



International Comparative Analysis of
Learning and Teaching in Math Lessons
in Several European Countries

Index

1	Introduction	5
2	Rationale and Background	7
3	Theoretical Framework	9
4	Concepts and Methods	13
4.1	Observation Instrument	13
4.2	Questionnaire	14
4.3	Training of Inspectors	15
4.4	Sample	15
5	Reliability	17
6	Validity	19
7	Results	23
7.1	Opportunity to Learn	25
7.2	Teaching	26
7.3	Monitoring Students' Progress	26
7.4	Measures Taken for Struggling Learners	27
7.5	Learning	27
7.6	Context	27
8	Discussion	31
	References	35
	Appendix	
I	Lesson Observation and Interview Form for Evaluating the Quality of Learning and Teaching	39
II	International Project Team	44

1 Introduction

The inspectorates of education from several countries in Europe have assessed, each in their own country, the quality of teaching and learning, using a shared framework of indicators and centrally trained inspectors. These inspectors have visited large national representative samples of primary schools. Within these schools they have observed one or two mathematics/arithmetic lessons in the grade with the largest proportion of 9-year-olds, which is the fourth grade in most countries. In this report, an international comparative analysis is made of the quality of teaching and learning in the mathematics/arithmetic lessons in these countries.

2 Rationale and Background

From 2002 onwards, initiatives and first steps for the project 'International Comparative Analysis of Learning and Teaching' (ICALT) were taken by the inspectorates of education in England, Flanders (Belgium), Lower Saxony (Germany), North Rhine-Westphalia (Germany) and the Netherlands. The Dutch Ministry of Education funded this project in 2002 and 2003. During these years, the inspectorates of education from these countries jointly developed an instrument to observe and analyse the quality of learning and teaching in primary schools in their countries. The results of the first stage of this project are reported in Van de Grift (2007). During the second stage of the project, some items were added to existing observation scales and a new observation scale about 'reflexivity and discursiveness' was constructed. Furthermore, a few interview items were added on 'opportunity to learn', 'monitoring of student progress' and a Likert scale was constructed for assessing 'special measures for struggling learners'. The aim of the second stage of the study was to deliver international benchmarks for the quality of several aspects of teaching and learning that can be used in the work of the national inspectorates and national educational policy agencies. Five European countries have decided to adopt this project and implement the final version of the developed instrument: Flanders (Belgium), Lower Saxony (Germany), the Slovak Republic, Scotland and the Netherlands.

3 Theoretical Framework

Both the Organisation for Economic Co-operation and Development (OECD) and the International Association for the Evaluation of Educational Achievement (IEA) regularly publish comparative studies on educational outcomes.

The recent OECD study from the Programme for International Student Assessment (PISA) showed remarkable differences in the mathematics competences between the average scores of the 15-year-old students in the countries participating in the ICALT project (OECD, 2007). The difference between the highest average score of the Dutch students and the lowest average score of the Slovakian students is 39 points, which is more than one third of a standard deviation.

Table 1. Mathematics scores in international comparative studies

	Mathematics scores 15-year-old students in 2006 (PISA, OECD, 2007)	Mathematics scores 10-year-old students in 2007 (TIMSS, Mullis, Martin & Foy, 2008)
England		541
Flanders (Belgium)	520	
Germany	504	525
Netherlands	531	535
Scotland		494
Slovak Republic	492	496
United Kingdom	495	

A recent IEA study from the Trends in International Mathematics and Science Study (TIMSS) programme shows a slightly different picture for the mathematics competences between the average scores of the 10-year-old students in the countries participating in the ICALT project (Mullis, Martin & Foy, 2008). Not every country participates with every age group in the PISA and TIMSS studies. In the TIMSS study in 2007, the Flemish students did not participate and the students of the United Kingdom in this study are divided into English and Scottish students. The difference between the highest scores of the average English student and the average score of the Scottish students is 47 points, which is almost half a standard deviation.

These differences in average scores between countries require explanations. Every year, several publications on comparative studies in education are published. The Organisation for Economic Co-operation and Development (OECD) for instance, releases 'Education at a Glance' annually and the European Commission publishes its 'Key Data on Education in Europe' every two years. These publications inform us about the international differences in quantitative facts like teacher salaries, contact hours, class sizes and so on, but there are no publications that offer reliable and valid information on international differences in the quality of teaching and learning strategies. This information is vital for international benchmarking on these important aspects of education and it is important for setting criteria for school (self-) evaluation as well.

Since 1961, a lot of observation studies have been carried out into the quality of teaching (Flanders, 1970; Stallings & Kaskowitz, 1974; Stallings, Needels &

Stayrook, 1979; Capie, Johnson, Anderson, Ellett & Okey, 1980; Florida Coalition for the Development of a Performance Measurement System, 1983; Veenman, 1986; Stringfield, Teddlie & Suarez, 1985; Virgilio, 1987; Virgilio & Teddlie, 1989; Teddlie, Virgilio & Oescher, 1990; Evertson, 1987; Evertson & Burry, 1989; Slavin, 1987; Van de Grift & Lam, 1998; Dutch Inspectorate of Education, 1998; HM Inspectorate of Education, 1999; 2001; Ofsted, 1995; Van de Grift 2007). However, most of these studies are small-scale national studies and the results are not adequate for the development of international benchmarks. Results and instruments of these observation studies were used for the development of indicators for the observation instrument which, in a further developed form, has been used in this project.

4 Concepts and Methods

One of the long-term aims of this study is to make preparations for new, combined studies like the PISA and TIMSS studies into the achievements of students and observational studies into the quality of teaching and learning in the participating countries. The basic research question of the study is: to what extent do teaching and learning in the year with the largest proportion of 9-year-olds at the beginning of the school year (the fourth grade in most countries) differ across four European countries? The idea of teaching and learning contains several core concepts: the opportunity students receive to learn the targets of the curriculum, the teaching practice, monitoring students' progress, the measures taken for struggling learners and the learning process of students. The concepts are operationalised in the instruments mentioned in Table 2.

Table 2. Instruments used

	Instrument	Number of items
Opportunity to learn		
Quality of the curriculum	Interview	3
Weekly minutes devoted to arithmetic	Interview	1
Percentage of students using textbooks and methods of 'younger grades'	Interview	1
Teaching		
Safe and stimulating learning climate	Observation	5
Clear and activating instruction	Observation	9
Efficient classroom management	Observation	4
Teaching learning strategies	Observation	9
Monitoring students' progress		
Times a year students' progress is monitored	Interview	1
Measures taken for struggling learners		
Adaptation of teaching to diverse needs of students	Observation	4
Curative measures for struggling learners	Likert scale	3
Learning		
Involvement of students	Observation	3
Reflexivity and discursiveness	Observation	4

Some of these aspects are suitable for observation, others for an interview or questionnaire.

4.1 Observation Instrument

During the first stage of this study, an observation instrument was developed in several pilot studies. England, Belgium Flanders, Lower Saxony and the Netherlands were involved in these pilots. The observation instrument was piloted for reliability, inter-rater reliability and validity based on more than 850 observations in the four countries mentioned. This pilot project has shown that the countries can be compared in a reliable and valid way, based on aspects like:

- safe and stimulating learning climate;
- clear instruction;
- adaptation of teaching;
- teaching learning strategies;
- classroom management;
- involvement of students (cf. Van de Grift, 2007).

A slight revision of this instrument was carried out in 2007. Some items were added to existing scales. One new scale 'reflexivity and discursiveness' was added with the help of Kaune (2006) and Cohors-Fresenborg (Cohors-Fresenborg & Kaune, 2007a). This includes three items concerning meta-cognitive activities (planning, monitoring and reflection) and one item concerning discursiveness (Cohors-Fresenborg & Kaune, 2007b), which is an indicator of to what degree students' verbal contributions to the classroom dialogue fit exactly to the topic discussed or to arguments other students have previously introduced. All scales and items are found in the appendix.

4.2 Questionnaire

The observation instrument is elaborated with a few interview items on 'opportunity to learn the minimum objectives of the curriculum', 'the frequency of monitoring of students' progress' and 'special curative measures for struggling learners'. After the inspectors made their observations they asked the teachers some questions about these three topics:

Students receive ample opportunity to learn the minimum objectives of the curriculum when:

- their teacher uses regular and modern arithmetic methods that cover the minimum objectives of the grade;
- enough weekly hours are spent on arithmetic;
- no students are lagging behind for more than a year.

The frequency of monitoring students' progress is determined by simply asking the amount of times teachers use standardised tests to gauge the achievements of their students. The quality of the curative measures for struggling learners is measured by three Likert-type questions:

- Does the teacher diagnose the learning problems of students at risk?
- Does the teacher have prescribed learning plans for students at risk?
- Does the teacher implement prescribed learning plans for students at risk?

This Likert scale has to fulfil the usual criteria for reliability and validity. The questions about opportunity to learn and monitoring are just low-inferential questions on factual matters.

4.3 Training of Inspectors

Two-day training sessions were organised for the inspectors in each of the participating countries. Several different DVDs were used during the training sessions. Some DVDs used lessons scoring around a p-value of .50 on the teaching and learning aspects of the instrument, other DVDs showed lessons scoring around .25 and .75. It is far easier to obtain high inter-observer agreement rates with high-quality lessons (score $>.75$) or weak lessons (score $<.25$) than with lessons with a score around the average (.50). The .25- and .75 DVDs were used during the training. The .50 DVDs were used for pre-measurement and post-measurement purposes in the training sessions. Inspectors were trained until they reached a consensus criterion of at least .80 on the 'difficult to judge' .50 DVDs.

4.4 Sample

In each country, a large representative sample was taken to ensure accurate cross-national comparisons. The representativeness of the samples is mostly guaranteed by several inspectorates, because the schools they visit are not selected according to special criteria. The schools visited were simply 'the next schools to visit', which is equivalent to a simple random sample. Other inspectorates (for instance in the Netherlands) have proportional strategies for the selection of schools to visit: schools in more challenging circumstances are visited more frequently than others. In order to solve this problem, the Netherlands Inspectorate of Education took a simple random sample. To make sure that accurate comparisons can be made, sample sizes were based on a sample precision of .05, which means that in several countries a national sample was drawn of about 300 observations. This succeeded in Flanders, Lower Saxony, Slovakia and the Netherlands, so the accuracy of these samples is about 5%. In Scotland, 94 lessons were observed. The accuracy of this Scottish sample is about 10%. The observations were made during one whole school year.

5 Reliability

The internal consistency of each scale within and across the four countries is analysed by means of the calculation of the Cronbach alpha coefficient. The results are presented in Table 3.

Table 3. Internal consistency of the observation scales and the Likert scale

Country/ cases	Flanders N=240	Lower Saxony N=286	Netherlands N=332	Slovak Republic N=324	Scotland N=94	Total N=1276
Safe and stimulating learning climate	.87	.89	.88	.88	.84	.88
Clear and activating instruction	.84	.88	.79	.89	.88	.87
Efficient classroom management	.83	.79	.83	.87	.85	.84
Adaptation of teaching to diverse needs of students	.86	.66	.78	.90	.83	.81
Teaching learning strategies	.89	.91	.70	.93	.92	.91
Involvement of students	.90	.82	.83	.87	.85	.86
Reflexivity and discursiveness	.89	.87	.81	.86	.86	.88
Curative measures for struggling learners	.78	.84	.71	-. ¹⁾	.68	.84

1) The Slovak inspectorate did not participate in this part of the study.

Most scales of the instrument were sufficiently homogeneous (Cronbach's alpha >.70) for our purpose in each of the countries involved. There are two small exceptions: the scale 'adaptations of teaching to diverse needs of students' has a Cronbach's alpha of .66 in Lower Saxony and the scale 'curative measures for struggling learners' has a Cronbach's alpha of .68 in Scotland.

Inter-rater reliability was calculated in an earlier study (Van de Grift, 2007). In that study, pairs of inspectors observed the same teachers during training sessions. The inter-rater reliability coefficients varied between .73 and .93 for the subscales, which is sufficiently high. For the second stage of this study, the observing inspectors were trained until they reached a consensus criterion of at least .80.

6 Validity

The construct validity of the observation instrument was analysed in several ways. The first procedure consisted in computing correlations between the aspects of teaching of the ICALT instrument and a global judgement on the quality of 'overall teaching'. The results are presented in Table 4.

'Quality of the mathematics curriculum', 'Weekly minutes devoted to mathematics' and 'Times a year students' progress is monitored' are in most countries normally based on decisions at school level and not at teacher level.

In Scotland however, teachers use their professional judgement to decide when a student is ready to sit a national assessment. Although this may occur once a year, Scottish teachers will also carry out other forms of assessment during the year. A relationship is thus not to be expected and correlations between these variables and 'Overall teaching' as observed in the mathematics lessons, are not expected to differ significantly from 0, as they are.

A low negative but significant correlation is found between 'Overall teaching' and the 'Percentage of students lagging behind'.

The teaching scales 'Safe and stimulating learning climate', 'Clear and activating instruction', 'Efficient classroom management' and 'Teaching learning strategies' are highly correlated with 'Overall teaching' as expected, which are important indications for construct validity.

'Measures taken for struggling learners' is expected to correlate moderately but significantly with 'Overall teaching', which was also found.

Table 4. Construct validity of the teaching scales

	Overall teaching
Opportunity to learn	
Quality of the curriculum	-.06
Weekly minutes devoted to arithmetic	.09
Percentage of students using textbooks and methods of 'younger grades'	-.10**
Teaching	
Safe and stimulating learning climate	.63**
Clear and activating instruction	.77**
Efficient classroom management	.70**
Teaching learning strategies	.74**
Monitoring students' progress	
Times a year students' progress is monitored	.04
Measures taken for struggling learners	
Adaptation of teaching to diverse needs of students	.58**
Curative measures for struggling learners	.28**

* significant at .05 level; ** significant at .01 level

Next, we computed the correlations between the aspects of learning of the ICALT instrument and a global judgement on the quality of learning. The results are presented in Table 5.

Table 5. Construct validity of the learning scales

	Overall learning
Learning	
Involvement of students	.72**
Reflexivity and discursiveness	.52**

* significant at .05 level; ** significant at .01 level

These correlations vary between .52 and .72 for the different parts of learning, which could be expected. This is an indication of good construct validity.

For the predictive validity, correlations were computed between the behaviour of teachers and the behaviour of students. We should expect that better teaching goes along with more involvement of students and more reflexivity and discursiveness.

'Opportunity to learn' as measured by the ICALT-instrument is mostly determined by school-level decisions ('Quality of the mathematics curriculum' and 'Weekly minutes devoted to mathematics') or variables highly influenced by contextual factors ('Percentage of students lagging behind'). That is why we expected only low or insignificant correlations with students' behaviour in classrooms ('Involvement of students' and 'Reflexivity and discursiveness'). This is in agreement with the findings in Table 6.

Table 6. Predictive validity of the teaching scales

	Involvement of students	Reflexivity and discursiveness
Opportunity to learn		
Quality of the curriculum	-.08	-.05
Weekly minutes devoted to arithmetic	.07	.01
Percentage of students using textbooks and methods of 'younger grades'	.16*	.08*
Teaching		
Safe and stimulating learning climate	.59**	.43**
Clear and activating instruction	.72**	.59**
Efficient classroom management	.63**	.42**
Teaching learning strategies	.66**	.67**
Monitoring students' progress		
Times a year students' progress is monitored	.03	.25**
Measures taken for struggling learners		
Adaptation of teaching to diverse needs of students	.51**	.48**
Curative measures for struggling learners	.34**	.31**

* significant on .05 level; ** significant on .01 level

The teaching scales 'Safe and stimulating learning climate', 'Clear and activating instruction', 'Efficient classroom management' and 'Teaching learning strategies' are highly correlated with student behaviour ('Involvement of students' and 'Reflexivity and discursiveness'), as expected. The better the teaching, the better the involvement of students and the better the reflexivity and discursiveness of the

students. These correlations varied between .42 and .72 for the different parts of teaching. This legitimates the decision to implement the construct 'reflexivity and discursiveness' in our study with a new scale.

'Monitoring students' progress' is not correlated with 'Involvement of students' as expected and is significantly correlated with 'Reflexivity and discursiveness', which was unexpected.

The same tendency is found for the measures taken for struggling learners and student behaviour. The more teachers adapt the teaching to the diverse needs of students and the better the curative measures for struggling learners, the better the involvement and the reflexivity and discursiveness of students. These correlations varied between .31 and .51 for the different parts of teaching.

We might observe that measurements based on school-level decisions ('Quality of the mathematics curriculum' and 'Weekly minutes devoted to mathematics') are not correlated with students' behaviour in classrooms.

'Times a year students' progress is monitored' is correlated with students' behaviour to a moderate degree.

These construct and predictive validity coefficients encourage the idea that the observation scales of the ICALT instrument measure what is intended to be measured.

7 Results

Some variables are measured in their own specific way: minutes, percentages, times and numbers. All scores on observation instruments and Likert scales were standardised by dividing the sum score by the product of the number of response categories and the number of items. Each scale thus varies between 0 and 1. An average score in table 6 lower than .25 means that the average teacher scores 'predominantly weak', a score between .25 and .50 means that the average teacher has 'more weaknesses than strengths', a score between .50 and .75 means that the average teacher has 'more strengths than weaknesses' and a score above .75 means that the average teacher scores 'predominantly strong'. These average scores are found in Table 6.

In this study, we deviate from the usual procedure used to ascertain whether differences found are significant or not. The main reason for this is that the significance of a difference says nothing about the practical relevance of a difference between countries (cf. Carver, 1978). We prefer to use 'effect sizes'. Effect sizes do tell us something about the relevance of an effect, without being conditional upon the size of a random sample. Furthermore, effect sizes are standardised, which enables us to compare the impacts of the different indicators. In this study effect sizes are computed by dividing the difference of a national average from the international average by the standard deviation. Effect sizes of .80 and more are considered large by Cohen (1988), effect sizes of .50 are moderate and effect sizes of .20 are small. Cohen's standardisation of effect sizes should be seen against the background of the fact that, in many educational experiments, effect sizes larger than .40 are seldom found. Effect sizes that are smaller than .15 will not be interpreted in this text. Effect sizes of .20 and more are considered relevant in educational terms. The national deviations from the international average, measured in effect sizes are found in Table 6a. Small effect sizes (between .20 and .50) are presented in *italics*. Moderate and large effect sizes (.50 and more) are presented in **bold** figures.

Table 6. Raw results

	Average	St. dev.	Flanders	Lower Saxony	Netherlands	Scotland	Slovak Republic
Opportunity to learn							
Quality of the curriculum	.98	.10	.97	.98	.99	.95	1.00
Weekly minutes devoted to arithmetic	264.77	44.24	306.49	235.21	295.38	293.06	225.00
Percentage of students using textbooks and methods of 'younger grades'	2.81	6.34	2.33	1.11	2.66	9.25	1.38
Teaching							
Safe and stimulating learning climate	.80	.13	.82	.77	.79	.89	.78
Clear and activating instruction	.74	.12	.73	.71	.72	.85	.75
Efficient classroom management	.78	.14	.82	.75	.79	.89	.75
Teaching learning strategies	.64	.16	.62	.55	.67	.73	.68
Monitoring students' progress							
Times a year students' progress is monitored	1.69	1.01	2.33	.63	2.03	.83	1)
Measures taken for struggling learners							
Adaptation of teaching to diverse needs of students	.65	.18	.61	.58	.70	.78	.64
Curative measures for struggling learners	.64	.41	.64	.31	.86	.88	2)
Learning							
Involvement of students	.76	.15	.74	.74	.77	.84	.76
Reflexivity and discursiveness	.58	.16	.57	.47	.62	.56	.67
Context							
Percentage of schools in urban regions	39.91	49.00	50.66	41.57	24.32	65.17	3)
Percentage of students from socially deprived families	16.10	17.34	20.97	15.58	11.88	16.68	3)
Percentage of students for which the language of instruction is not their native language	12.07	23.07	14.16	16.26	8.63	1.51	3)
Number of students at school	218.17	118.63	205.48	227.14	225.23	201.50	192.4
Number of children in classroom	20.27	5.18	18.86	19.77	21.49	21.39	22
Percentage of combined classes	27.0	44.4	10	31	32	40	0
Percentage of beginning teachers	10.6	30.8	10	2	15	22	3)
Percentage of male teachers	18.7	39.0	20	20	23	1	19.38

1) In the Slovak Republic, this year no standardised national tests were used, though they are being developed. However schools and teachers do check the students on a regular basis.

2) This is not measured in the Slovak Republic but struggling learners receive individually targeted assistance from the teacher.

3) These data are not available for the Slovak Republic.

	Flanders	Lower Saxony	Netherlands	Scotland	Slovak Republic
Opportunity to learn					
Quality of the curriculum	-.10	.00	.10	-.30	.20
Weekly minutes devoted to arithmetic	.94	-.67	-.69	.64	-.90
Percentage of students using textbooks and methods of 'younger grades'	-.08	-.27	-.02	1.02	-.23
Teaching					
Safe and stimulating learning climate	.15	-.23	-.08	.69	-.15
Clear and activating instruction	-.08	-.25	-.17	.92	.08
Efficient classroom management	.29	-.21	.07	.79	-.21
Teaching learning strategies	-.13	-.56	.19	.56	.25
Monitoring students' progress					
Times a year students' progress is monitored	.63	-1.05	.34	-.85	¹⁾
Measures taken for struggling learners					
Adaptation of teaching to diverse needs of students	-.22	-.39	.28	.72	.06
Curative measures for struggling learners	.00	-.80	.54	.59	²⁾
Learning					
Involvement of students	-.13	-.13	.07	.53	.00
Reflexivity and discursiveness	..06	-.69	.25	-.13	.56
Context					
Percentage of schools in urban regions	.22	.03	-.32	.52	³⁾
Percentage of students from socially deprived families	.28	-.03	-.24	.03	³⁾
Percentage of students for which the language of instruction is not their native language	.09	.18	-.15	-.46	³⁾
Number of students at school	-.11	.08	.06	-.14	-.22
Number of children in classroom	-.27	-.10	.24	.22	.33
Percentage of combined classes	-.38	.09	.11	.29	.61
Percentage of beginning teachers	-.02	-.28	.14	.37	³⁾
Percentage of male teachers	.03	.03	.11	-.45	.02

1) In the Slovak Republic, this year no standardised national tests were used, though they are being developed. However schools and teachers do check the students on a regular basis.

2) This is not measured in the Slovak Republic but struggling learners receive individually targeted assistance from the teacher.

3) These data are not available for the Slovak Republic.

7.1 Opportunity to learn

Opportunity to learn has different aspects: the quality of the curriculum (i.e., the textbooks and methods), the amount of time offered to reach the targets of the curriculum, and last but not least, the amount of students that lag behind so far that they do not even have a chance to reach the minimum targets. The quality of the curriculum is measured by simple questions like: are the textbooks and methods regular and modern textbooks and do they cover the minimum targets for arithmetic? Almost all schools in Flanders, Lower Saxony, the Netherlands, the Slovak Republic and Scotland use regular and modern textbooks that cover the minimum targets for arithmetic. Differences between countries seem to be very small, with Scotland on the lowest side of the dimension and Slovak Republic on the highest side.

The time devoted to arithmetic is quite different in the countries researched. In Flanders, the Netherlands and Scotland, about 5 hours are devoted to arithmetic weekly. In Lower Saxony and the Slovak Republic the weekly time for arithmetic is less than 4 hours weekly.

On average 2 to 3% of the students lag behind for more than a year. The same average is found in Flanders and the Netherlands. In Lower Saxony and the Slovak Republic, the percentage of students requiring support is a bit lower (1.11% and 1.38%, respectively) but is much higher in Scotland (9.25%). Students requiring support may for instance use texts designed for younger students during a short period of time in order to revise a specific aspect of the curriculum. Later on they may move back to their normal textbook and will be able to reach minimum targets. In Scotland children are taught in mixed ability classes in primary school so there may be students with additional support needs in many classes. In some cases the observation may have taken place in a class where the children work in ability groups for maths. If the observation took place in a lower ability group then this may have skewed the data.

7.2 Teaching

The quality of the learning climate does not seem to differ very much between Flanders, Lower Saxony, Slovak Republic and Netherlands. These countries all score within a narrow range around the international average of .80, with Lower Saxony on the low side. The average score of the Scottish teachers is much higher though. The Scottish teachers differ from the national average with an effect size of .69.

The quality of instruction does not differ very much between Flanders, Slovak Republic and Netherlands either. These countries score around .74. In Lower Saxony the quality of instruction is slightly lower (effect size -.25) and in Scotland much higher (effect size .92).

Teachers in the Netherlands score for classroom management around the international average (.78). Teachers in Lower Saxony and Slovak Republic score slightly lower than the international average (effect size -.21) and Flemish teachers slightly better (effect size .29) than the international average. Scottish teachers score much better on classroom management than the international average (effect size .79).

Teachers in the Netherlands and Flanders score on teaching learning strategies around the international average (.64). Teachers in Slovak Republic score slightly higher (effect size .25) and in Scotland the scores are clearly higher than the

international average on teaching learning strategies (effect size .56). Teachers in Lower Saxony on the other hand score clearly lower than the international average on learning strategies (effect size -.56).

7.3 Monitoring students' progress

Monitoring students' progress is researched by a simple question about the frequency of use of standardised tests. Flemish teachers monitor students' progress more than twice a year. Dutch teachers twice a year and teachers in Lower Saxony and Scotland monitor students' progress less than once a year. As mentioned, in Scotland there is continuous assessment of various types but national assessments are used once a year.

7.4 Measures taken for struggling learners

Slovak teachers score around the international average for adaptation of teaching to the diverse needs of students. Dutch teachers score slightly better (effect size .28) and teachers in Scotland score much better (effect size .72), while teachers in Flanders and Lower Saxony score slightly lower than the international average (effect sizes -.22 and -.39, respectively).

Curative measures taken for struggling learners is at the international level in Flanders, but much more on the low side in Lower Saxony (effect size -.80) and more on the high side in the Netherlands (effect size .54) and in Scotland (effect size .59).

7.5 Learning

Involvement of students is about equal in Flanders, Lower Saxony, the Slovak Republic and the Netherlands. Teachers in these countries score around the international average of .76 on involvement of students. In Scotland, involvement of students is better (effect size .53).

Reflexivity and discursiveness is around the international average in Flanders and Scotland. In Lower Saxony the quality of reflexivity and discursiveness is lower (effect size -.69) and in the Slovak Republic and the Netherlands much higher (effect size .56 and .25).

7.6 Context

The countries differ in the amount of schools found in urban regions. The international average in our samples is about 40%. The same percentage is found in the sample of Lower Saxony. In the Dutch sample, only 24% of the schools are found in urban regions, in the Flanders' sample 51% and in the Scottish sample 65%.

There was not so much difference found in the percentage of students from socially deprived families. The international average in the samples is about 16 per cent.

About the same percentages are found in Lower Saxony and Scotland. In the Netherlands, this is lower (12%). It is higher in Flanders (21%).

The countries differ more in the amount of students for which the language of instruction is not their native language. In Scotland this is less than 2 per cent, in the Netherlands 9%, Flanders 14% and in Lower Saxony about 30%.

Average school size and average classroom size does not seem to differ very much in the four counties. School size and classroom size are in all countries on average about 218 and 20 students, respectively.

In Lower Saxony the percentage of beginning teachers is two. In Flanders, the Netherlands and Scotland this is 10, 15 and 22%, respectively.

The percentage of male teachers is about 20 in Flanders, Lower Saxony, the Slovak Republic and the Netherlands and 1 per cent in Scotland.

Differences in contextual variables might influence results on teaching and learning. We therefore computed the averages scores on the scales measuring the quality of learning and teaching after correction for these contextual variables. Table 7 shows the results on opportunity to learn, teaching, monitoring students' progress, measures taken for struggling learners and learning, after correction for the contextual variables.

A quick comparison between Tables 6 and 7 makes clear that the differences between raw scores (Table 6) and context corrected scores (Table 7) are negligible.

Table 7. Results corrected for context

	Flanders	Lower Saxony	Netherlands	Scotland	Slovak Republic ¹⁾
Opportunity to learn					
Quality of the curriculum	.98	.99	.98	.95	1.00
Weekly minutes devoted to arithmetic	304.89	234.91	294.50	292.89	225.26
Percentage of students using textbooks and methods of 'younger grades'	1.80	1.21	2.95	11.31	.08
Teaching					
Safe and stimulating learning climate	.83	.77	.79	.86	.78
Clear and activating instruction	.73	.71	.72	.84	.75
Efficient classroom management	.82	.74	.78	.87	.75
Teaching learning strategies	.62	.56	.66	.70	.68
Monitoring students' progress					
Times a year students' progress is monitored	2.23	.70	2.07	.83	-
Measures taken for struggling learners					
Adaptation of teaching to diverse needs of students	.60	.58	.70	.79	.64
Curative measures for struggling learners	.67	.26	.85	.88	-
Learning					
Involvement of students	.74	.72	.77	.84	.76
Reflexivity and discursiveness	.57	.47	.62	.53	.67

1) The data for the Slovak Republic are only corrected for available variables: school size, class size and percentage of male teachers.

8 Discussion

The average score of the teachers in Flanders is very often located around the international average. There is one exception; Flemish teachers monitor the progress of their students more than twice (2.3) a year. This is more often than the international average (1.7).

The teachers in the Lower Saxony devote less than 4 hours weekly to arithmetic while their colleagues in most other European countries devote about 5 hours a week to arithmetic. Teachers in Lower Saxony score more often below the international average on: 'teaching learning strategies', 'monitoring students' progress', and 'curative measures for struggling learners' and the students in Lower Saxony score more often below the international average on 'reflexivity and discursiveness'. The weaker evaluation of some criteria in Lower Saxony is in accordance with the judgements of the regular inspections in Lower Saxony. The development of general process competencies like learning strategies, has been established with a major commitment in Germany by the implementation of national curriculum standards, starting in 2004; it is not yet sufficiently found in the practice of schools. Matching teaching strategies and students' abilities or learning problems in an individualised curriculum, which is planned systematically and over the longer term and supported by special materials for individual students, is little practised. This is clearly demonstrated in Table 7. This experience already becomes evident during the training of the inspectors. On the question on the observation form 'How many students use textbooks and methods of 'younger' grades, relatively few observations were expected for schools in Lower Saxony. Students with large learning problems frequently transfer to a lower class to repeat the curriculum or to schools for students with special needs. Concerning the concepts of meta-cognition and discursiveness, we assume that the lower value is caused by a more precise view, due to more intensive training on the concepts, which had already been accomplished by the developers of the concept, Professors Kaune and Cohors Fresenborg, in Lower Saxony. Standardised tests are rarely used on the initiative of individual schools and instructors in Germany, which is shown by the IGLU and PISA results. From grade 1 to 4 in Lower Saxony, only once, in grade 3 is a written mathematics test taken in all schools.

Dutch teachers often score around the international average. There is one positive exception. Dutch teachers more often offer 'curative measures for struggling learners'.

Scottish teachers often score higher than their European colleagues. They more often have a 'safe and stimulating climate', a 'clear and activating instruction', a 'efficient classroom management', better 'teaching learning strategies', they adapt their teaching more often to the diverse needs of students and have more curative measures for struggling learners. The involvement of students is better than in the other participating countries. In Scotland however, more students are using the textbooks and methods of 'younger grades', and students' progress is monitored half as much (.83) as on average in Europe (1.69). Nevertheless we should be rather careful with the interpretation of the Scottish results. We have to keep in mind that the Scottish sample is much smaller than the samples in the other European

countries. Therefore, the precision of this Scottish sample is about 10%, while in the other European countries, the sample precision is about 5%. This means that another Scottish sample might deliver different results.

The teachers in the Slovak Republic devote less than 4 hours weekly to arithmetic while their colleagues in the other European countries devote about 5 hours a week to arithmetic. The average scores of Slovak teachers are around the international average in all teaching aspects. The 'reflexivity and discursiveness' of Slovak students is better than the international average.

It is tempting to compare these figures about teaching and learning with the figures on student's achievements from PISA and TIMSS studies. We prefer not to make such premature analyses. It is more important to make preparations for combined studies into the achievements of students like the PISA and TIMSS studies and observational studies into the quality of the teaching and learning in the participating countries.

References

- Capie, W., Johnson, C.E., Anderson, S.J., Ellett, C. and Okey, LR. (1980). *Teacher performance assessment instruments*. Athens, GA: University of Georgia.
- Carver, R. P. (1978). The Case Against Statistical Significance Testing. *Harvard Educational Review*, 48 (3), 378-399.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohors-Fresenborg, E. & Kaune, C. (2007a). Modelling Classroom Discussions and Categorising Discursive and Meta-cognitive Activities. *Proceedings of CERME 5*, 1180-1189.
- Cohors-Fresenborg, E. & Kaune, C. (2007b). Kategoriensystem für metakognitive Aktivitäten beim schrittweise kontrollierten Argumentieren im Mathematikunterricht. *Arbeitsbericht Nr. 44*. Osnabrück: Forschungsinstitut für Mathematikdidaktik.
- Evertson, C. (1987). *Classroom observation record: observation record for project STAR*. Nashville, TN: Vanderbilt University.
- Evertson, C. & Burry, J. (1989). Capturing classroom context: the observation instrument as lens for assessment. *Journal of Personnel Evaluation in Education*, 2, 297-320.
- Flanders, N. (1970). *Analyzing teaching behavior*. Reading (Mass.): Addison-Wesley.
- Florida Coalition for the Development of a Performance Measurement System (1983). *Domain: Knowledge base of the Florida performance measurement system*. Tallahassee, FL: Office of teacher education, certification and in-service staff development.
- HM Inspectorate of Education (1999). *Quality in Secondary Schools: mathematics 1999*. HM Inspectorate of Education: Edinburgh.
- HM Inspectorate of Education (2001). *Standards and Quality in Primary Schools: Mathematics 2001*. HM Inspectorate of Education: Edinburgh.
- Inspectorate of Education The Netherlands (1998). *Integral school supervision 1999*. Utrecht: Inspectorate of Education The Netherlands
- Kaune, C. (2006). Reflection and Meta-cognition in Mathematics Education - Tools for the Improvement of Teaching Quality. *Zentralblatt für Didaktik der Mathematik*, 38 (4), 350-360.
- Mullis, I.V.S., Martin, M.O., Gonzales, E.J. & Chrostowski, S.J. (2004). *TIMSS 2003 International Mathematics Report*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I.V.S., Martin, M.O. & Foy, P. (2008). *TIMSS 2007 International Mathematics Report*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

OECD (2006). *PISA 2006 Science competencies for tomorrow's world. Volume 1 and 2*. Paris: OECD.

Ofsted (1995). *Guidance on the inspection of nursery & primary schools*. London: HMSO.

Slavin, R.E. (1987). Ability grouping and achievement in elementary schools. *Review of Educational Research*, 57, 293-336.

Stallings, J. & Kaskowitz, D. (1974). *Follow-through classroom observation evaluation (1972-1973)*. Menlo Park, CA: SRI International.

Stallings, J., Needels, M. & Stayrook, N. (1979). *How to change the process of teaching basic reading skills in secondary schools*. Menlo Park, CA: SRI International.

Stringfield, S., Teddlie, C. & Suarez, S. (1985). Classroom interaction in effective and ineffective schools: preliminary results from phase HI of the Louisiana

School Effectiveness Study. *Journal of Classroom Interaction*, 20, 31-37.

Teddlie, C., I. Virgilio & J. Oescher (1990). Development and validation of the Virgilio Teachers Behavior Instrument. *Educational and Psychological Measurement* 50,421-430.

Van de Grift, W. & J.F. Lam (1998). Het didactisch handelen in het basisonderwijs. [Instruction in primary education.] *Tijdschrift voor Onderwijsresearch* 23(3), 224-241.

Van de Grift, W. (2004). *Quality of teaching in England, Belgium Flanders, Lower Saxony (Germany) and the Netherlands*. Utrecht: The Netherlands Inspectorate of Education.

Van de Grift, W. (2007). Quality of teaching in four European countries: a review of the literature and application of an assessment instrument *Educational Research*, 49(2), 127-152.

Veenman, S., Lem, P., Voeten, B., Winkelmolen, B. & H. Lassche (1986). *Onderwijs in combinatieklassen*. [Education in multi-graded classrooms.] 's- Gravenhage: SVO.

Virgilio, I. (1987). *An examination of the relationships among school effectiveness in elementary and junior high schools*. Doctoral dissertation, University of New Orleans.

Virgilio, I. & Teddlie, C. (1989). *Technical Manual for the Virgilio Teacher Behavior Inventory*. Unpublished manuscript, University of Orleans.

Appendix I

Lesson Observation and Interview Form for Evaluating the Quality of Learning and Teaching

School		Teacher:	M / F
Age group:		Beginning teacher:	Y / N
Activity/subject matter: arithmetic		Mixed age group:	Y / N
Name of inspector:		# of students in classroom:	...
% students from socially deprived families	...%	# of students in school:	...
% students for which the language of instruction is not their native language	...%	# of residents in community:	...

Ask the teacher the following questions:			No=0 Yes=1
Opportunity to learn the minimum objectives	1	Are textbooks and methods for arithmetic frequently used by other schools?	0 1
	2	Do textbooks and methods for arithmetic cover the minimum objectives of the grade?	0 1
	3	Are the textbooks and methods for arithmetic obsolete or rarely used?	0 1
	4	How many weekly hours are spent on arithmetic?	...
	5	How many students use textbooks and methods of 'younger' grades?	...
Monitoring	6	How many times a years are the achievements of students tested with standardised tests?	...
Curative measures for struggling learners	7	Does the teacher diagnose the learning problems of students at risk?	0 1
	8	Does the teacher have prescribed learning plans for students at risk?	0 1
	9	Does the teachers implement prescribed learning plans for students at risk?	0 1

Observe the following events:					
Rate ¹ Please circle the correct answer: 1= predominantly weak; 2= more weaknesses than strengths 3= more strengths than weaknesses; 4= predominantly strong.					
Observed ² Please circle (voluntary) the correct answer: 0= no, I didn't observe this; 1= yes, I have observed this.					
Indicator: The teacher ...			Rate ¹	Good practice examples: The teacher ...	Observed ²
Safe and stimulating learning climate	10	...ensures a relaxed atmosphere	1 2 3 4	...addresses the children in a positive manner	0 1
				...reacts with humour and stimulates humour	0 1
				...allows children to make mistakes	0 1
				...demonstrates warmth and empathy toward all students	0 1
	11	...shows respect for the students in behaviour and language use	1 2 3 4	...allows students to finish speaking	0 1
				...listens to what students have to say	0 1
				...makes no role-confirming remarks	0 1
	12	...promotes the mutual respect and interest of students	1 2 3 4	...encourages children to listen to each other	0 1
				...intervenes when children are being laughed at	0 1
				...takes (cultural) differences and idiosyncrasies into account	0 1
				...ensures solidarity between students	0 1
	13	...supports the self-confidence of students	1 2 3 4	...ensures that events are experienced as group events	0 1
...feeds back on questions and answers from students in a positive way				0 1	
...pays students compliments on their results				0 1	
...honours the contributions made by children				0 1	

	14	...encourage students to do their utmost	1 2 3 4	... praises students for efforts towards doing their utmost	0 1
				...makes clear that all students are expected to do their utmost	0 1
				...expresses positive expectations to students about what they are able to take on	0 1
Clear and activating instruction	15	...clarifies the lesson objectives at the start of the lesson	1 2 3 4	...informs students at the start of the lesson about the aims of the lesson	0 1
				...clarifies the aim of the assignment and what the students will learn from it	0 1
	16	...evaluates whether the objectives have been achieved at the end of the lesson	1 2 3 4	...checks the students' achievements	0 1
				...verifies or evaluates whether the aims of the lesson have been achieved	0 1
	17	...gives clear instructions and explanations	1 2 3 4	...activates the children's prior knowledge	0 1
				...explains in sequential stages	0 1
				...asks questions that are understood by the students	0 1
				...summarises the lesson materials from time to time	0 1
	18	...gives clear explanations of the learning materials and the assignments	1 2 3 4	...ensures that every child knows what he has to do	0 1
				...explains how assignments are aligned to the aims of the lesson	0 1
				...clearly indicates the materials that can be used as learning aids	0 1
	19	...involves all students in the lesson	1 2 3 4	...gives assignments that stimulate students into active involvement	0 1
				...gives turns to and/or involves those students who do not voluntarily participate in classroom activities	0 1
				...ensures that students listen carefully and keep on working	0 1
				...waits sufficiently long to allow children to reflect after posing a question	0 1
				...gives the opportunity to respond to students who don't put their hands up	0 1
	20	...makes use of teaching methods that activate the students	1 2 3 4	...makes use of conversational forms and discussion forms	0 1
				...provides graduated exercises	0 1
				...permits working in corners/groups	0 1
				...makes use of ICT	0 1
				...uses a variety of instruction strategies	0 1
				...varies assignments	0 1
				...varies lesson materials to be used	0 1
			...uses material and examples from the students' daily life	0 1	
			...asks lots of questions	0 1	
21	...poses questions which encourage thinking	1 2 3 4	...uses pauses long enough to let all students think following questions	0 1	
			...encourages students to ask one another questions	0 1	
			...asks students to explain their understanding of topics to each other	0 1	
			...regularly checks for understanding	0 1	
			...poses questions that elicit feedback	0 1	
22	...checks whether students understand the lesson content	1 2 3 4	...poses questions that initiate reflection	0 1	
			...regularly checks for understanding	0 1	
23	...gives feedback on answers of students	1 2 3 4	...makes explicitly clear whether an answer is correct or not	0 1	
			...makes explicitly clear why an answer is correct or not	0 1	
			...gives feedback on the way students arrive at their answers	0 1	

	24	...checks whether students are completing the assignments correctly	1 2 3 4	...checks whether students have understood what they have to do	0 1
				...gives feedback on the social functioning involved in the completion of tasks	0 1
Classroom management	25	...gives a well structured lesson	1 2 3 4	...ensures clearly recognisable components in the lesson, with well-managed transition points	0 1
				...presents the lesson with a logical flow, moving from simple to complex	0 1
				...gives assignments clearly related to what students learned during instruction	0 1
				...ensures sufficient variation between instruction, teacher-led practice, and making assignments	0 1
	26	...ensures the orderly progression of the lesson	1 2 3 4	Entering and leaving the classroom takes place in an orderly manner	0 1
				...intervenes in a timely and appropriate way to any order disruptions	0 1
				...acts as a 'watchdog' for agreed codes of behaviour and rules	0 1
				...makes sure that all students are involved in learning activities until the end of the lesson	0 1
				...makes sure that there is clarity about when and how students can obtain help to do their work in class	0 1
				...makes sure that there is clarity about what options are available when students have finished their assignments	0 1
	27	...uses learning time efficiently	1 2 3 4	...starts lessons on time	0 1
				There is no loss of time at the start, during or at the end of the lesson	0 1
There are no dead moments				0 1	
The children are not kept waiting				0 1	
28	...ensures efficient classroom management	1 2 3 4	...makes clear which lesson materials should be used	0 1	
			The lesson materials are ready to use	0 1	
			Lesson materials are adapted to the level and the experience of the students	0 1	
Adaptation of teaching to diverse needs of students	29	...adapts the instruction to the relevant differences between students	1 2 3 4	...allows students who need less instruction to commence with the work	0 1
				...gives extra instruction to small groups or individual students	0 1
				...does not direct himself exclusively to the middle bracket	0 1
	30	...adapts the assignments and processing to the relevant differences between students	1 2 3 4	...makes a distinction in the scope of the assignments between individual children	0 1
				...does not give all children the same time to complete the assignment	0 1
				...allows some children to make use of auxiliary materials	0 1
	31	...offers struggling learners extra learning or instruction time	1 2 3 4	...offers struggling learners extra learning time	0 1
				...offers struggling learners extra instruction time	0 1
				...offers struggling learners extra time for practice	0 1
				...offers struggling learners pre instruction before formal instruction starts	0 1
				...offers struggling learners extra instruction when the formal instruction has ended	0 1
	32	...supports the self-confidence of struggling learners	1 2 3 4	...feeds back on questions and answers from struggling learners in a positive way	0 1
...expresses positive expectations to struggling learners about what they are able to take on				0 1	
...pays struggling learners compliments on their results				0 1	
...honours the contributions made by struggling learners				0 1	

INSPECTIE VAN HET ONDERWIJS

Teaching learning strategies	33	...stimulates students to think about solutions	1 2 3 4	...gives students clues about solutions	0 1
				...teaches students search and reference strategies	0 1
				...teaches students the use of organisation resources	0 1
				...offers checklists for solving problems	0 1
	34	...let students speak aloud while thinking	1 2 3 4	...offers students opportunities to give solutions	0 1
				...ask students to put their solutions into words	0 1
	35	...teaches students how to break down complicated problems	1 2 3 4	...teaches students how to simplify problems	0 1
				...teaches students how to break down complicated problems into simple ones	0 1
				...teaches students how to bring order to complicated problems	0 1
	36	...teaches students how to check solutions	1 2 3 4	...teaches students how to estimate outcomes	0 1
				...teaches students to predict outcomes	0 1
				...lets students relate outcomes to practical situations	0 1
	37	...ensures that the teaching materials are orientated towards transfer	1 2 3 4	...promotes the conscious use of what has been learned in other areas	0 1
				...teaches students how to use solutions in other (different) contexts	0 1
				...relates problems to earlier solved problems	0 1
	38	...fosters critical thinking in students	1 2 3 4	...asks students to identify the reasons why specific activities take place	0 1
				...invites students to give their opinions	0 1
				...asks the students to reflect on solutions or answers given	0 1
				...invites students to give their own examples	0 1
	39	...invites students to use strategies that can help them solve different types of problems	1 2 3 4	...invites students to explain the steps of the problem-solving strategy they use	0 1
...explicitly provides instruction in problem-solving strategies				0 1	
...invites students to explain the advantages and disadvantages of solutions				0 1	
40	...stimulates the use of control activities	1 2 3 4	...gives attention to estimatory calculation/anticipatory reading	0 1	
			...makes solutions relate to the context	0 1	
			...stimulates the use of alternative solutions	0 1	
41	...provides interactive instruction and activities	1 2 3 4	...facilitates mutual interaction between students	0 1	
			...ensures interaction between students and the teacher	0 1	
Indicator: Students			Rate ¹	Good practice examples: Students ...	Observed ²
Reflexivity and discursiveness	42	...plan their own learning process	1 2 3 4	...plan a sequence of cognitive tools (computing steps, constructing steps)	0 1
				...plan a sequence of intermediate results	0 1
				...plan meta-cognitive activities	0 1
	43	...watch over their own thinking and learning process	1 2 3 4	...control calculation	0 1
				...control terminology, notation and argumentation	0 1
				...control the reference to facts and aims	0 1
				... practise self-monitoring concerning own calculations, expressions or terminology	0 1

	44	...reflect on the appropriateness of solutions and methods used	1 2 3 4	...reflect on each other's solutions and methods used	0 1
				...reflect the wilful choice of a representation (formula, graphic, term, text passage)	0 1
	45	...practise discursiveness	1 2 3 4	...link a discourse contribution (naming of reference points or persons)	0 1
				...are sure of the conversation basis/linking of the following or preceding statement	0 1
				...are revealing the rules for the discourse	0 1
	Involvement of students	46	There is good individual involvement by the students	1 2 3 4	...are attentive
...take part in learning/group discussions					0 1
...work on the assignments in a concentrated and task-focused way					0 1
47		...are interested	1 2 3 4	...listen to the instructions actively	0 1
				...ask questions	0 1
48		...are active learners	1 2 3 4	...ask 'deeper' questions	0 1
				...take responsibility for their own learning process	0 1
				...work independently	0 1
				...take initiatives	0 1
Final judgement		49	The overall quality of teaching I assess as:	1 2 3 4	
	50	The overall quality of the involvement of students I assess as:	1 2 3 4		

Appendix II

International Project Team

Wim van de Grift (project leader and author),
Inspectorate of Education, Utrecht, The Netherlands
and the University of Groningen, The Netherlands
Phone: +31 6 1299 2069
E-mail: w.j.c.m.van.de.grift@owinsp.nl

Christa Kaune and Elmar Cohors-Fresenburg (observation instrument: Reflexivity
and Discursiveness), University of Osnabrück, Germany

Ineke van Kampen (data analysis),
Inspectorate of Education, Utrecht, The Netherlands

Jaqui Horsburg (data co-ordination Scotland),
HM Inspectorate of Education, Dundee, Scotland

Zuzana Lukackova (data co-ordination Slovak Republic),
State School Inspection, Bratislava, Slovak Republic

Matthias Boeckermann and Norbert Sommer (data co-ordination Lower Saxony),
School Inspection of Lower Saxony, Bad Iburg, Germany

Yvan Verbauwhede (data co-ordination Flanders),
Inspectorate of Education, Brussels, Flanders, Belgium

Many thanks to all the inspectors who participated in the ICALT-project.