of the baseline assessment score of OBIS in Year 1. A relatively large part of the children from low social families originate from ethnic minorities. There is no clear explanation why their progress after a relative successful start at entry does not continue in the same way in later years. A possible explanation may be found in the difficulties second-language learners (L2) encounter when a stronger appeal to their linguistic skills is made (Tabors & Snow, 2001).

Pupil's profile and Target setting

A quite different explanation may be sought in the educational context. The combination of teacher's perception, expectations and target setting may have a sustaining impact on their pupils' cognitive progress, either inhibiting or stimulating one depending on the specifications of the pupil's cognitive and social profile. The teacher's perception of the pupil's cognitive capacities is generally quite accurate, but in their view there is apparently no cause for individual target setting. When asked about each

pupil individually, the teachers reported hardly any adjustment in teaching or targets after they assessed a pupil. Although target setting is considered as a key facilitator of school improvement, it seems that teachers take differences in learning performance as a natural matter of fact.

Teaching style

This chapter also stresses the importance of identifying differences between classes, despite the fact that most of the variation in learning achievement occurs within schools or classes rather than between classes or groups. From the perspective that in the event particular teaching styles favour learning progress, all pupils in the group will benefit, it was investigated which teaching practices matters.

After controlling for differences between pupils within groups, it appeared that differences between groups can be explained by literature-based classroom practices. Those teachers who focus on 'book orientation', 'ideas of reading' and 'alphabetical principles' achieve better results with their pupils than teachers who spend less attention to these practices. Clearly, these teaching styles refer to a lot more than just knowing concepts of print, rhyming or letter and sound recognition. This type of instruction is representative for a wide range of teaching contexts and classroom practices that encourage and facilitate communication, and social interaction around literature based activities. While looking for cross-level interaction effects between teaching style and pupil characteristics, we could not identify a differential effect that indicates that teachers adjust their instruction in favour of e.g. disadvantaged pupils.

Experimental and control group

Teachers in the experimental group were interested in our meetings about monitoring progress and target setting as a means to facilitate and impede learning progress. Although some significant differences between teachers in the experimental and control group in literature-based teaching practices were found, it was not possible to prove that this was the result of the experimental design of the study. It is sufficient to say that the teaching practices vary and that this supports our assumption that the variation in progress on group level is the result of already existing classroom activities or perhaps acquired during the study.

Conclusions and recommendations

Summing up, several conclusions can be drawn from this study and recommendations can be made. Regarding the monitoring of progress it has been demonstrated that over a period of several years, baseline assessment is the best predictor of learning achievement. As we live in an unfair world, it is clearly demonstrated that in the course of time there is generally more progress for the more able children at entry.

It also has become clear that the social background factor has several pathways of influencing learning progress. On one hand, the progress children of low social families make during their first year at school is considerable; on the other hand over time the progress of children from a low social background is a sustaining factor hindering their progress in later years. Our recommendation is that because of the stigmatising role of social background as a predictor of learning achievement, which leads to confusing 'messages' for teachers, information about the *actual* and *value-added* progress is exactly the information teachers need to know about their pupils (Fitz-Gibbon, 1996, 1997; Tymms, 1999).

Based on our results, we conclude that there is no practice such as target setting in the lower years of primary school, either for individuals or for groups. These findings are in accordance with the conclusion of Van der Wel & Krooneman (2003) that schools are not familiar with the use of intermediate targets. The main conclusion of their study on the implementation of intermediate targets is that no systematic effect of target setting could be found. As we are convinced that target setting helps teachers to adjust their curriculum for the benefit of their pupils, we recommend that target setting should be introduced on a large and systematic scale.

As for the benefit of literature-based teaching styles for whatever pupils it concerns, we cannot emphasise enough that in the lower years of primary education the basis is laid for later reading skills – decoding and comprehensive reading. As it is known that good reading skills are essential for many other learning subjects and areas of life, we recommend that teachers are encouraged to provide all children with the strongest possible literacy foundations in the first years of their school career.

7.6 References

Bowles, S. & Gintis, H. (1976). *Schooling in capitalist America. Educational reform and contradictions of economic life.* New York: Basic books.

Fitz-Gibbon, C. T. (1996). *Monitoring Education: Indicators, Quality and Effectiveness.* London:Cassell.

Fitz-Gibbon, C. T. (1997). The Value-Added National Project: Feasibility Studies for a National System of Value Added Indicators (Final report. London: School Curriculum and Assessment Authority.

Hoeven-van Doornum, A.A. van der (1990). Effecten van leerlingbeelden en streefniveaus op schoolloopbanen. Nijmegen: ITS.

Hoeven-van Doornum, A.A. van der, Voeten, M.J.M., & Jungbluth, P. (1989). The effect of aspiration levels set by teachers for their pupils on learning achievement.

In B. Creemers, T. Peters & D. Reynolds (Eds.), *School Effectiveness and school Improvement* (pp. 231-239). Lisse: Swets & Zeitlinger.

Hoeven-van Doornum, A.A. van der, Voeten, M.J.M., & Jungbluth, P. (1990a). De invloed van klas- en schoolkenmerken op leerprestaties en vervolgadvies in het basisonderwijs. In C.A.C. Klaassen & P.L.M. Jungbluth (Eds.), *Onderwijs Research Dagen 1990. Onderwijs en Samenleving.* Nijmegen: ITS.

Hoeven-van Doornum, A.A. van der, Voeten, M.J.M., & Jungbluth, P. (1990b). Het effect van leerkrachtverwachtingen en streefniveaus op schoolloopbanen. *Tijdschrift voor Onderwijsresearch, 15,* 23-42.

Hoeven-van Doornum, A.A. van der, Voeten, M.J.M., & Jungbluth, P. (1993). The influence of parents' expectations on school careers. *Tijdschrift voor Onderwijsresearch, 18, 369-379.*

Hox, J. (2002). *Multilevel analysis. Techniques and applications.* London: Lawrence Erlbaum Associates, Publishers.

Jöreskog, K.G. & Sörbom, D. (2001) LISREL 8.50: User's Reference Guide. Chicago: Scientific Software International.

Kline, R.B. (1998). *Principles and practice of Structural Equation Modeling*. New York/London: The Guilford Press.

Kreft, I. & De Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage Publications.

Lee, V.E., Bryk, A.S. & Smith, J.B. (1993). The organization of effective secondary schools. In L. Darling-Hammond (Ed.), *Review of Research in Education, vol. 19 (*pp.171-267). Washington, D.C.: American Educational Research Association.

MLwin (2001). J. Rasbah, *Multilevel Models Project, version 1.10.* Institute of Education. University of London.

Onderwijsraad (2003). *Wat scholen toevoegen.* Den Haag: Onderwijsraad.

Scheerens, J. & Bosker, R.J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Snijders, T. & Bosker, R. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modelling*. London: Sage.

Tabors, P.O., & Snow, C.E. (2001). Young bilingual children and early literacy development. In S.B. Neumann & D.K. Dickinson (Eds.), *Handbook of Early Literacy Research.* New York /London: The Guilford Press.

Tymms, P. (1999). *Baseline Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-added Indicators*. London: David Fulton Publishers.

Wel, J.J. van der, & Krooneman, P.J. (2003). *De implementatie van 'Tussendoelen beginnende geletterdheid'*. Amsterdam: Regioplan beleidsonderzoek.

Samenvatting en aanbevelingen

Dit boek beschrijft het onderzoek naar de kwaliteiten van de OBIS-toets, het Onderbouwinformatiesysteem dat de voorschoolse en de daarna op school verworven kennis van 4- tot 6-jarige kinderen meet. OBIS is afgeleid van een Engelse toets, de PIPS[1] Baseline Assessment wordt al jarenlang op grote schaal toegepast in het Verenigd Koninkrijk. Het onderzoek aan OBIS zoals hier wordt beschreven, omvat twee delen. In het eerste deel gaat het om vragen die de constructie van OBIS betreffen en de betrouwbaarheid en validiteit van OBIS. In het tweede deel wordt OBIS gebruikt om enkele specifiek onderwijskundige vraagstellingen te beantwoorden zoals:

- het effect van regelmatig testen op de leerprestaties van kinderen;
- het voorspellen van leerprestaties uit enerzijds de OBIS testresultaten en anderzijds het sociaal-economische milieu;
- het gebruik van OBIS om kinderen met leerachterstand, die speciaal onderwijs nodig hebben, op te sporen.

Deel I, Inhoud en structuur van OBIS

In hoofdstuk 1 worden het doel en de achtergrond van de studie uiteengezet en worden de onderzoeksvragen geformuleerd. Aansluitend volgen beschrijvingen van de studieopzet en de gebruikte methodes. Tenslotte worden de begrippen 'toegevoegde waarde' en 'meting van de voortgang', besproken omdat ze van belang zijn bij het gebruik van OBIS.

PIPS en dus ook OBIS zijn bedoeld voor het meten van cognitieve vaardigheden, in het bijzonder de geletterdheid en de gecijferdheid bij kinderen van 4 tot 6 jaar. Verder bevatten de toetsen een onderdeel om de houding van het kind ten opzichte van school te bepalen. Beide toetsen zijn bijzonder kindvriendelijk, ze maken gebruik van aantrekkelijke plaatjes en de tijd die nodig is voor een toetsafname bedraagt slechts ca. 20 minuten. De toetsen worden individueel afgenomen, als regel door de eigen leerkracht. Dit heeft tot gevolg dat de leerkracht al doende direct inzicht krijgt in de capaciteiten van het kind.

Hoofdstuk 2 geeft uitvoerig rekenschap van OBIS en het gebruik daarvan in de school. De correlaties van de verschillende reken- en taalonderdelen tussen PIPS en OBIS zijn heel hoog. OBIS voldoet aan alle criteria van betrouwbaarheid en validi-

1 PIPS staat voor Performance Indicators in Primary Schools.

teit. Na de beschrijving van deze karakteristieken volgt een quasi-experimenteel, meerjarig onderzoek waarin OBIS wordt gebruikt. Aan dit onderzoek deden 450 kinderen van 11 verschillende scholen mee. Naast gebruik van OBIS in de leerjaren 1 en 2 werd hierbij ook de vervolgversie V-OBIS voor de leerjaren 3 en 4 ingezet. Het regelmatig testen leidt tot betere leerprestaties van de kinderen en een grotere professionaliteit van de leerkrachten zoals tot uitdrukking komt in hun aanpak van het voorbereidend leesonderwijs. De OBIS toetsresultaten blijken een betere voorspeller van toekomstige leerprestaties te zijn dan de sociaal-economische en etnische achtergrond van het kind. Aansluitend bij recente discussies in de Nederlandse onderwijspolitiek wordt besproken in hoeverre OBIS toetsresultaten bruikbaar zijn als criterium voor de toewijzing van extra middelen aan scholen.

OBIS is afgeleid van de Engelse PIPS Baseline Assessment en in dat kader worden de verschillen tussen het Nederlandse en het Engelse onderwijssysteem besproken. In beide landen is uitvoerig gedebatteerd over de manier waarop jonge kinderen zouden kunnen worden getoetst. De uitkomsten van dit debat in de twee landen verschillen nogal en hebben ertoe geleid dat in het Verenigd Koninkrijk op grote schaal jonge kinderen worden getoetst, terwijl het toetsen in Nederland nog in de kinderschoenen staat.

Bij het de constructie van OBIS uit de PIPS Baseline Assessment behoefden de rekenonderdelen weinig meer dan te worden vertaald uit het Engels. De correlatie tussen de Engelse en Nederlandse rekenopgaven was dan ook hoog (.90). De taalonderdelen vroegen om meer aanpassingen, maar ook hier zijn hoge correlaties van de Nederlandse en Engelse taalonderdelen bereikt, te weten 0.80 voor woordenschat en 0.90 voor letterkennis.

Hoofdstuk 3 vormt de overgang van het beschrijvende naar het verklarende deel van de studie. Allereerst wordt de dataverzameling in detail besproken, te beginnen met een beschrijving van de deelnemende scholen. Vervolgens worden de relaties tussen de meetwaarden voor de verschillende lees- en rekenonderdelen van de toets berekend en worden vergelijkingen tussen groepen, ingedeeld naar leeftijd, geslacht en sociaal-economisch milieu, gemaakt.

PIPS en OBIS zijn opgezet met het doel goede voorspellers te zijn van latere leerprestaties voor taal en rekenen. Inderdaad blijkt dat OBIS scores van de beginmeting hoog correleren met de scores één (.78) en 2 jaar (.68) later. De onderdelen letter- en cijferkennis blijken de beste voorspellers te zijn. Uit de hoge test-hertest correlaties, .92 voor rekenen en .97 voor taal, blijkt dat de betrouwbaarheid van OBIS hoog is. We hebben OBIS ook vergeleken met bestaande toetsen, n.l. Taal alle Kleuters en Ordenen, beide van het CITO. De correlaties van OBIS met deze toetsen zijn hoog.

De eerder genoemde V-OBIS is een vervolgtoets op OBIS bedoeld voor het 3de leerjaar. De toets werd verkregen uit de Engelse toets PIPS Year 1. V-OBIS meet naast de leerprestaties ook de houding ten aanzien van school, het algemene ontwikkelingsniveau, de woordenschat en het ruimtelijk redeneren. De laatst genoemden zijn uitstekende voorspellers van schoolvorderingen en zijn geschikt om vast te stellen of het kind de verwachte voortgang maakt. Deze zogenaamde vergelijkende toegevoegde waarde van de school kan voor ieder kind worden bepaald, ongeacht of eerdere OBIS toetsresultaten beschikbaar zijn.

Deel 2, Studies met OBIS

OBIS is gebruikt in een reeks vervolgstudies die in de volgende hoofdstukken worden beschreven. De dataverzameling voor het hoofdonderzoek vindt plaats wanneer de kinderen voor het eerst op school komen en vervolgens 1 en 2 jaar later zodat de voortgang van hun leerprestaties kan worden gemeten. Aan het hoofdonderzoek doen 450 kinderen uit 11 verschillende scholen mee, na 2 jaar is het aantal deelnemende kinderen gedaald tot 400.

In hoofdstuk 4 wordt onderzocht of een regelmatige toetsafname met OBIS stimulerend werkt op de leerprestaties en of het leidt tot verhoging van de professionaliteit bij de leerkrachten. Het onderzoek werd uitgevoerd als een quasi experimentele studie waarbij de scholen werden verdeeld over 3 behandelingsgroepen. In deze opzet kregen de leerkrachten in groep 1 behalve de OBIS uitslagen alle relevante informatie en terugkoppeling op leerling- en klasniveau ter stimulering van hun lesgeven. In groep 2 kregen de leerkrachten ook OBIS uitslagen en terugkoppeling, maar geen gedetailleerde informatie over de interpretatie daarvan. Groep 3, de controlegroep, ontving alleen de OBIS uitslagen.

De onderzoeksresultaten bevestigen dat OBIS een betrouwbaar instrument is voor beginmetingen. De reacties van de leerkrachten op het gebruik van de toets waren in de regel zeer positief. De beginscores van kinderen uit etnische minderheidsgroepen waren aanzienlijk lager dan die van hun autochtone leeftijdsgenoten. Na een jaar op school te hebben gezeten bleken de allochtone kinderen relatief echter meer te zijn vooruit gegaan dan hun leeftijdsgenoten uit hoger SES. Hieruit zou kunnen worden afgeleid dat de terugkoppeling van de individuele toetsresultaten aan de leerkracht een positief effect heeft. Dit effect kwam het sterkst tot uitdrukking bij de leesvaardigheid van de kinderen. Opvallend genoeg waren er geen verschillen in de grootte van het testeffect tussen de groepen 2 en 3. Dit lijkt te worden veroorzaakt door een zogenaamd Hawthorne effect waarbij de controlegroep 3, louter door deelname aan het experiment, beter gaat presteren.

Ook in dit onderzoek blijkt het sociaal-economische milieu een goede voorspeller te zijn van toekomstige leerprestaties. Bij nadere analyse in een multi-level model blijkt echter dat het effect van SES nagenoeg verdwijnt wanneer rekening wordt gehouden met de OBIS beginmeting.

Het volgende hoofdstuk (Hfd. 5) beschrijft OBIS als het instrument bij uitstek om toekomstige leerprestaties te voorspellen. Groepskenmerken, zoals het sociaal-economische en etnisch milieu, komen voor dit doel op de tweede plaats. Deze resultaten zijn van belang in de steeds terugkerende politieke discussie over de criteria voor toewijzing van extra gelden aan de scholen ter bestrijding van achterblijvende leerprestaties. Tot nu toe gebeurt dit in Nederland op grond van het sociaal-economische en etnisch milieu.

OBIS kan ook worden gebruikt voor longitudinale metingen van de leerprestaties. Op deze manier kunnen niet alleen de schoolvorderingen van individuele leerlingen in kaart worden gebracht, maar ook de toegevoegde waarde van de school in de tijd worden bepaald. De OBIS scores over een periode van 3 jaar laten zien dat het aandeel van slecht presterende kinderen uit laag sociaal-economisch milieu constant blijft. Bij de allochtone kinderen uit laag sociaal-economisch milieu daalt dit aantal echter van 54% naar 44%. In overeenstemming daarmee neemt het aandeel goed presterende allochtone kinderen toe van 14% naar 23%. Deze resultaten maken duidelijk dat de sociale achtergrond, vergeleken met de leerprestatie, een weinig betrouwbare maat is voor het vaststellen van leerachterstanden.

In hoofdstuk 6 wordt OBIS gebruikt voor het opsporen van risicoleerlingen.Vroegtijdige herkenning van deze leerlingen biedt de mogelijkheid om ze extra aandacht te geven of door te verwijzen naar het speciaal onderwijs. In de afgelopen jaren werd een toenemend aantal kinderen verwezen naar het speciaal onderwijs. Om deze trend te keren heeft het ministerie de richtlijn WSNS (Weer Samen Naar School) uitgevaardigd. De analyse van OBIS scores over een periode van drie jaren leidt in dit verband tot interessante conclusies. Ten eerste bleek dat de groep van kinderen met de 16% laagste OBIS scores, hogere scores hadden zowel in het voorafgaande jaar als in het volgende jaar. Dit betekent dat de samenstelling van de laagst scorende groep niet constant is en dat er voortdurend kinderen in en uit deze groep gaan. In de tweede plaats bleek dat de kans om een risicoleerling te zijn, dat wil zeggen deel uit te maken van de groep met de 16% laagste scores, het best werd voorspeld door de OBIS score in jaar 1. Deze score wordt 2 maanden na aanvang van het schooljaar gemeten en representeert dus in hoofdzaak niet-schoolse kennis. Deze ogenschijnlijk tegengestelde bevindingen leiden tot de conclusie dat een kind het risico label niet altijd permanent hoeft te dragen. Dit wordt duidelijk wanneer we het risico label niet koppelen aan de score in een bepaald jaar, maar aan de toegevoegde waarde, d.w.z. de relatieve

vooruitgang van de leerling in opeenvolgende jaren. Zo kunnen kinderen met een gelijke OBIS score in jaar 3 toch kunnen verschillen in risico status, afhankelijk van hun scores in jaar 1.

In het laatste hoofdstuk (7) worden de argumenten bijeen gebracht om leerprestaties vanaf de binnenkomst op school te meten. De centrale vraag hierbij is hoe goed een begintoets, in dit geval OBIS, toekomstige cognitieve leerprestaties voorspelt. De daarbij behorende vraag is met welke school- dan wel klaskenmerken rekening moet worden gehouden. De resultaten van de longitudinale LISREL analyses laten zien dat vergeleken met de sociale achtergrond en de verwachtingen van de leerkracht van de capaciteiten van het kind, de begintoets de beste voorspeller is van toekomstige leerprestaties. De multilevel analyses tonen aan dat kinderen voordeel hebben van een op taal gerichte aanpak in de klas. Kruiseffecten tussen de onderwijsstijl en de leerlingkenmerken konden niet worden aangetoond. De leerkrachten zeiden bij hun instructie geen rekening te houden met de toetsresultaten van hun leerlingen.

Conclusies en aanbevelingen

Uit de hier beschreven studies kunnen verschillende belangrijke conclusies worden getrokken en aanbevelingen worden gedaan.

Sociale achtergrond, leerkrachtverwachtingen en begintoets

In de afgelopen 10 jaar, en tot op heden is de begintoets onderwerp van discussie. Het belangrijkste argument van tegenstanders is dat er andere informatie beschikbaar is om schoolsucces te voorspellen. Daarbij gaat het om de sociale achtergrond en de verwachtingen die de leerkracht koestert ten aanzien van de cognitieve capaciteiten van de leerling. De voorstanders daarentegen gaan ervan uit dat een beginmeting de beste manier is om de voortgang van leerlingen te vergelijken en om de toegevoegde waarde van de school te bepalen.

De huidige resultaten laten zien dat de begintoets een betere voorspeller is van toekomstige leerprestaties dan de sociale achtergrond en de inschatting van de leerkracht. Ook is gebleken dat de sociale achtergrond het leerproces langs verschillende wegen beïnvloedt. Enerzijds boeken de kinderen uit lagere sociale milieus aanzienlijke vooruitgang in hun eerste jaar op school, anderzijds is dit milieu over een langere periode een factor die hun vooruitgang belemmert.
Vanwege de stigmatiserende rol van de sociale achtergrond bij het voorspellen van leerresultaten is onze aanbeveling dat de leerkracht moet kunnen beschikken over de werkelijke voortgang van hun leerlingen en de toegevoegde waarde van de leerprestaties.

Onderwijs doelen
Op grond van de resultaten moeten we concluderen dat er in de eerste leerjaren van de basisschool geen expliciete onderwijsdoelen worden geformuleerd. Dit sluit aan bij de bevindingen van Van der Wel & Krooneman (2003) dat scholen niet gewend zijn aan het gebruik van tussendoelen. Het verdient aanbeveling om tussendoelen systematisch in het onderwijs te introduceren.

Onderwijsstijl
In de eerste jaren van het primair onderwijs wordt de basis gelegd voor toekomstige leesvaardigheden, het decoderen en het begrijpend lezen. Goede leesvaardigheden zijn van doorslaggevende betekenis voor het verkrijgen van kennis op tal van terreinen. Het verdient krachtige aanbeveling om alle kinderen gedurende de eerste jaren van hun schoolloopbaan een zo goed mogelijke basis van leesvaardigheden mee te geven.

Risicoleerlingen: het meten van achterstand versus het meten van vooruitgang
Na een tijdperk waarin de politiek er op gericht was om de ontwikkelingskansen van achterstandskinderen te verbeteren mag de vraag worden gesteld of we ons blijvend moeten bezig houden met het bepalen van achterstand of juist over zouden moeten gaan naar het meten van vooruitgang. Vanuit een economisch perspectief kan de politiek aanvoeren dat er geen reden is om alle bij kinderen de onderwijsvooruitgang of achterstand in de eerste jaren van de basisschool te volgen. Voortgang kan echter alleen maar worden vastgesteld in vergelijking met de gemiddelde leerprestatie. Om stagnatie van het leerproces in een vroeg stadium te kunnen vast te stellen moeten we informatie hebben over de voortgang van de leerprestaties in de tijd. Risicoleerlingen kunnen alleen maar worden herkend door te vergelijken met leeftijdsgenoten, over een langere periode van tijd.

In this book the author gives an account of the evaluation of OBIS, a test measuring the cognitive development of 4-6 year-old children. OBIS is derived from the British PIPS Baseline Assessment. The use of OBIS to measure the value added indicators, the role of OBIS to predict future learning achievements and other practical applications of OBIS, are dealt with. Therefore, this book is of particular significance for teachers, students, researchers, and policy makers as well.

Anneke van Doornum (1946) was 18 years when she started as a teacher in primary school. She married and after the birth of her 3 children she took a degree in education (1986). Since then she has worked in the field of educational research. Her present position is senior scientist in the Institute for Applied Social Sciences (ITS) of the Radboud University of Nijmegen. She was one of the first to demonstrate the sustaining effect of teachers' expectations on school careers of their pupils (PhD thesis: Effecten van leerlingbeelden en streefniveaus op schoolloopbanen. ITS, 1990). Later on she published a book on childrens' perception of their environment (Kinderen over het milieu. ITS, 1992). Following recent changes in the curriculum of secondary schools that were based on the hypothesis of transfer of knowledge, she took up the challenge to investigate this hypothesis in a field study (Vaststellen van Transfervermogen. ITS, 1999).

ISBN 90 5554 286 5
NUR 840