 Journal of Classroom Interaction, ISSN 0749-4025. © 2016, Vol 51.2, pp. 50-72

Differentiation Practices in Grade 2 and 3: Variations in Teacher Behavior in Mathematics and Reading Comprehension Lessons

Evelien S. Ritzema
Marjolein I. Deunk
Roel J. Bosker
University of Groningen, The Netherlands

ABSTRACT

This study focused on the differentiation practices of second- and third-grade teachers in mathematics and reading comprehension lessons. Preconditions for differentiation, classroom organization, and how teachers dealt with students of different ability levels were investigated through observations, using a time-sampling instrument. Data of 43 teachers, from 18 schools, show the importance of taking context factors into account. The study also focused on how students of four different performance levels were addressed by their teachers. It was found that teachers mostly adapted teaching to the relatively weak performing students in their class by addressing them more often, in a content-related way. Relatively advanced students received additional attention less often.

INTRODUCTION

Every teacher is confronted with students who differ in their cognitive abilities. As a consequence, the teacher has to adapt his/her teaching to the diverse student needs (Corno, 2008; Vogt & Rogalla, 2009), thereby creating a responsive educational environment. Differentiation provides an educational tool for arranging such educational conditions. Or, as Bosker (2005) puts it, without differentiation, teaching will not be adaptive: differentiation is the essential part of a teaching process that optimally tailors instruction to students' needs. Tomlinson et al. (2003, p. 121) define differentiation as an "approach to teaching in which teachers proactively modify curricula, teaching methods, resources, learning activities, and student products to address the diverse needs of individual students and small groups of students to maximize the learning

opportunity for each student in a classroom". This definition pertains to a broad array of student differences (e.g., student performance levels, interests, and learning styles); however, the current study focused solely on how teachers deal with differences in student performance levels.

In recent years, there have been concerns about the quality of the differentiation practices of Dutch teachers (Educational Inspectorate, 2008b, 2010, 2012). Only 48% of the schools in a representative sample were judged to sufficiently differentiate on all items used to measure differentiation practices (Educational Inspectorate, 2013a). Further, it was reported that flexible adjustment of instruction to performance levels poses problems for Dutch teachers: because they have been found to insufficiently analyze, interpret, and act on student performance data, it is questionable whether the additional instruction that is provided optimally fits students' needs (Educational Inspectorate, 2010; 2013a). Given the importance of alignment between instruction and students' proficiency level (Moon, 2005), the effectiveness of additional instruction that is not based on performance data is doubtful.

The reports of the Dutch Inspectorate provide a general picture of how much teachers differentiate, but they do not provide detailed information on what happens in classrooms. In many (international) studies it also remains unclear how and how often teachers actually address learner differences with respect to instruction, learning time, and the difficulty of tasks. For instance, researchers who investigate grouping practices regularly draw conclusions that are not based on observations. Instead, teacher self-reports of grouping practices are used to provide information about the relation between these practices and student outcomes (Condron, 2008; MacIntyre & Ireson, 2002; Nomi, 2009; Tach & Farkas, 2006). Other researchers, some of whom use observational data for checking program fidelity, only focus on the grouping practice with respect to specific groups of students (Pierce et al., 2011; VanTassel-Baska, Zuo, Avery, & Little, 2002; Vaughn et al., 2003); they do not discuss what happened to the other students in the class. In short, naturally occurring differentiation practices are often not described in detail, leaving unanswered the questions what actually happens in class and how context factors might influence these practices. In the current study we aim to provide, albeit in an explorative manner, such a picture of how teachers differentiate in heterogeneous classes in daily practice. Differentiation practices in second- and third-grade mathematics and reading comprehension lessons were explored, specifically taking into account teacher behavior towards students of different performance levels and the influence of context factors - heterogeneity of the class, multi- or single-gradedness of the class, and the subject area.

DIFFERENTIATION PRACTICES

One way of differentiating is to provide additional learning time to specific groups of learners. However, as Houtveen, van de Grift, and Creemers (2004) showed, it is not the mere provision of additional time, but the additional instruction that helps weak students forward. Such additional instruction might be organized through within-class grouping for specific subject areas. By establishing temporary homogeneous groups, teachers can manage heterogeneity by adapting their instruction to the ability level of the (small) group. Slavin (1987) and Lou et al. (1996) have shown that within-class ability grouping has positive effects on student performance. However, the formation of ability groups only facilitates effective differentiation if the teachers really adapt instruction and materials to students' performance levels (Lou et al., 1996).

Flexible grouping is also stressed in this respect: it decreases the important disadvantages of grouping, like stigmatizing low achievers and offering these students different opportunities-to-learn (Nomi, 2009). Students should thus only temporarily be assigned to a group in order to master specific skills (Slavin, 1987); they should also be

able to learn from their more able classmates during whole-class teaching (see Corno, 2008) or within heterogeneous small groups.

These considerations are in line with some of the features of effective differentiation described by Tomlinson (2005): a) the use of classroom organization in such a way that the lesson includes time for whole-class, small group, and individual attention; b) the use of (formative) assessment to base instruction on; and c) the flexible use of time, space, materials, and instructional strategies. It is thus essential for effective differentiation to be fully embedded in the teaching practice (Bosker, 2005).

Given the embeddedness of differentiation in the broad array of activities that comprise effective teaching, its appropriate implementation requires rather complex teacher skills (Slavin, 1987; Whitburn, 2001). In order to adjust teaching to the diverse needs of students and to put differentiation into practice, teachers not only have to be able to make well-informed instructional decisions, but they also need to possess good organizational skills. For instance, while tutoring a small group, the teacher has to make sure that the other students in the class not only work on relevant tasks, but also stay on task, without disturbing each other or the teacher.

Teachers also have to manage time adequately, making sure that all students are addressed sufficiently and that there is enough time for whole-class teaching and evaluative moments at the end of the lesson, leaving room for reflection on the key lesson objectives for all students (Muijs & Reynolds, 2011). Further, regular monitoring of students' activities during seatwork, thereby informally assessing students' levels of understanding, provides the teacher with relevant information that can be used to ensure alignment of additional instruction.

Context factors like the subject domain, the heterogeneity of the class and its single- or multi-gradedness might influence how well teachers apply differentiation practices. For instance, in a heterogeneous class the teacher is confronted with a vast number of different performance levels. As a result, the teacher might feel the need to use more homogeneous performance groups than in a rather homogeneous class. In a multi-grade class heterogeneity is evident, since the teacher has to deal with two (or more) year groups. The subject domain might also influence differentiation practices: the (cumulative) nature of a subject area and the way it is structured in the textbook might facilitate the use of differentiation practices.

CONTEXT FACTORS

Heterogeneity. Some classes contain more homogeneous ability levels than others. As several authors have argued (e.g., Hanushek & Wößmann, 2006; Huang, 2009), groups with a rather homogeneous ability level might facilitate teaching, since targeted instruction can be provided that matches all students' needs sufficiently.[1] However, when confronted with a class that contains a large degree of heterogeneity in ability levels, a teacher might feel forced to form ability groups. Working with larger numbers of ability groups might affect the amount of time students have to work on their own, resulting in less direct and intense support (Wilkinson & Hamilton, 2003). Wilkinson and Hamilton (2003) referred in this respect to the organizational constraints that influence the number of small groups and the diversity within these groups. They stated that teachers feel frustrated when working with six or more ability groups.

Multi-grade classes. Some classes contain more homogeneous ability levels than others. As several authors have Multi-grade classes are not a typical Dutch phenomenon; they can be found in many countries, like Great Britain, Northern Ireland, Finland, Norway, Switzerland, Canada, Australia, USA, Nepal, and Peru (Little, 2004; Mulryan-

[1] Despite the possibility of providing more targeted instruction, Huang (2009), Hanushek and Wößmann (2006), and Ireson and Hallam (2001) argued that homogeneous grouping has differential effects, leading to greater inequalities at the expense of low-achieving students. Due to a slower pace and lower instructional quality (Huang, 2009; Ireson & Hallam, 1999), homogeneous grouping often negatively affects low achievers.

Kyne, 2007). As mentioned above, multi-grade classes are heterogeneous by nature. Since they contain at least two year groups, they also contain at least two grade-specific learning objectives. Several American studies have found that multi-grade students were positively selected, meaning that students with higher academic ability and more positive and independent behavior were assigned to multi-grade classes (Burns & Mason, 2002; Thomas, 2012). This is assumed to facilitate teaching, since teachers can stick to two curricula and need not use further ability-grouping in the specific grades to address all students' needs. Yet these selection processes were not found to be present in the Netherlands (Veenman, 1997).

It is widely acknowledged that teaching multi-grade classes is more difficult than teaching single-grade classes, because of the two curricula, more preparation time, and organizational challenges (Mason & Burns, 1997; Veenman, 1997). Students are further left to work independently for longer, due to the two curricula that are taught (Mason & Burns, 1997). With respect to grouping practices, the Dutch Inspectorate of Education has found that teachers adapt their instruction less often to students of different cognitive levels in multi-grade classes than in single-grade classes (63% and 68%, respectively), possibly as a result of the high-quality classroom management skills required (Educational Inspectorate, 2013b). Especially in small schools, where multi-grade classes consisting of more than two grades are common, teachers seem to experience difficulties in adapting teaching to students' needs (Educational Inspectorate, 2013a; 2013b).

Subject area. Another context factor that might play a role in the implementation of differentiation practices is the subject area at hand. In the differentiation literature, findings are often reported that only pertain to one subject area, like mathematics (e.g., Houtveen et al., 2004; Leonard, 2001; Pierce et al., 2011; Tieso, 2005; Whitburn, 2001) or reading (e.g., Condron, 2008; Connor et al., 2009; Nomi, 2009; Reis, McCoach, Little, Muller, & Kaniskan, 2011; Tach & Farkas, 2006). However, the nature of the subject area (being either cumulative and well-structured or multidimensional and less systematic) might influence the use of differentiation practices. A study on differences between teacher behavior in mathematics and reading lessons, showed that math lessons tended to be more structured (Wiley, Good, & McCaslin, 2008).

A recent study by Nurmi, Viljaranta, Tolvanen, and Aunola (2012) on whether Finnish teachers adapt their instruction based on the performance level of first-grade students also took both subject domains into account. Here, the subject domain appeared to affect teacher behavior. Although students' low achievement levels led to a greater amount of active instruction in both reading and math, the teaching of the subjects was different. Unlike Wiley et al. (2008), Nurmi et al. (2012) found that Finnish reading lessons consisted of well-structured instruction and teacher-directed activities, whereas the mathematics lessons were more freely structured and contained more variation in instruction.

Comparing Dutch teachers' uses of differentiation practices in these two subject domains it was found that these practices are used more frequently in mathematics than in reading lessons. During math lessons, 70% of the teachers differentiates in task-difficulty and/or provides extended instruction to the low performers (Educational Inspectorate, 2010). Differentiation practices were found to a lesser extent during reading comprehension lessons. These lessons consist mainly of whole-class teaching, and specific attention to struggling readers through extended instruction is rare (Houtveen, 2002). Teacher self-reports on how they organize the classroom during reading comprehension lessons confirm this picture: in grades 2 and 3, 40% of teachers provide only whole-class teaching, whereas 50% only tailor tasks to ability levels during seatwork (van Berkel et al., 2007). This difference between the two subject domains might (partly) be explained by a difference in curricular textbooks: the math

curricular textbooks offer suggestions for grouping and task-selection, whereas reading comprehension textbooks often do not (Educational Inspectorate, 2008a; 2010; Kuiper & Palma, 2012).

In investigating the use of differentiation practices, it thus seems to be important to acknowledge that context factors can influence the findings. Furthermore, a detailed description of differentiation can only be made when teaching practices directed at different types of students are taken into account. Information on teacher guidance of students of different performance levels provides a more fine-grained picture of differentiation practices in (heterogeneous) classes.

DIFFERENTIATION PRACTICES TARGETING SPECIFIC TYPES OF STUDENTS

Several studies have been carried out to investigate how teachers deal with students of different performance levels. Reezigt, Houtveen, and van de Grift (2002) examined the implementation and effects of adaptive teaching. To investigate its implementation, 69 observations were conducted in the final year of kindergarten and grade 1. During each lesson, three types of students were observed: low-performers, high-performers, and students with behavioral problems. The observations showed that teachers did not address weak and strong performers differently regarding the tasks at hand, students' setting, and the number and types of utterances to the students.

Teachers thus had similar interactions with weak and strong students. Only the students with behavioral problems received more teacher attention, mostly referring to their behavior. Jurik and colleagues (2013) found a different pattern in a more recent, German study. They observed how different types of students were engaged during science lessons, and found that teachers interacted more with high-performing students who were highly interested. These findings confirm earlier results from Good and Brophy (2003), who reported that teachers engage high-performing students to a higher degree in their lessons. However, Nurmi et al. (2012) found the opposite. In their study, based on teacher logs, they found that teachers mostly adapted their teaching to poor-performing students by giving them more active instruction. In the Dutch context, differentiation was also found to mostly focus on raising the bar for struggling students, leading to more time and instruction for these low performers (Educational Inspectorate, 2013a). Summarizing these findings, it can be concluded that, up to now, no clear picture exists of how much and in what ways teachers actually address students of different performance levels.

RESEARCH QUESTION

It is widely accepted that differentiation is needed in order to help students of different performance levels reach their potential. However, using differentiation is not an easy task, especially in very heterogeneous groups, like multi-grade classes. Furthermore, differentiation is not a generic skill: its implementation may be influenced by the learning domain at hand and may differ for students of different performance levels. These considerations led to the following question:

In what ways and under what conditions do teachers use differentiation practices in their mathematics and reading comprehension lessons in grades 2 and 3?

To answer this question, two sub-questions were formulated:

1. *To what extent does the use of differentiation practices relate to context factors, such as heterogeneity of performance levels, type of class (multi- or single-grade), and subject domain?*

2. *How do teachers differentiate between students of varying performance levels?*

We expected mathematics and reading comprehension lessons to be structured differently: reading comprehension lessons were expected to contain more whole-class teaching and less extended instruction (Educational Inspectorate 2010; van Berkel et al., 2007). We hypothesized that this would result in less differentiation during reading comprehension lessons. We further expected less differentiation in multi-grade classes, because teachers already have to manage two or more year groups in these classes, and are, therefore, less likely to create (multiple) ability groups within the individual year groups (Educational Inspectorate, 2013a). Despite the contradictory findings on the relation between student performance and the amount of teacher attention given, the fact that the Dutch Inspectorate found convergent differentiation in 82% of schools (Educational Inspectorate, 2010) led us to hypothesize that struggling students would be addressed most frequently and would receive most additional instruction, compared to average and high-performing students.

Differentiation was examined at two levels: class level and student level. At class level, preconditions for the implementation of differentiation, being a well-managed classroom, providing a content-rich lesson, and informal assessment and opportunity to provide additional instruction by walking through the classroom during seatwork, were investigated, as was classroom organization related to differentiation (grouping practices and the provision of extended instruction); contextual factors were taken into account. At student level, the activities different types of students were engaged in, as well as the type and amount of teacher talk provided to these types of students, were investigated.

METHOD

PARTICIPANTS

In order to avoid selecting a group of teachers who did not use any relevant differentiation practices, we selected teachers who were likely to use these practices, at least to some extent. Therefore, the observational data on naturally occurring differentiation practices were collected in the context of a professional development program (PDP) on data use. One of its components, the use of performance goals, is supposed to make teaching more targeted, since teachers know what performance levels they are aiming at. Given the different aims for their students, the participating teachers were expected to adapt instruction, opportunity to learn, time on task, and task difficulty, thereby tailoring the educational environment to their students' educational needs. In the current observational study, it was explored how this selective group of teachers, who were expected to be able to implement differentiation because of their participation in a PDP in which relevant preconditions for differentiation were fostered, used differentiation practices in their lessons.

The PDP, including the observation study as it is presented in this article, had been piloted in 7 schools. In the pilot, 18 teachers were observed during a mathematics lesson. Of these 18 teachers 11 teachers were observed during a reading comprehension lesson. In the current study, 43 teachers of grades 2 and 3 from 18 schools in the northern part of the Netherlands were systematically observed during both a mathematics and a reading comprehension lesson. Teachers in this part of the country seem to master complex skills, like adaptive teaching, to a lesser extent than elsewhere in the Netherlands (Educational Inspectorate, 2013b). This is deemed problematic, since the number of small schools containing multi-grade classes is relatively high in this area, and is expected to rise even further as a consequence of demographic contraction.

Classroom observations took place in April/May 2012. Forty-one teachers were observed during a math lesson. Of these 41 math lessons, 23 observations were directed

at grade 2 and 18 observations were directed at grade 3. Since not all of the participating teachers taught reading comprehension, the reading comprehension lessons of 32 teachers were observed. Here, 17 observations concentrated on grade 2 and 15 on grade 3. Information on the classes is given in Appendix A, in which an overview of the schools, teachers, classes, and numbers of students in class is provided. The latter is specified both by the total number of students that were present in class during the observation as well as the total number of students in the observed grades. Differences between these two numbers of students are caused by the multi-or single-gradedness of the class: the sample consisted of 48 multi-grade classes and 25 single-grade classes. The multi-grade classes contained either the grades 1 and 2 (, and 3), the grades 2 and 3, or the grades 3 and 4. Although these multi-grade classes contained two or more grades, all lesson observations focused on only one specific grade and the students therein.

Seven schools in the sample were categorized as small schools, with a student population of less than 100 children. Small schools generally have multi-grade classes, sometimes containing more than two grades. In these small schools, class size is generally smaller than the average Dutch class size (22.6 students in 2011-2012). Most classes in our sample had a class size comparable to this average Dutch class size: The average number of students present in class during the observations was 21.7 (*SD*: 6.0), whereas the average number of students in the observed grades was smaller, 14.3 (*SD*: 8.5). Sixteen percent of the classes had 15 students or less, and seven percent had more than 30.

VARIABLES AND INSTRUMENTS

Observation instrument. A time-sampling instrument, based on Kooiman et al. (2005), was used to conduct the observations; discrete teacher behavior and student activity were observed. Every two minutes, observations were conducted in blocks of one minute, with gaps of one minute to code what was observed. During every coding moment, four teacher variables were scored: a) teacher talk, b) the specific student/group of students that was addressed by the teacher talk, c) the position of the teacher, and d) classroom organization. In addition, the lesson activities of four selected students of different achievement levels were scored. Table 1 (see next page) provides an overview of the categorical variables that were measured using the time-sampling instrument. The observers were requested to record the observed categories per variable by using the numbers mentioned in Table 1. In the online version of this article, 6 columns are added to Table 1 in which the raw counts (per grade, per subject and per type of class) can be retrieved. These raw counts were used in the analyses.

The results for the three variables teacher talk, position of the teacher, and classroom organization were used to answer sub question 1. Sub question 2 was answered using information on the variables teacher talk, student(s) addressed by the teacher talk, and lesson activities of four selected students.

For all variables, the situation at the start of the observation minute was recorded by the observer, also when the situation changed during the minute. Although the main variables 2 to 5 in Table 1 were scored at the beginning of the observation minute, an exception was made for the variable *teacher talk*: During the observation minute, the first utterance belonging to categories 1-3 was recorded (that is, task at hand, explanation, content-related questioning and organization – see Table 1). The category *other* was scored only when, during the observation minute, no teacher utterances could be scored as subcategory 1-3. For scoring the observations a recording scheme was used, which is presented in Appendix B. An additional column, *Elaboration*, was added to the recording scheme. Here, the observers wrote down the words or sentence used by the teacher based on which the variable *teacher talk* was scored.

Both in the pilot study and in the current study five subcategories were used to score the category teacher talk. In this way, during the PDP feedback could be provided to the teachers on their observed teaching practice, such as largely focusing on organizational

Table 1

Overview of Variables Used in the Time-Sampling

Variable	Categories	Explanation
1. Teacher talk	1. Task at hand	Teacher refers to the task at hand (e.g., "We'll start with exercise 1, page 14", "We are going to do some mental arithmetic now")
	2A. Explanation pertaining to content	Teacher provides information on the task, strategies, and solutions (e.g., "18 times 6. To solve this, you can take two steps. First, you calculate 10*6 and then 8*6")
	2B. Content-related questioning	Teacher asks for information on the task, strategies, and solutions (e.g., "How much is 6*8?"/"How did you solve it?")
	3. Organization	Teacher refers to the general sequence of the lesson or conditions for working (e.g., "Maria, please pay attention" or "You can come to me after the whole-class instruction")
	4. Other	Other teacher behavior
2. Position of the teacher	1. In front of the class	Teacher is standing or sitting in front of the students
	2. At a student's table or a group of tables	Teacher is standing or sitting with a small group of students or a single student
	3. Walking around	Teacher goes round the class
	4. At the desk	Teacher sits at the desk
	5. Other	Teacher position is not 1-4 (e.g., the teacher teaches the other year group in a multi-grade class or is outside of the class)
3. Classroom organization	1. Whole-class instruction	Whole class is taught by the teacher
	2. Extended instruction	Some of the students receive extended instruction, the other students do seatwork
	3. Seatwork	Everybody does exercises on their own (individually, in pairs, or small groups)
4. Student who is addressed during teacher talk	1. Very weak student	Selected student, minimum level
	2. Weak student	Selected student, basic level
	3. Average student	Selected student, proficient level
	4. Advanced student	Selected student, advanced level
	5. Other student	Non-selected student
	6. Group of students/ whole class	The whole class or a (small) group of students
	7. Other	A colleague/students in the other grade in a multi-grade class setting
5. Activity very weak student (the same for weak, average, and advanced students)	1. Whole-class teaching	Student is engaged in whole-class learning
	2. Extended instruction	Student receives additional instruction in a small group
	3. Individual teacher instruction	Student receives additional, individual instruction or is working individually with the teacher
	4. Without teacher guidance	Student works on his own or with a peer
	5. Other	Student is outside the classroom or is working on exercises from a different subject area.

issues during the lesson or regularly activating students by inviting them to come up with an answer. One of the aims of the current observational study was the focus on content or organization by the teachers in the three different contexts outlined, that is, in the mathematics and reading comprehension lessons, in the multi- and single-grade lessons and in homogeneous or heterogeneous classes. As a result, the 5 original categories were merged into three broader categories. Categories 3 and 4 (*organization* and *other*) differed from categories 2A and 2B (*explanation* and *content-related utterance*) in the sense that the former were not content-related, whereas the latter were. The category *task-related* seemed to be unique in nature. Teacher talk that refers to the task at hand is task-related, but is not really related to content and may also have an organizational character, as in "We'll start with exercise 1, page 14". Therefore, task-related remained a

separate category. The three categories of teacher talk – content-related, organizational, and task-related - are used throughout the remainder of this article.

The observations were conducted on site. Per lesson observation one observer was seated in the back of the classroom, registering the recording scheme immediately per observation block. A stop watch was used to maintain the determined observation period (that is, a block of one minute per timeframe of two minutes). The observations were conducted by four observers: two researchers and two research assistants who had been trained prior to the actual observations. In the pilot study the observations were conducted by the two researchers. Both for the pilot study and for the current study the observers were trained. Training in the use of the time-sampling instrument consisted of discussions on the categories and conventions, and work on video material. A manual was prepared for this purpose. This consisted of both instructions for recording and clear descriptions and examples of the variables and categories. In some cases the descriptions did not provide sufficient guidance: especially as regards teacher talk or teacher position some indistinctness in the categorization was encountered. These cases were thoroughly discussed by the observers and decisions were recorded in an addendum to the manual.[2] After being trained, the inter-rater reliability of the four observers was considered sufficient (Cohen's kappa = .82). The inter-rater reliability of the two observers in the pilot study was somewhat lower, but also considered sufficient (Cohen's kappa = .74).

Selected students. Part of the observation instrument concentrated on teacher talk directed towards students of four different achievement levels and their activities. For this purpose, four students in each class were selected based on their prior performance: a very weak, weak, average, and advanced student. As indicators for prior performance, students' scores on the standardized assessment for mathematics or reading comprehension (RC) were used. The level of prior performance of a selected student was relative to the prior performance levels of their classmates – absolute scores were not a decisive criterion.

The selection of the four students took place before the lesson observations started: The researchers had received prior assessment information for all participating classes at the beginning of the overall PDP. Before going to the schools, this assessment information was used to determine which students would be observed.[3] Teachers were not aware that specific students were explicitly observed, as we wanted to make sure that teacher activities directed towards all students were business as usual. To enable the observers to identify the selected students in the class, teachers were asked to provide a schematic overview of the students in classroom before the lesson started. As previous achievement scores on mathematics or reading comprehension were used for selecting the student and as student's performance on both subjects may differ, the students selected for math and reading comprehension were not necessarily the same.

Selection criterion: Previous achievement on standardized assessments. The selection procedure was conducted using the proficiency-classification associated with the standardized assessments for math and reading comprehension, used by the Netherlands Institute for Educational Measurement ("Cito"). These standardized assessments are part of the Cito LOVS assessment system that is used in 85% of Dutch schools and that are used throughout primary school (grades 1-6). Both the math and RC assessments are considered to have a good validity and overall reliability in all grades. For mathematics, Cronbach's alpha is at least .91 (Janssen, Verhelst, Engelen, & Scheltens, 2010); for RC, Cronbach's alpha is at least .89 (Feenstra, Kleintjes, Kamphuis, & Krom, 2010). For selecting the four students, the end-of-the-year assessment scores of the previous school year were used (May/June 2011).

[2] Readers who are interested in replicating this observation study are requested to contact the first author for more information on the manual and the addendum.

[3] There were three exceptions as regards assessment scores on reading comprehension. No achievement information, i.e., no scores on the standardized assessment, was available for the students of three teachers, as their schools did not use this specific assessment prior to the PDP. In these three classes, the teachers were asked to name two students for all four performance levels (very weak, weak, average, and advanced). From each pair of students, one student was, on the spot, randomly selected and observed.

Table 2

Scoring Intervals Belonging to the Nationally Used Classification for the Standardized Assessments for Mathematics and Reading Comprehension (June assessment of grade 1 and 2, respectively)

	Mathematics[a]					Reading Comprehension[b]				
Assessment	**E**	**D**	**C**	**B**	**A**	**E**	**D**	**C**	**B**	**A**
Proficiency ranges end of grade 1-test	0-16	17-24	26-33	35-43	45<	> -23	-22- -14	-13- -3	-2-8	9<
Proficiency ranges end of grade 2-test	0-37	38-46	47-56	57-65	66<	> -6	-5-2	3-13	15-23	24<

[a] Math test scores for all assessments (grade 1 to 6) are associated with a single proficiency scale, ranging from 0 to 169. [b] The (arbitrary) proficiency scale of the standardized assessment for reading ranges from –87 up to + 147 (grade 1 to 6).

The proficiency-classification associated with the Cito standardized assessments is norm-referenced. The classification consists of five categories, ranging from category A to E. In Table 2, per assessment used in the selection process, the ranges of proficiency scores belonging to the A-E categories are depicted. Category A represents the proficiency scores of the highest performing Dutch students in a specific subject domain, while category E represents the proficiency scores of the lowest performers.

For the purpose of selecting the four students the Cito-classification was somewhat adjusted. In the PDP, to which the observations were connected, four performance levels were identified and teachers were encouraged to base their instructional decisions on these four performance levels. Accordingly, it was decided to observe students of only four performance levels by splitting up the C-category. C-scores normally correspond to the 25% just below the average performance on the assessment (the lowest 25% to 50% scores). In the current study the C scores were divided in low and high C-scores, splitting the range in proficiency scores of the C-category in two.[4]

The classification in four categories was then as follows. Students were considered *very weak* if their proficiency scores corresponded to that of the lowest 10% of Dutch students (an E-score), they were considered *weak* if their scores corresponded to that of approximately the lowest 40% of Dutch students (D- and low C-scores), excluding the lowest 10% (the E-scores). Students having a proficiency score similar to that of the 25% highest scoring Dutch students were selected as *advanced* students (an A-score), and *average* students were assumed to perform at the level between the lowest 40% and highest 25% of proficiency levels (high C- or B-scores).

By distinguishing students in the four broad proficiency categories, the categories, by definition, identified broad groups of students. Although the adjusted Cito proficiency-classification was the core of the selection process, it is notable that this selection always took the context of the class into account. From the perspective of the teachers, the A-scores and E-scores represent extreme proficiency levels. If the teachers adapt their instruction to different ability levels, then this adaptive teaching probably at least affects these extreme scoring students. But, such A- or E-students were not present in each of the observed classes. In cases where the students in class did not neatly fit the descriptions for very weak, weak, etc., the students closest to these descriptions were selected, assuming that in each class there are relative differences between students and that teachers are supposed to adapt their teaching to these relative differences. Thus, although a student might not perform at a very weak level in an absolute sense, he/she does so in a relative sense: For the teacher this student still is the weakest student, with the most instructional needs.

This relative selection procedures had consequences for the characteristics of the selected groups of students. In Table 3, the ranges of proficiency scores of the four observed types of students are depicted, both for the mathematics and the RC assessment.

[4] Dividing the C-scores into two categories is not uncommon in the Netherlands, as teachers in primary education regularly use these classifications, naming their C-students ”low C student” or ”high C student”.

From Table 3 it can be derived that the mean score differs per type of student - that is, very weak, weak, average and advanced. As expected, the observed very weak students have a lower mean score than the weak students, etcetera. This holds for both subjects and grades. However, a consequence of the relative selection procedure was that the score intervals of the four types of students overlapped considerably, reflecting the diversity of the classes that were included in our study.

Comparing the information in Table 2 and Table 3, two relevant conclusions can be drawn. First, both for reading comprehension and mathematics, the mean scores of the very weak students appear to be (almost) at C-category level. Second, there a is large variety in scores per type of student: In a specific grade 2-class the very weak student's math score was in the A-category, whereas in another grade 2-class the advanced student only scored in the C-category. Given the upward pattern of the mean scores per type of student, these two examples obviously represent the extreme scores. Yet, similar patterns hold, to a greater or lesser extent, across grades and subjects. Hence, although mean achievement scores across grades and subjects raise per type of student - as was expected -, it is emphasized that the results should be interpreted with care, due to the indistinctiveness of the four types of students from an absolute perspective.

Other variables: Context factors. In order to answer the first sub question, regarding the relation between differentiation practices and context factors, three context factors were taken into account: subject domain, single- or multi-gradedness of the class, and heterogeneity of the class. *Subject domain* was coded as a dummy variable, using mathematics as the reference group. *Single-* or *multi-gradedness* of the class was also coded using a dummy variable (single-grade class being the reference group). In multi-grade classes, only one of the two (or more) grades was observed. *Class heterogeneity* was calculated by taking the standard deviation of the pretest scores on the standardized mathematics or reading assessment from the Cito standardized assessment system (Janssen, Verhelst, Engelen, & Scheltens, 2010; Feenstra, Kleintjes, Kamphuis, & Krom, 2010). This standard deviation was based on the pretest scores of students in the observed class, entailing that in a multi-grade class the standard deviation was based only on the results of the students in the observed grade.

ANALYSES

In order to investigate the relations between differentiation practices and the context factors, the raw counts were aggregated at teacher level and nonparametric tests (Kruskal Wallis, Mann Whitney U, Wilcoxon Signed Rank) and Spearman's rank correlations (ρ) were used. Nonparametric testing was decided on because of the skewed distribution of the data (interested readers are invited to contact the first author for more information

Table 3
Descriptives of the Standardized Assessment Scores of the Observed Four Types of Students

Types of students	Very Weak		Weak		Average		Advanced	
	M (SD)	Range	*M (SD)*	Range	*M (SD)*	Range	*M (SD)*	Range
Mathematics								
Math scores - grade 2	28.52 (11.31)	1-52	38.43 (7.33)	24-50	46.00 (7.70)	28-60	65.68 (15.20)	36-88
Math scores - grade 3	45.82 (10.15)	24-67	54.13 (8.33)	43-70	63.84 (7.54)	52-78	83.61 (17.66)	59-113
Reading Comprehension								
RC scores - grade 2	-13.94 (10.82)	-36 - 6	-1.94 (8.59)	-21 - 15	7.07 (10.44)	-19- 24	34.81 (10.46)	18-60
RC scores - grade 3	4.00 (14.14)	-11 - 32	13.31 (10.36)	0–32	21.23 (10.51)	2-46	41.83 (15.71)	13-68

on the skewness the data). The second research question, whether the types of teacher talk were related to the four types of students, was answered using multilevel analyses. In line with the structure of the data, students nested in teachers, multilevel multiple regression analyses were conducted, using MLwiN software (Rasbash, Browne, Healy, Cameron, & Charlton, 2012).

In these multilevel analyses, the amount of a certain type of teacher talk was regressed on the four types of students, while controlling for the following three covariates: type of class (multi- or single-grade), subject domain (mathematics or reading comprehension), and year group (grade 2 or 3). The normality of the residual distributions at student level was checked. As the normality of the residual distributions might be considered debatable, the dependent variables were transformed into normal scores and additional multilevel analyses were fitted using these normal scores. It was found that the predicted values in the multilevel models were robust against the normal scores transformations. For interpretative reasons the multilevel models presented in the Results section contain the original data; for information on the normal scores analyses, the interested reader is referred to the corresponding author. In this article, the results of the nonparametric tests, correlations, and multilevel analyses are described and interpreted using a critical value of $p = .05$.

RESULTS

This section is structured as follows. First, the preconditions for differentiation and classroom organization are discussed by taking into account all teacher activities during the complete lessons. This means that teacher activities could be directed to any of the students present in the observed class. After this general overview of teachers' differentiation practices, we focus on the observed teacher behavior towards the four selected students in each class. Such a closer examination provides us with information on how different types of students were actually addressed during the lesson.

CONTENT-RELATEDNESS OF TEACHER TALK

In total, 1,865 teacher utterances were scored during all observed lessons. The mean percentages of teacher talk that was related to task, content, or organization are presented in Table 4. Of all observed teacher talk, 44% was content-related, and 41% was organizational. The latter means that teachers spoke to their students in class in an organizational way, did not address any of the students in the observed grade (for instance, in multi-grade classes where the observed grade was doing seatwork, while the teacher instructed the non-observed grade), or only addressed them by making empty remarks, like "Good job, Sarah". The columns under SD show the differences among teachers' scores. Since the standard deviation is rather large, it can be concluded that there was substantial variation between teachers: some teachers mainly focused on content while others referred to organizational matters more often.

Nonparametric significance testing showed that teachers used significantly more content-related teacher talk in the RC lessons ($Mdn = 49.19$) than in the math lessons ($Mdn = 38.46$, $Z = -2.44$, $p = .015$). Teachers further used significantly more organizational teacher talk during their math lessons than in the RC lessons ($Mdn_{Math} = 46.43$, $Mdn_{RC} = 33.37$, $Z = -2.75$, $p = .006$). Regarding the type of class, no significant differences were found between teachers' reference to content in single-grade classes and multi-grade classes ($Mdn_{single} = 46.67$, $Mdn_{multi} = 41.66$, $U = 483.5$, $Z = -1.36$, $p = .18$), whereas teachers used significantly more organizational talk in multi-grade classes than in single-grade classes ($Mdn_{multi} = 46.55$, $Mdn_{single} = 33.33$, $U = 429.5$, $Z = -1.98$, $p = .047$).

The heterogeneity of the class also was found to play a role in teachers' content-related and organizational talk. Teachers' talk focused more on content ($\rho = .39$, $p < .05$) and

Table 4
Mean Percentages of Type of Teacher Talk for the two Subject Domains and Class Types

		Task-related (1)		Content-related (2A+B)		Organizational (3+4)	
	n	*M*	*(SD)*	M	*(SD)*	M	*(SD)*
Mathematics	41	13.8	10.2	39.8	16.9	46.4	18.2
Reading comprehension	32	15.7	10.6	49.8	17.0	34.4	17.6
Single-grade	25	16.9	10.2	47.7	14.2	35.3	14.1
Multi-grade	48	13.6	10.3	42.4	19.0	44.2	20.4
Total	73	14.63	10.3	44.2	17.6	41.2	18.8

less on organization ($\rho = -.32$, $p < .05$) in classes with high heterogeneity than in classes that were more homogeneous.

TEACHER POSITION: WALKING THE ROUNDS

Our findings on the amount of walking around, a precondition for well-aligned differentiation, showed that, on average, the teachers walked around for 19% of the time. Again, there were large differences between teachers: some did not walk around at all (n=9) and some did so for over 40% of the time (n=5). When we compared subject domains, we saw that teachers on average walked around 17% of the time during reading comprehension (*Mdn* = 16.33) and 21% of the time during mathematics lessons (*Mdn* = 19.36). This difference was not significant ($Z = -1.86$, $p = .063$). Although the median time spent on walking around was 23% in single-grade classes and 14% in multi-grade classes, this difference was also not significant ($U = 444$, $Z = -1.82$, $p = .067$). Heterogeneity of the class was not significantly related to the percentage of time spent on making rounds either ($\rho = -.16$, $p > .05$).

CLASSROOM ORGANIZATION

Having investigated the preconditions for differentiation, we examined classroom organization, specifically the use of extended instruction. The mean percentages of the lesson spent on whole-class instruction, small-group instruction, and seatwork are presented in Table 5. On average, teachers spent almost 60% of the lesson on whole-class teaching and 10% on providing extended instruction to a small group of students; for 30% of the lesson, all students worked alone on math/reading tasks. The structures of the mathematics and reading comprehension lessons differed. In RC lessons, significantly more whole-class teaching took place than in math lessons ($Mdn_{RC} = 67.20$, $Mdn_{Math} = 50$, $Z = -3.85$, $p < .001$). The mean percentages in Table 4 show that teachers spent on average more time on extended instruction during the math lessons (13%) than during the RC lessons (6%). This difference was significant ($Mdn_{Math} = 0$, $Mdn_{RC} = 0$, $Z = -2.07$, $p = .039$). There is also significantly more time spent on seatwork in the math lessons compared to the RC lessons ($Mdn_{Math} = 38.89$, $Mdn_{RC} = 22.83$, $Z = -2.53$, $p = .011$). Single- and multi-grade classes did not differ significantly regarding the amount of

Table 5
Classroom Organization: Lesson Phases (% of Lesson Time)

		Whole-class instruction		Extended instruction		Seatwork	
	n	*M*	*SD*	M	*SD*	M	*SD*
Mathematics	41	49.9	16.0	12.9	17.4	37.2	19.5
Reading comprehension	32	70.1	16.4	5.8[a]	12.4	24.2	15.7
Single-grade	25	61.1	15.9	13.2	15.9	25.6	15.2
Multi-grade	48	57.5	20.5	8.0[a]	15.5	34.5	20.1
Total	73	58.7	19.0	9.8	15.7	31.5	18.9

[a] Due to the skewed distribution of the data, SDs were found that were larger than the mean scores. The alternative central tendency, *Mdn*, equaled zero, which would have provided little insight in the observed occurrences of extended instruction in the math and RC lessons.

time spent on whole-class teaching and extended instruction. However, they did differ significantly in the amount of time spent on seatwork: Students in multi-grade classes were left to work on their own for longer than in single-grade classes (Mdn_{multi}= 36.71, Mdn_{single} = 23.08, U = 422, Z = -2.07, p = .038).

Correlations between classroom organization and heterogeneity show that there was no significant relation between the amount of extended instruction and heterogeneity, but the amount of whole-class teaching and seatwork was significantly related to heterogeneity. In more heterogeneous classes, the percentage of time spent on whole-class teaching was higher than in more homogeneous classes (ρ = .24, p < .05), whereas more time was spent on seatwork in more homogeneous classes (ρ = -.39, p < .05).

A more detailed look at the occurrence of extended instruction shows that teachers provided extended instruction in 34% of all observed lessons (Table 6). In these lessons, extended instruction lasted on average almost 30% of the lesson, but again, teachers differed strongly in the amount of time spent on this instructional form. Regarding the two subject domains, a significant difference was found: teachers provided extended instruction to small groups of students more often in math lessons than in reading comprehension lessons (χ^2= 3.87, df = 1, p = .049). Yet in those lessons in which teachers provided extended instruction, the mean time spent on it was comparable for the math and reading lessons (t = .48, df = 23, p = .64).

Comparing the provision of extended instruction in single- and multi-grade classes, we saw that it was provided in 13 multi-grade classes and in 12 single-grade classes. This difference was not significant (χ^2 = 3.19, df = 1, p = .074). In those classes in which extended instruction was provided, the mean percentage of time spent on extended instruction did not differ significantly for multi- and single-grade classes either (t = -.37, df = 23, p = .72). As regards the heterogeneity of the class, it was found that, if teachers provided extended instruction, then the teachers tended to give such additional instruction more often in more heterogeneous classes (ρ = .29, p < .05). The percentage of lesson time spent on it did not relate to class heterogeneity.

TEACHER BEHAVIOR TOWARDS THE FOUR SELECTED STUDENTS

Knowing how teachers organize their classroom only provides a general picture of how teachers create (differentiated) learning opportunities. A closer look at the kinds of activities different types of students in class are engaged in and how teachers address these students gives a more complete overview of what happened with whom.

When specifically investigating the activities that students of relative different performance levels were engaged in, we saw a slight tendency that, the weaker the students were the more extended instruction they received and the less they worked on their own during seat work (ρ = -.25, p < .05 and ρ = .16, p < .05 respectively). In Table 7 (see next page), the average percentages of instructional activities that students

Table 6
Number of Lessons with Extended Instruction

	No. of lessons	No. of lessons containing extended instruction	Extended instruction[a] (% of the lesson)	
	n	*n*	*M (SD)*	Range
Mathematics	41	18	29.4 (14.1)	11.5-60.7
Reading comprehension	32	7	26.5 (12.7)	9.7-45.2
Single-grade	25	12	27.5 (10.9)	9.7-44.0
Multi-grade	48	13	29.6 (16.0)	11.5-60.7
Total	73	25	28.6 (13.5)	9.7-60.7

[a] Percentages based on lessons that contained extended instruction

Table 7
Setting of the Selected Students

		Whole-class instruction		Extended instruction		Individual guidance		Seatwork	
	N[a]	%	*SD*	%	*SD*	%	*SD*	%	*SD*
Very weak	70	57.3	19.5	7.6	13.4	1.4	7.3	31.9	17.9
Weak	71	58.2	18.9	5.4	10.2	0.4	2.1	35.0	18.8
Average	71	57.7	19.0	3.5	11.0	0.3	1.3	37.7	19.1
Advanced	72	55.6	21.2	1.2	4.6	0.5	2.3	41.0	22.1

Note. As the category "other" is not presented in Table 5, the sum of the percentages per type of student may deviate slightly from 100. This category was omitted as it was hardly observed.

[a] As not all classes contained at least 4 students, the number of classes in which a certain type of student was available differed slightly.

of different performance levels were engaged in, are provided. When comparing these activities, no significant differences for the four types of students were found regarding the amount of whole-class teaching, individual guidance, and seatwork they were engaged in. As can be derived from Table 6, the mean percentages of time the students were engaged in extended instruction seemed to vary, raising from 1% of the lesson time for the advanced students to 7% for the very weak students. Formal testing, using the Kruskal Wallis test, showed that there were indeed significant differences between the four types of students in terms of the percentage of time they were engaged in extended instruction ($\chi^2 = 18.23$, $df = 3$, $p = .000$). For an explorative interpretation of these differences, post hoc contrast testing was conducted. Contrasts were tested by carrying out six Mann-Whitney U tests using a critical value of $p = .008$ in order to take into account the problem of multiple comparisons. These tests showed that the percentage of time being engaged in extended instruction indeed seemed to vary for the four types of students: The advanced students received significantly less extended instruction than the weak and very weak students ($p = .001$ and $p = .000$, respectively).

The amount of teacher talk directed to the four selected students is presented in Figure 1. Teachers differed significantly regarding the amount of talk they directed to the different types of students ($\chi^2 = 12.82$, $df = 3$, $p = .005$). Again, six post hoc contrast tests - Mann Whitney U tests using a critical value of p = .008 - were conducted to determine in an explorative way which groups of students received higher or lower amounts of teacher talk. Of the selected students, the very weak students were individually addressed significantly more than the average and advanced students ($p = .003$ and $p = .003$ respectively). Apparently, very weak students received more attention from the teacher than students of higher ability, whether during whole-class teaching, small-group instruction, or individual guidance.

Ultimately, we sought to determine whether the level of student performance predicted the type of teacher attention during the whole lesson. For this purpose, we conducted a multilevel regression analysis, taking into account the variance at two

Figure 1. *Selected Students, Individually Addressed by the Teachers (%)*

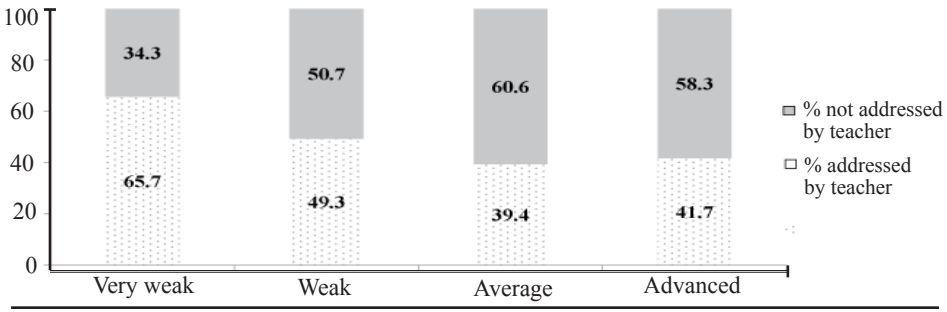

Table 8
Multilevel Analyses Predicting Three Types of Teacher Talk

	Task				Content				Organization			
	Model A1		Model A2		Model B1		Model B2		Model C1		Model C2	
	β	*SE*	β	*SE*	β	*SE*	β	*SE*	*β*	*SE*	*β*	*SE*
Response: Teacher talk												
Fixed part												
Intercept	.09	.04	.18	.05	.28	.13	.59	.16	.25	.07	.28	.08
Grade	.05	.04	.05	.04	.24	.12	.25*	.12	-.02	.06	-.02	.06
Multi-grade	-.04	.04	-.04	.04	.40*	.12	.40*	.13	.02	.06	-.02	.06
Subject domain	-.02	.04	-.02	.04	-.11	.12	-.11	.12	-.16*	.06	-.16*	.06
Performance level (very weak=reference)												
Weak			-.12*	.06			-.38*	.15			-.07	.08
Average			-.15*	.06			-.44*	.15			-.02	.08
Advanced			-.10	.06			-.46*	.15			-.02	.08
Random part												
Class	.00	.00	.00	.00	.05	.05	.06	.05	.01	.02	.02	.01
Students	.11	.01	.11	.01	.81	.08	.76	.07	.20	.02	.20	.02
-2*log		175.09		167.24		759.93		747.53		368.55		367.64
No of students		284		284		284		284		284		284

*$p<.05$

levels – teachers and the selected students. Since we were interested in whether teachers addressed students with varying performance levels differently during the lessons, we took all selected students into account.

Six models are presented in Table 8. Models A1 and A2 related to whether the student performance level predicted the amount of task-related teacher talk.[5] In model A1, the covariates multi-grade class, subject, and year group were added, but did not significantly predict task-related teacher talk. In Model A2, the student performance levels were added. Deviance testing to assess model fit showed that Model A2 fitted the data significantly better than model A1 ($p = .049$; using a chi-square distribution with $df = 3$, the outcome of the deviance test (7.85) was slightly higher than the critical value (7.81) for $p = .05$). It can be concluded that the weak and average students received significantly less task-related teacher talk than the very weak students, but this was not the case for the advanced students.

Model B1 and B2 focused on whether student performance levels predicted content-related teacher talk. Model B1 presents a model containing the three covariates. In this model it was found that the selected students in the multi-grade classes received significantly more content-related teacher utterances. In model B2, the performance levels were added to the model. The asterisked negative regression coefficients for the weak, average, and advanced students indicate that these types of students received significantly less content-related teacher talk than the very weak students. Inclusion of the performance levels significantly increased the fit of the model ($p = .006$; the deviance decreased with 12.41, which is higher than the critical level for $p = .05$ [7.81] in a chi-square distribution with $df = 3$). In models C1 and C2 the organizational teacher talk is predicted. Apart from subject domain, none of the variables predicted organizational talk, leading to the conclusion that the amount of organizational talk was not associated with the performance levels of students.

[5] The types of teacher talk represent the full amount of *teacher talk* - either related to task, content, or organization - directed to the four types of selected students during the math or reading comprehension lesson.

CONCLUSION AND DISCUSSION

In this study, the differentiation practices of 43 second- and third-grade teachers were explored using low-inference observations. The main aim was to explore how teachers in daily practice adapt their teaching to students of different performance levels during mathematics and reading comprehension lessons, by looking at preconditions for differentiation, classroom organization, and the ways teachers dealt with students of relative different performance levels. The findings of this study showed the importance of taking context factors into account, since differentiation practices can be influenced by subject domain, heterogeneity of the class, and the single- or multi-gradedness of the class.

Two central questions were addressed. First, we investigated to what extent the use of differentiation practices was related to the three context factors. This question was answered by examining teacher talk, walking around, and classroom organization, including the provision of extended instruction. We found that teachers referred more to content in reading comprehension lessons, whereas they used more organizational talk in math lessons. There was more organizational talk in multi-grade classes than in single-grade classes, and there seemed to be more content-related talk in more heterogeneous than in more homogeneous classes.

Regarding teachers' making the rounds in order to formatively assess students' performance, the three context factors did not play a significant role. However, the context factors were related to classroom organization, that is, the proportion of time spent on whole-class teaching, extended instruction, and seatwork. Teachers used more whole-class teaching during the reading lessons, whereas more time was spent on seatwork and providing extended instruction in the math lessons. In multi-grade classes and in more homogeneous classes the amount of time spent on seatwork by all students was also higher. In heterogeneous classes teachers tended to provide more whole-class instruction. When we specifically looked at whether teachers provided small-group instruction, we found that they gave extended instruction more often during the math lessons and in more heterogeneous classes. Overall, these findings provided evidence for the assumption that context factors play a role in the way differentiation practices are used. As a result, we recommend that they should not be neglected when investigating differentiation practices in natural settings.

The second question referred to the ways teachers dealt with students of four relative different performance levels (very weak, weak, average, and advanced). It was found that, when teachers provided small-group instruction, it was hardly directed to the relatively advanced students, but mostly to the relatively (very) weak performers. Teachers also individually addressed the relatively very weak students most; most content-related teacher talk was directed to these very weak students.

The results confirm two of the three hypotheses that were formulated. First, we found a difference between lessons depending on subject domain. Compared with reading comprehension lessons, less whole-class teaching, more extended instruction, and more seatwork was provided during math lessons, confirming our first hypothesis. These findings are in line with Nurmi (2012), who stated that there is more variation in instruction in mathematics lessons. Nevertheless, extended instruction was relatively scarce in both math and reading lessons: it was provided in a quarter of the observed reading comprehension and half of the observed math lessons. This means that in quite a lot of classes, weak students were not given additional support in order to reduce their arrears; nor did advanced students receive additional support that stimulated and challenged them. The provision of additional instruction by creating small homogeneous groups was not related to the type of class: extended instruction was provided in a comparable manner in single- and multi-grade classes. Findings on "making the rounds", a precondition for differentiation, also did not show differences between multi- and single-grade classes: teachers spent an equal amount of time making the rounds

in both types of classes. This means that our hypothesis that differentiation would be used more often in single-grade classes was not confirmed. Finally, concerning the third hypothesis that the weakest students would receive the most attention, we found that relatively very weak students were addressed most often, mainly using content-related questioning or explanations. The relatively advanced students were not targeted by additional teacher guidance and teachers did not shorten the length of whole-class teaching for these students. These findings are in line with convergent differentiation, where minimum learning objectives are supposed to be attainable for all students, if necessary with additional teacher support. The third hypothesis, that the weak students would receive most attention, was thus confirmed. This leads to the conclusion that during mathematics and reading comprehension lessons there seems to be room for improvement in supporting and challenging all students, especially the advanced ones.

LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

The current findings not only provide general information on how teachers adapted their teaching to different student needs, for example, through their classroom organization, but they also show how different types of students in class are approached. This student perspective adds to the general picture in that it revealed in what ways students were treated differently. The addition of a qualitative aspect, like the type of teacher talk, gives low-inference information about the task-focus in class, also with respect to different types of students. However, a limitation is that some qualitative aspects remain unclear, like the quality of the extended instruction provided. As mentioned by Lou et al. (1996), small-group instruction can be interpreted in a loose sense, meaning that students are merely physically placed in a small group. It can also be interpreted in a strict sense, where teachers' instructional behavior is adapted to the needs of the students in the small group, for instance, by using different instructional strategies or specific materials. More qualitative information on what teachers actually did while they provided extended instruction in small groups would have provided essential information on the quality of the differentiation practices.

Another issue that needs further investigation is the large amount of seatwork during the mathematics lessons. During this seatwork, students were left to work on their own without help from the teacher for quite some time. It would be interesting to investigate whether and how the time spent on individual seatwork related to student achievement and to study whether its effectiveness differed for groups of students. Linking observed differentiation practices to student outcomes would increase our understanding of the effectiveness of these practices. However, as mentioned above, such relations would gain in value when both general teaching quality and the quality of the differentiation practices are taken into account.

A second limitation is that this study only focused on how teachers used differentiation practices in a natural setting. However, a sole focus on the teacher only covered part of differentiation and its effects. Differentiation is assumed to help improve students' learning. Taking an interactive perspective and taking students' (re)actions into account would have led to a better, more detailed understanding of the effects of the differentiation activities performed by the teacher.

A third limitation that should be brought forward is the selection procedure that was used in this study, in which four types of student – very weak, weak, average and advanced – were selected relative to the performance levels of their classmates. The researchers purposefully used this selection procedure, as the main aim of the study was to explore teachers' differentiation practices in daily classroom practice, meaning that teachers base their instructional decisions on the (levels of performance of their) students in their class.

The relative selection procedure resulted in the expected upward pattern in average initial performance levels. However, it also led to considerable overlap in initial performance levels across the four types of students. The findings should therefore be

interpreted cautiously, as no clear connection can be made to the absolute performance levels of the four types of students. Basing the selection process on absolute performance levels would solve this problem and would make the interpretation of results more clear and straightforward. However, the sole use of absolute criteria might disregard relevant information used for instructional decision making by teachers who have to deal with small or rather homogeneous classes. Future studies in which both a relative and an absolute selection procedure are taken into account would probably lead to a more comprehensive picture on how teachers deal with students of different performance levels.

The main aim of this study was to investigate teacher variability in the use of differentiation practices. The findings show that, when investigating differentiation practices, it is important not only to consider teachers' general differentiation behavior, but also to take into account more detailed behavior towards students of different performance levels. Further, it can be concluded that it is not possible to generalize over subject domains, the single- or multi-gradedness of the class, or the heterogeneity of the class when investigating differentiation practices. In order to examine how teachers really differentiate between students in their lessons, these context factors should be taken into account. ■

REFERENCES

Bosker, R. J. (2005). *De grenzen van gedifferentieerd onderwijs. [The borders of differentiated education]*. Groningen: GION.

Burns, R. B., & Mason, D. A. (2002). Class composition and student achievement in elementary schools. *American Educational Research Journal, 39(1)*, 207-233.

Condron, D. J. (2008). An Early Start: Skill grouping and unequal reading gains in the elementary years. *Sociological Quarterly, 49(2)*, 363-394.

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., et al. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher, 38(2)*, 85-99.

Corno, L. (2008). On teaching adaptively. *Educational Psychologist, 43(3)*, 161-173.

Feenstra, H., Kleintjes, F., Kamphuis, F., & Krom, R. (2010). *Wetenschappelijke verantwoording begrijpend lezen groep 3 t/m 6. [Scientific account for the reading comprehension tests grade 1 to 4]*. Arnhem, the Netherlands: CITO.

Good, T. L., & Brophy, J. E. (2003). *Looking in classrooms.* (9th edition). Boston: Pearson/Allyn and Bacon.

Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences- in-differences evidence across countries. *The Economic Journal, 116 (510)*, C63-C76.

Houtveen, A. A. M. (2002). *Effecten van adaptief onderwijs. Evaluatie van het landelijk project schoolverbetering. [Effects of adaptive teaching. Evaluation of the LPS-project]*. Alphen aan de Rijn, The Netherlands: Samson.

Houtveen, A. A. M., van de Grift, & Creemers, B. P. M. (2004). Effective school improvement in mathematics. *School Effectiveness and School Improvement, 15(3-4)*, 337-376.

Huang, M. (2009). Classroom homogeneity and the distribution of student math performance: A country-level fixed-effects analysis. *Social Science Research, 38(4)*, 781-791.

Educational Inspectorate (2008a). *Basisvaardigheden rekenen-wiskunde in het basisonderwijs. Een onderzoek naar het niveau van rekenen-wiskunde in het basisonderwijs en naar verschillen tussen scholen met lage, gemiddelde en goede reken-wiskunderesultaten. [Mathematical skills in primary education. A study on the mathematics level in primary education and on differences between schools having low, average and good mathematical results]*. Utrecht: Educational Inspectorate.

Educational Inspectorate (2008b). *De staat van het onderwijs. Onderwijsverslag 2006/2007. [The state of education. Educational year report 2006/2007]*. Utrecht: Educational Inspectorate.

Educational Inspectorate (2010). *Opbrengstgericht werken in het basisonderwijs. Een onderzoek naar opbrengstgericht werken bij rekenen-wiskunde in het basisonderwijs. [Data-driven decision making in primary education. A study on data-driven decision making in mathematics education]*. Utrecht: Educational Inspectorate.

Educational Inspectorate (2012). *De staat van het onderwijs. Onderwijsverslag 2010/2011. [The state of education. Educational year report 2010/2011]*. Utrecht: Inspectie van het Onderwijs.

Educational Inspectorate (2013a). *De staat van het onderwijs. Onderwijsverslag 2011-2012. [The state of*

Dutch education. Educational year report 2011-2012]. Utrecht: Educational Inspectorate.
Educational Inspectorate (2013b). *De kwaliteit van basisscholen en bestuurlijk handelen in het noorden van Nederland. Ontwikkelingen in de periode 2008-2012. [The quality of primary schools and school board activities in the northern part of the Netherlands. Developments in the period 2008-2012]*. Utrecht: Educational Inspectorate.
Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS rekenen-wiskunde voor groep 3 tot en met 8. [Scientific account of the mathematics tests for grades 1 to 6]*. Arnhem: Cito.
Jurik, V., Gröschner, A., & Seidel, T. (2013). How student characteristics affect girls' and boys' verbal engagement in physics instruction. *Learning and Instruction, 23(0)*, 33-42.
Kooiman, M. C., Hofman, R. H., Doolaard, S., & Guldemond, H. (2005). *Adaptief onderwijs in scholen voor speciaal basisonderwijs. [Adaptive teaching in schools for special education]*. Groningen: GION.
Kuiper, R. J., & Palma, M. G. (2012). *Aanwijzingen voor differentiatie in de handleiding van de lesmethode: Een inhoudsanalyse en vergelijking voor rekenen en begrijpend lezen. [Suggestions for differentiation in textbook manuals: a content analysis and comparison of mathematics and comprehensive reading textbooks]*. Groningen: Master-thesis University of Groningen.
Leonard, J. (2001). How group composition influenced the achievement of sixth-grade mathematics students. *Mathematical Thinking and Learning, 3(2-3)*, 175-200.
Little, A. (2004). *Learning and teaching in multi-grade settings.* Background paper prepared for the Education for All Global Monitoring Report 2005 The Quality Imperative. Paris, France: UNESCO.
Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66(4)*, 423-458.
MacIntyre, H., & Ireson, J. (2002). Within-class ability grouping: Placement of pupils in groups and self-concept. *British Educational Research Journal, 28(2)*, 249-263.
Mason, D. A., & Burns, R. B. (1997). Reassessing the effects of combination classes. *Educational Research and Evaluation, 3(1)*, 1-53.
Moon, T. R. (2005). The role of assessment in differentiation. *Theory into Practice, 44(3)*, 226-233.
Muijs, D., & Reynolds, D. (2011). *Effective teaching: Evidence and practice.* London: SAGE.
Mulryan-Kyne, C. (2007). The preparation of teachers for multi-grade teaching. *Teaching and Teacher Education, 23(4)*, 501-514.
Nomi, T. (2009). The effects of within-class ability grouping on academic achievement in early elementary years. *Journal of Research on Educational Effectiveness, 3(1)*, 56-92.
Nurmi, J., Viljaranta, J., Tolvanen, A., & Aunola, K. (2012). Teachers adapt their instruction according to students' academic performance. *Educational Psychology, 32(5)*, 571-588.
Pierce, R. L., Cassady, J. C., Adams, C. M., Neumeister, K. L. S., Dixon, F. A., & Cross, T. L. (2011). The effects of clustering and curriculum on the development of gifted learners' math achievement. *Journal for the Education of the Gifted, 34(4)*, 569-594.
Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2012). *MLwiN version 2.23*. Bristol, England: Centre for Multilevel Modeling.
Reezigt, G. J., Houtveen, A. A. M., & Van de Grift, W. (2002). *Ontwikkelingen in en effecten van adaptief onderwijs in de klas en integrale leerlingenzorg op schoolniveau. [Developments in and effects of adaptive teaching in classroom and integral student care at school level]*. Groningen: GION.
Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal, 48(2)*, 462-501.
Sammons, P., Taggart, B., Sylva, K., Melhuish, E., Siray-Blatchford, I., Barreau, S., et al. (2005). *Variations in teacher and pupil behaviors in year 5 classes.* No. EPPE 3-11. London: University of Education.
Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best evidence synthesis. *Review of Educational Research, 57*, 293.
Tach, L. M., & Farkas, G. (2006). Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. *Social Science Research, 35(4)*, 1048-1079.
Thomas, J. L. (2012). Combination classes and educational achievement. *Economics of Education Review, 31(6)*, 1058-1066.
Tieso, C. (2005). The effects of grouping practices and curricular adjustments on achievement. *Journal for the Education of the Gifted, 29(1)*, 60-89.
Tomlinson, C. A. (2005). Grading and differentiation: Paradox or good practice? *Theory into Practice, 44(3)*, 262-269.
van Berkel, S., Krom, R., Heesters, K., Van der Schoot, F., & Hemker, B. (2007). *Balans van het leesonderwijs halverwege de basisschool 4. Uitkomsten van de vierde peiling in 2005. [Overview of reading education midway primary education 4. Results of the fourth sounding in 2005]*. Arnhem: Cito.
VanTassel-Baska, J., Zuo, L., Avery, L. D., & Little, C. A. (2002). A curriculum study of gifted-student

learning in the language arts. *Gifted Child Quarterly, 46(1),* 30-44.
Vaughn, S., Linan-Thompson, S., Kouzekanani, K., Pedrotty Bryant, D., Dickson, S., & Blozis, S. A. (2003). Reading instruction grouping for students with reading difficulties. *Remedial and Special Education, 24(5)*, 301-315.
Veenman, S. (1997). Combination classrooms revisited. *Educational Research and Evaluation, 3(3)*, 262-276.
Vogt, F., & Rogalla, M. (2009). Developing adaptive teaching competency through coaching. *Teaching and Teacher Education, 25(8)*, 1051-1060.
Whitburn, J. (2001). Effective classroom organisation in primary schools: Mathematics. *Oxford Review of Education, 27*, 411.
Wiley, C. H., Good, T. L., & McCaslin, M. (2008). Comprehensive school reform instructional practices throughout a school year: The role of subject matter, grade level, and time of year. *Teachers College Record, 110(11)*, 2361-2388.
Wilkinson, I. A. G., & Hamilton, R. J. (2003). Learning to read in composite (multi-grade) classes in New Zealand: Teachers make the difference. *Teaching and Teacher Education*, *19*(*2*), 221-235.

Correspondence regarding this article should be directed to Evelien S. Ritzema from the University of Groningen. Email may be sent to e.s.ritzema@rug.nl

APPENDIX A

Table A1

Overview of the Observed Teachers, Classes and Number of Observed Students

School	Teacher	Class	Math				Reading			
			Grade	Multi-grade	No of students present in class	Observed students	Grade	Multi-grade	No of students present in class	Observed students
School 1	1A	1-1	2	x	22	10	2	x	22	10
	1B	1-1	2	x	22	10	2	x	24	11
	1C	1-2	3	x	17	9	3	x	17	9
School 2	2A	2-1	3		18	18	3		18	18
	2B	2-1	3		18	18	3		18	18
School 3	3A	3-1	2		22	22	2		22	22
	3B	3-2	3		26	26	3		26	26
School 4	4A	4-1	2	x	24	15	2	x	24	15
	4B	4-1	2	x	24	15	2	x	24	15
	4C	4-2	3	x	14	9	3	x	14	9
School 5	5A	5-1	3	x	10	2	3	x	10	2
School 6	6A	6-1	2		23	23	2		25	25
	6B	6-1	2		23	23				
	6C	6-2	2		25	25	2		25	25
	6D	6-3	2	x	16	13	3	x	16	3
	6E	6-4	3	x	26	11	3	x	26	11
School 7	7A	7-1	2	x	19	11	2	x	19	11
	7B	7-2	3	x	15	6	3	x	15	6
School 8	8A	8-1	3		20	20	3		20	20
	8B	8-2	2		33	33	2		33	33
	8C	8-2	2		33	33				
School 9	9A	9-1	2	x	20	9	2	x	20	9
School 10	10A	10-1	2	x	21	8	2	x	21	8
	10B	10-1	2	x	24	8				
	10C	10-2	2		26	26				
	10D	10-3	3		26	26	3		26	26
School 11	11A	11-1	3	x	23	5	3	x	23	5
	11B	11-2	3	x	20	5	2	x	20	15
School 12	12A	12-1	3	x	30	22	3	x	31	23
	12B	12-2	3	x	26	18				
School 13	13A	13-1	2	x	22	8	2	x	22	6
School 14	14A	14-1	3		31	31	3		31	31
	14B	14-2	2	x	28	8				
	14C	14-3	3	x	24	6				
	14D	14-4					2		27	27
	14E	14-5					2		28	28
School 15	15A	15-1	2	x	12	5	2	x	8	5
School 16	16A	16-1	2	x	18	10				
	16B	16-1	2	x	18	10				
	16C	16-2	3	x	9	6				
School 17	17A	17-1	2	x	8	3				
School 18	18A	18-1	2	x	21	12	2	x	21	12
	18B	18-2	3	x	14	9	3	x	14	9
Total					871	587			690	493

Note. The numbers of (observed) students used in this study are based on the actual presence of students in class during the observation. For organizational reasons, the observations for math and reading were regularly conducted on different days. Moreover, part time teachers teach the same class on different weekdays. Hence, the number of students mentioned for both subject domains and for the two part time teachers can slightly differ: Some students were absent during the observations, for instance due to illness.

APPENDIX B

Table B1

Time-Sampling Scheme used during the Observed Lessons

Class	Teacher				Student Activity					
Time	***Phase***	***Position***	***Teacher talk (TT)***	***Student addressed during TT***	***Very Weak***	***Weak***	***Average***	***Advanced***	***Elaboration***
0									
2									
4									
6									
8									
10									
12									
14									
16									
18									
20									
22									
24									
26									
28									
30									
32									
34									
36									
38									
40									
42									
..									
60									