MASTER THESIS

MEASURING DIFFERENTIATED INSTRUCTION FROM DIFFERENT PERSPECTIVES

Nienke Bach Kolling BSc

FACULTY OF BEHAVIOURAL, MANAGEMENT AND SOCIAL SCIENCES

ELAN TEACHER DEPARTMENT OF TEACHER DEVELOPMENT

EXAMINATION COMMITTEE

Dr. M. van Geel

Prof. Dr. A.J. Visscher

JANUARY 2020

Table of contents

Abstract

The process of adapting education to student differences is called differentiation. Previous research shows that students are more involved, more motivated and better performing when teachers differentiate their education. Therefore the Dutch Ministry of Education and the primary education counsel have made differentiation skills a basic requirement for all teachers in 2020. As a result, it is important that instruments are developed to assess teachers' differentiation skills. There are three common methods for measuring differentiation: classroom observations by external observers, teacher self-assessments and student surveys. This research examined the degree of agreement between these three rater-groups on how they assess teachers' differentiation behaviour. To compare the different perspectives, an instrument was developed that includes similar items for all three rater-groups. The participants were nine primary school teachers (grade 8 to 10), their students (n = 171) and an external observer. They assessed the degree of differentiation during a math instruction for students in three instruction groups on 12 items. Cohen's kappa was used to determine the degree of agreement between the different raters. Moderate to almost perfect agreement was found on 10 out of 96 items. The observer and the teachers agreed on six items. Teachers and students, and observer and students, agreed on only two items. Overall, the scores of the various rater-groups show little agreement. There is almost no agreement between rater-groups in the basic instruction group and fair agreement in the enrichment group and intensive group. The results show that when measuring differentiation, it is important to consider which measurement method to use, as the scores will differ depending on the chosen method.

1. Introduction

Schools have been grouping students by age for years, attempting to create homogeneous groups of students that are appropriate for teacher-centred one-size-fits-all teaching. Although part of children's development is indeed related to their age, there are other influencing factors as well. This means that even within classes all sorts of cognitive, socio-economic, socio-emotional and cultural differences between students will exist. These differences influence students' educational needs and teachers wanting to adapt their education to these differences face growing challenges as student differences increase.

In the Netherlands this increase in students' differences is caused by the introduction of laws requiring every child to be taught in regular classrooms. As a result, a large group of children with special needs (e.g. mild intellectual disabilities, learning disabilities and behavioural problems) is now following education in regular classrooms, instead of being referred to special education schools. There has also been an increase of socio-cultural diversity in the Netherlands over the last years, creating more differences in culture, language development and socio-economic status among students. Also, the increased attention for talent development and gifted students urges teachers to not only adapt their education to students in need of intensive support, but to the high-achieving, gifted students as well.

The process of adapting education to student differences is called *differentiation*. The Dutch Inspectorate of Education highly values differentiated teaching and yearly measures the quality of teachers' differentiation skills. For years they have been reporting on teachers struggling to differentiate their education: "it looks like adapting to student differences is too big of a task for part of the teachers, especially in the instruction phase of a lesson" (Inspectie van het Onderwijs, 2016). This is striking, as research shows that students are more involved, more motivated and better performing when teachers master differentiation skills (Inspectie van het Onderwijs, 2012, 2015; Rock, Gregg, Ellis, & Gable, 2008; Tomlinson, 2004). The Dutch Ministry of Education and the primary education counsel (PO-raad) have agreed that differentiation skills are essential for quality education and made possessing these skills a basic requirement for all teachers in 2020 (*Bestuursakkoord voor de sector primair onderwijs*, 2014).

This agreement and the Inspectorate reports recommend the availability of valid instruments for measuring teachers' differentiation skills and emphasizes their importance. This study explored existing differentiation instruments and answered the research question *What are differences and similarities between existing instruments measuring differentiation in terms of the aspects of differentiation they measure?* Based on existing instruments a new instrument has been developed in the second part of this study. This instrument combines the three most used methods to measure differentiation: classroom observation, teacher self-assessment and student surveys. It has been used to answer the research question *What is the degree of agreement between teachers, students and an external observer regarding how they assess teachers' differentiation behaviour?* This could be useful information for the development of future differentiation instruments.

2. Theoretical framework

2.1 Differentiation

A considerable amount of literature has been published on the topic of differentiation. Nevertheless, it seems that there is still no agreement on the definition of differentiation. Most studies agree on the broad definition of differentiation being *the process of adapting education to differences between students*. While this definition focusses on making educational adaptations, Roy, Guay and Valois (2013, p. 1187) add to this: "using systematic procedures for academic progress monitoring and data-based decision making". This suggests that differentiation includes two components: analysing data to inform adaptations (monitoring) and the implementation of these adaptations. Van Geel et al. (2018) made a differentiation skill hierarchy and found four chronological differentiation stages: (1) preparing a lesson period, (2) preparing a lesson, (3) teaching a lesson and (4) evaluating a lesson. They found that teachers use monitoring as well as adaptation skills in each of these stages: when preparing a lesson, teachers use their evaluation of the previous lesson as well as their preparation of the lesson period. Also, during the lesson they continuously monitor the progress and achievement of their students. The adaptations that teachers make for their students during a lesson are planned in their lesson period and lesson preparation and evaluated during and after each lesson. Although monitoring and adapting seem inseparably linked, they are hardly described together in differentiation research, where most research focusses on the adaptation strategies. Research on monitoring can be found in the field of data-based decision making in education. Both components will be further elaborated.

2.1.1 Monitoring

Monitoring includes systematically collecting, analysing and interpreting student data. When teachers use this information to decide which instructional adaptations to apply, these adaptations are more likely to be effective (Inspectie van het Onderwijs, 2016; Roy et al., 2013). The process of data use has three steps (Coburn & Turner, 2011). Data use begins as people *notice* the data or patterns in the data. After noticing the data, it should be *interpreted*: what does these data mean? The third step is the *construction of implications for action*, which will inform what people do in response to the data. Monitoring includes these three steps. Coburn and Turner argue that interpretation is a central part of the data use process, "playing a role in how individuals notice data in the first place, how they make meaning of it, and how they come to understandings about the implications of the data for action" (Coburn & Turner, 2011, p. 177). Interpretations are always influenced by peoples' pre-existing beliefs, experiences and knowledge, making proper data use a complex skill.

Which data teachers collect or use depends on the *goals* they want to achieve with their education, the *target group* they see for differentiation and the *student differences* they see as relevant. Teachers can differentiate either to reach a minimum performance level with all students, or to help every student reach his maximum potential (Deunk, Doolaard, Smale-Jacobse, & Bosker, 2015). Striving for the first is called convergent differentiation and will decrease differences between students (Bosker, 2005). Divergent differentiation on the other hand is likely to maintain differences or increase them. Among the advocates for divergent differentiation is Tomlinson, who argues that every student should have the opportunity to perform at his best (Tomlinson, 2003).

Differentiation is sometimes seen by researchers and teachers as a way of primarily serving special needs or low ability students (Roiha, 2014; Ruys, Defruyt, Rots, & Aelterman, 2013). This view is connected to the concept of convergent differentiation. However, most recent literature stresses the

importance of adapting to the learning needs of all students, implying divergent differentiation (Prast, Van de Weijer-Bergsma, Kroesbergen, & Van Luit, 2015; Roy et al., 2013; Tomlinson, 1999, 2000).

Data on various student differences can be collected, but the relevance of these differences depends partly on the frequency they occur in. In a survey among 773 Dutch primary school teachers, most teachers reported a large degree of cognitive differences (96.4%), differences in learning style (90%) and socio-emotional development (88.9%) between their students (Jettinghoff & Grootscholte, 2014). The same teachers reported a limited degree of socio-economic and cultural differences. Student readiness (the current level of knowledge and skills in a subject area) is a cognitive difference and is frequently used in literature together with students' learning profile and interests (e.g. Prast et al., 2015; Tomlinson et al., 2003). According to Vygotksy (1978), students learn when they undertake activities that are slightly more difficult than what they are capable of independently. Whether activities fall within this so-called zone of proximal development, depends on students' readiness. According to Reezigt (1993), teachers adapt their education mainly to performance differences.

In this study the following definition of differentiation will be used: *differentiation is the process of adapting education to cognitive differences between all students, using systematic procedures for academic progress monitoring and data-based decision making*. This definition emphasizes the importance of adapting for all students, not only special needs or low ability students. It also emphasizes student performance as student difference to which education should be adapted.

2.1.2 Adapting instruction

Teachers can make a wide range of adaptations in their teaching, which makes it infeasible to present an exhaustive list. The Dutch Inspectorate of Education distinguishes four aspects of education that can be adapted: content, instruction, learning tasks and time (Inspectie van het Onderwijs, 2016). Often cited is Tomlinson's division in content (what students learn), process (how they learn it), product (how students demonstrate their learning) and learning environment (e.g. Rock et al., 2008; Roy et al., 2013; Tomlinson, 2001; Tomlinson et al., 2003).

Although there is a lot of research regarding the various ways in which teachers can adapt their instruction, it is still not clear which strategies are actually effective (i.e. lead to better performance of students). The reason for this is that most research on adapting instruction can be grouped into one of three types. The first type of studies develop and evaluate a differentiation instrument. The results of these studies are primarily focused on the quality of the instrument, not on the effectivity of the listed strategies (see for example Gentry & Owen, 2004; Nelson, Demers, & Christ, 2014; Rock et al., 2008).

The second type of research has an experimental design: teachers in the experimental group differentiate their education, generally after having received a training, while teachers in the control group do not. The effects of the intervention – often on student performance or motivation – are measured by means of a pre- and post-test (see for example Chamberlin & Powers, 2010; Kamminga, 2014; Van der Scheer, Glas, & Visscher, 2016). Any measured positive effects can be attributed to the intervention, however, the intervention typically consists of the implementation of multiple adaptation strategies simultaneously. This makes it very difficult to conclude which strategy contributed most to the measured positive effect or which strategy had no effect at all.

The third type of research aims to identify frequently used differentiation strategies among teachers. This is often done by asking teachers to rate the frequency of their use of a number of possible strategies (see for example Graham et al., 2008; Prast et al., 2015; Whipple, 2012). The results will reveal popular and less frequent used strategies. So did Prast et al. (2015) find a relatively low usage

of differentiation for high-achieving students in primary school math lessons (e.g. additional instruction and curriculum compacting), although offering enrichment tasks for high-achieving students ranked among the most used strategies, together with giving additional instruction to low-achieving students. Whipple (2012) concluded that most teachers in her study differentiate in the content they teach and by adapting to students' interest. Graham et al. (2008) compared the use of spelling strategies between weak and strong spellers and found that teachers conference more with weak spellers and their parents, as well as they re-teached them skills and strategies more often. They concluded that 42% of the teachers indicated they made few or no adaptations for weaker spellers. The limitation of studies of this type is that the results only concern the strategies listed in the survey. That this is not very likely to be exhaustive, is shown by Graham et al. (2008) when they asked teachers in a open-ended question to identify additional adaptations not previously mentioned in the survey and received 190 unique adaptations.

Concluding, most studies into differentiation strategies do not investigate effective strategies at all or give a limited view because they investigate a combination of strategies or their own selection of strategies. There is little research that shows which differentiation strategies are effective and so it remains uncertain what high-quality adaptations are. This implicates that differentiation instruments are not yet able to measure the quality of teachers' differentiation skills. An instrument can indicate how often a teacher makes an instructional adaptation, but without research showing that more adaptations lead to better educational quality, teachers' differentiation quality cannot be properly measured.

2.2 Measuring differentiation

There are three common methods for measuring differentiation: classroom observations, teacher self-assessments and student surveys. Each method has its own strengths and weaknesses. These are briefly mentioned here.

2.2.1 Classroom observations

Classroom observations are commonly used to evaluate teachers and for teachers to receive feedback on their performance from principals or peers. Usually an observation form is used for classroom observations. These forms consist of a list of predefined statements that an observer uses to assess teachers during an observation. The advantages of such a form are that it is clear for both teacher and observer where the focus of the observation lies and it avoids teacher self-report (Hill, Charalambous, & Kraft, 2012). Also, when observers are trained in the interpretation of the items, it increases consistent assessment among observers, making the results of an observation more objective (Dobbelaer, 2019). A disadvantage of using classroom observations is that an observation usually offers information on only one or few lessons and an observer may not be able to observe the entirety of the instructional process, e.g. how cognitive differences between students affect instructional decisions of the teacher (Williams et al., 2014). Classroom observations could also be a time-consuming and costly method when using multiple (trained) observers who each observe the same teacher over several lessons, which is needed to obtain a valid score (Dobbelaer, 2019).

2.2.2 Teacher self-assessments

Collecting teacher perceptions through questionnaires is another common method for measuring differentiation. These self-assessments usually include a list of adaptation strategies on which teachers must indicate how often they applied the listed strategies. Using teacher perceptions is a cost-efficient method, as no external raters are needed. It is also a good way to gain insight into

teachers' underlying thought and intentions that inform their decisions (Dobbelaer, 2019; Goe, Bell, & Little, 2008). A disadvantage of using self-assessments is the risk that teachers overestimate their skills and behaviour, for example because they give socially desirable answers or they are unable to recognize their underperformance (Goe et al., 2008). Also, teachers could interpret items differently, leading to less valid results (Dobbelaer, 2019).

2.2.3 Student surveys

Differentiation can also be assessed by students filling in a survey with questions or statements about their teachers' differentiation behaviour. There are several advantages to using student surveys: student perceptions are relatively easy to collect with multiple respondents providing their opinion at the same time, they are formed over longer periods of time, they offer multiple perspectives on differentiation as it is experienced by different students and allow for insight into the way students experience differentiation as they are the target group of teachers' adaptations (Nelson et al., 2014; Van der Scheer et al., 2016). Despite these advantages, Van der Scheer, Glas and Visscher (2016) quote that "student perceptions are rarely used in primary education for collecting information about how (well) teachers teach". The main reason might be that student perceptions can easily be influenced by external factors: teachers can behave differently to different students, students may have different expectations from teachers and students may interpret teacher behaviour differently (Van der Scheer et al., 2016). Also, not all students can fill in a questionnaire (e.g. young children) and students might have different interpretations of items, making the results less valid (Dobbelaer, 2019).

2.3 Research questions

With at least three methods to measure differentiation, the question arises how these methods relate to each other. Therefore, study 1 is an explorative study into the differences and similarities between existing differentiation instruments (e.g. which constructs are measured by the items, how are the items formulated, how can the items be scored and which method is chosen). It answers the first research question: *What are differences and similarities between existing instruments measuring differentiation in terms of the aspects of differentiation they measure?*

Usually, one method is used to assess differentiation behaviour. If there is little agreement between the different measurement methods, the results will depend on the method being used. If there is a lot of agreement, it is possible to choose the most suitable method for each situation, for example a more cost-efficient method (Dobbelaer, 2019). Therefore study 2 aims at answering the second research question: *What is the degree of agreement between teachers, students and an external observer regarding how they assess teachers' differentiation behaviour?* The results of the first study are used to develop a new instrument that can be simultaneously used by teachers, students and observers to assess teachers' differentiation behaviour.

3. Study 1: Comparing differentiation instruments

3.1 Method

3.1.1 Aim

This is an explorative study, aiming to compare existing differentiation instruments to see whether these instruments measure the same aspects of differentiation. The central research question is: *What are differences and similarities between existing instruments measuring differentiation in terms of the aspects of differentiation they measure?*

3.1.2 Procedure

A literature search was conducted in the scientific databases ERIC and Google Scholar to find existing differentiation instruments. The databases were first searched with different combinations of the general keywords *differentiation* or *differentiated instruction* and *instrument* or *measurement*. A second, more specific search also included the keywords *observation*, *survey*, *self-assessment* and *student perception*. The citation lists of relevant studies also have been used to search for other articles.

Relevant studies include an instrument used to measure differentiation. In order to compare instruments, the full instrument or a list of items should be available. Studies focus preferably on primary education, but secondary education is also admissible. Instruments may measure differentiation during a specific subject as well as during teaching in general. Following the definition of differentiation used in this study that differentiation is adapting education to the learning needs of all students, studies that investigated differentiation as a way of only serving a specific group of students (e.g. high or low achieving students) have been excluded. Instruments include preferably only differentiation items, but instruments with only one or more sub scales on differentiation are also admissible.

3.1.3 Data analysis

ATLAS.ti is a program for qualitative data analysis and was used to compare the differentiation instruments. Two types of analyses were carried out using ATLAS.ti. For the first analysis all instruments were grouped based on specific characteristics, such as measuring method (observation form, teacher self-assessment, student survey), level (primary/secondary education), subject and item format. This grouping allowed for a comparison per measurement method.

For the second analysis all items per instrument were coded. In the first coding round, all items were determined as a way of monitoring and collecting data or as a differentiation strategy. Some items did not fit in either category because they are conditional to differentiation, such as 'I know the opportunities for differentiation offered by the curriculum'. Or they included a strategy that does not specifically apply to differentiation, such as 'The teacher gives positive feedback about commitment' or 'Students and the teacher work together to solve problems'. Items specifically aimed at differentiation for weaker or stronger students and subject specific items (math or spelling) were also identified in this coding round.

The more than two hundred items identified as a differentiation strategy were coded in more detail in the second coding round. The items could be distinguished in strategies adapting learning goals, choice, materials, ICT, assessment, instruction, time, questions, pace, assignments, grouping and

adaptations for stronger and weaker students. The items identified as monitoring were in this coding round coded monitoring before, during or after the lesson.

3.2 Results

Appendix A presents an overview of the 17 instruments used for his study. There are seven observation instruments, six self-assessment reports for teachers and five student surveys. Nine of these instruments are developed for use in math lessons. Five instruments include only one or a few subscales on differentiation. Two studies did not include a full outline of their instrument and as a result only part of these items is known, or the exact wording of these items is unknown. First, the examined instruments will be discussed per measurement method (classroom observation, teacher self-assessment, student survey), followed by a comparison of adaptation items and monitoring items from all instruments.

3.2.1 Comparison of instruments per measurement method

Classroom observation instruments

Seven observation instruments have been examined with an average of 16 differentiation items per instrument (SD = 19). All instruments were developed for observations in primary classrooms and are no older than ten years. Two instruments measure teaching quality instead of differentiation, but consider differentiation as a part of teaching quality, and therefore include a subscale on differentiation. Five instruments have been used to observe math lessons, although only three of them actually include math specific items, like 'The teacher makes the connection between the different levels of action'. Five instruments use Likert scales as a way of scoring items. These Likert scales vary from 1-3 (1=ineffective, 3=effective) to 1-5 (1=unsatisfactory, 5=excellent). Another instrument requires that items are scored on either observed or not observed and again another instrument provides the observer per item with a written explanation of the meaning of each score. Three instruments present items supported by good practice examples: the item 'The teacher adapts the practice of the learning content to relevant differences between students' is accompanied by the good practice example 'The teacher differs the complexity of assignments between students'. These examples can be scored on their occurrence during a lesson.

When looking at the subscales of the observation instruments, there is one instrument entirely focused on differentiated instruction, with the subscales Whole group instruction, Extra instruction for struggling students and Instruction for high-achieving students. Another instrument also distinguishes items for high- and low-achieving students. Three instruments have subscales on classroom management. Other frequently used subscales concern differentiation of instruction, practice, content, task and process.

The instrument developed by VanTassel-Baska et al. (2008) is notable because it includes items to observe teachers which are also translated into items to observe students. For example, the teacher item 'The teacher provided opportunities for independent or group learning' matches the student item 'Students worked on projects individually or in pairs/groups'.

Self-assessment instruments

Six instruments for teachers to assess their own differentiation skills have been examined with an average of 32 items per instrument (SD = 16). In five instruments teachers do this by indicating per item how often they do the above, on a 4-point, 5-point or 7-point Likert scale (e.g. 1=never, 5=always). With the sixth instrument, teachers have to indicate per item whether they find the position a strong or a weak point of themselves. With the exception of one instrument, the

instruments are not older than ten years. Five instruments have been used in research into primary school teachers, one instrument has been tested with the help of mathematics teachers in secondary education. This instrument also includes math-specific items, such as 'I'll allow students to solve a sum very differently.' Two other instruments have also been used by teachers to assess their differentiation skills in the math domain, but only one of these instruments also includes math specific items. Of all instruments studied, there is one instrument that was specifically designed for differentiation in spelling education, with items such as 'How do you encourage students to use invented spellings?'

The instrument of Roy, Guay and Valois (2013) has two scales: 'instructional adaptations', with items concerning the adjustments that teachers can do in their classroom and 'academic progress monitoring', with items on the systematic approach to data collection to base the adjustments on. Most instruments make such a distinction. Similar scales to 'academic progress monitoring' are 'identification of educational needs', 'analyse learner' and 'assessment'. Similar to the instructional adaptations scale are 'differentiated practice', 'craft instruction' and 'content/process/product'. Van 't Riet (1995) uses a different approach. In addition to two scales on differentiation behaviour, his instrument also includes a scale with non-differentiation behavioural items that are scored in the opposite direction, such as the item 'I let all pupils work through the same material'. This instrument also has items in both a teacher and pupil variant. For example, a teacher item is 'I discuss the differences between students' solutions with the whole class' and the accompanying student item is 'This teacher discusses the differences between students' solutions with the whole class'.

A different instrument is that of Graham et al. (2008). The items include, in contrast to the other instruments, no examples of differentiation behaviour, but all kinds of (spelling specific) lesson activities, such as repeating material, using games to practice and collaboration. Differentiation behaviour of the teacher must be apparent from the way in which the items are scored: how often does the teacher do this with good and how often with weak spellers. A teacher who differentiates well will make many adjustments for both groups of pupils and not only for the weak spellers. It is noteworthy that there is no third response category for average spellers. This makes it somewhat unclear whether teachers have to omit the average spellers in answering the questions or divide the class into two groups, with the average spellers falling under the 'good spellers'. Because of its different form it is not possible to divide the items of this instrument into the used categories, but because of the interesting way of scoring the instrument is mentioned here.

Three other instruments include items specifically focused on weak or strong student, e.g. 'I set extra challenging goals for high-achieving students'. All studied self-assessment instruments are about the daily practice of teachers, not about a specific lesson. The instrument to be developed in the second study will thus deviate on this point because the results will be compared to a single observation.

Student surveys

Five student surveys have been examined with an average of 12 differentiation items per instrument (SD = 5). What is striking is that only two instruments were actually designed to measure differentiation. A third instrument has a scale on differentiation as part of a questionnaire about the perceptions of students of their learning environment. Two instruments do not have a scale that explicitly deals with differentiation, but there are certainly items that point to differentiation behaviour. The questionnaire of Van der Scheer et al. (2016) for example, includes the scales 'clear instruction' and 'challenge students'. When a teacher differentiates well, it is plausible that both strong, average and weak students find the instruction clear and feel challenged. The student questionnaire of Gentry and Owen (2004) also includes a scale 'challenge'. Their scale 'choice' has also been included: when teachers want to connect to the interests of their students, they will offer them

more choice. This topic is also reflected in many other instruments and this scale determines whether students also experience these options.

What is also striking is that most instruments have been designed for or used in secondary education. Only the instrument of Van der Scheer et al. (2016) has asked fourth grade students (group 6 of the Dutch primary school) to assess their teacher. An explanation could be that researchers expect older students to be more able to assess their teacher. The instrument to be developed will therefore focus on the oldest students of primary schools (Dutch group 6-8).

3.2.2 Comparison of items on instructional adaptations

The instructional adaptations that teachers can make during their lessons are grouped into 15 categories. These categories are listed below per phase of the lesson in which they occur: during the entire lesson, the instruction phase, or the practice phase.

Adaptations during the entire lesson

Adapting the *learning goals* and corresponding expectations to differences between students is mentioned in four instruments. Most of these items are formulated rather vague, like 'The teacher adapts the learning goals/expectations to relevant differences between students'. From this item it does not become clear which differences are considered to be relevant. Other items on adapting learning goals take student performance as a relevant difference to adapt goals to. One instrument distinguishes between 'extra challenging goals for high-achieving students' and 'well-considered minimum goals for very low-achieving students'.

Providing students with options is a way to better connect education with students. Students can be given a *choice* in instruction, practice and grouping. It is assumed that students make choices based on their educational needs and occasionally on the basis of interests. An example item is 'The teacher offers the students options that enable them to better align aspects of instruction and practice with their educational needs'.

Teachers may use a variety of *materials* to support their lessons. Half of the reviewed instruments include items on adapting these materials. Three types of items can be distinguished. Items on the use of extra materials: 'The teacher allows some children to make use of auxiliary materials'. Items on the use of different materials: 'I use a variety of materials other than the standard text'. And items that specify the student differences to adapt materials to: 'My teacher helps me pick books or materials that are on my level'. In addition to adapting materials to students' level, adapting to their learning needs and interests is mentioned.

Only three instruments include items on differentiating through *ICT*. The instruments of Graham et al. (2008) and Prast et al. (2015) include some items on ICT as a way to learn new things, practicing weaker skills and offering challenge to stronger students. The research by Nieboer (2015) focuses on differentiation by means of the app Snappet, where the assignments that students receive are adapted to their previous answers. Her observation tool includes items on the use of Snappet by teachers. It seems somewhat striking that most recent differentiation instruments have no items on ICT, whereas nowadays ICT is often used in schools to adapt assignments to the level of students. In a review of several studies into ICT-based differentiation, Deunk et al. (2015) suggest that ICT can certainly contribute to the performance of students.

Adapting *assessment* to students' needs is mentioned in four instruments. Items such as 'I provide a variety of assessments tasks' and 'Grades are assigned based on individual student growth' show the adjustments that can be made in assessing students. Other items distinguish between the use of formative and summative assessment.

Adaptations during the instruction phase

The most frequently mentioned instructional adaptation is *more instruction*, usually for weaker students by means of pre-teaching prior to whole-class instruction or extended instruction at the end of the whole-class instruction ('The teacher offers struggling learners pre-instruction before formal instruction starts'). Additional instruction for stronger students focuses on instructing on another level or different subject matter ('The teacher gives stronger students additional instruction on the enrichment matter').

Varying the *content of the instruction* for different groups of students is frequently mentioned within the context of instruments that specifically focus on differentiation during math lessons. This concerns in particular the use of strategies and step-by-step plans, for example, 'The teacher determines the preferred strategy with weaker students' and 'The teacher discusses the advantages and disadvantages of strategies with stronger students'. Other math-specific items concern instruction at various levels of acting (informal – formal) and giving time to think after introducing a math problem.

The third most frequently mentioned strategy is adapting the *questions* asked to students during the instruction phase. Teachers could ask open-ended questions that can be answered by all students ('I deliberately ask open-ended questions during whole-class instruction') or ask questions on different levels to reach all students. One instrument distinguishes the type of questions that should be asked to weaker and stronger students: questions to weaker students about remembering, understanding and applying (lower order thinking skills) and questions to stronger students about analysing, evaluating and creating (higher order thinking skills).

Four instruments have items on adapting the *pace* of the instruction to students. Other frequently mentioned items on adaptation during the instruction phase include explaining the learning content in different ways, like visually or verbally, at different abstraction levels or with more or less support.

Most of the items about adapting instruction are specifically aimed at weaker or stronger students. For weaker students extended instruction, pre-teaching, a slower pace of instruction, more repetition and visual support are mentioned. Three instruments also have items on adapting the communication for weaker students: teachers should give them positive feedback, compliments and express positive expectations towards weaker students. Items for stronger students focus on providing enrichment materials with enrichment instruction, challenging questions, a faster pace of instruction and more process-focussed feedback.

Adaptations during the practice phase

The practice phase comes right after the instruction and is the phase in which students process the learning content, usually by making assignments. Most differentiation instruments include items about adapting practice to students. In addition to general items, such as 'Adapts practicing the learning content to relevant differences between students', there are many items that mention concrete adaptations. Most often mentioned is adapting the *difficulty* of the assignments ('Vary the complexity of assignments to match students' abilities'), the *amount* of assignments ('The teacher gives different groups of students a different number of assignments'), the type of assignments ('I especially let weaker students make repeating assignments') and the *time* that students get for their assignments ('The teacher gives weaker students extra time for the processing of the assignments').

Adapting assignments based on needs, skills and interests of students is most often mentioned. Items aiming at stronger students suggest to give these students more assignments, to put them to work more quickly, to let them do a lot of work independently and at their own pace and to offer them

enrichment material. To accommodate weaker students, teachers can provide more help, offer repetition assignments and select only the most important assignments.

Half of the instruments include items about *grouping* students during the practice phase. Most often these items suggest to vary different working and grouping forms, such as working individually, working together in pairs or in a larger group ('Let students work both individually and as a group'). Some items emphasize that grouping should be flexible: 'Flexible grouping is used'. Research has shown that flexibility is an important characteristic of groups (Deunk et al., 2015; Prast et al., 2015). There are also items in which the emphasis is on the student characteristics that groups should be formed on, such as level, working pace, previously achieved results, interests and learning preferences ('I group students for learning activities based on readiness, interests, and/or learning preferences').

3.2.3 Comparison of items on monitoring

In order to make instructional adaptations based on the needs and interests of students, these student characteristics must be mapped. Therefore, teachers must gather enough information about their students, analyse this and use the results in their lesson preparation and during their lesson. At the end of each lesson, they must also examine whether the adjustments made have produced the desired effect and if students are making sufficient progress. Monitoring students is therefore ideally a continuous process. The items coded as monitoring items can be divided into three categories, following the differentiation stages of Van Geel et al. (2018): before the lesson (preparing lesson period/lesson), during the lesson and after the lesson (evaluating lesson). These three moments of monitoring all have a different goal. Weak items only mention monitoring activity, but not to which end the monitoring takes place.

Before the lesson

Before any lesson or when preparing a period of lessons, it is important to collect data on relevant student characteristics. This makes it easier to plan lessons, cluster students into groups with the same educational needs and set goals. The item 'Readiness, interests, preferences, strengths, learning needs, stages of learning are evaluated' shows multiple student characteristics for teachers to collect data on. Most differentiation instruments focus on collecting data on the level or performance of students as these items show: 'My teacher knows what subjects or skills are easier for me' and 'I analyse the answers on curriculum-based tests to assess students' educational needs'.

During the lesson

Monitoring students should be a continuous process and should therefore also take place during the lesson: 'The teacher checks the progress of students during the practice phase of the lesson' and 'I assess specific students' educational needs based on (informal) observations during the math lesson'. Monitoring during the lesson gives teachers the opportunity to adjust their education right away and gives them feedback for further lessons.

After the lesson

At the end of the lesson and afterwards, teachers evaluate whether the set goals have been achieved and whether the chosen approach matched students' needs. This includes items such as 'I assess at the end of the lesson to determine knowledge acquisition', 'I mainly check the homework of weak students' and 'I evaluate whether a specific type of instruction was effective for specific students'.

3.3 Conclusion

3.3.1 Differences and similarities

This study aimed to answer the question *What are differences and similarities between existing instruments measuring differentiation in terms of the aspects of differentiation they measure?* There are some similarities between the examined instruments:

- Most instruments use Likert scales as a way of scoring items;
- Most instruments are specifically designed to be used by either an observer, teachers or students. A single instrument has been designed for usage by multiple parties, such as a teacher and his students;
- Most instruments include similar differentiation scales with respect to adaptations during the entire lesson, during the instruction phase, during the practice phase or monitoring;
- Most classroom observation instruments and teacher self-assessments are designed to measure differentiation, unlike most student surveys that measure differentiation as a part of teaching quality or classroom environment;
- Many items are formulated rather vague, like 'The teacher adapts the learning goals/expectations to relevant differences between students'. From this item it does not become clear which differences are considered to be relevant.

The examined instruments differ in a number of ways:

- Some instruments are designed for classroom observations in general, others specifically for use in math lessons;
- There is a big difference in the number of items per instrument, varying from less than 10 items to more than 50 items;
- There is a large difference in the number of items per measurement method: student surveys include on average 12 items (SD = 5), observation instruments on average 16 items (SD = 19) and self-assessment instruments on average 32 items (SD = 16);
- Some instruments are intended to assess a specific lesson and other instruments are intended to assess teaching across multiple lessons during a period or school year;
- Some instruments have items specifically focused on differentiation for stronger or weaker students, while other instruments do not specify this;
- Observation and self-assessment instruments generally aim to assess differentiation directly, while student surveys assess differentiation as part of measuring the teaching quality;
- The examined observation instruments are designed for use in elementary classrooms, while the student surveys have been used in secondary classrooms. The self-assessment instruments have been used in both types of schools.

The examined instruments include approximately three hundred differentiation-related items. These items have been divided into two categories: *instructional adaptations* (during the entire lesson, during the instruction, during the practice phase) and *monitoring*. The instructional adaptations category covers most items and was therefore further broken down into learning goals, choice, materials, ICT, assessment, instruction, time, questions, pace, assignments, grouping and adaptations for stronger and weaker students. Van Geel et al. (2018) conducted a similar study of comparing differentiation instruments. They grouped all scales and factors into six categories: curriculum, identifying instructional needs, setting challenging goals, monitoring and diagnosing, adapting instruction and general teaching quality dimensions. They found that most items belong to the category *adapting instruction* and coded these items into fairly similar categories: grouping, materials,

assignments and tasks, pace and provided learning time, questions, classroom activities, and instruction.

All items in these subcategories are about adaptations that teachers can make, but those adaptations are only effective if they meet the needs of the students. Van Geel et al. (2018) rightly point out that there are hardly any items that assess this match: “Although items such as ‘Attended appropriately to students who struggle with learning’ in Tomlinson et al.’s (2008) observation scheme or “I adapt the level of abstraction of instruction to the needs of the students” from the self-assessment instrument by Prast et al. (2015) appear to relate to this match, it is unclear how the observer or teacher would be able to indicate the appropriateness of the adaptation(s)”.

3.3.2 Implications

This study has provided an overview of existing differentiation instruments and their characteristics. Implications of the results of this study for the development of a new differentiation instrument are discussed below.

As mentioned before, most instruments include many examples of differentiation strategies. However, crucial to successful differentiation is not making adaptations, but making adaptations that match the needs of the students. For the instrument to be developed, it is therefore important that items not only list possible strategies, but focus on adapting those strategies to the needs of the students (e.g., ‘The teacher gave extra instruction to students who needed this’ instead of ‘The teacher gave extra instruction’). Items should be formulated in such a way that external observers and students can also assess them, so their perspectives can be compared.

In the theoretical framework it was already concluded that there is little research on which differentiation strategies are effective. As long as a widespread definition of high-quality differentiation is missing, instruments measuring differentiation can assess which strategies the teacher uses, but not whether this is high-quality differentiation. The instrument to be developed will therefore have to include relevant aspects of differentiation, but cannot claim to measure its quality.

For the instrument to be developed, it is important that the same items can be assessed by observers, teachers and students, so that their perspectives can be compared. The instrument to be developed will therefore focus on differentiation during the instruction phase, as this phase is clearly visible to all involved, in contrast to, for example, the preparation of the teacher or monitoring student results after the lesson. In order to view the instruction phase more specifically, the items in the instrument to be developed will focus on math instruction. A large part of the examined instruments focus on math education and also include math-specific items. This is useful input for the items to be developed and can help to formulate the items as explicitly as possible.

For the student variant of the instrument to be developed, it is important that items focus on the individual perception of students and therefore be formulated personally, for example ‘The teacher explains well to me’ instead of ‘The teacher explains well’. Students may have different needs and expectations of their teacher, teacher behaviour may be interpreted differently by students, and teachers may approach different students differently, but by asking students about their personal experiences this could be overcome (Van der Scheer et al., 2016).

3.3.3 Limitations

Seventeen existing differentiation instruments were reviewed and compared in this explorative study. However, the psychometric quality of these instruments has not been taken into account. For more information on this topic, the reader should consult the concerning research publication.

4. Study 2

4.1 Method

4.1.1 Aim

This study aimed at answering the following research question: *What is the degree of agreement between teachers, students and an external observer on how they assess teachers' differentiation behaviour?* To compare the different perspectives, an instrument has been developed that included similar items for all three rater-groups. This instrument has been tested in multiple classrooms.

4.1.2 Instrumentation

The first study provided useful insights into existing instruments to measure differentiation. Based on these insights, a list of criteria has been formulated for the instrument to be developed. This instrument should:

- Measure relevant aspects of differentiation, both monitoring and adapting. Relevant aspects are in line with the definition of differentiation as used in this study and are frequently mentioned in the instruments that were studied;
- Measure to what extent teachers differentiate their instruction, instead of judging how well they differentiate, as scientific literature does not fully agree on what high-quality differentiation is;
- Focus on teacher behaviour rather than on student behaviour, as the teacher does the monitoring and adapting;
- Include similar items for teachers, students and observers, so their perspectives can be compared;
- Include items which can be scored by an observer during a single classroom observation;
- Include math specific items, as around half of the studied instruments show that subject specific items can be more concrete and the majority of these items focus on math, giving useful examples;
- Allow raters to score a teacher's differentiation behaviour for different instruction groups, as teachers may adapt their instruction differently for different groups of students;
- Allow raters to score to what extent they agree with the items, using a Likert scale;
- Be used and understood by students and therefore, target students in the highest grades of primary school as older students are more likely to be able to express themselves.

To determine relevant aspects of differentiation, **Fout! Verwijzingsbron niet gevonden.** has been drawn based on the results of the first study. It shows the number of items for each category for each of the three methods. It is apparent from this table that student surveys have a relatively small number of items and – because of this – cover less categories, compared to the other two methods. Most items concern 'adapting through giving choices' and 'monitoring during the lesson'. 'Adapting the whole-class instruction' and 'subgroup instruction' are the most popular categories in observation instruments. Teacher self-assessment instruments include a relatively large number of monitoring items, as this might not be as visible to students and observers as making adaptations is and because students and observers will only be able to score monitoring that takes place during the lesson.

Items classified as 'non-specific' are broadly formulated items on differentiation, like 'The teacher adapts the instruction to the relevant differences between students'. It clearly is a differentiation item, but it does not provide information on which differences are assumed as relevant and how to adapt

instruction accordingly. Other items in these categories have only been mentioned once and therefore have no separate category.

Table 1. Items from the instruments studied per category and measurement type

Category	Observation form	Teacher self-assessment	Student survey	Total
Adaptations during entire lesson				
Choice	3	9	8	20
Materials*	8	5	1	14
Goals*	6	5	1	12
Assessment	-	3	2	5
ICT*	-	4	-	4
Non-specific/other	4	10	-	14
Adaptations during instruction				
Low-achieving*	23	3	-	26
High-achieving*	13	-	-	13
Whole-class*	14	3	1	18
Time*	18	5	1	24
Grouping	2	3	-	5
Pace	2	3	-	5
Non-specific/other	1	5	8	14
Adaptations during practice				
Difficulty*	8	6	2	16
Grouping	8	6	1	15
Time	6	-	2	8
Amount*	3	4	-	7
Non-specific/other	7	12	2	21
Monitoring				
Before the lesson	1	19	-	20
During the lesson*	8	9	4	21
After the lesson	-	11	-	11
Non-specific/other	-	2	2	4
Total	135	127	35	297

*Relevant categories

Relevant categories

Relevant categories that should be included in the new instrument, are in line with the definition of differentiation used, are frequently used in all three types of instruments and take into account the before stated criteria. They are shown with an asterisk in Table 1. For overall adaptations these are the use of materials, goals and ICT. Although there are only four items on ICT, ICT can offer many differentiation possibilities which makes this an interesting category to include. There is a large number of items on whole-class instruction and subgroup instruction and therefore, they can be considered relevant, although current student instruments have no items on it. When it comes to

adaptations made during the practice phase, changing the difficulty and the number of assignments will often be related, for example when a teacher compacts a student's assignments (number) in order to replace them with assignments on another level (difficulty). From the monitoring items, only the ones during the lesson can be scored by all three rater-groups.

Non-relevant categories

Non-relevant categories are not in line with the definition of differentiation used, not frequently used in all three types of instruments and/or do not take into account the before stated criteria. Choice seems a relevant category to include, but was not included because from the items on choice in the instruments studied, it is not clear which teacher behaviour can be observed when a teacher offers choice. The items are too vague to expect agreement between raters. Grouping during the practice phase is not included, because from the items on grouping used in the studied instruments, it is not clear whether students with different or similar levels should be grouped. Also, following the used definition of differentiation, grouping is not so much an adaptation that teachers make in order to meet the needs of different students, but more because people in general can learn from each other and students should learn how to work together.

Items have been constructed for each marked category and can be found in Table 2. The criteria stated in the previous have been taken into account in formulating the items. The original items were constructed in Dutch and can be found in Appendix B.

Table 2. Constructed items

Category	Observation item	Teacher item	Student item
Adapting			
Goals	The teacher has, in addition to the basic lesson goal, one or more additional lesson goals set for specific (groups of) students.	I have, in addition to the basic lesson goal, one or more additional lesson goals set for specific (groups of) students.	I learned this lesson exactly the same as the other children in my class.
Materials	The teacher uses suitable material for students who need it.	I have used suitable material for students who need it.	The fact that my teacher used material to explain something to me helped me to better understand it.
ICT	The teacher uses ICT to let students practice at their own level.	I have used ICT to let students practice at their own level.	I have practiced math assignments on the computer or tablet that I find difficult.
Adapting instruction			
Low-achieving and high-achieving	The teacher gives extra instruction at the level of the student.	I have given extra instruction at the level of the student.	The extra instruction helped me to better understand.

Whole-class	The teacher makes the connection between the levels at which the students act and the level at which they must act during the lesson.	I have made the connection between the levels at which the students act and the level at which they must act during the lesson.	My teacher has explained in different ways (for example by telling, showing a model, letting us do something, etc.)
Time	The teacher adjusts the instruction time to what each student needs to achieve the lesson goal.	I have adjusted the instruction time to what each student needs to achieve the lesson goal.	I have received enough instruction to understand the lesson (i.e. not too much and not too little instruction).
Adapting practice			
Difficulty and amount	The teacher adjusts the assignments (amount and difficulty) to what each student needs to achieve the lesson goal.	I have adjusted the assignments (amount and difficulty) to what each student needs to achieve the lesson goal.	The assignments I had to do were at my level (i.e. not too difficult and not too easy). AND I have received enough assignments to practice (i.e. not too many and not too few assignments).
Monitoring			
Monitoring during the lesson	The teacher checks the students' understanding during the basic instruction.	I have checked the students' understanding during the basic instruction.	During the instruction to the whole class, my teacher knew whether I understood the instruction.
	The teacher checks the students' understanding during the extended instruction.	I have checked the students' understanding during the extended instruction.	During the extra instruction, my teacher knew if I understood the instruction.
	The teacher checks the students' understanding during the enrichment instruction.	I have checked the students' understanding during the enrichment instruction.	During the extra instruction, my teacher knew if I understood the instruction.
	The teacher checks the students' progress during the practice phase.	I have checked the students' progress during the practice phase.	My teacher checked while working if I understood.

At the end of the lesson, the teacher evaluates whether all students have achieved the lesson goal.	At the end of the lesson, I have evaluated whether all students have achieved the lesson goal.	My teacher now knows how well I understood the lesson.

The items in Table 2 form an instrument with three versions: an observation form for an external observer, a self-assessment for teachers and a survey for students. The observation form and self-assessment consist of the same 12 items, only differently formulated, e.g. "The teacher has…" or "I have…". Ten items have to be scored for each instruction group (basic, intensive and enrichment), two items are only applicable for one instruction group ('The teacher checks students' understanding during the intensive instruction'). The items in the student survey are more personally formulated and include more explanation, e.g. 'I have received the right amount of instruction to understand this lesson (so not too much and not too little)'. To be able to compare students' ratings with those of the teacher and observer, it is necessary to know for each student which instruction group they are assigned to.

Each item can be scored on a 4-point Likert scale (1=strongly disagree, 4=strongly agree). The advantage of a 4-point Likert scale compared to a 5-point Likert scale is that it prevents respondents from scoring a 3 every time when they are in doubt. Some items can also be scored 'not applicable', e.g., when there was no extra instruction for students in the intensive group, it cannot be assessed whether the extra instruction matched the level of the students.

Two teachers and five students have given feedback on the draft versions of the instruments. Based on their comments small changes have been made in the layout of the instrument and in the formulation of the items. Appendix C, D and E show the final versions of the instrument.

4.1.3 Procedure

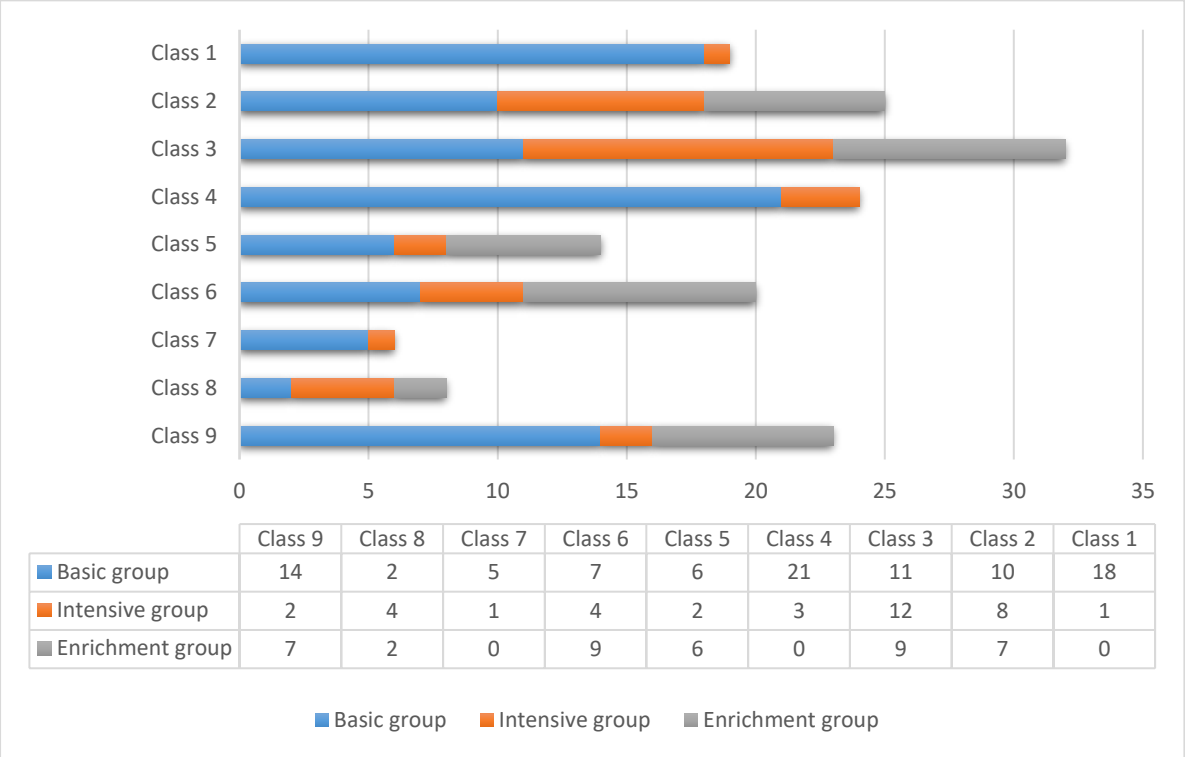
Participating teachers were visited during a single math lesson with an instructional phase. The teachers were asked to share the instruction groups they would use for that lesson prior to the visit, so that the observer could score a teacher's differentiation behaviour separately for each instruction group. Most teachers did this by providing the names of the students per instruction group and indicating where these children were in the class during the visit. During the lesson, the researcher would fill in the observation form. At the end of the lesson, the teacher was asked to leave the classroom and fill in the self-assessment, while the researcher would administer the student survey with the students. The researcher has read the items from the student survey aloud to the students and explained them when necessary. The students were told to answer the questions based on the lesson they had just received.

4.1.4 Participants

Nine teachers from five different primary schools in Enschede, Deventer and Eindhoven (the Netherlands) agreed to participate. The teachers (56% female) teach grade 8 to grade 10, which are the highest grades in Dutch primary education. Class sizes ranged from 6 to 32 students per class, with a mean of 19 students (SD = 8 students). This is not representative of the Dutch population of teachers and class size, but that is not a problem as the purpose of this study is to explore whether the different perspectives match. Figure 1 shows that teachers in six classrooms divided their students into three instruction groups: a basic group for around average students, an intensive group for low performing students and an enrichment group for high performing students. In three classrooms there were only

two instruction groups; these classes did not have an enrichment group. Teachers' differentiation behaviour was scored for each instruction group separately, forming a total of 24 cases. Cases varied from 3 raters per case (1 observer, 1 teacher, 1 student) to 23 raters (1 observer, 1 teacher, 21 students).

A total of 171 students participated by filling in the student survey after their lesson. Students were between 9 and 12 years old. The most students belonged to the basic group (94). There were similar numbers of students in the intensive group (37) and enrichment group (40).



	Class 9	Class 8	Class 7	Class 6	Class 5	Class 4	Class 3	Class 2	Class 1
Basic group	14	2	5	7	6	21	11	10	18
Intensive group	2	4	1	4	2	3	12	8	1
Enrichment group	7	2	0	9	6	0	9	7	0

Figure 1. Number of students per class per instruction group

4.1.5 Data analysis

The statistical programme SPSS has been used to analyse the data. Frequency tables have been used to inspect the data, as well as mean scores. Next, Cohen's kappa was run to determine if there was agreement between two groups of raters on each of the items on differentiation per instruction group, taking into account the possibility of agreement occurring by chance. As kappa measures agreement between two raters, the mode of all student scores per instruction group was used. The mode is the value that occurs most often and therefore gives a good idea of what most students have scored. Kappa coefficients have a value between 0 (no agreement) and 1 (perfect agreement).

First, the kappa for all scores was calculated (K_Overall). K_Overall shows to what extent the raters agreed to give a score of 1, 2, 3, 4 or 'not applicable' on each item. Second, for items that could be scored 'not applicable', kappa was calculated for all times raters agreed about whether the item is applicable or not (K_Applicable). Items that could not be scored 'not applicable' have no kappa coefficients. Third, kappa was calculated for all applicable scores (K_Judgement). Items with only 'not applicable' scores are presented as 'n = 0' in this column.

In some cases all raters within a rater-group gave the same score to an item (e.g. the observer scored the item on extra instruction for all cases in the basic group with 'not applicable'). SPSS considers these variables as constant and calculates kappa as zero. However, this does not necessarily mean that there is no agreement between raters. Therefore the 'Raw Agreement' was calculated by dividing the number of agreements in scores by the total (n). As with the kappa coefficients, Raw Agreement has a value between 0 (no agreement) and 1 (perfect agreement), only Raw Agreement is not adjusted for chance.

4.2 Results

4.2.1 Descriptive statistics

Observer

All cases were scored by the same observer. From the data in Figure 2, it is apparent that the observer scored few items with a 2 (5.6%). Scores 1 (20.1%) and 4 (22.2%) occur almost as often, followed by score 3 (15.6%). Score 5 (not applicable) is given by the observer to one third of all items (36.5%). For the enrichment group this even applies to almost half of the items (48.6%). This is lower for the intensive group, with 25.0% of the items. It seems that the observer sees more differentiation for students in the intensive group than for students in the enrichment group. Items related to the intensive and enrichment group are most often scored with a 4. Items related to the basic group most often with a 1. This could suggest that more differentiation took place for students in the intensive and enrichment group.

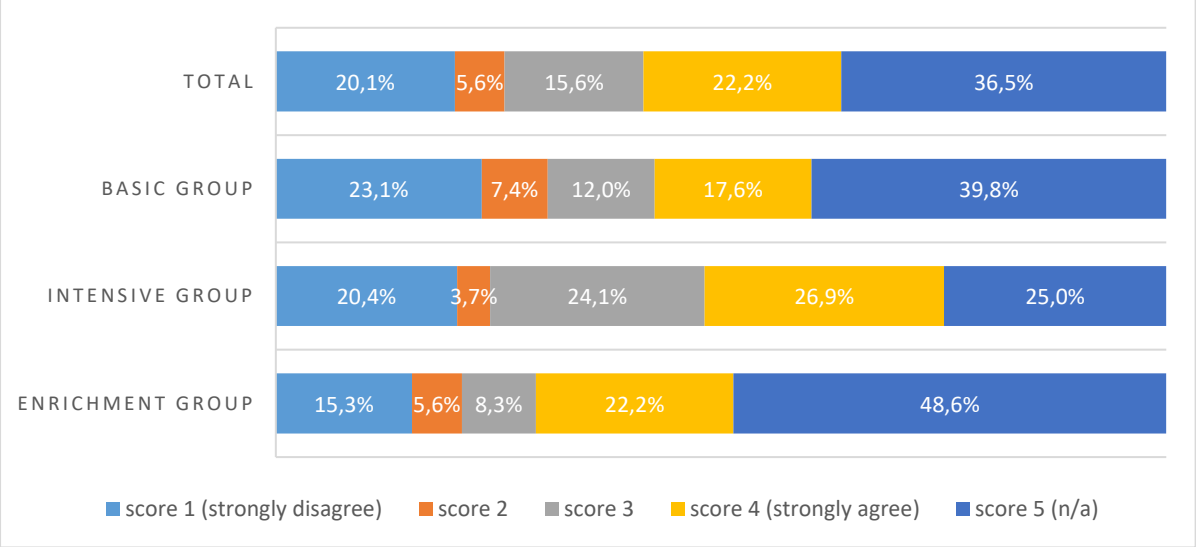


Figure 2. Frequency of scores given by observer

Table 3 shows the mean scores per question given by the observer for each instruction group. Some items could also be scored as 'not applicable'. These scores are not included in the calculation of the mean score. The table shows how many scores the mean score is based on (n). The observer judged that no extra instruction took place in the basic group and the enrichment group and no material was used (n = 0). For six cases the observer gave a score for checking understanding during the extended

instruction. In all other cases for the other instruction groups, the observer has indicated that an extended or enriching instruction did not take place (n = 0).

The mean scores on an item in the different instruction groups differ by a maximum of 0.94 from each other. This could mean that teachers differentiate for the different instruction groups to a reasonably similar extent. However, for most items the standard deviation is quite high. This suggests that there is a lot of variation between the given scores per item. The items could be scored on a 4-point Likert scale and most items have a minimum score of 1 and a maximum score of 4. This indicates that the observer has observed a great deal of variation between teacher behaviour within the same instruction group.

Table 3. Mean scores given by observer

	Basic group		Intensive group		Enrichment group	
Items	Mean (SD)	n	Mean (SD)	n	Mean (SD)	n
Goals	1.00 (.000)	9	1.33 (1.000)	9	1.83 (1.329)	6
Instruction time	2.89 (1.364)	9	3.67 (.500)	9	3.50 (.837)	6
Extra instruction		0	2.57 (1.134)	7		0
Level of action	2.88 (1.126)	8	3.00 (.866)	9	3.00 (1.414)	2
Materials		0	4.00 (.000)	2		0
Assignments	2.89 (1.453)	9	2.89 (1.453)	9	3.83 (.408)	6
ICT	2.50 (1.732)	4	2.50 (1.732)	4	3.00 (1.732)	3
Understanding basic	2.75 (.463)	8	2.87 (.354)	8	2.00 (.000)	2
Understanding intensive		0	3.67 (.516)	6		0
Understanding enrichment		0		0		0
Monitoring progress	2.56 (1.130)	9	3.11 (.928)	9	3.00 (1.095)	6
Evaluation	1.89 (1.364)	9	1.89 (1.364)	9	1.50 (1.225)	6

Teachers

The data in Figure 3 shows that teachers score their own differentiation least often with a 1 (12.2%) and most often with a 3 (25.7%) or 4 (23.6%). A striking exception are the scores for the enrichment group, were teachers score a third of the items with a 2 (31.9%). On average teachers score 19.4% of the items as "not applicable". This is lower for students in the intensive group (13.9%) and higher for students in the enrichment group (25.0%). This could mean that teachers differentiate more for students in the intensive group. Compared to the scores of the observer and the students, teachers score relatively little "not applicable".

Table 4 shows the mean scores per question given by the teachers for each instruction group. Some items could also be scored as 'not applicable'. These scores are not included in the calculation of the mean score. The table shows how many scores the mean score is based on (n). Most teachers have scored all items as ‘applicable’ by giving them a score. The items ‘understanding intensive’ and ‘understanding enrichment’ were seen by most teachers as being applicable to the instruction group in question (e.g. whether a teacher has checked the students' understanding during the extended

instruction has been scored by the teachers only for students in the intensive group, because they were probably the only students who received this instruction).

For most items, the standard deviation is quite high. This suggests that there is a lot of variation between the given scores per item. The items could be scored on a 4-point Likert scale and most items have a minimum score of 1 and a maximum score of 4. This indicates a great deal of variation between teacher behaviour within the same instruction group.

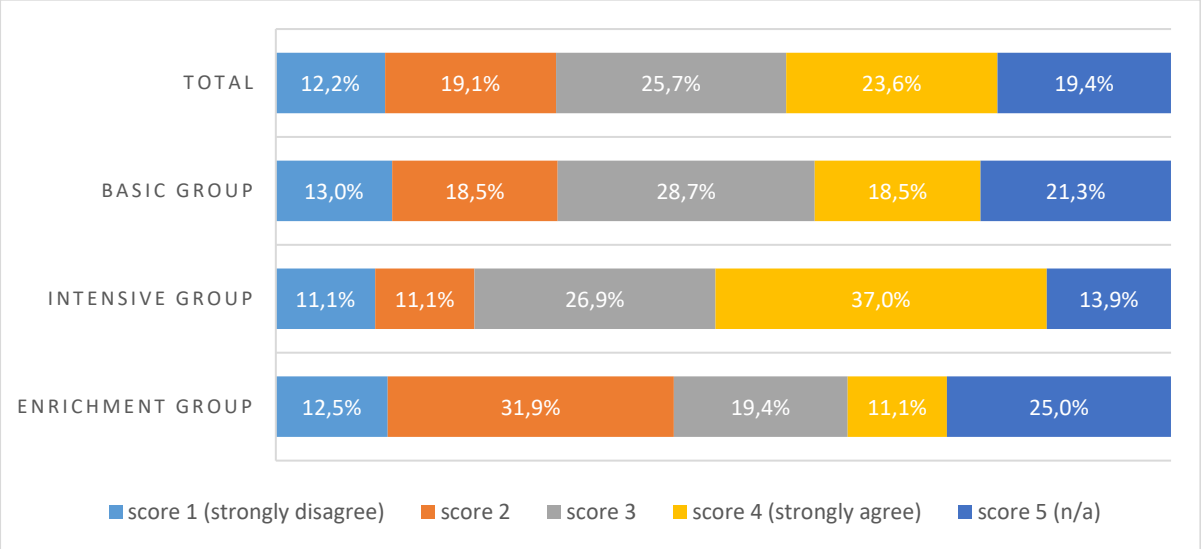


Figure 3. Frequency of scores given by teachers

Table 4. Mean scores given by teachers

	Basic group		Intensive group		Enrichment group	
Items	Mean (SD)	n	Mean (SD)	n	Mean (SD)	n
Goals	2.00 (.707)	9	2.78 (.972)	9	2.00 (.707)	5
Instruction time	2.89 (1.054)	9	3.11 (1.269)	9	2.50 (1.049)	6
Extra instruction	2.75 (.886)	8	3.44 (.527)	9	2.00 (1.000)	3
Level of action	2.78 (.833)	9	2.89 (.782)	9	1.80 (.447)	5
Materials	3.29 (.951)	7	3.29 (.951)	7	2.25 (.500)	4
Assignments	2.78 (.972)	9	2.78 (1.302)	9	3.50 (.548)	6
ICT	2.14 (1.464)	7	2.14 (1.464)	7	2.80 (1.643)	5
Understanding basic	2.56 (1.014)	9	3.00 (1.225)	9	1.80 (.447)	5
Understanding intensive		0	3.71 (.488)	7		0
Understanding enrichment		0		0	2.67 (.577)	3
Monitoring progress	2.89 (1.167)	9	3.67 (.500)	9	2.50 (1.049)	6
Evaluation	2.67 (1.000)	9	2.67 (1.000)	9	2.17 (.753)	6

For some items there is a big difference between the mean scores in the different instruction groups. Teachers scored the item on 'extra instruction' for the intensive group on average with 3.44, while they gave an average score of 2.00 for the enrichment group. This could suggest that teachers do give extra instruction to help below average students, but do not challenge above average students with instruction on their level. The items 'understanding basic' and 'monitoring progress' also have a high average score for the intensive group and a much lower average score for the enrichment group. It seems that teachers judge that they differentiate more for below average students.

Students

Students score few items with a 1 (7.8%) and 2 (6.7%), as can be seen in Figure 4. Most often they score items with a 4 (28.2%). This applies to students in all instruction groups. In about 40% of the items, students indicate that this does not apply to them. For students in the enrichment group this is the case for almost half of the items (47.1%). This is considerably less for students in the intensive group (28.4%). This could mean that students in the intensive group experience more differentiation than students in the enrichment group.

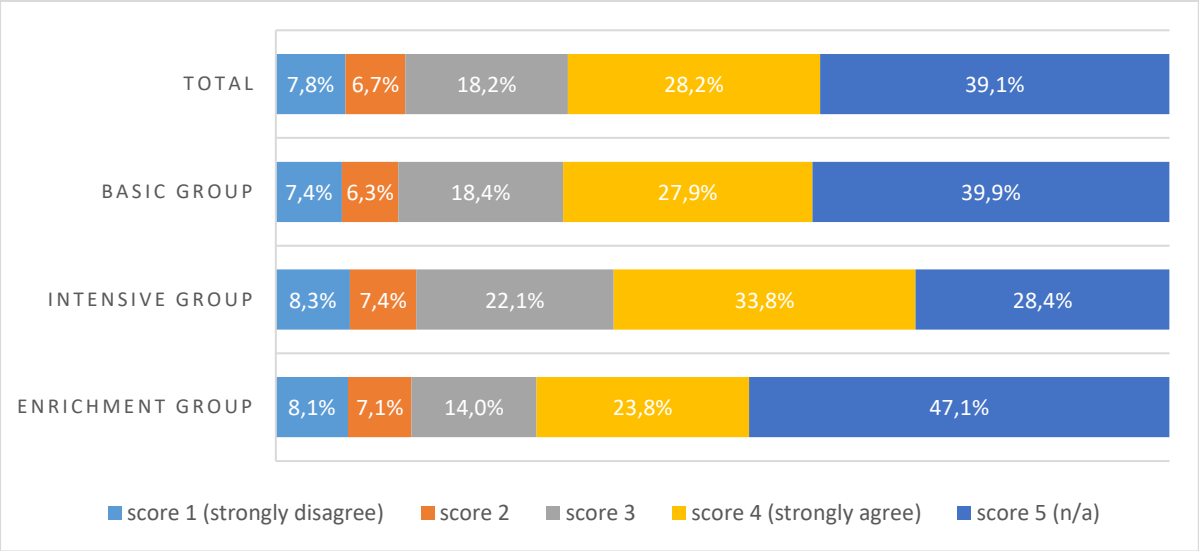


Figure 4. Frequency of scores given by students

Table 5 shows the mean scores per question given by the students for each instruction group. Some items could also be scored as 'not applicable'. These scores are not included in the calculation of the mean score. The table shows how many scores the mean score is based on (n). Most students have scored all items as 'applicable' by giving them a score. Most students in all three instruction groups judged that there was no extra instruction and no materials were used.

Just as with the observer and teachers scores, the standard deviation for most of the mean student scores is quite high. This suggests that there is a lot of variation between the given scores per item. The items could be scored on a 4-point Likert scale and most items have a minimum score of 1 and a maximum score of 4. This indicates that students within the same instruction group experienced their teacher behaviour quite differently.

The mean student scores are relatively close to each other, compared to the observer and teacher scores. This could mean that students in the different instruction groups experience a similar degree of differentiation.

Table 5. Mean scores given by students

	Basic group		Intensive group		Enrichment group	
Items	Mean (SD)	n	Mean (SD)	n	Mean (SD)	n
Goals	3.19 (.627)	94	2.78 (.854)	37	2.88 (.966)	40
Instruction time	3.64 (.716)	94	3.27 (.990)	37	3.00 (1.177)	40
Extra instruction	4.00 (.000)	6	3.60 (.764)	25	4.00 (.000)	2
Level of action	2.82 (1.073)	61	3.29 (.976)	28	2.50 (1.000)	12
Materials	3.57 (.535)	7	2.40 (1.517)	5	3.00 (.000)	1
Assignments	3.40 (.794)	94	3.59 (.644)	37	3.42 (.844)	40
ICT	1.96 (.957)	49	2.14 (1.108)	21	2.15 (1.226)	20
Understanding basic	3.27 (.892)	85	3.00 (1.069)	36	3.25 (.967)	20
Understanding intensive		0	3.00 (1.085)	18		0
Understanding enrichment		0		0	3.50 (.707)	2
Monitoring progress	2.63 (1.287)	94	3.14 (1.110)	37	2.87 (1.305)	40
Evaluation	3.26 (1.026)	94	3.32 (.852)	37	3.27 (.902)	37

4.2.2 Inter-rater reliability

Table 6 to table 14 show the kappa coefficients and Raw Agreement scores for each item, first for observers and teachers in each of the three instruction groups, second for teachers and students and last for observers and students. The results from these tables will be presented here, first for the agreement on the application of items and second for the agreement on the given scores.

Agreement on the application of items

A moderate to almost perfect agreement has been found for 28 out of 51 items on whether or not an item is applicable (16 times *k* or RA > .81, 7 times *k* or RA > .61 and 5 times *k* or RA > .41). The observer and students agreed on the application of most items, with agreement on 14 items (10 times *k* or RA > .81, 3 times *k* or RA > .61 and 1 time *k* or RA > .41). The observer and teachers agreed on 7 items (4 times *k* or RA > .81, 1 time *k* or RA > .61 and 2 times *k* or RA > .41), the same number as for teachers and their students (2 times *k* or RA > .81, 3 times *k* or RA > .61 and 2 times *k* or RA > .41).

Some items could not be assessed as not applicable. This applied to the items 'goals', 'instruction time', 'assignments', 'monitoring progress' and 'evaluation'. There is a lot of agreement between observer and teachers, and teachers and students about whether or not the items 'extra instruction', 'level of action', 'understanding basic' and 'understanding enrichment' are applicable. Between observer and students this also applies to the items 'materials' and 'ICT'. Items for which (almost) no agreement has been found are 'materials' (only between observer and students), 'ICT' (only between observer and students) and 'understanding intensive'.

Agreement on the given scores

When raters agreed that an item was applicable to score, they had moderate to almost perfect agreement on 10 out of 96 items (7 times *k* or RA > .81 and 3 times *k* or RA > .41). The observer and teacher agreed on the score of most items, with agreement on 6 items (5 times *k* or RA > .81 and 1 time *k* or RA > .41). The teachers and students agreed on 2 items, the same number of items as for observer and students (1 time *k* or RA > .81 and 1 time *k* or RA > .41). There is only sufficient agreement on the given scores for the items 'materials', 'ICT', 'understanding basic' and 'understanding intensive'. This agreement was found in the intensive and enrichment instructional group. No sufficient agreement on the given scores was found for the other eight items.

For the items 'goals', 'instruction time, 'assignments', 'monitoring progress' and 'evaluation' only kappa coefficients and Raw Agreement scores lower than 0.4 were found on both aspects of agreement.

Table 6. Inter-rater reliability of observer and teachers in the basic group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	9	.000	.222						
Instruction time	9	.250							
Extra instruction	9	.000	.111	9	.000	.111	0	-	
Level of action	9	.143		9	.000	.889***	8	.167	
Materials	9	.000	.222	9	.000	.222	0	-	
Assignments	9	.297							
ICT	9	.542*		9	.372		4	1.000***	
Understanding basic	9	.033		9	.000	.889***	8	.043	
Monitoring progress	9	.390							
Evaluation	9	.074							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

Table 7. Inter-rater reliability of observer and teachers in the intensive group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	9	.113							
Instruction time	9	.000							
Extra instruction	9	.053		9	.000	.778**	7	.152	
Level of action	9	.105		9	.000	1.000***	9	.105	
Materials	9	.237		9	.151		2	.000	1.000***
Assignments	9	.167							
ICT	9	.542*		9	.372		4	1.000***	
Understanding basic	9	.038		9	.000	.889***	8	.049	
Understanding intensive	9	.471*		9	.182		5	1.000***	
Monitoring progress	9	.063							
Evaluation	9	.074							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

Table 8. Inter-rater reliability of observer and teachers in the enrichment group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	6	.032							
Instruction time	6	.071							
Extra instruction	6	.000	.500*	6	.000	.500*	0	-	
Level of action	6	.071		6	.182		2	.333	
Materials	6	.000	.333	6	.000	.333	0	-	
Assignments	6	.333							
ICT	6	.520*		6	.333		3	1.000***	
Understanding basic	6	.000		6	.182		2	.000	.500*
Understanding enrichment	6	.000	.500*	6	.000	.500*	0	-	
Monitoring progress	6	.111							
Evaluation	6	.032							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

Table 9. Inter-rater reliability of teachers and students in the basic group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	9	.071							
Instruction time	9	.036							
Extra instruction	9	.000		9	.031		1	.000	.000
Level of action	9	.018		9	.000	.667**	6	.000	
Materials	9	.115		9	.069		7	.120	
Assignments	9	.085							
ICT	9	.129		9	.372		4	.000	
Understanding basic	9	.182		9	.000	.889***	8	.184	
Monitoring progress	9	.069							
Evaluation	9	.297							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

Table 10. Inter-rater reliability of teachers and students in the intensive group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	9	.200							
Instruction time	9	.100							
Extra instruction	9	.050		9	.000	.778**	7	.061	
Level of action	9	.100		9	.000	.778**	7	.160	
Materials	9	.390		9	.250		3	1.000***	
Assignments	9	.036							
ICT	9	.129		9	.372		4	.000	
Understanding basic	9	.033		9	.000	1.000***	9	.033	
Understanding intensive	9	.091		9	.286		5	.091	
Monitoring progress	9	.053							
Evaluation	9	.115							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

Table 11. Inter-rater reliability of teachers and students in the enrichment group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	6	.250							
Instruction time	6	.241							
Extra instruction	6	.000	.500*	6	.000	.500*	0	-	
Level of action	6	.032		6	.077		1	.000	.000
Materials	6	.000	.333	6	.000	.333	0	-	
Assignments	6	.000							
ICT	6	.357		6	.333		3	.500*	
Understanding basic	6	.091		6	.333		3	.000	.000
Understanding enrichment	6	.000	.500*	6	.000	.500*	0	-	
Monitoring progress	6	.000							
Evaluation	6	.154							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

Table 12. Inter-rater reliability of observers and students in the basic group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	9	.000	.222						
Instruction time	9	.154							
Extra instruction	9	.000	.889***	9	.000	.889***	0	-	
Level of action	9	.045		9	.400		6	.034	
Materials	9	.000	.889***	9	.000	.889***	0	-	
Assignments	9	.102							
ICT	9	.481*		9	1.000***		4	.000	
Understanding basic	9	.206		9	1.000***		8	.077	
Monitoring progress	9	.274							
Evaluation	9	.036							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

Table 13. Inter-rater reliability of observers and students in the intensive group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	9	.000							
Instruction time	9	.313							
Extra instruction	9	.060		9	.286		5	.167	
Level of action	9	.050		9	.000	.778**	7	.000	
Materials	9	.743**		9	.727**		2	.000	1.000***
Assignments	9	.019							
ICT	9	.481*		9	1.000***		4	.000	
Understanding basic	9	.068		9	.000	.889***	8	.143	
Understanding intensive	9	.031		9	.364		4	.333	
Monitoring progress	9	.045							
Evaluation	9	.161							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

Table 14. Inter-rater reliability of observers and students in the enrichment group

Item	n	K_Overall	RA	n	K_Applicable	RA	n	K_Judgement	RA
Goals	6	.200							
Instruction time	6	.368							
Extra instruction	6	.000	1.000***	6	.000	1.000***	0	-	
Level of action	6	.250		6	.571*		1	.000	.000
Materials	6	.000	1.000***	6	.000	1.000***	0	-	
Assignments	6	.100							
ICT	6	.750**		6	1.000***		3	.500*	
Understanding basic	6	.250		6	.667**		2	.000	.000
Understanding enrichment	6	.000	1.000***	6	.000	1.000***	0	-	
Monitoring progress	6	.077							
Evaluation	6	.059							

Note. K_Overall = Cohen's Kappa of the overall scores; K_Applicable = Cohen's Kappa if raters agree whether the items are applicable or not; K_Judgement = Cohen's Kappa of the applicable scores; RA = Raw Agreement.

k = **k* > .41 (moderate); ***k* > .61 (substantial); ****k* > .81 (almost perfect).

RA = **RA* > .41 (moderate); ***RA* > .61 (substantial); ****RA* > .81 (almost perfect).

4.3 Conclusion and discussion

Based on the review in the first study, an instrument was developed in the second study to measure differentiation from the three most used perspectives: teachers, observers and students. The second study answered the following research question: *What is the degree of agreement between teachers, students and an external observer regarding how they assess teachers' differentiation behaviour?* For the degree of agreement between the rater-groups, two aspects were considered: 1) do raters agree with each other on whether an item was applicable during the lesson and could therefore be scored and 2) if an item could be scored, do raters agree on the given score.

4.3.1 Agreement on the application of items

In addition to a score between 1 and 4, some items could also be assessed as 'not applicable'. An example is the item "I have checked the students' understanding during the extended instruction". If no extended instruction took place during the lesson, this item should be scored as 'not applicable'. The observer scored a total of 36.5% of the items as not applicable, compared to 39.1% of the students and only 19.4% of the teachers. So in many cases, the observer and the students concluded that an item could not be scored, but the teacher did give a score (other combinations occurred as well). When looking at the kappa coefficients and the Raw Agreement scores within all relationships, a moderate to almost perfect agreement has been found for 28 out of 51 items on whether or not an item is applicable. This applies to twice as many items between the observer and the students than for the other relationships.

No reasonable degree of agreement was found for about half of the scored items. Various explanations for this are possible. First of all, it is possible that raters have interpreted items differently. For example, the item 'materials' ("The teacher uses suitable material for students who need it") was only scored by the observer when the teacher had students work with math-related materials, such as fractional circles or number racks. This was only the case in 2 of the 24 cases, according to the observer. Nevertheless, teachers in 18 cases scored the item as applicable. It could be that these teachers scored the use of other materials, such as their own use of the smart board during the lesson. A different formulation or an explanation of the item might prevent this.

In addition to the formulation of the items, the formulation of the answer options could also explain the difference in agreement between raters. It could be that for some raters, the difference between scoring a '1' and scoring 'not applicable' was not clear. For example, if a teacher has had all students work with the same material during an instruction, while students in the enrichment group no longer needed the material, then material has been used (so the item does apply to the lesson) but the material is not adjusted to the level of all students. In this situation the item should have been scored with a '1' for the enrichment group. It could be that some raters have scored this situation as 'not applicable' because the teacher did not differentiate his use of materials. An attempt was made to clarify this by underlining the last part of the item ("... for students who need it") and thereby emphasizing this part, and by adding a good and less good example. To improve the instrument, the 'not applicable' option should be further clarified for raters, or replaced with the option 'no materials used'.

Another possible explanation could be that the participating teachers have scored their actions in general, rather than just their actions during the observed lesson (i.e. "I usually do this, so I score myself a 3 or 4", instead of "I didn't do this today, so I score the item as not applicable"). All studied teacher self-assessment instruments in the first study had teachers assess their daily practice, instead of a specific lesson. Teachers were notified by the researcher at the start of the observation that they

had to assess the observed lesson. To improve the instrument it could be included in the formulation of the items that it specifically concerns the observed lesson ("During this lesson I have ...").

4.3.2 Agreement on the given scores

When a rater assessed an item as applicable to the lesson, a score between 1 and 4 could be given, where 1 represents the same approach for all students and therefore little or even no differentiation and 4 represents a highly differentiated approach. Cohen's kappa was used to determine the degree of agreement between the given scores of the different raters. The observer and the teachers have moderate agreement or higher on 6 of the 32 items (in all three instruction groups). This applies to only two items for the relationship between teachers and students, and observer and students. Overall, the scores of the various rater-groups show little agreement. Looking at the different instruction groups, there is almost no agreement in the basic instruction group (on only one item) and fair agreement in the enrichment group (on four items) and intensive group (on five items). The results show that when measuring differentiation, it is important to consider which measurement method to use, as the scores will differ depending on the chosen method.

Some possible explanations for these results have already been mentioned and also apply here. In addition to that, the difference between the scores 1 to 4 may have been unclear to raters, leading to different interpretations. The items can be scored on a Likert scale from 1 to 4. To clarify the difference between the scores, scores 1 and 4 contain a description. This description gives an example for each item of non-differentiated behaviour (score 1) and differentiated behaviour (score 4). It is possible that raters could not make the link with the example in situations other than described in the examples. Score 2 and score 3 do not contain an example and it could be that therefore these scores have been interpreted differently more often.

Another explanation is that teachers have given socially desirable answers, because they know that they are expected to differentiate their instruction. This was also suggested by Dobbelaer (2019) in her research. To prevent this as much as possible, teachers could complete the self-observation anonymously and they were the only ones who were able to view the results for their class afterwards.

An observer can only assess what he sees during an observation. Teachers who differentiate their instruction, are known to analyse and plan their instructional adaptations before, during and after the lesson. Most of this is not visible to an observer. It is plausible that the observer in this study was not fully able to interpret every differentiation behaviour correctly. Giving observers and teachers the possibility to explain their score per item, could improve the instrument.

A final explanation for the limited degree of agreement are the students. When discussing the student survey as a method to measure differentiation in section 2.2.3, it was already concluded that student perceptions are not often used in primary education to collect information about teachers (Van der Scheer et al., 2016). It was mentioned that students' opinions could be easily influenced by external factors, such as having different expectations from the teacher or interpreting teacher behaviour differently. Also, the researcher noticed many questions from students while conducting the student survey. These questions mainly concerned the interpretation of questions in the survey. This suggests that the formulation of the questions was not clear to students and that students have interpreted questions differently.

A striking result is that no significant degree of agreement has been found for the items 'goals', 'instruction time', 'assignments', 'monitoring progress' and 'evaluation'. These are the only items in the instrument that raters could not score as 'not applicable'. It is unclear whether this is the cause for the lack of agreement between raters on these items. It is possible that raters found the items not

applicable to the lesson, but because they did not have this option they gave a random score. However, the observer was convinced that these items could have been assessed in all cases. To improve the instrument, these items should be viewed critically. The items may need to be formulated differently or maybe even left out of the instrument if it might be (too) difficult for some rater-groups to assess these items.

4.3.3 Implications

The results of this study show that it is difficult to find agreement between the perceptions of teachers, students and observers on teachers' differentiation behaviour. Due to a lack of agreement, it could be concluded that all three rater-groups should be present when measuring differentiation in the future. However, which conclusions can be drawn from this, if each method leads to different results? Another possibility is that every rater-group only assesses part of the items. For example, a student might be better able to assess if he has gotten sufficient instruction time or instruction at his own level, than an external observer who does not know the student. On the other hand, assessing whether a teacher makes the connection between the level at which the student acts and the level at which he must act during the lesson is more difficult for students to assess than it is for teachers. By dividing the items over the rater-groups, all three perspectives are included and the advantages of the different methods are taken into account.

4.3.4 Representativeness of respondents

The choice of respondents influences the results of the study. For this study it proved difficult to find a sufficient number of teachers who wanted to participate. Therefore, it has not been possible to select a representative group of teachers. Of the participating teachers in this study, 44% are male and class sizes vary from 6 to 32 students. In 2017, only 15.6% of teachers in Dutch primary education were male, which is a considerable difference with the percentage of men in this study (Onderwijs in Cijfers, n.d.). More than half of the Dutch primary school classrooms in 2018 had between 20 and 26 students (Ministerie van Onderwijs Cultuur en Wetenschap, 2019). This means that four of the nine classrooms in this study are smaller than average. It should be noted that three of these classrooms are so-called 'combination groups', with students from different grades in the same class. Not all these grades belonged to the target group of this study and therefore only part of the students participated. The actual class size of these classrooms is higher. However, the respondents in this study are not representative of Dutch primary education and that may have influenced the results.

The teachers and students in this study are not trained in interpreting and scoring the items in the instrument they were asked to use. As discussed before, different interpretations of an item can lead to a lesser degree of inter-rater reliability. The external observer has been the same person in all cases and this should strengthen the reliability of these observations.

4.3.5 Limitations

Cohen's kappa was used to analyse the data in this study. This analysis has some limitations. First, this study has a relatively small number of respondents with nine teachers, 171 students and one observer. As a result, some items were scored by few raters and this could lead to lower kappa coefficients. Second, kappa is a measure for inter-rater reliability and only takes into account actual agreement between scores, not the degree of agreement. For example, it could be argued that there is more agreement between raters when they score the same item '1' and '2', instead of '1' and '4', but Cohen's kappa does not make this distinction. Third, kappa compares the scores of two raters. This is not a problem for the agreement between the observer and teacher in the same case, but for the agreement between a teacher and his students it would require a different kappa coefficient for each student. To overcome this, the mode of the student scores for every instruction group per class was

used. Although the mode is the value that appears most often, it does not include all student scores. For future research it would be interesting to use more complex analysis techniques that take into account the limitations mentioned here.

4.3.6 Future research

During the literature review, it became clear that different operationalisations of differentiation are being used in research. Also, many studies do not clearly indicate how they have operationalised differentiation. For example, studies define differentiation as adapting education to differences between students, without indicating to which student differences teachers should adapt to. For future research into differentiation, it is suggested that researchers should be more aware of the different interpretations of the definition and more specifically record how they operationalise differentiation. In addition, much research into differentiation focuses on the adaptations teachers can make for their students. Several researchers also emphasize the importance of monitoring, to base the adaptations on. At the time of this literature review, these two aspects of differentiation were hardly described together in differentiation research. In view of the different interpretations of the concept of differentiation, future research should also include both aspects of differentiation.

In the second study an instrument was developed to measure the degree of agreement between observers, teachers and students. One of the criteria for the instrument was that it should measure to what extent teachers differentiate their instruction, without judging how well they differentiate. The reason for this was the limited number of studies that show which differentiation strategies are effective. Therefore is remains uncertain what high-quality adaptations are, making it not yet possible to properly measure the quality of teachers' differentiation skills. This is an interesting area for future research, with the increasing attention for teachers' differentiation skills and the ongoing development of instruments claiming to measure differentiation.

References

Bestuursakkoord voor de sector primair onderwijs. (2014).

Bosker, R. J. (2005). *De grenzen van gedifferentieerd onderwijs*. Rijksuniversiteit Groningen.

Chamberlin, M., & Powers, R. (2010). The promise of differentiated instruction for enhancing the mathematical understandings of college students. *Teaching Mathematics and Its Applications*, *29*, 113–139. https://doi.org/10.1093/teamat/hrq006

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 173–206. https://doi.org/10.1080/15366367.2011.626729

Deunk, M., Doolaard, S., Smale-Jacobse, A., & Bosker, R. J. (2015). *Differentiation within and across classrooms: A systematic review of studies into the cognitive effects of differentiation practices*. Groningen.

Dobbelaer, M. J. (2019). *The quality and qualities of classroom observation systems* (Dissertation). https://doi.org/10.3990/1.9789036547161

Gentry, M., & Owen, S. V. (2004). Secondary student perceptions of classroom quality: Instrumentation and differences between advanced/honors and nonhonors classes. *The Journal of Secondary Gifted Education*, *16*(1), 20–29.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*. Retrieved from www.tqsource.org

Graham, S., Morphy, P., Harris, K. R., Fink-Chorzempa, B., Saddler, B., Moran, S., & Mason, L. (2008). Teaching spelling in the primary grades: A national survey of instructional practices and adaptations. *American Educational Research Journal*, *45*(3), 796–825. https://doi.org/10.3102/0002831208319722

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Inspectie van het Onderwijs. (2009). *International comparative analysis of learning and teaching in math lessons in several European countries*.

Inspectie van het Onderwijs. (2012). *De staat van het onderwijs. Onderwijsverslag 2010-2011*.

Inspectie van het Onderwijs. (2015). *De staat van het onderwijs. Onderwijsverslag 2013-2014*.

Inspectie van het Onderwijs. (2016). *De staat van het onderwijs. Onderwijsverslag 2014-2015*.

Jettinghoff, K., & Grootscholte, M. (2014). *Omgaan met verschillen in de klas: Onderzoek naar de professionaliserings- en ondersteuningsbehoeften van leraren bij het omgaan met verschillen tussen leerlingen in de klas*.

Kamminga, A. (2014). *Differentiatie in de rekenles: Zelfinschatting en observatie van leerkrachtgedrag*. Universiteit Utrecht, Utrecht.

Ministerie van Onderwijs Cultuur en Wetenschap. (2019). *Groepsgrootte en leerling-leraarratio in het po 2018*. Retrieved from https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/kamerstukken/2019/06/17/kamerbrief-over-groepsgrootte-en-leerling-leraarratio-in-basisonderwijs-2018/kamerbrief-

over-groepsgrootte-en-leerling-leraarratio-in-basisonderwijs-2018.pdf

Mulder, Q. (2014). *The effect of differentiated instruction on student mathematics achievement in primary school classrooms*. University of Twente.

Nelson, P. M., Demers, J. A., & Christ, T. J. (2014). The Responsive Environmental Assessment for Classroom Teaching (REACT): The dimensionality of student perceptions of the instructional environment. *School Psychology Quarterly*, *29*(2), 182–197. https://doi.org/10.1037/spq0000049

Nieboer, S. (2015). *Literatuurverslag Snappet-onderwijs en gedifferentieerde instructie.*

Onderwijs in Cijfers. (n.d.). Percentage mannen in het onderwijs. Retrieved December 31, 2019, from https://www.onderwijsincijfers.nl/themas/onderwijspersoneel/mannen-in-het-onderwijs/percentage-mannen-in-het-onderwijs

Prast, E. J., Van de Weijer-Bergsma, E., Kroesbergen, E. H., & Van Luit, J. E. H. (2015). Readiness-based differentiation in primary school mathematics: Expert recommendations and teacher self-assessment. *Frontline Learning Research*, *3*(2), 90–116.

Reezigt, G. J. (1993). *Effecten van differentiatie op de basisschool*. RUG/GION.

Riet, S. P. van 't. (1995). *Het didactisch handelen van wiskundedocenten met betrekking tot interne differentiatie*. Vrije Universiteit, Amsterdam.

Rock, M. L., Gregg, M., Ellis, E., & Gable, R. A. (2008). REACH: A Framework for Differentiating Classroom Instruction. *Preventing School Failure: Alternative Education for Children and Youth*, *52*(2), 31–47. https://doi.org/10.3200/PSFL.52.2.31-47

Roiha, A. S. (2014). Teachers' views on differentiation in content and language integrated learning (CLIL): Perceptions, practices and challenges. *Language and Education*, *28*(1), 1–18. https://doi.org/10.1080/09500782.2012.748061

Roy, A., Guay, F., & Valois, P. (2013). Teaching to address diverse learning needs: Development and validation of a Differentiated Instruction Scale. *International Journal of Inclusive Education*, *17*(11), 1186–1204. https://doi.org/10.1080/13603116.2012.743604

Ruys, I., Defruyt, S., Rots, I., & Aelterman, A. (2013). Differentiated instruction in teacher education: A case study of congruent teaching. *Teachers and Teaching*, *19*(1), 93–107. https://doi.org/10.1080/13540602.2013.744201

Tomlinson, C. A. (1999). Mapping a Route Toward Differentiated Instruction. *Educational Leadership*, *57*(1), 12.

Tomlinson, C. A. (2000). Reconcilable Differences? Standards-Based Teaching and Differentiation. *Educational Leadership*, *58*(1), 6–11.

Tomlinson, C. A. (2001). Grading for Success. *Educational Leadership*, *58*(6), 12–15.

Tomlinson, C. A. (2003). Deciding to Teach Them All. *Educational Leadership*, *61*(2), 6–11.

Tomlinson, C. A. (2004). Sharing responsibility for differentiating instruction. *Roeper Review*, *26*(4), 188–189. https://doi.org/10.1080/02783190409554268

Tomlinson, C. A., Brighton, C., Callahan, C. M., Moon, T. R., Brimijoin, K., Conover, L. a, & Reynolds, T. (2003). Differentiating Instruction in Response to Student Readiness, Interest, and Learning Profile in Academically Diverse Classrooms: A Review of Literature. *Journal for the Education of the Gifted*, *27*(2/3), 119–145. https://doi.org/10.1177/016235320302700203

Tomlinson, C. A., Brimijoin, K., & Narvaez, L. (2008). *The Differentiated School: Making Revolutionary Changes in Teaching and Learning*. Retrieved from https://books.google.com/books?hl=nl&lr=&id=cW5PQ7l38lEC&pgis=1

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. https://doi.org/10.1080/00131880701369651

Van de Weijer-Bergsma, E., Van Luit, J. E. H., Prast, E. J., Kroesbergen, E. H., Kaskens, J., Compagnie-Rietberg, C., … Logtenberg, H. (2016). *Differentieren in het rekenonderwijs: Hoe doe je dat in de praktijk?* Doetinchem: Graviant scientific & educational books.

Van der Scheer, E. A., Glas, C. A. W., & Visscher, A. J. (2016). *The effects of a data-based decision making intervention on students ' perceptions of teaching quality*.

van Geel, M., Keuning, T., Frèrejean, J., Dolmans, D., van Merriënboer, J., & Visscher, A. J. (2018). Capturing the complexity of differentiated instruction. *School Effectiveness and School Improvement*, *30*(1), 51–67. https://doi.org/10.1080/09243453.2018.1539013

Van Nus, M. (2015). *De psychometrische kwaliteiten van het STIP instrument: Differentiatie in het basisonderwijs*. University of Twente, Enschede.

VanTassel-Baska, J., Feng, A. X., Brown, E., Bracken, B., Stambaugh, T., French, H., … Bai, W. (2008). A study of differentiated instructional change over 3 years. *Gifted Child Quarterly*, *52*(4), 297–312. https://doi.org/10.1177/0016986208321809

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

Whipple, K. A. (2012). *Differentiated instruction: A survey study of teacher understanding and implementation in a southeast Massachusetts school district*. Northeastern University, Boston.

Williams, R. T., Swanlund, A., Miller, S., Konstantopoulos, S., Eno, J., van der Ploeg, A., & Meyers, C. (2014). Measuring Instructional Differentiation in a Large-Scale Experiment. *Educational and Psychological Measurement*, *74*(2), 263–279. https://doi.org/10.1177/0013164413507724

Appendix A. Overview of instruments for measuring differentiation

Instrument	Type	Level	Subject	Items	Item format	Scales	Scales name	Reference
Differentiatie in Mathematische Instructie (DMI)	Observation	Elementary	Differentiation in math lessons	BA: 5 Other: ?	5-point Likert scale	4	Breed aanbod (BA), Klassenmanagement, Differentiatie voor zwakke rekenaars, Differentiatie voor sterk rekenaars	(Kamminga, 2014)
Kijkwijzer Differentiëren in de rekenles	Observation	Elementary	Differentiation in math instruction	52	5-point Likert scale	3	Klassikale instructie; Verlengde instructie, preteaching en begeleide inoefening; Subgroepinstructie voor leerlingen met sterke rekenvaardigheden	(Van de Weijer-Bergsma et al., 2016)
International Comparative Analysis of Learning and Teaching (ICALT)	Observation	Elementary	Learning and teaching in math lessons	32 (**4**)	4-point Likert scale	5 (**1**)	Safe and stimulating learning climate, Clear and activating instruction, Classroom management, **Adaptation of teaching to diverse needs of students**, Teaching learning strategies	(Inspectie van het Onderwijs, 2009; Van de Grift, 2007)
Observatie-instrument voor afstemmen van instructie en verwerking op verschillen	Observation	Elementary (groep 4)	Differentiation in math instruction	7	4-point Likert scale	X	X	(Mulder, 2014)
STIP	Observation	Elementary (groep 5-8)	Differentiation in math lessons	33	Observed or not observed	4	Differentiatie in inhoud, Differentiatie in taak, Differentiatie in proces, Differentiatie in leeromgeving	(Van Nus, 2015)

Observatieschema	Observation	Elementary	Differentiation with Snappet	8 (**6**)	5-point Likert scale	4	Differentiatie in instructie en verwerking, Differentiatie in instructie, Differentiatie in verwerking, Flexibele klasorganisatie	(Nieboer, 2015)
Classroom Observation Scale-Revised (COS-R) Student Observation Scale (SOS)	Observation	Elementary	Teachers' instructional practice Student engagement behaviors	2x 25 (**4**)	3-point Likert scale	6 (**1**)	Curriculum planning and delivery, **Accommodation for individual differences**, Problem solving strategies, Critical thinking strategies, Creative thinking strategies, Research strategies	(VanTassel-Baska et al., 2008)
Differentiation Self-Assessment Questionnaire (DSAQ)	Teacher self-assessment	Elementary	Differentiation in math lessons	33	5-point Likert scale	5	Identification of educational needs, Differentiated goals, Differentiated instruction, Differentiated practice, Evaluation of progress and process	(Prast et al., 2015)
Differentiated Instruction Scale (DIS)	Teacher self-assessment	Elementary	Differentiation in math and French lessons	12	5-point Likert scale	2	Instructional adaptations, Academic progress monitoring	(Roy et al., 2013)
Interne differentiatie – lerarenversie (ID-LK) + leerlingenversie (ID-LL)	Teacher self-assessment + Student survey	Secondary	Differentiation in math lessons	45	5-point Likert scale	3	Procesgeoriënteerd interne-differentiatiegedrag, Productgeoriënteerd interne-differentiatiegedrag, Non-differentiatiegedrag	(Riet, 1995)
REACH	Teacher self-assessment	Elementary	Differentiation	56	Strength or Need	5	Teacher (reflect), Content (evaluate), Learner (analyze), Instruction (craft), Assessment (hone)	(Rock et al., 2008)

Differentiated Instruction Survey	Teacher self-assessment	Elementary	Differentiation	26	4-point Likert scale	6	Student interest, Assessment, Lesson planning, Content, Process, Product	(Whipple, 2012)
Survey	Teacher self-assessment	Elementary	Instructional practices and adaptations in spelling	21	7-point Likert scale	X	X	(Graham et al., 2008)
Responsive Environmental Assessment for Classroom Teaching (REACT)	Student survey	Secondary	Student perceptions of the classroom teaching environment	27 (**5**)	4-point Likert scale	6 (**1**)	Positive reinforcement, Instructional presentation, Goal setting, **Differentiated instruction**, Formative feedback, Instructional enjoyment	(Nelson et al., 2014)
Classroom Survey	Student survey	Secondary	Student perceptions of differentiation in math lessons	14	6-point Likert scale	X	X	(Chamberlin & Powers, 2010)
Student perception questionnaire	Student survey	Elementary	Student perceptions of teaching quality	36 (**15**)	5-point Likert scale	5 (**2**)	Classroom climate, Classroom management, **Clear instruction**, **Challenging students**, Goal-orientation	(Van der Scheer et al., 2016)
Student Perceptions of Classroom Quality (SPOCQ)	Student survey	Secondary	Student perceptions of class activities	38 (**14**)	5-point Likert scale	5 (**2**)	Appeal, **Challenge**, **Choice**, Meaningfulness, Self-efficacy	(Gentry & Owen, 2004)

Appendix B. Dutch items from instrument

Category	Observation item	Teacher item	Student item
Adapting			
Goals	De leerkracht heeft, naast het basislesdoel, één of meerdere aanvullende lesdoelen opgesteld voor specifieke (groepen) leerlingen.	Ik heb, naast het basislesdoel, één of meerdere aanvullende lesdoelen opgesteld voor specifieke (groepen) leerlingen.	Ik heb deze les precies hetzelfde geleerd als de andere kinderen uit de klas.
Materials	De leerkracht zet geschikt materiaal in bij leerlingen die dit nodig hebben.	Ik heb geschikt materiaal ingezet bij leerlingen die dit nodig hadden.	Dat de juf/meester materiaal heeft gebruikt om mij iets uit te leggen, heeft mij geholpen om het beter te begrijpen.
ICT	De leerkracht zet ICT in om leerlingen op hun eigen niveau te laten oefenen.	Ik heb ICT ingezet om leerlingen op hun eigen niveau te laten oefenen.	Ik heb rekensommen geoefend op de computer of tablet die ik moeilijk vind.
Adapting instruction			
Low-achieving and high-achieving	De leerkracht geeft extra instructie op het niveau van de leerling.	Ik heb extra instructie gegeven op het niveau van de leerling.	De extra uitleg heeft mij geholpen om het beter te begrijpen.
Whole-class	De leerkracht legt de verbinding tussen de handelingsniveaus waarop de leerlingen handelen en het niveau waarop tijdens de les gehandeld moet gaan worden.	Ik heb de verbinding gelegd tussen de handelingsniveaus waarop de leerlingen handelen en het niveau waarop tijdens de les gehandeld moest gaan worden.	Mijn juf/meester heeft op verschillende manieren uitgelegd (bijvoorbeeld door te vertellen, een model te laten zien, ons iets te laten doen, enz.)
Time	De leerkracht stemt de instructietijd af op wat elke leerling nodig heeft om het lesdoel te bereiken.	Ik heb de instructietijd afgestemd op wat elke leerling nodig heeft om het lesdoel te bereiken.	Ik heb genoeg uitleg gekregen om de les te kunnen begrijpen (dus niet te veel en niet te weinig uitleg).
Adapting practice			
Difficulty and amount	De leerkracht stemt de opdrachten (hoeveelheid en moeilijkheid) af op wat elke leerling nodig heeft	Ik heb de opdrachten (hoeveelheid en moeilijkheid) afgestemd op wat elke leerling	De opdrachten die ik moest maken waren van mijn niveau (dus niet te moeilijk en niet te makkelijk).

	om het lesdoel te bereiken.	nodig heeft om het lesdoel te bereiken.	EN Ik heb genoeg opdrachten gekregen om te kunnen oefenen (dus niet te veel en niet te weinig opdrachten).
Monitoring			
Monitoring during the lesson	De leerkracht controleert het begrip van de leerlingen tijdens de basisinstructie.	Ik heb het begrip van de leerlingen gecontroleerd tijdens de basisinstructie.	Mijn juf/meester had tijdens de uitleg aan de hele klas door of ik het wel of niet begreep.
	De leerkracht controleert het begrip van de leerlingen tijdens de verlengde instructie.	Ik heb het begrip van de leerlingen gecontroleerd tijdens de verlengde instructie.	Mijn juf/meester had tijdens de extra uitleg door of ik het wel of niet begreep.
	De leerkracht controleert het begrip van de leerlingen tijdens de verrijkingsinstructie.	Ik heb het begrip van de leerlingen gecontroleerd tijdens de verrijkingsinstructie.	Mijn juf/meester had tijdens de extra uitleg door of ik het wel of niet begreep.
	De leerkracht controleert de voortgang van de leerlingen tijdens de verwerking van de opdrachten.	Ik heb de voortgang van de leerlingen gecontroleerd tijdens de verwerking van de opdrachten.	Mijn juf/meester kwam tijdens het werken langs om te kijken of ik het begrijp.
	De leerkracht evalueert aan het einde van de les of alle leerlingen het lesdoel hebben behaald.	Ik heb aan het einde van de les geëvalueerd of alle leerlingen het lesdoel hebben behaald.	Mijn juf/meester weet nu of ik de les goed of slecht heb begrepen.

Appendix C. Instrument observer version

Groep:	School:	Datum:
Leerkracht:		# lln:

Instructiegroepen:	# lln	Leerling geobserveerd:

Opmerkingen:

Observatieformulier

STELLING 1: De leerkracht heeft, naast het basislesdoel, één of meerdere <u>aanvullende lesdoelen</u> opgesteld voor specifieke (groepen) leerlingen.

VOORBEELD De leerkracht introduceert het basislesdoel aan het begin van de les. Voor de intensieve groep geldt hetzelfde lesdoel. De verrijkingsgroep werkt elke dag zelfstandig aan verrijkingsstof uit het plusboek, maar hiervoor heeft de leerkracht geen lesdoel opgesteld.	1 — 2 — 3 — 4	VOORBEELD De leerkracht introduceert het basislesdoel aan het begin van de les. Voor leerlingen die uitvallen op een specifiek onderdeel is een reparatiedoel opgesteld. Voor de verrijkingsgroep heeft hij een verrijkend (of verdiepend) lesdoel, passend bij de stof uit het plusboek.

Basisgroep	❍	❍	❍	❍
Intensieve groep	❍	❍	❍	❍
Verrijkingsgroep	❍	❍	❍	❍

STELLING 2: De leerkracht stemt de instructietijd af op wat elke leerling nodig heeft om het lesdoel te bereiken.

VOORBEELD		VOORBEELD
De leerkracht laat alle leerlingen de basisinstructie bij een herhalingsles volgen. De verrijkingsgroep werkt zelfstandig aan de verrijkingsstof, hier krijgen ze geen instructie bij.	1 — 2 — 3 — 4	De verrijkingsgroep hoeft maar een deel van de basisinstructie te volgen, daarna gaan zij zelfstandig aan het werk. Zij krijgen later een extra instructie over de verrijkingsstof. De leerkracht geeft enkele leerlingen die moeite hebben met het lesdoel een verlengde instructie.

Basisgroep	❍	❍	❍	❍	
Intensieve groep	❍	❍	❍	❍	
Verrijkingsgroep	❍	❍	❍	❍	

STELLING 3: De leerkracht geeft extra instructie <u>op het niveau van de leerling</u>.

VOORBEELD		VOORBEELD
De verlengde instructie is een herhaling van de basisinstructie, waarbij de leerkracht en leerling(en) de sommen samen maken.	1 — 2 — 3 — 4	De leerkracht leert leerlingen in de intensieve groep het gebruik van één strategie aan, gebruikt concreet materiaal en geeft hen veel positieve feedback en complimenten. Met leerlingen in de verrijkingsgroep bespreekt hij verrijkings-stof, hij leert hen meerdere strategieën en richt zich vooral op het rekenproces.

Basisgroep	❍	❍	❍	❍	❍ geen extra instructie gegeven
Intensieve groep	❍	❍	❍	❍	❍ geen extra instructie gegeven
Verrijkingsgroep	❍	❍	❍	❍	❍ geen extra instructie gegeven

STELLING 4: De leerkracht legt de verbinding tussen de handelingsniveaus waarop de leerlingen handelen en het niveau waarop tijdens de les gehandeld moet gaan worden.

VOORBEELD		VOORBEELD
In een les over vermenigvuldigen waarin de leerlingen kale keersommen moeten maken, bespreekt de leerkracht alleen het formeel handelen.	1 —— 2 —— 3 —— 4	In een les over vermenigvuldigen waarin de leerlingen kale keersommen moeten maken (formeel handelen), laat de leerkracht ook de getallenlijn uit een eerdere les nog een keer zien, waarop herhaald opgeteld kan worden (abstract handelen) omdat hij weet dat sommige leerlingen nog moeite hebben met de kale sommen. Hij maakt een duidelijke link tussen beide situaties.

Basisgroep	❍	❍	❍	❍	❍ geen instructie gegeven
Intensieve groep	❍	❍	❍	❍	❍ geen instructie gegeven
Verrijkingsgroep	❍	❍	❍	❍	❍ geen instructie gegeven

STELLING 5: De leerkracht zet geschikt materiaal in <u>bij leerlingen die dit nodig hebben</u>.

VOORBEELD		VOORBEELD
De leerkracht laat alle leerlingen instructiegeld gebruiken bij een les waarin het rekenen met geld wordt herhaald.	1 —— 2 —— 3 —— 4	De leerkracht gebruikt instructiegeld tijdens de verlengde instructie voor een leerling die moeite heeft met rekenen met geld.

Basisgroep	❍	❍	❍	❍	❍ geen materiaal gebruikt
Intensieve groep	❍	❍	❍	❍	❍ geen materiaal gebruikt
Verrijkingsgroep	❍	❍	❍	❍	❍ geen materiaal gebruikt

STELLING 6: De leerkracht stemt de opdrachten (hoeveelheid en moeilijkheid) af op wat elke leerling nodig heeft om het lesdoel te bereiken.

VOORBEELD		VOORBEELD
Alle leerlingen maken de opdrachten uit de basisstof.	1 — 2 — 3 — 4	De leerkracht heeft voor de verrijkingsgroep de basisstof gecompact en geeft hen verdiepende of verrijkende opdrachten (moeilijkheid). Voor de leerlingen in de intensieve groep heeft hij de basisopdrachten geselecteerd, zodat zij zich kunnen focussen op het behalen van het lesdoel (hoeveelheid).

Basisgroep	❍	❍	❍	❍	
Intensieve groep	❍	❍	❍	❍	
Verrijkingsgroep	❍	❍	❍	❍	

STELLING 7: De leerkracht zet ICT in om leerlingen op hun eigen niveau te laten oefenen.

VOORBEELD		VOORBEELD
De leerkracht laat alle leerlingen met dezelfde software/app werken. Hij heeft voor alle leerlingen dezelfde opgaven klaargezet en het programma is niet adaptief.	1 — 2 — 3 — 4	De leerkracht laat de leerlingen met verschillende software/apps werken. Voor sommige leerlingen heeft hij specifieke opgaven klaargezet, andere leerlingen werken met een adaptief programma dat opgaven biedt op het niveau van de leerling.

Basisgroep	❍	❍	❍	❍	❍ geen ICT gebruikt
Intensieve groep	❍	❍	❍	❍	❍ geen ICT gebruikt
Verrijkingsgroep	❍	❍	❍	❍	❍ geen ICT gebruikt

STELLING 8: De leerkracht controleert het begrip van de leerlingen tijdens de <u>basisinstructie</u>.

VOORBEELD		VOORBEELD
De leerkracht geeft tijdens zijn instructie twee leerlingen de beurt om het goede antwoord op een oefensom te geven.	1 — 2 — 3 — 4	De leerkracht laat de leerlingen tijdens de instructie oefensommen uitwerken op een wisbordje, waardoor hij snel kan zien welke leerlingen het begrepen hebben en welke leerlingen niet.

Basisgroep	❍	❍	❍	❍	❍ n.v.t.
Intensieve groep	❍	❍	❍	❍	❍ n.v.t.
Verrijkingsgroep	❍	❍	❍	❍	❍ n.v.t.

STELLING 9: De leerkracht controleert het begrip van de leerlingen tijdens de <u>verlengde instructie</u>.

VOORBEELD		VOORBEELD
De leerkracht legt uit, de leerlingen luisteren en gaan daarna zelfstandig aan het werk.	1 — 2 — 3 — 4	De leerkracht legt uit en stelt tussendoor veel vragen, zodat hij een goed beeld heeft van wat de leerlingen al beheersen en wat nog niet.

Intensieve groep	❍	❍	❍	❍	❍ n.v.t.

STELLING 10: De leerkracht controleert het begrip van de leerlingen tijdens de <u>verrijkingsinstructie</u>.

VOORBEELD		VOORBEELD
De leerkracht legt uit, de leerlingen luisteren en gaan daarna zelfstandig aan het werk.	1 — 2 — 3 — 4	De leerkracht legt uit en stelt tussendoor veel vragen, zodat hij een goed beeld heeft van wat de leerlingen al beheersen en wat nog niet.

Verrijkingsgroep	❍	❍	❍	❍	❍ n.v.t.

STELLING 11: De leerkracht controleert de voortgang van de leerlingen tijdens de verwerking van de opdrachten.

VOORBEELD		VOORBEELD
De leerkracht komt tijdens het werken alleen aan tafel bij de leerlingen die een vraag hebben.	1 — 2 — 3 — 4	De leerkracht loopt tijdens het werken meerdere rondes door de klas en kijkt bij elke leerling even mee, om een beeld te krijgen van zijn/haar voortgang.

Basisgroep	❍	❍	❍	❍
Intensieve groep	❍	❍	❍	❍
Verrijkingsgroep	❍	❍	❍	❍

STELLING 12: De leerkracht evalueert aan het einde van de les of alle leerlingen het lesdoel hebben behaald.

VOORBEELD		VOORBEELD
De leerkracht sluit de les af door te zeggen dat de rekenspullen opgeruimd mogen worden en met een compliment over hoe rustig er gewerkt is.	1 — 2 — 3 — 4	De leerkracht doet samen met de leerlingen nog een som die bij het lesdoel past, waarin de leerlingen de geleerde strategie nog eens toepassen. De leerkracht controleert welke leerlingen in staat zijn de som op de juiste manier op te lossen en wie niet.

Basisgroep	❍	❍	❍	❍
Intensieve groep	❍	❍	❍	❍
Verrijkingsgroep	❍	❍	❍	❍

Appendix D. Instrument teacher version

Groep: School: Datum:

**Je gegevens worden anoniem verwerkt, maar moeten wel gekoppeld kunnen worden aan de antwoorden van jouw klas.*

12 stellingen over de rekenles die je zojuist hebt gegeven

Geef per stelling voor elke instructiegroep aan, welke score (1 - 4) het beste past bij je handelen tijdens de rekenles. Bij score 1 en 4 hoort steeds een voorbeeld.

Let op: beoordeel de ***STELLING*** *(niet hoe goed het voorbeeld bij je past).*

STELLING 1: Ik heb, naast het basislesdoel, één of meerdere aanvullende lesdoelen opgesteld voor specifieke (groepen) leerlingen.

VOORBEELD bij score 1		VOORBEELD bij score 4
De leerkracht introduceert het basislesdoel aan het begin van de les. Voor de intensieve groep geldt hetzelfde lesdoel. De verrijkingsgroep werkt elke dag zelfstandig aan verrijkingsstof uit het plusboek, maar hiervoor heeft de leerkracht geen lesdoel opgesteld.	1 — 2 — 3 — 4	De leerkracht introduceert het basislesdoel aan het begin van de les. Voor leerlingen die uitvallen op een specifiek onderdeel is een reparatiedoel opgesteld. Voor de verrijkingsgroep heeft hij een verrijkend (of verdiepend) lesdoel, passend bij de stof uit het plusboek.

In hoeverre geldt bovenstaande stelling voor de leerlingen in de:

Basisgroep	❍	❍	❍	❍
Intensieve groep	❍	❍	❍	❍
Verrijkingsgroep	❍	❍	❍	❍

STELLING 2: Ik heb de instructietijd afgestemd op wat elke leerling nodig heeft om het lesdoel te bereiken.

VOORBEELD		VOORBEELD
De leerkracht laat alle leerlingen de basisinstructie bij een herhalingsles volgen. De verrijkingsgroep werkt zelfstandig aan de verrijkingsstof, hier krijgen ze geen instructie bij.	1 — 2 — 3 — 4	De verrijkingsgroep hoeft maar een deel van de basisinstructie te volgen, daarna gaan zij zelfstandig aan het werk. Zij krijgen later een extra instructie over de verrijkingsstof. De leerkracht geeft enkele leerlingen die moeite hebben met het lesdoel een verlengde instructie.

Basisgroep	❍	❍	❍	❍	
Intensieve groep	❍	❍	❍	❍	
Verrijkingsgroep	❍	❍	❍	❍	

STELLING 3: Ik heb extra instructie gegeven <u>op het niveau van de leerling</u>.

VOORBEELD		VOORBEELD
De verlengde instructie is een herhaling van de basisinstructie, waarbij de leerkracht en leerling(en) de sommen samen maken.	1 — 2 — 3 — 4	De leerkracht leert leerlingen in de intensieve groep het gebruik van één strategie aan, gebruikt concreet materiaal en geeft hen veel positieve feedback en complimenten. Met leerlingen in de verrijkingsgroep bespreekt hij verrijkings-stof, hij leert hen meerdere strategieën en richt zich vooral op het rekenproces.

Basisgroep	❍	❍	❍	❍	❍ geen extra instructie gegeven
Intensieve groep	❍	❍	❍	❍	❍ geen extra instructie gegeven
Verrijkingsgroep	❍	❍	❍	❍	❍ geen extra instructie gegeven

STELLING 4: Ik heb de verbinding gelegd tussen de handelingsniveaus waarop de leerlingen handelen en het niveau waarop tijdens de les gehandeld moest gaan worden.

VOORBEELD		VOORBEELD
In een les over vermenigvuldigen waarin de leerlingen kale keersommen moeten maken, bespreekt de leerkracht alleen het formeel handelen.	1 — 2 — 3 — 4	In een les over vermenigvuldigen waarin de leerlingen kale keersommen moeten maken (formeel handelen), laat de leerkracht ook de getallenlijn uit een eerdere les nog een keer zien, waarop herhaald opgeteld kan worden (abstract handelen) omdat hij weet dat sommige leerlingen nog moeite hebben met de kale sommen. Hij maakt een duidelijke link tussen beide situaties.

Basisgroep	❍	❍	❍	❍	❍ geen instructie gegeven
Intensieve groep	❍	❍	❍	❍	❍ geen instructie gegeven
Verrijkingsgroep	❍	❍	❍	❍	❍ geen instructie gegeven

STELLING 5: Ik heb geschikt materiaal ingezet <u>bij leerlingen die dit nodig hadden</u>.

VOORBEELD		VOORBEELD
De leerkracht laat alle leerlingen instructiegeld gebruiken bij een les waarin het rekenen met geld wordt herhaald.	1 — 2 — 3 — 4	De leerkracht gebruikt instructiegeld tijdens de verlengde instructie voor een leerling die moeite heeft met rekenen met geld.

Basisgroep	❍	❍	❍	❍	❍ geen materiaal gebruikt
Intensieve groep	❍	❍	❍	❍	❍ geen materiaal gebruikt
Verrijkingsgroep	❍	❍	❍	❍	❍ geen materiaal gebruikt

STELLING 6: Ik heb de opdrachten (hoeveelheid en moeilijkheid) afgestemd op wat elke leerling nodig heeft om het lesdoel te bereiken.

VOORBEELD		VOORBEELD
Alle leerlingen maken de opdrachten uit de basisstof.	1 — 2 — 3 — 4	De leerkracht heeft voor de verrijkingsgroep de basisstof gecompact en geeft hen verdiepende of verrijkende opdrachten (moeilijkheid). Voor de leerlingen in de intensieve groep heeft hij de basisopdrachten geselecteerd, zodat zij zich kunnen focussen op het behalen van het lesdoel (hoeveelheid).

Basisgroep	❍	❍	❍	❍	
Intensieve groep	❍	❍	❍	❍	
Verrijkingsgroep	❍	❍	❍	❍	

STELLING 7: Ik heb ICT ingezet om leerlingen op hun eigen niveau te laten oefenen.

VOORBEELD		VOORBEELD
De leerkracht laat alle leerlingen met dezelfde software/app werken. Hij heeft voor alle leerlingen dezelfde opgaven klaargezet en het programma is niet adaptief.	1 — 2 — 3 — 4	De leerkracht laat de leerlingen met verschillende software/apps werken. Voor sommige leerlingen heeft hij specifieke opgaven klaargezet, andere leerlingen werken met een adaptief programma dat opgaven biedt op het niveau van de leerling.

Basisgroep	❍	❍	❍	❍	❍ geen ICT gebruikt
Intensieve groep	❍	❍	❍	❍	❍ geen ICT gebruikt
Verrijkingsgroep	❍	❍	❍	❍	❍ geen ICT gebruikt

STELLING 8: Ik heb het begrip van de leerlingen gecontroleerd tijdens de <u>basisinstructie</u>.

VOORBEELD		VOORBEELD
De leerkracht geeft tijdens zijn instructie twee leerlingen de beurt om het goede antwoord op een oefensom te geven.	1 — 2 — 3 — 4	De leerkracht laat de leerlingen tijdens de instructie oefensommen uitwerken op een wisbordje, waardoor hij snel kan zien welke leerlingen het begrepen hebben en welke leerlingen niet.

Basisgroep	❍	❍	❍	❍	❍ n.v.t.
Intensieve groep	❍	❍	❍	❍	❍ n.v.t.
Verrijkingsgroep	❍	❍	❍	❍	❍ n.v.t.

STELLING 9: Ik heb het begrip van de leerlingen gecontroleerd tijdens de <u>verlengde instructie</u>.

VOORBEELD		VOORBEELD
De leerkracht legt uit, de leerlingen luisteren en gaan daarna zelfstandig aan het werk.	1 — 2 — 3 — 4	De leerkracht legt uit en stelt tussendoor veel vragen, zodat hij een goed beeld heeft van wat de leerlingen al beheersen en wat nog niet.

Intensieve groep	❍	❍	❍	❍	❍ n.v.t.

STELLING 10: Ik heb het begrip van de leerlingen gecontroleerd tijdens de <u>verrijkingsinstructie</u>.

VOORBEELD		VOORBEELD
De leerkracht legt uit, de leerlingen luisteren en gaan daarna zelfstandig aan het werk.	1 — 2 — 3 — 4	De leerkracht legt uit en stelt tussendoor veel vragen, zodat hij een goed beeld heeft van wat de leerlingen al beheersen en wat nog niet.

Verrijkingsgroep	❍	❍	❍	❍	❍ n.v.t.

STELLING 11: Ik heb de voortgang van de leerlingen gecontroleerd tijdens de verwerking van de opdrachten.

VOORBEELD		VOORBEELD
De leerkracht komt tijdens het werken alleen aan tafel bij de leerlingen die een vraag hebben.	1 —— 2 —— 3 —— 4	De leerkracht loopt tijdens het werken meerdere rondes door de klas en kijkt bij elke leerling even mee, om een beeld te krijgen van zijn/haar voortgang.

Basisgroep	❍	❍	❍	❍
Intensieve groep	❍	❍	❍	❍
Verrijkingsgroep	❍	❍	❍	❍

STELLING 12: Ik heb aan het einde van de les geëvalueerd of alle leerlingen het lesdoel hebben behaald.

VOORBEELD		VOORBEELD
De leerkracht sluit de les af door te zeggen dat de rekenspullen opgeruimd mogen worden en met een compliment over hoe rustig er gewerkt is.	1 —— 2 —— 3 —— 4	De leerkracht doet samen met de leerlingen nog een som die bij het lesdoel past, waarin de leerlingen de geleerde strategie nog eens toepassen. De leerkracht controleert welke leerlingen in staat zijn de som op de juiste manier op te lossen en wie niet.

Basisgroep	❍	❍	❍	❍
Intensieve groep	❍	❍	❍	❍
Verrijkingsgroep	❍	❍	❍	❍

Dit waren de stellingen. Hartelijk dank voor het invullen!

Heeft u interesse in:

❍ de uitslagen van uw klas, en/of

❍ de uitslagen van het hele onderzoek

Vul dan hier u e-mailadres in:

_____

Appendix E. Instrument student version

Naam:

Groep:

Naam juf/meester:

Vragen over de rekenles

Beantwoord elke vraag door het ❍ in te kleuren dat het beste bij jou past.

De vragen gaan over **<u>jou</u>**

1.	**Ik heb deze les precies hetzelfde geleerd als de andere kinderen uit de klas.**			
❍ klopt helemaal niet	❍ klopt een beetje niet	❍ klopt een beetje	❍ klopt helemaal	

2.	**Mijn juf/meester had tijdens de <u>uitleg aan de hele klas</u> door of ik het wel of niet begreep.**				
❍ klopt helemaal niet	❍ klopt een beetje niet	❍ klopt een beetje	❍ klopt helemaal	❍ Er was geen uitleg aan de hele klas

3. Ik heb <u>extra</u> uitleg gekregen.

❍ Nee *(ga naar vraag 4)*

❍ Ja:

De extra uitleg heeft mij geholpen om het beter te begrijpen.

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

Mijn juf/meester had tijdens de extra uitleg door of ik het wel of niet begreep.

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

4. Mijn juf/meester heeft op verschillende manieren uitgelegd, <u>bijvoorbeeld</u>:

- **kale sommen uitleggen**
- **model laten zien →**
- **ons iets laten doen**

❍ Nee *(ga naar vraag 5)*

❍ Ja:

Dat de juf/meester op verschillende manieren heeft uitgelegd, heeft mij geholpen om het beter te begrijpen.

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

5. Ik heb genoeg uitleg gekregen om de les te kunnen begrijpen (dus niet te veel en niet te weinig uitleg).

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

6. Mijn juf/meester heeft materiaal gebruikt om mij iets uit te leggen.

❍ Nee *(ga naar vraag 7)*

❍ Ja:

Dat de juf/meester materiaal heeft gebruikt om mij iets uit te leggen, heeft mij geholpen om het beter te begrijpen.

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

7. De opdrachten die ik moest maken waren van mijn niveau (dus niet te moeilijk en niet te makkelijk).

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

8. Ik heb genoeg opdrachten gekregen om te kunnen oefenen (dus niet te veel en niet te weinig opdrachten).

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

9. Ik heb rekensommen geoefend op de computer of tablet.

❍ Nee *(ga naar vraag 10)*

❍ Ja:

Ik heb rekensommen geoefend op de computer of tablet die ik moeilijk vind.

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

10. Mijn juf/meester kwam tijdens het werken langs om te kijken of ik het begrijp.

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

11. Mijn juf/meester weet nu of ik de les goed of slecht heb begrepen.

❍	❍	❍	❍
klopt helemaal niet	klopt een beetje niet	klopt een beetje	klopt helemaal

Dat waren de vragen!