Learning and Instruction 54 (2018) 22–34

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Differentiated instruction in primary mathematics: Effects of teacher professional development on student achievement

Emilie J. Prast*, Eva Van de Weijer-Bergsma, Evelyn H. Kroesbergen, Johannes E.H. Van Luit

Utrecht University, Department of Pedagogical and Educational Sciences, P.O. Box 80140, 3508 TC Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Differentiated instruction
Adaptive teaching
Professional development
Achievement
Mathematics

ABSTRACT

This large-scale study examined the effects of a teacher professional development (PD) programme about differentiated instruction on students' mathematics achievement. Thirty primary schools (N = 5658 students of grade 1–6) divided over three cohorts participated: Cohort 1 received the PD programme in Year 1, Cohort 2 in Year 2, and Cohort 3 was control. During the PD, teachers learned how to adapt mathematics education to diverse educational needs using within-class ability groups. In Year 1, the PD had a significant small positive effect on student achievement growth. The effect size was similar for low-achieving, average-achieving and high-achieving students. In Year 2, no significant effects were demonstrated. In sum, teacher PD about differentiation has the potential to promote the achievement of all students. However, implementing differentiation is not straightforward and future research is necessary to unravel which factors make PD about differentiation succeed.

1. Introduction

Primary school classrooms are traditionally diverse in terms of the academic ability and achievement level of the students. With the current movement towards inclusion of children with special educational needs in general education classrooms, the range of ability and achievement levels is continuously increasing, as are the specific educational needs associated with these. Differentiation, i.e. the adaptation of instruction to students' different educational needs, is often promoted as a solution for responding to this diversity. In this study, we investigate whether teacher professional development (PD) about differentiation has a positive effect on student achievement in primary school mathematics.

1.1. Definitions: differentiation, ability grouping, and adaptive teaching competency

Roy, Guay, and Valois (2013, p.1187) define differentiated instruction as 'an approach by which teaching is varied and adapted to match students' abilities using systematic procedures for academic progress monitoring and data-based decision-making.' Thus, the focus is on differentiation based on students' current achievement level, also called cognitive or readiness-based differentiation. According to this definition, teachers should monitor students' academic progress to identify students' educational needs and then adapt instruction to these needs. The way in which progress is monitored and the nature of instructional adaptations can vary substantially, and various organisational formats can be used (e.g. individual or group-based; see Prast, Van de Weijer-Bergsma, Kroesbergen, and Van Luit (2015) for a discussion of this issue).

One frequently used way to organise differentiation is homogeneous within-class ability grouping (hereafter: ability grouping), in which students of similar academic ability or (current) achievement level are placed together in subgroups within the heterogeneous classroom (Tieso, 2003). Ability grouping is not synonymous to differentiation: it is an organisational format that can be used to implement differentiation, provided that instruction and practice are indeed adapted to the educational needs of the different ability groups.

A related term for adapting instruction to students' educational needs is adaptive teaching. A distinction is made between macro-adaptations (planned adaptations, e.g. pre-designed tasks at various levels of difficulty for low-achieving and high-achieving students) and micro-adaptations (spontaneous adaptations in direct response to students' needs; Corno, 2008). The term 'differentiation' seems to be more commonly used for macro-adaptations, whereas 'adaptive teaching' is more commonly used for micro-adaptations. However, the construct of 'adaptive teaching competency' (Vogt & Rogalla, 2009) does include both adaptive planning competency (teachers' capacity to plan adaptations beforehand; macro-adaptivity) and adaptive implementation competency (teachers' capacity for making adaptations on the spot; micro-adaptivity). In this article, we use 'differentiation' to refer to the process of monitoring progress and making instructional adaptations as

[*] Corresponding author.
E-mail addresses: e.prast@uu.nl (E.J. Prast), e.vandeweijer@uu.nl (E. Van de Weijer-Bergsma), e.h.kroesbergen@uu.nl (E.H. Kroesbergen), j.e.h.vanluit@uu.nl (J.E.H. Van Luit).

https://doi.org/10.1016/j.learninstruc.2018.01.009
Received 22 April 2016; Received in revised form 23 May 2017; Accepted 28 January 2018
Available online 10 February 2018

defined by Roy et al. (2013). In line with Vogt and Rogalla (2009), we use ‘adaptive teaching competency’ to refer to teachers' capacities for making both planned and spontaneous adaptations to students' identified educational needs. We focus on planned adaptations based on students' current achievement level, but acknowledge that teachers should also be able to make adaptations on-the-fly in direct response to students' needs.

1.2. *Achievement effects of ability grouping*

Reviews about the effects of ability grouping have shown that positive effects can be obtained if instruction is tailored to the needs of the students in the subgroups and if the grouping arrangement is flexible (Kulik & Kulik, 1992; Lou et al., 1996; Slavin, 1987; Tieso, 2003). In contrast, slight negative effects of within-class ability grouping in primary school were found across three studies in which variations in instructional treatment were not explicitly described (Deunk, Doolaard, Smale-Jacobse, & Bosker, 2015).

An unresolved issue is the potential existence of differential effects depending upon achievement level. While Slavin (1987) reported a higher median effect size for low-achieving students than for average-achieving and high-achieving students, other reviews have found different patterns with smaller (Kulik & Kulik, 1992; Lou, Abrami, & Spence, 2000) or even negative effects (Deunk et al., 2015) for low-achieving students. Previously reported negative effects of ability grouping for low-achieving students have been ascribed to stigmatization and lower educational quality in low-ability groups (Gamoran, 1992). However, it has also been argued that these negative conditions can be prevented: negative stigma may be overcome by ensuring that the subgroups are within-class and flexible (Tieso, 2003) and by promoting a growth mindset rather than a fixed mindset of ability level (Dweck, 2000; i.e. participation in additional instruction should be communicated as an opportunity to learn, rather than as a sign of fixed low ability). Moreover, when ability grouping is used as a means to adapt education to the specific needs of the students in the groups, this may enhance (rather than reduce) educational quality for low-achieving students because the instruction can be better attuned to their needs (Gamoran, 1992). In an experimental study in which different types of ability grouping were compared and coupled with systematically prescribed instructional differentiation, Tieso (2005) found positive effects of flexible within-class grouping for all subgroups (low-achieving, average-achieving, and high-achieving).

1.3. *Achievement effects of differentiation*

A recent comprehensive literature review about the effects of differentiation on student achievement demonstrated that high-quality research about this topic is scarce (Deunk et al., 2015). For primary schools, only sixteen studies met the inclusion criteria, and most of these were still either too narrow (ability grouping only, without information about whether instructional adaptations were made; e.g. Leonard, 2001) or too broad (interventions in which differentiation was one of many components; e.g. Success for All; Borman et al., 2007) to specifically evaluate the effects of differentiation. However, promising findings were obtained with the five remaining studies, which demonstrated significant positive effects of two technological applications for differentiation. Individualizing Student Instruction (McDonald Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; McDonald Connor et al., 2011a; McDonald Connor et al., 2011b) provides the teacher with recommendations about the amount and type of literacy instruction needed by individual students based on their scores on a computerised test. Accelerated Math (Ysseldyke & Bolt, 2007; Ysseldyke et al., 2003) continuously monitors students' progress and adapts practice tasks to students' individual skill level. While the review thus yielded evidence for the effectivity of technological applications for individual differentiation, studies in which (group-based) differentiation is mainly implemented by the teacher are scarce and often suffer from methodological limitations - most importantly small sample size and lack of a control group. Nevertheless, case studies of individual teachers and their classes (Brimijoin, 2002; Brown & Morris, 2005; Grimes & Stevens, 2009) do suggest that teachers may enhance the achievement of their students by implementing differentiation, although the generalisability of these findings may be limited due to the small sample size. In sum, there is some evidence to suggest that differentiation may promote student achievement in primary schools, especially when technological applications are used. However, there is still a need for large-scale studies in which differentiation is primarily in the hands of the teacher. While technological applications can be valuable for quantitative differentiation, teachers are still necessary for refined, qualitative diagnosis and adaptations.

1.4. *Adaptive teaching competency*

Teachers have an important role in enhancing student achievement: students of effective teachers achieve more (Nye, Konstantopoulos, & Hedges, 2004). According to the dynamic model of teacher effectiveness (Kyriakides, Creemers, & Antoniou, 2009), the most effective teachers distinguish themselves by the application of differentiation. Such teachers are skilled at adapting education to the needs of their students: they possess ‘adaptive teaching competency’ (Vogt & Rogalla, 2009). This requires extensive subject matter knowledge as well as advanced diagnostic, didactical, pedagogical, and classroom management skills (Smeets, Ledoux, Regtvoort, Felix, & Mol Lous, 2015; Vogt & Rogalla, 2009). For teachers with less-developed knowledge and skills, implementing differentiation can be difficult. Many teachers feel that initial teacher education did not sufficiently prepare them for implementing differentiation (Inspectorate of Education, 2015). Therefore, a need for PD about differentiation has been identified (Royal Dutch Academy of the Sciences, 2009; Schram, Van der Meer, & Van Os, 2013).

1.5. *Differentiation in mathematics using the cycle of differentiation*

Against this background, project GROW (in Dutch, this is an acronym for differentiated mathematics education) was launched with the goal of developing and evaluating an effective PD programme for differentiation in primary school mathematics. We focused exclusively on mathematics, since domain-specific guidelines may provide teachers with more concrete advice for practical application than general guidelines. To ensure strong links between theory and practice, we collaborated intensively with a consortium of educational consultants and teacher trainers with expertise in mathematics. In the first stage of the project, we sought consensus among these experts about what teachers should do in daily practice to implement differentiation successfully. This resulted in the cycle of differentiation displayed in Fig. 1 (see also Prast et al., 2015).

The cycle of differentiation starts with the identification of educational needs. First, the teacher should analyse the students' current skill level and divide the students over homogeneous achievement groups (typically low-achieving, average-achieving, and high-achieving). These achievement groups are used part of the time, besides whole-class instruction and individual practice and feedback, to cater specifically for the educational needs of the different subgroups. Students should be able to switch between groups based on changes in their educational needs (Tieso, 2003). In addition to achievement tests, ongoing and refined diagnostic measures such as the analysis of daily work and diagnostic interviews should be used to signal changes in educational needs and to determine qualitative educational needs (i.e. why a student struggles with a particular type of sums and what the student needs to overcome this problem). In the second step, the teacher sets differentiated goals which should be challenging but realistic for the students in the different subgroups (Csikszentmihalyi, 1990).

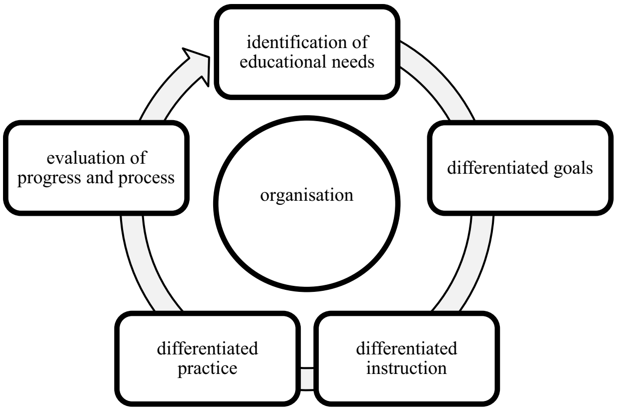

identification of educational needs
differentiated goals
differentiated instruction
differentiated practice
evaluation of progress and process
organisation

Fig. 1. Cycle of differentiation (Prast et al., 2015; reprinted with permission).

Third, the teacher differentiates instruction through broad whole-class instruction which engages students of diverse achievement levels, subgroup instruction tailored to the needs of that subgroup, and individual adaptations. One important way to differentiate instruction in mathematics is to use the stages from concrete to abstract mathematical reasoning (Gal'perin, 1969; Van Groenestijn, Borghouts, & Janssen, 2011). In subgroup instruction for low-achieving students, teachers need to spend more attention on concrete reasoning in order to build the understanding which underlies abstract reasoning. High-achieving students also need specific guidance and feedback, especially when they are working on appropriately challenging tasks (VanTassel-Baska & Stambaugh, 2005). Fourth, the practice tasks should be differentiated both quantitatively and qualitatively. For the low-achieving subgroup, the crucial tasks that are crucial for mastery of the goals for low-achieving students should be selected. For the high-achieving subgroup, the regular material should be compacted and enriched with challenging tasks which stimulate higher-level thinking (Rogers, 2007). Fifth, the teacher should evaluate whether the students have met the goals and whether the applied adaptations of instruction and practice had the desired effect using both formal (i.e. achievement tests) and informal measures (i.e. analysis of daily work). The evaluation phase informs the teacher about students' current achievement level and about instructional approaches that work for these students, completing the cycle and serving as new input for the identification of educational needs.

1.6. Research questions and hypotheses

The cycle of differentiation described above represents best practice as recommended by experts based on their experiential knowledge. However, as we have argued, quantitative empirical evidence proving that differentiation has positive effects on student achievement is scarce and there is still a need for large-scale studies in which differentiation is primarily in teachers’ hands. In this article, we examine the effect of the PD programme developed for project GROW - in which teachers learn how to differentiate their mathematics lessons using the cycle of differentiation - on student achievement.

First, we investigate whether there is an overall effect of the PD programme on student achievement in the total sample. We expect a positive overall effect on achievement, because the PD programme should enable teachers to meet the educational needs of their students better.

Second, we examine whether the effects of the PD are similar or different for students of different achievement levels (differential effects). We hypothesise that the direction of effects is positive for students of all achievement levels, including low-achieving students. As we have argued, we expect that potential negative consequences of ability grouping for low-achieving students can be overcome by grouping students flexibly based on their current achievement level and by using this grouping structure to adapt education to the educational needs of the students (Gamoran, 1992; Slavin, 1987; Tieso, 2003). The PD programme should provide teachers with the knowledge and skills to make appropriate adaptations for students of diverse achievement levels, thereby using ability grouping as a means to differentiate instruction. Most previous reviews about within-class ability grouping have also yielded positive effect sizes for all achievement groups (Kulik & Kulik, 1992; Lou et al., 2000; Slavin, 1987), with exception of the review by Deunk et al. (2015) in which it was unclear whether and how the instruction was adapted to the needs of the students in the group. Besides the direction of the effects of the PD, we also explore whether the magnitude of effects differs between achievement groups (i.e. bigger or smaller effects for low-achieving or high-achieving students). Since previous reviews have been inconsistent about this (Deunk et al., 2015; Kulik & Kulik, 1992; Lou et al., 2000; Slavin, 1987), we do not formulate specific hypotheses regarding the relative magnitude of effects.

2. Method

2.1. Design

The design of the study is shown in Table 1. Participating schools were randomly assigned to one of three cohorts. In each cohort, data were collected across two schoolyears (i.e. all schools provided data on all measurement occasions), but the timing of the intervention differed between the cohorts: Cohort 1 participated in the PD programme in Year 1 and was a follow-up condition in Year 2, Cohort 2 was a control condition in Year 1 and participated in the PD programme in Year 2, and Cohort 3 served as a control condition in both years (but was offered to participate in the PD programme in the following schoolyear). Thus, we could examine the short-term effect of the intervention in two independent cohorts (Cohort 1 in Year 1 and Cohort 2 in Year 2) as well as the long-term effect (Cohort 1 in Year 2).

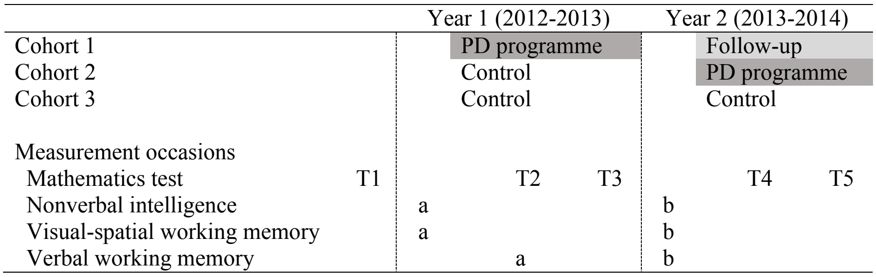
2.2. Participants

Schools were recruited with advertisements and flyers, with the proposed deal of free participation in the PD programme in combination with two years of data collection. Schools that were willing to participate could register themselves on a project website and we selected the first 32 schools that had registered. In the course of the project, two of these schools dropped out. The first school (assigned to Cohort 1), dropped out after the first measurement occasion because it perceived the project as too intensive. The second school (assigned to Cohort 2), quit with the PD programme in the course of Year 2 after identifying other priorities for PD. Since the experimental condition of this school was neither purely control nor purely experimental, data collected at this school were disregarded. Thus, thirty schools spread across the Netherlands participated. These schools were diverse in terms of school size (M = 209 students per school, range 52–550) and mathematics curriculum used (five different curricula in different versions). Fifteen schools (50%) used single-grade classes. Nine schools (30%) used multi-grade classes (typically two adjacent grades within one classroom). Six schools (20%) used a combination of single-grade and multi-grade classes.

Data from all students in grade 1 through 6 were analysed (students who entered grade 1 in Year 2 only provided data in Year 2, students who left primary school in Year 2 only provided data in Year 1). The sample consisted of 196 classes in Year 1 and 186 classes in Year 2 (average class size: 24 students). In total, 5658 students (50.8% male) participated.

Table 2 provides descriptive information about the participating students and their teachers, split by year and cohort. In Year 1, student age differed significantly between the cohorts, F (2, 4748) = 3.80, p = .023, partial η^2 = .002. Pairwise comparisons indicated that students of Cohort 2 were significantly younger than students of Cohort 1

Table 1
Research design.



		Year 1 (2012-2013)			Year 2 (2013-2014)		
Cohort 1			PD programme			Follow-up	
Cohort 2			Control			PD programme	
Cohort 3			Control			Control	
Measurement occasions							
Mathematics test	T1		T2	T3		T4	T5
Nonverbal intelligence		a			b		
Visual-spatial working memory		a			b		
Verbal working memory			a		b		

Note. a = students in grade 1–6 in Year 1; b = students who enter grade 1 in Year 2.

and 3 ($p < .05$ with Bonferroni correction). However, the effect size was very small and might be explained by students' grade level. That is, although grade levels were approximately equally represented in all cohorts (with 15.2–18.2% of students in each grade), Cohort 2 had relatively many students in grade 1 (18.2%) and relatively few students in grade 6 (15.2%) in Year 1. In Year 2, no age differences were found, F (2, 4683) = 1.60, $p = .202$, partial $\eta^2 = .001$. All subsequent analyses were controlled for grade level.

At the beginning of the study, teachers had an average of about fifteen years of teaching experience, with a broad range from zero to forty years. In Year 1, the mean number of years of experience of the teachers did not differ significantly across cohorts (F (2, 235) = 1.84, $p = .160$, partial $\eta^2 = .016$). In Year 2, teachers of Cohort 2 had significantly fewer years of experience than teachers of Cohort 1 and 3 (F (2, 242) = 4.73, $p = .010$, partial $\eta^2 = .038$; pairwise comparisons for Cohort 2 versus 1 and 3 were significant ($p < .05$ with Bonferroni correction)). The fact that this difference was only significant in Year 2 and not in Year 1 may be explained by the relatively large percentage of teachers who were new at the school in Year 2 in Cohort 2.

2.3. Measures

2.3.1. Mathematics achievement

Mathematics achievement was measured using the Cito Mathematics Tests (CMT; Janssen, Scheltens, & Kraemer, 2005a). These are national Dutch tests which are commonly administered at the middle and end of each schoolyear to monitor students' progress in mathematics throughout primary school. For each grade level, different versions with developmentally appropriate tasks for both the middle and end of the schoolyear have been developed (mid grade 1 through mid grade 6 – at the end of grade 6, a general end-of-primary-school test is nationally administered instead of the CMT). In all versions, five main domains are covered: (a) numbers and number relations, covering the structure of the number line and relations between numbers, (b) addition and subtraction, (c) multiplication and division, (d) complex math applications, often involving multiple mathematical manipulations, and (e) measuring (e.g., weight and length). From mid grade 2 to mid grade 6, the following domains are added successively: (f) estimation, (g) time, (h) money, (i) proportions, (j) fractions, and (k) percentages. The raw score on each grade-level test is converted into a mathematics competence score (for each raw score on each grade-level test, the CMT manual lists the corresponding competence score; thus, a competence score of 50 refers to the same competence level, regardless of which grade-level test was used). This competence score increases from grade 1 (minimum score: 0) through grade 6 (maximum score: 169) and can be used to assess growth in mathematics competence over time (Janssen, Scheltens, & Kraemer, 2005b). The reliability coefficients of the different versions range from .91 to .97 (Janssen, Verhelst, Engelen, & Scheltens, 2010). Based on a large sample which is representative for the Dutch population, norms are provided for each measurement occasion (Keuning et al., 2015). These include the mean competence score and its standard deviation for each grade level and timepoint (middle or end of the year).

Table 2
Information about participants, split by Year and Cohort.

	Cohort 1	*Cohort 2*	*Cohort 3*	*Total*
Students				
N Year 1	1514	1370	1867	4751
N Year 2	1494	1408	1790	4692
Age Year 1 (*M*, *SD*)	8.96 (1.82)	8.79 (1.82)	8.94 (1.86)	8.90 (1.84)
Age Year 2 (*M*, *SD*)	8.89 (1.80)	8.79 (1.83)	8.88 (1.83)	8.86 (1.82)
Gender Year 1 (% boys)	49.9%	50.3%	53.1%	51.3%
Gender Year 2 (% boys)	49.3%	49.6%	52.5%	50.6%
Teachers				
N Year 1	101	81	115	297
N Year 2	98	82	111	292
Years of experience Year 1 (*M*, *SD*)	16.54 (10.82)	13.17 (10.35)	14.80 (10.35)	15.11 (10.58)
Years of experience Year 2 (*M*, *SD*)	17.81 (10.91)	12.91 (9.49)	16.93 (10.82)	16.01 (10.75)
New at the school in Year 2 (*N*, %)	11 (11.2%)	13 (15.9%)	13 (11.7%)	37 (12.7%)

2.3.2. Nonverbal intelligence

Since (nonverbal) intelligence has been shown to be an important predictor of mathematics achievement (Deary, Strand, Smith, & Fernandes, 2007; Geary, 2011), nonverbal intelligence was measured to be included in the model as a covariate. To this end, the Raven Standard Progressive Matrices (SPM; Raven, Court, & Raven, 1996) was administered. Validity and reliability of the SPM as a measure of nonverbal, fluid intelligence are well-established (Schweizer, Goldhammer, Rauch, & Moosbrugger, 2007; Strauss, Sherman, & Spreen, 2006). Moreover, the Raven SPM has demonstrated good internal consistency and predictive validity in the same sample as the current study (Van de Weijer-Bergsma, Kroesbergen, Jolani, & Van Luit, 2016).

The SPM consists of five series of 12 diagrams or designs with one part missing. Students have to select the correct part which logically completes the designs. The difficulty level progressively increases over the test. A proportion correct score was calculated by dividing the total number of correct answers by the total number of items completed (students with missings on more than five items were treated as missing on the whole SPM). To control for the linear and quadratic effects of age, ageresidualised scores were created by regressing the proportion correct score on age and age-squared and saving the unstandardised residuals.

2.3.3. Working memory

Working memory – another important predictor of mathematics achievement (Friso-Van den Bos, Van der Ven, Kroesbergen, & Van Luit,

2013) – was also measured to be included as a covariate. Working memory was assessed with two online tasks suitable for self-reliant administration: the Lion game and the Monkey game. The Lion game is a visual–spatial complex span task (Van de Weijer-Bergsma, Kroesbergen, Prast, & Van Luit, 2015). Students are presented with a 4 × 4 matrix on the computer screen. In each trial, eight lions of different colours are consecutively presented at different locations in the matrix. Students have to remember the last location where a lion of a certain colour has appeared.

The Monkey game is a backward word span task (Van de Weijer-Bergsma et al., 2016). Students hear a number of spoken words, which they have to remember and recall backward by clicking on the words presented visually in a 3 × 3 matrix. For example, if students hear 'moon – fish – rose', they should click 'rose – fish – moon'. Both tasks consist of five levels in which working memory load is manipulated by increasing the number of lions or words (one through five) that students have to remember. A mean proportion correct score indicating the proportion of lions or words recalled in the correct serial position was calculated and subsequently converted into an ageresidualised score.

Both tasks have demonstrated excellent internal consistency ($\alpha = .90$ for the Lion game and $\alpha = .87$ for the Monkey game) and have been shown to predict mathematics achievement ($\beta = .15$ for the Lion game and $\beta = .18$ for the Monkey game, $p < .001$) in the same sample as that of the current study (Van de Weijer-Bergsma et al., 2015; Van de Weijer-Bergsma et al., 2016). In addition, the Lion game has been shown to correlate ($r = .51$ - $.59$, $p < .001$) with the individually administered Automated Working Memory Assessment (Alloway, Gathercole, Kirkwood, & Elliott, 2008; Van de Weijer-Bergsma et al., 2015).

2.3.4. Evaluation questionnaire for teachers

At the end of the PD programme, teachers were asked to complete an evaluation questionnaire. In 15 items, teachers were asked to rate on a five-point Likert scale what they learned (based on the steps of the cycle of differentiation), whether they used what they learned in their daily mathematics teaching, and whether they perceived positive effects on their students' motivation and achievement. A sample item is: 'In the PD, I learned how to (better) diagnose my students' educational needs'. All items are provided in Table 3 (see section 3.1).

2.4. Procedure

Mathematics achievement was measured five times (see Table 1): at the middle and end of Year 1 and Year 2 and a baseline measurement at the end of the year before the study started (because the CMT is only administered at the middle and end of the schoolyear, it could not be administered at the beginning of Year 1). The CMT was administered by the classroom teacher. The SPM was group-administered in the classroom under supervision of a research assistant at the beginning of Year 1. A one-hour time limit was applied. The working memory tasks were administered online: teachers were asked to make sure that their students completed the task self-reliantly within a specified time frame. The Lion game was administered at the beginning of Year 1. The Monkey game was still in development at that time so it was administered at the middle of Year 1. Students who entered grade 1 in Year 2 completed both working memory tasks and the SPM at the beginning of Year 2.

2.5. Professional development programme

Following the characteristics of effective teacher PD as summarised in a literature review by Borko, Jacobs, and Koellner (2010), the PD programme was designed to:

- connect to daily teaching practice and focus on students' learning
- include models of preferred instructional practice
- offer opportunities for active teacher learning
- stimulate collaboration and exchange between teachers
- offer multiple contexts, including classroom practice, for teacher learning
- be long-term, intensive and sustainable.

The PD programme consisted of three main components: PD for all teachers, an additional training for internal project coaches, and active involvement of the principal.

2.5.1. PD for all teachers

Ten three-hour team meetings spread across the schoolyear were provided for all teachers within the school. Six of these meetings were led by professional educational consultants who had collaborated in designing the PD programme as members of the consortium. The other four meetings were provided by the school's own project coaches (see below). During the team meetings, teachers learned about the cycle of differentiation and strategies for each step of the cycle. Attention was also spent on prerequisite knowledge, such as knowledge about the diverse solution procedures students use to solve particular types of problems and common mistakes. Various formats were used, including interactive lectures and application of the strategies in practical exercises. Lesson Study (Murata, 2011) was also applied in adapted form: teachers collectively prepared a mathematics lesson with specific attention for differentiation, one teacher taught the lesson and videotaped it, and the group evaluated the lesson afterwards. Besides active participation in the team meetings, teachers were required to read selected literature and to apply certain strategies for differentiation in their mathematics lessons.

On the continuum from highly specified to highly adaptive approaches to PD (Koellner & Jacobs, 2015), we tried to find a balance between specification of the programme and adaptation to the needs and interests of specific schools and teachers. While the cycle of differentiation represented the common core of the PD programme, schools and teachers could also determine their own focus in consultation with the external educational consultant. To facilitate this adaptivity, the materials for the PD programme were organised like a toolbox consisting of a Prezi presentation, practical application exercises, and articles about the cycle of differentiation in general and practical strategies for each step. The educational consultants were asked to spend attention on each step of the cycle over the course of the year, but to select the most relevant exercises and literature based on the school's needs.

2.5.2. Project coaches

At each school, at least two team members were trained to be a project coach. The role of the project coach was to function as a change leader (Fullan, 2002) by coaching teachers in the process of implementing differentiated instruction. Project coaches were prepared for this role in five additional meetings which were organised regionally together with the project coaches of other participating schools. Meetings covered topics such as the analysis of the baseline situation and progress regarding differentiation within a school, the implementation of Lesson Study, and how to carry out classroom observations. Also, project coaches were required to read additional literature and write a paper about a self-selected aspect of differentiation relevant for their school. During the PD programme, project coaches gradually assumed more responsibility. Project coaches led four of the team meetings - during which teaching teams discussed new schoolwide policies for differentiation and engaged in Lesson Study - and observed lessons of individual teachers to provide formative feedback about their application of differentiation. After the PD programme ended, project coaches were still available to coach and support their colleagues in further implementation of differentiation. To enhance continued implementation, project coaches received a follow-up package which they could use for continued PD with the team and a

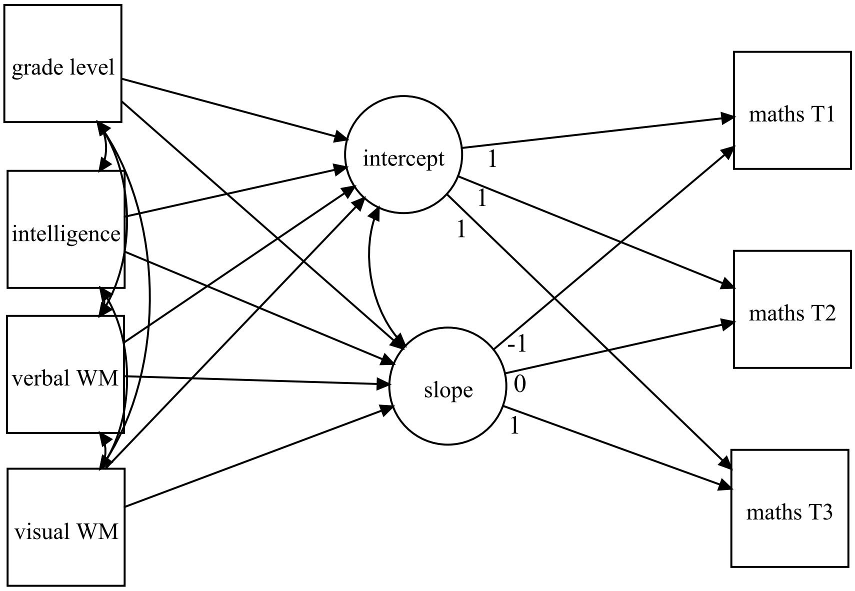


Fig. 2. Model 1 for Year 1. WM = working memory.

train-the-trainer package to train new project coaches if necessary.

2.5.3. Involvement of the principal

Since administrative support is vital for successful implementation of new instructional practices (Klingner, Ahwee, Pilonieta, & Menendez, 2003), principals were actively involved. In an intake meeting, the educational consultant and the principal discussed the current situation in the school regarding differentiation and expectations about the PD programme. The roles and responsibilities of the principal, project coaches, and teachers were made explicit and attention was spent on how the principal could facilitate the PD programme. Principals were expected to be present at the team meetings. During the schoolyear, two two-hour intervision meetings were organised for principals and project coaches to discuss progress, identify barriers to implementation, and make plans to facilitate implementation. Based on this, principals had to write a school-level plan for the continued implementation of differentiation in mathematics.

2.6. Analyses

The data were analysed with latent growth curve models using M*plus* version 7.31 (Muthén & Muthén, 1998–2012). First, the general effect of the intervention was evaluated. Subsequently, multiple-group models were used to evaluate whether the effect differed between achievement groups.

For the overall analysis, two models were estimated. Model 1 consisted of a latent growth curve model of mathematics achievement with control for covariates (see Fig. 2). The analyses were carried out separately for each year of the study to enable separate evaluation of the effects in the intervention and post-intervention year[1]: the Year 1 model included T1, T2 and T3, with T2 specified as the intercept (since verbal working memory task - a predictor of the intercept - was administered at T2). The right-hand side of Fig. 2 specifies the linear growth model for Year 1. The left-hand side of the figure lists the covariates, which were specified as predictors of the intercept and slope. The Year 2 model was analogous and included T3, T4 and T5 (T3 was used as the beginning point in this model because the CMT is not administered at the beginning of the schoolyear).

In Model 2, dummy variables representing the experimental conditions were added to the model to evaluate the effect of the PD programme. For the Year 1 analysis, the variable 'PD in Year 1' (coded as 1 for students in Cohort 1 and 0 for students in Cohort 2 and 3) was specified as an additional predictor of the intercept and slope to evaluate the short-term effect of the intervention on students in Cohort 1. For the Year 2 analysis, the variable 'PD in Year 2' (1 = Cohort 2, 0 = Cohort 1 and 3) was similarly added to evaluate the short-term effect of the intervention on students in Cohort 2. In addition, the variable 'PD in Year 1' was retained in the Year 2 analysis to evaluate the long-term effect of the PD on students in Cohort 1. In the interpretation of the results, we focus on the effect of the PD on the slope (rate of achievement growth). Effects on the intercept (level of achievement) are hard to interpret because the intercept is influenced by all timepoints in the model and, therefore, these analyses do not clarify whether any differences in level of achievement were already present at baseline or emerged over the course of the year as a result of the PD. Thus, the effect of the PD on the intercept was only included in the model to enable a more pure evaluation of the effect of PD on the slope (controlling for any differences between the cohorts in level of achievement) but the effect on the intercept itself was not interpreted. To test whether baseline mathematics achievement differed significantly between the cohorts, an additional ANOVA of the CMT scores at T1 with control for grade level was performed (see section 3.2).

Third, the full model (i.e. Model 2 from the overall analysis) was estimated as a multiple-group model for students of three achievement groups. Students were divided over three groups based on their CMT score at the first timepoint of the analysis (T1 for Year 1, T3 for Year 2). The multiple-group model was estimated for students of Grade 2–6 only, because students of Grade 1 had not yet entered primary school when this test was administered. Z-scores comparing students' competence score to the national norms (*M* and *SD* on each grade-level test) were computed. To create three approximately equally sized groups, students with z-scores below −0.5 were assigned to the low-achieving group, z-scores above 0.5 to the high-achieving group, and z-scores between −0.5 and + 0.5 to the average-achieving group. Wald tests were used to evaluate whether the parameters estimating the effect of the PD on the intercept and slope were significantly different between achievement groups, which would be an indication of differential effects.

In all analyses, the 'type = complex' option in M*plus* was used to control for the nesting of students within classes. This method ensures

[1] Rather than in a piecewise growth model for both years together, because the sample partly differed between years due to students entering grade 1 or leaving grade 6 and because students were not necessarily nested in the same classes in both years.

Table 3
Evaluation of the PD by participating teachers.

	Cohort 1 (Year 1)		Cohort 2 (Year 2)	
	M	*SD*	*M*	*SD*
In the PD, I extended my knowledge about mathematics education in general (e.g. didactics)	3.69	0.78	3.56	0.97
In the PD, I learned how to (better) …				
… Diagnose my students' educational needs	3.66	0.76	3.47	0.86
… Set differentiated goals	3.78	0.72	3.34	0.98
… Broaden whole-class instruction	3.35	0.89	3.17	1.00
… Adapt instruction for low-achieving students	3.17	0.97	3.50	1.05
… Adapt practice for low-achieving students	3.15	0.82	3.27	0.96
… Adapt instruction for high-achieving students	3.56	0.80	3.36	0.99
… Adapt practice for high-achieving students	3.58	0.87	3.23	0.99
… Evaluate whether my chosen way of teaching was effective for my students	3.15	0.86	3.09	1.00
… Organise differentiation in practice (e.g. working with subgroups)	3.33	0.99	3.29	1.13
… Apply (more) differentiation in my mathematics lessons	3.53	0.76	3.36	1.05
I can use what I learned in the PD for the preparation and teaching of my mathematics lessons	3.84	0.84	4.00	0.79
I actually use what I learned in the PD for the preparation and teaching of my mathematics lessons	3.90	0.78	3.86	0.87
Applying (more) differentiation in my mathematics lessons has a positive effect on my students' motivation	3.87	0.73	3.70	0.83
Applying (more) differentiation in my mathematics lessons has a positive effect on my students' achievement	3.47	0.78	3.57	0.69

Note. 1 = fully disagree, 5 = fully agree.

that standard errors are corrected for the clustered data structure without building a full multilevel model (McNeish, Silverman, & Stapleton, 2017). In our case, multilevel modeling was complicated since grade level was neither purely an individual-level variable nor purely a class-level variable due to the existence of multigrade classes. Single-level analysis methods with cluster-robust standard errors (such as type = complex) are an appropriate and computationally less demanding alternative for multilevel modeling (McNeish, Stapleton, & Silverman, 2017). Model fit was evaluated using the chi-square statistic, the comparative fit index (*CFI*), the Tucker-Lewis Index (*TLI*), the root mean squared error of approximation (*RMSEA*), and the standardised root mean square residual (*SRMR*). Due to the large sample size, the chi-square statistic was expected to be significant. The models were judged to have a good fit if they had values above .95 for the *CFI* and *TLI* and values below .06 and .08 for the *RMSEA* and *SRMR*, respectively (Hu & Bentler, 1999).

3. Results

3.1. Teacher participation in and evaluation of the PD

In Year 1, 81 teachers of Cohort 1 (81.0%) obtained their certificate for participation, indicating presence at least eight out of ten team meetings. In Year 2, 72 teachers of Cohort 2 (90%) obtained their certificate. Although reasons for absence were not always known to us, teachers who missed many team meetings often had reasons such as having left or entered the school in the course of the year, long-term illness, maternity leave, or a part-time job (i.e. teachers were asked to attend the team meetings that were planned on days they did not work, but this was not always possible).

The teacher evaluation questionnaire about the PD programme was completed by 76 teachers of Cohort 1 at the end of Year 1 and 73 teachers of Cohort 2 at the end of Year 2. As can be seen in Table 3, teachers were moderately positive about what they learned in the PD, with scores above the midpoint of the scale for all questions. Teachers indicated that they had learned about all steps in the cycle of differentiation. Moreover, the majority of teachers (76.3% and 76.7% of teachers who completed the questionnaire in Cohort 1 and 2, respectively) mostly or fully agreed that they actually used what they had learned in the PD for the preparation and teaching of their mathematics lessons. Teachers also perceived positive effects of implementing (more) differentiation on students' motivation and achievement.

3.2. Descriptive statistics and missing data

Descriptive statistics of students' scores on the mathematics tests, the nonverbal intelligence test, and the two working memory measures are displayed in Table 4. An ANOVA comparing the raw competence scores of the cohorts on the mathematics test at T1 showed a significant but very small effect of cohort (F (2, 3511) = 3.15, p = .043, partial η^2 = .002; pairwise comparisons not significant). However, after controlling for grade level, these differences disappeared (F (2, 3510) = 1.80, p = .165, partial η^2 = .002; pairwise comparisons not significant). Thus, students of Cohort 1 (estimated mean[2] = 77.88, SE = 0.41), Cohort 2 (estimated mean = 77.41, SE = 0.43), and Cohort 3 (estimated mean = 76.84, SE = 0.37) had comparable baseline scores.

Grade level was uniformly distributed with approximately 17% of students in each grade level. The other variables approximated the normal distribution, but some skewness and kurtosis was present. Therefore, the Maximum Likelihood Robust estimator, which is robust to deviations from normality, was used in all subsequent analyses.

The percentage of available data - and, conversely, the percentage of missing data - is provided in the last column of Table 4. Most of the missing data on the mathematics test are missing by design because the CMT is neither administered before the start of grade 1 nor at the end of grade 6. Remaining causes for missingness are absence on the day of testing and - in case of the working memory tasks - technical problems with the games and lack of systematic administration by some teachers. M*plus* can handle missing data well by making flexible use of all relevant available information for each parameter. To enable the inclusion of cases with missing values on one or more covariates (which are, by default, completely removed from the analysis in M*plus*), we specified the variances of the covariates as parameters to be estimated in all models.

3.3. Overall analysis year 1

Model 1 had a good fit: *RMSEA* = .024, *CFI* = 0.999, *TLI* = 0.998, *SRMR* = .012. As expected, the chi-square test was significant: χ^2 (5) = 18.13, p = .003. The growth model explained over 95% of the variance in the observed variables. Model results are displayed in

[2] After controlling for grade level; raw means and standard deviations are reported in Table 4.

Table 4
Descriptive statistics.

	Cohort 1	Cohort 2	Cohort 3	Total	*Min.*	*Max.*	*n (%)*
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*			
Maths T1	76.78 (24.80)	76.18 (24.46)	78.59 (24.96)	77.33 (24.78)	7.00	143.00	3514 (73.96)[b]
Maths T2	77.34 (30.87)	74.87 (29.77)	76.90 (29.90)	76.45 (30.18)	0.00	154.00	4448 (93.62)[b]
Maths T3	77.95 (25.33)	74.39 (25.99)	74.24 (25.50)	75.48 (25.65)	0.00	149.00	3523 (74.06)[b,c]
Maths T4	75.13 (28.92)	75.20 (29.11)	75.91 (28.70)	75.44 (28.89)	0.00	154.00	4048 (86.27)[c]
Maths T5	78.89 (28.92)	77.43 (24.72)	77.95 (24.67)	78.06 (24.44)	0.00	164.00	3358 (71.57)[c]
Nonverbal intelligence[a]	−0.78 (7.47)	0.21 (7.76)	0.36 (7.75)	−0.04 (7.68)	−30.36	23.16	4998 (88.33) [d]
Verbal WM[a]	0.00 (0.14)	0.01 (0.14)	−0.01 (0.15)	0.00 (0.14)	−0.54	0.44	4618 (81.62) [d]
Visual-spatial WM[a]	0.01 (0.15)	0.00 (0.17)	−0.01 (0.17)	0.00 (0.16)	−0.72	0.41	4763 (84.18) [d]

Note. WM = working memory.
[a] Ageresidualised score.
[b] Percentage of students in Year 1 (*N* = 4751).
[c] Percentage of students in Year 2 (*N* = 4692).
[d] Percentage of total number of students (*N* = 5658).

Table 5. Regarding the prediction of the latent variables, all covariates had a significant positive effect on the intercept and this effect was largest for grade level. Only grade level had a significant effect on the slope, and this effect was negative (i.e. students in lower grade levels acquired new knowledge and skills at a faster pace). Taken together, grade level, nonverbal intelligence, visual-spatial working memory and verbal working memory explained 88% of the variance of the intercept and 16% of the variance of the slope of mathematics achievement.

Model 2, in which the effect of the intervention was added, had a good fit: χ^2 (10) = 19.89, *p* = .030, *RMSEA* = .014, *CFI* = 1.000, *TLI* = 0.999, *SRMR* = .017. PD in Year 1 had a significant but small positive effect on the slope: β = 0.15, *p* = .007. Thus, students in Cohort 1 gained about 2.5 points *more* on the CMT in the course of Year 1 than students in the other cohorts (average growth is 14.4 points). Adding the effect of the intervention to the model explained an additional 2% of the slope variance. In sum, in line with our hypothesis, the intervention had a positive short-term effect on the slope of mathematics achievement in Cohort 1.

Table 5
Overall model Year 1 (N = 4751).

Parameter	Model 1			Model 2		
	Estimate	*SE*	*p*	Estimate	*SE*	*p*
Predictors of the intercept[a]						
Grade level	0.87	0.01	< .001	0.87	0.01	< .001
Nonverbal intelligence	0.20	0.01	< .001	0.20	0.01	< .001
Verbal WM	0.09	0.01	< .001	0.09	0.01	< .001
Visual-spatial WM	0.07	0.01	< .001	0.07	0.01	< .001
PD in Year 1	n/a			0.04	0.01	< .001
Predictors of the slope[a]						
Grade level	−0.40	0.08	< .001	−0.40	0.08	< .001
Nonverbal intelligence	−0.03	0.03	.338	−0.02	0.03	.600
Verbal WM	0.07	0.04	.069	0.07	0.04	.082
Visual-spatial WM	0.01	0.03	.719	0.00	0.03	.992
PD in Year 1	n/a			0.15	0.05	.007
Correlations[a]						
Intercept with slope	0.00	0.05	.937	−0.02	0.05	.744
Nonverbal intelligence with grade level	0.07	0.02	.004	0.07	0.02	.004
Verbal WM with grade level	0.08	0.03	.001	0.08	0.03	.001
Verbal WM with nonverbal intelligence	0.40	0.02	< .001	0.40	0.02	< .001
Visual-spatial WM with grade level	0.08	0.02	.001	0.08	0.02	.001
Visual-spatial WM with nonverbal intelligence	0.37	0.02	< .001	0.37	0.02	< .001
Visual-spatial WM with verbal WM	0.35	0.02	< .001	0.35	0.02	< .001
Intercepts[b]						
Intercept	76.54	0.33	< .001	75.78	0.38	< .001
Slope	7.53	0.18	< .001	7.21	0.22	< .001
Residual variances[b]						
Maths T1	28.34	5.22	< .001	28.50	5.15	< .001
Maths T2	42.82	4.66	< .001	41.75	2.61	< .001
Maths T3	28.82	4.66	< .001	29.32	4.41	< .001
Intercept	104.35	3.49	< .001	103.15	3.37	< .001
Slope	8.50	2.23	< .001	8.20	2.16	< .001
Explained variances						
Maths T1	0.97	0.01	< .001	0.97	0.01	< .001
Maths T2	0.95	0.00	< .001	0.95	0.00	< .001
Maths T3	0.97	0.01	< .001	0.97	0.01	< .001
Intercept	0.88	0.01	< .001	0.88	0.01	< .001
Slope	0.16	0.06	.015	0.18	0.06	.006

Note. WM = working memory. For parsimony, the means (all close to 0 due to centering) and variances of the covariates are omitted from the table.
[a] Standardised.
[b] Unstandardised.

3.4. Multiple-group model year 1

The multiple-group model, in which the full model was estimated separately for three achievement groups, initially yielded two negatively estimated residual variances (for mathematics T1 and T3) in the average-achieving group. This problem was solved by fixing the residual variance of mathematics T1 to 0 in this group. This solution was deemed acceptable, since the model generally explained a very large proportion of the variance in the observed mathematics scores (leaving little residual variance) and since the variance was likely to be smaller within the groups because they were created based on mathematics achievement at T1. After fixing this residual variance to 0, the model had a good fit: χ^2 (31) = 102.89, *p* < .001, *RMSEA* = .044, *CFI* = 0.997, *TLI* = 0.995, *SRMR* = .035. As can be seen in Table 6, the results were largely similar to the overall model, although the effects of the covariates and their correlations differed somewhat. In addition to the previously found effects, nonverbal intelligence had a significant positive effect on the slope within all achievement groups and verbal working memory had a significant positive effect on the slope within the average-achieving and high-achieving group.

PD in Year 1 had a significant positive effect on the slope of mathematics achievement for average-achieving students (β = 0.10, *p* = .040) and high-achieving students (β = 0.12, *p* = .036). For low-achieving students, the effect was similar in size but did not reach significance (β = 0.12, *p* = 0.051). However, Wald tests demonstrated that the effect of PD on achievement growth was not significantly different between achievement groups (low-achieving vs. average-achieving students: *W* = 0.19, *p* = .667; low-achieving vs. high-achieving: *W* = 0.01, *p* = .913; average-achieving vs. high-achieving: *W* = 0.08, *p* = 0.776). Thus, the fact that the effect did not reach significance in the low-achieving group probably does not reflect a

Table 6
Multiple-group model Year 1.

Parameter	Low-achieving *n* = 989			Average-achieving *n* = 1300			High-achieving *n* = 1225		
	Estimate	*SE*	*p*	Estimate	*SE*	*p*	Estimate	*SE*	*p*
Predictors of the intercept[a]									
Grade level	0.92	0.01	< .001	0.96	0.00	< .001	0.93	0.01	< .001
Intelligence	0.09	0.01	< .001	0.05	0.01	< .001	0.11	0.01	< .001
Verbal WM	0.06	0.02	< .001	0.03	0.01	< .001	0.06	0.01	< .001
Visual-spatial WM	0.04	0.02	.010	0.02	0.01	.083	0.03	0.01	.021
PD in Year 1	0.04	0.01	.002	0.02	0.01	.052	0.05	0.01	< .001
Predictors of the slope[a]									
Grade level	−0.34	0.07	< .001	−0.29	0.05	< .001	−0.24	0.07	< .001
Intelligence	0.15	0.05	.003	0.13	0.04	.001	0.12	0.05	.007
Verbal WM	0.07	0.04	.099	0.11	0.04	.001	0.14	0.05	.003
Visual-spatial WM	0.08	0.05	.161	0.06	0.04	.084	−0.06	0.05	.197
PD in Year 1	0.12	0.06	.051	0.10	0.05	.040	0.12	0.06	.036
Correlations[a]									
Intercept with slope	0.45	0.09	< .001	0.62	0.03	< .001	0.15	0.08	.048
Intelligence with grade level	0.08	0.04	.052	0.06	0.04	.075	0.03	0.04	.379
Verbal WM with grade level	0.07	0.04	.083	0.07	0.04	.063	0.03	0.04	.524
Verbal WM with intelligence	0.26	0.03	< .001	0.27	0.03	< .001	0.33	0.03	< .001
Visual-spatial WM with grade level	0.16	0.04	< .001	0.03	0.03	.063	0.01	0.05	.923
Visual-spatial WM with intelligence	0.38	0.04	< .001	0.23	0.03	< .001	0.19	0.03	< .001
Visual-spatial WM with verbal WM	0.31	0.04	< .001	0.23	0.03	< .001	0.24	0.04	< .001
Intercepts[b]									
Intercept	69.63	0.44	< .001	84.31	0.27	< .001	95.12	0.35	< .001
Slope	8.17	0.29	< .001	7.24	0.23	< .001	5.36	0.29	< .001
Residual variances[b]									
Maths T1	17.00	6.31	.007	0.00 [c]	n/a	n/a	17.00	6.31	.007
Maths T2	55.21	4.27	< .001	40.49	3.16	< .001	55.21	4.27	< .001
Maths T3	23.13	6.72	< .001	0.67	4.15	.872	23.13	6.72	.001
Intercept	41.14	3.05	< .001	26.16	1.56	< .001	41.14	3.05	< .001
Slope	13.97	3.16	< .001	15.40	1.43	< .001	13.97	3.16	< .001
Explained variance									
Maths T1	0.98	0.01	< .001	1.00	n/a	n/a	0.96	0.01	< .001
Maths T2	0.92	0.01	< .001	0.91	0.01	< .001	0.88	0.01	< .001
Maths T3	0.99	0.02	< .001	0.99	0.01	< .001	0.95	0.02	< .001
Intercept	0.90	0.01	< .001	0.94	0.01	< .001	0.90	0.01	< .001
Slope	0.16	0.04	< .001	0.13	0.03	< .001	0.11	0.04	.004

Note. WM = working memory, intelligence = nonverbal intelligence. For parsimony, the means (all close to 0 due to centering) and variances of the covariates are omitted from the table.

[a] Standardised.

[b] Unstandardised.

[c] Fixed to 0.

different effect size but may be a consequence of the slightly smaller sample size of the low-achieving subsample. Therefore we conclude that, in Year 1, the intervention had a positive effect on mathematics achievement growth for all achievement groups, in line with our hypothesis. Since these effects were similar across achievement groups, we found no evidence for differential effects.

3.5. *Overall analysis year 2*

Model 1 of the Year 2 analysis had a good fit: χ^2 (5) = 37.29, p < .001, *RMSEA* = .037, *CFI* = 0.999, *TLI* = 0.996, *SRMR* = .015. As can be seen in Table 7, the results of Model 1 in Year 2 resembled the results of Model 1 in Year 1. The effects of the covariates on the intercept and slope were similar and, taken together, explained 88% of the intercept variance and 22% of the slope variance. The fit of Model 2 was good as well: χ^2 (15) = 23.66, p = .071, *RMSEA* = .011, *CFI* = 1.000, *TLI* = 1.000, *SRMR* = .017. However, adding the effect of the intervention did not explain additional variance. In contrast to the Year 1 findings, participation in the PD programme in Year 2 did not have a significant short-term effect on the slope (β = 0.03, p = .640). Regarding the long-term effect of the intervention, PD in Year 1 had no significant effect on the slope of students in Cohort 1 in Year 2 (β = −0.06, p = .665). In sum, in contrast to our hypothesis, neither short-term nor long-term effects of the intervention on mathematics achievement growth could be demonstrated in Year 2.

3.6. *Multiple-group model year 2*

In the Year 2 multiple-group model, two residual variances (mathematics T3 and mathematics T5) were initially negatively estimated in the low-achieving and average-achieving group and were fixed to 0. After this, the multiple-group model had a good fit: χ^2 (49) = 136.392, p < .001, RMSEA = .039, CFI = 0.996, TLI = 0.995, SRMR = .035. Again, the results were similar to the overall model, although the predictive value of the covariates and the correlations between them varied somewhat between the achievement groups (see Table 8). Similar to the overall model, the multiple group model demonstrated no significant short-term or long-term effect of PD on the slope in any of the achievement groups. Wald tests confirmed that these parameters were similar across achievement groups. Thus, in contrast to our hypothesis, neither long-term nor short-term effects on the mathematics achievement growth could be demonstrated in any of the achievement groups in Year 2.

4. Discussion

This large-scale study investigated the effects of a PD programme about differentiation on student achievement growth in mathematics. We hypothesised that the PD programme would have a positive effect on student achievement and that this would be true for students of all achievement levels. Our results provide partial support for these

Table 7
Overall model Year 2 (N = 4692).

Parameter	Model 1			Model 2		
	Estimate	*SE*	*p*	Estimate	*SE*	*p*
Predictors of the intercept[a]						
Grade level	0.87	0.01	< .001	0.87	0.01	< .001
Intelligence	0.19	0.01	< .001	0.20	0.01	< .001
Verbal WM	0.10	0.01	< .001	0.10	0.01	< .001
Visual-spatial WM	0.08	0.01	< .001	0.07	0.01	< .001
PD in Year 1[c]	n/a			0.03	0.01	.015
PD in Year 2[d]	n/a			0.00	0.01	.896
Predictors of the slope[a]						
Grade level	−0.47	0.08	< .001	−0.46	0.08	< .001
Intelligence	−0.01	0.04	.804	−0.02	0.04	.655
Verbal WM	−0.03	0.03	.342	−0.03	0.03	.340
Visual-spatial WM	0.01	0.04	.788	0.02	0.04	.665
PD in Year 1[c]	n/a			−0.06	0.06	.356
PD in Year 2[d]	n/a			0.03	0.07	.640
Correlations[a]						
Intercept with slope	0.09	0.06	.095	0.10	0.05	.059
Intelligence with grade level	0.07	0.02	.002	0.07	0.02	.002
Verbal WM with grade level	0.04	0.03	.090	0.04	0.02	.090
Verbal WM with intelligence	0.39	0.02	< .001	0.40	0.02	< .001
Visual-spatial WM with grade level	0.04	0.02	.087	0.04	0.02	.088
Visual-spatial WM with intelligence	0.37	0.02	< .001	0.37	0.02	< .001
Visual-spatial WM with verbal WM	0.36	0.02	< .001	0.36	0.02	< .001
Intercepts[b]						
Intercept	76.67	0.33	< .001	76.07	0.48	< .001
Slope	7.35	0.20	< .001	7.42	0.36	< .001
Residual variances[b]						
Maths T3	36.42	4.88	< .001	36.07	4.82	< .001
Maths T4	43.05	2.86	< .001	43.29	2.84	< .001
Maths T5	23.86	4.89	< .001	23.44	4.88	< .001
Intercept	103.82	3.71	< .001	102.88	3.66	< .001
Slope	9.53	2.29	< .001	9.66	2.29	< .001
Explained variance						
Maths T3	0.96	0.01	< .001	0.96	0.01	< .001
Maths T4	0.95	0.00	< .001	0.95	0.00	< .001
Maths T5	0.97	0.01	< .001	0.97	0.01	< .001
Intercept	0.88	0.01	< .001	0.88	0.01	< .001
Slope	0.22	0.07	.002	0.22	0.07	.002

Note. WM = working memory, intelligence = nonverbal intelligence. For parsimony, the means (all close to 0 due to centering) and variances of the predictors are omitted from the table.

[a] Standardised.

[b] Unstandardised.

[c] Cohort 1: long-term effect.

[d] Cohort 2: short-term effect.

hypotheses: the PD had positive effects on students of all achievement levels in Year 1, but these effects could not be replicated in Year 2.

In Year 1, the overall analysis demonstrated a small but significant positive effect of the PD programme on student achievement growth in mathematics. The multiple-group analysis demonstrated that the direction of effects was positive for all achievement groups, as hypothesised, and that the effect size was similar for low-achieving, average-achieving and high-achieving students. Thus, we found no evidence for differential effects depending upon achievement level. Our findings contrast with some previous studies of naturally occurring ability grouping - without information about differentiation - in which negative effects of being placed in a low-ability group were found (Condron, 2008; Nomi, 2010; reviewed by; Deunk et al., 2015).

In line with previous reviews about ability grouping which stressed the importance of adapting instruction to the specific needs of the groups (Kulik & Kulik, 1992; Lou et al., 1996; Slavin, 1987), we believe that an important success factor in our project was that teachers were provided with the skills and knowledge to use ability grouping as a *means* to differentiate instruction rather than as an end in itself. In the PD programme, attention was spent on all four dimensions of adaptive teaching competency, which has been shown to relate positively to student achievement (Vogt & Rogalla, 2009): knowledge about mathematics (e.g. the sequence in which children learn mathematical concepts and skills), diagnostic competence (i.e. how to monitor progress and identify educational needs), teaching methods (e.g. how to vary the level of abstraction in response to students' needs) and classroom management (e.g. how to organise within-class ability grouping). In the evaluation questionnaire, teachers indicated that they had learned about all steps in the cycle of differentiation (identification of educational needs, differentiated goals, differentiated instruction, differentiated practice, and evaluation of progress and process; Prast et al., 2015). Moreover, the majority of teachers indicated that they actually used what they had learned in their daily mathematics teaching. We speculate that the positive effects of the PD programme on student achievement can be explained by an increase in teachers' competence for and actual implementation of differentiation, which enabled teachers to better meet their students' educational needs. However, a limitation of this study is that we did not directly investigate the classroom processes underlying the achievement effects since we focused on the final outcome of student achievement. Also, it cannot be determined whether specific components of the intervention were particularly effective. This would require very extensive studies in which specific aspects of the intervention would be systematically varied across multiple experimental conditions. However, due to the interdependence of the steps of the cycle of differentiation, it seems more likely that all aspects of the cycle of differentiation work together than that one isolated component would be effective by itself. In future research, mixed methods could be used to examine in more depth how the PD affects classroom processes and, in turn, student achievement.

In contrast to our hypothesis, the positive effects of the PD in Year 1 could not be replicated in Year 2. One possible explanation is that schools in Cohort 2 were less motivated for the PD programme than schools in Cohort 1 due to the design of the study. When schools registered for the study, most schools were eager to participate in the PD programme. Possibly, schools in Cohort 1, in which the PD programme immediately started, were ready and motivated, whereas schools in Cohort 2 had to wait for one year during which motivation or priorities for PD may have changed. Indeed, one school from Cohort 2 dropped out and several schools in Cohort 3 declined participation in the PD programme when it was offered to them after Year 2. This shows that a school's needs and priorities are dynamic and that a PD programme which suits the needs of a school in one year may not be (as) interesting for the school one or two years later.

Another possible explanation for the smaller effects in Cohort 2 is that teachers of Cohort 2 on average had fewer years of teaching experience at the start of the intervention. Moreover, relatively many teachers were new at the school in the year of the PD. For less experienced teachers and for teachers who just started at a new school, it may be more challenging to implement differentiation since they may need to spend attention first on more basic issues such as classroom management and (new) everyday routines. In addition, relatively many schools in Cohort 2 started to use another mathematics curriculum during the course of the study. This may have drawn teachers' attention towards implementation of the new curriculum rather than to the implementation of differentiation (although school administrators themselves generally viewed it as an asset that these could be combined). These explanations illustrate that this study was situated in the dynamic context of daily practice in schools. This is both a strength and a limitation: while it promotes the practical validity of the findings, it diminishes the experimental control.

Table 8
Multiple-group model Year 2.

Parameter	Low-achieving n = 1029			Average-achieving n = 1159			High-achieving n = 1285		
	Estimate	*SE*	*p*	Estimate	*SE*	*p*	Estimate	*SE*	*p*
Predictors of the intercept[a]									
Grade level	0.92	0.01	< .001	0.96	0.01	< .001	0.93	0.01	< .001
Intelligence	0.10	0.02	< .001	0.05	0.01	< .001	0.11	0.01	< .001
Verbal WM	0.04	0.01	.002	0.01	0.01	.253	0.04	0.01	.002
Visual-spatial WM	0.05	0.01	< .001	0.03	0.01	.003	0.03	0.01	.012
PD in Year 1[c]	0.04	0.02	.035	0.01	0.01	.475	0.03	0.02	.123
PD in Year 2[d]	0.00	0.02	.864	0.00	0.01	.825	−0.01	0.02	.728
Predictors of the slope[a]									
Grade level	−0.32	0.06	< .001	−0.40	0.05	< .001	−0.24	0.08	.002
Intelligence	0.09	0.05	.050	0.15	0.04	< .001	0.10	0.04	.022
Verbal WM	0.05	0.04	.313	0.08	0.04	.039	0.01	0.05	.847
Visual-spatial WM	−0.01	0.04	.845	−0.02	0.04	.695	0.11	0.05	.023
PD in Year 1 [c]	−0.02	0.07	.819	0.02	0.06	.780	−0.06	0.07	.341
PD in Year 2 [d]	0.02	0.06	.765	0.01	0.06	.887	0.02	0.08	.791
Correlations[a]									
Intercept with slope	0.25	0.05	< .001	0.63	0.03	< .001	0.40	0.08	< .001
Intelligence with grade level	0.01	0.04	.725	0.12	0.04	.002	0.07	0.04	.077
Verbal WM with grade level	0.12	0.04	.002	0.07	0.04	.101	0.03	0.04	.381
Verbal WM with intelligence	0.21	0.03	< .001	0.25	0.03	< .001	0.33	0.03	< .001
Visual-spatial WM with grade level	0.13	0.05	.004	0.08	0.04	.045	0.02	0.03	.528
Visual-spatial WM with intelligence	0.35	0.04	< .001	0.23	0.03	< .001	0.21	0.03	< .001
Visual-spatial WM with verbal WM	0.30	0.04	< .001	0.24	0.03	< .001	0.23	0.03	< .001
Intercepts[b]									
Intercept	66.87	0.67	< .001	82.77	0.33	< .001	94.69	0.50	< .001
Slope	8.50	0.51	< .001	7.46	0.37	< .001	5.82	0.41	< .001
Residual variances[b]									
Maths T1	0.00 [e]	n/a	n/a	0.00 [e]	n/a	n/a	31.00	5.20	< .001
Maths T2	43.05	2.52	< .001	45.46	2.42	< .001	54.88	3.73	< .001
Maths T3	0.00 [e]	n/a	n/a	0.00 [e]	n/a	n/a	18.78	6.34	.003
Intercept	58.10	3.94	< .001	25.30	1.69	< .001	40.68	2.46	< .001
Slope	21.03	1.93	< .001	16.05	1.34	< .001	15.16	2.38	< .001
Explained variance									
Maths T1	1.00	n/a	n/a	1.00	n/a	n/a	0.94	0.01	< .001
Maths T2	0.92	0.01	< .001	0.90	0.01	< .001	0.88	0.01	< .001
Maths T3	1.00	n/a	n/a	1.00	n/a	n/a	0.96	0.02	< .001
Intercept	0.89	0.01	< .001	0.94	0.01	< .001	0.90	0.01	< .001
Slope	0.11	0.04	.007	0.18	0.04	< .001	0.08	0.04	.047

Note. WM = working memory, intelligence = nonverbal intelligence. For parsimony, the means (all close to 0 due to centering) and variances of the covariates are omitted from the table.
[a] Standardised.
[b] Unstandardised.
[c] Cohort 1: long-term effect.
[d] Cohort 2: short-term effect.
[e] Fixed to 0.

4.1. Implications and future research

This study was designed to have strong links to educational practice. Therefore, the PD programme was designed in collaboration with experienced teacher trainers who could bridge theory and practice. Moreover, this was the first large-scale study to investigate achievement effects of a PD programme about differentiation in mathematics. This question has large practical relevance because, although differentiation and PD about this topic are often promoted by policy makers, little was known about the effects of such interventions. The results show that PD about differentiation can improve student achievement, but that such achievement effects are not guaranteed.

Probably, much depends on whether teachers are able to apply what they learned during the PD in daily practice. We noticed during the PD programme that, while most teachers already implemented some aspects of differentiation such as tiered tasks if those were provided by the mathematics curriculum, the challenge of the PD programme was to increase the *quality* of differentiation by (1) implementing differentiation more systematically, using the full cycle of differentiation for students of all achievement levels and (2) improving the match between diagnosed educational needs and instructional adaptations. This required substantial mathematical knowledge, for example regarding the typical sequence of learning mathematical concepts and operations (enabling teachers to move back to more fundamental steps if necessary). While in-service teacher education may be a way to develop such knowledge, pre-service teacher education could also strive to equip teachers with more systematic knowledge about mathematics and didactics of mathematics before they enter the workplace. PD for in-service teachers could then focus on more advanced components of adaptive teaching competency for which this knowledge is required, such as how to use refined diagnostics to find the most appropriate instructional adaptations for a particular student. Besides adaptive teaching competency, the implementation of differentiation is also influenced by contextual factors such as the availability of appropriate instructional materials and preparation time (Roiha, 2014). School administrators could support their teachers in the implementation of differentiation by facilitating such practical aspects (Puzio, Newcomer, & Goff, 2015).

A limitation of the current study is that we did not examine directly how teacher and student learning processes influenced students' learning outcomes. While previous case studies have reported about the process of starting to implement differentiation (e.g. Brimijoin, 2002), future research could examine the effects of PD on adaptive teaching competency, implementation of differentiation, student learning processes and student achievement jointly and investigate how these effects interact over time. Small-scale studies using both quantitative and qualitative measures may be suitable to unravel such processes.

Another issue is whether the achievement effects in the current study were practically significant. The effect sizes were quite small but if this modestly higher achievement growth could be sustained over multiple years, the cumulative effect would be substantial. However, the higher achievement growth was not sustained in Cohort 1 after the PD programme had ended. This may require prolonged PD (c.f. VanTassel-Baska et al., 2008). To increase the effect sizes, future research could also investigate how technological applications for differentiation such as Accelerated Math (Ysseldyke et al., 2003) and PD about differentiation could be combined. Technological applications could be used to support and relieve teachers wherever possible, complemented with PD to develop teachers' competencies in qualitative analysis and refined instructional adaptations.

In conclusion, the results of this study show that PD about differentiation in mathematics has the potential to raise the achievement of all students. This is consistent with educational theories including the zone of proximal development (Vygotsky, 1978), aptitude-treatment interaction (Cronbach & Snow, 1977), and adaptive teaching (Corno, 2008) which propose that educational needs vary based on achievement level and that adapting education to those diverse needs leads to more effective learning. Our results indicate that schoolwide PD about systematic implementation of differentiation using the cycle of differentiation may have positive effects over and above the spontaneous adaptations that many teachers already make by themselves. Despite the drawbacks discussed above, we think that these results are sufficiently promising to continue this line of research.

Acknowledgements

This work was financed by the Netherlands Organisation for Scientific Research, grant number 411-10-753. The NWO was not involved in the collection, analysis, interpretation, or reporting of the data. We thank Elien van Es, Merel Plat and Tanja van Veldhuizen for their great work in data collection. We thank dr. Rens van de Schoot for his useful statistical advice. We thank the consortium members for their valuable input in the PD programme. Last but not least, we thank all teachers and students for their participation.

References

Alloway, T., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2008). Evaluating the validity of the automated working memory assessment. *Educational Psychology, 28*, 725–734. http://dx.doi.org/10.1080/01443410802243828.

Borko, H., Jacobs, J., & Koellner, K. (2010). Contemporary approaches to teacher professional development. In P. Peterson, E. Baker, & B. McGaw (Eds.). *International encyclopedia of education* (pp. 548–556). Oxford, UK: Elsevier.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal, 44*, 701–731. http://dx.doi.org/10.3102/0002831207306743.

Brimijoin, K. (2002). *Expertise in differentiation: A preservice and inservice teacher make their way.* Charlottesville, VA: University of Virginia.

Brown, J., & Morris, D. (2005). Meeting the needs of low spellers in a second-grade classroom. *Reading & Writing Quarterly, 21*, 165–184. http://dx.doi.org/10.1080/10573560590915969.

Condron, D. J. (2008). An early start: Skill grouping and unequal reading gains in the elementary years. *The Sociological Quarterly, 49*, 363–394. http://dx.doi.org/10.1111/j.1533-8525.2008.00119.x.

Corno, L. (2008). On teaching adaptively. *Educational Psychologist, 43*, 161–173. http://dx.doi.org/10.1080/00461520802178466.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York, NY: Irvington.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience.* New York, NY: HarperPerennial.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13–21. http://dx.doi.org/10.1016/j.intell.2006.02.001.

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, The Netherlands: GION.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development.* Philadelphia, PA: Psychology Press.

Friso-Van den Bos, I., Van der Ven, S. H. G., Kroesbergen, E. H., & Van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review, 10*, 29–44. http://dx.doi.org/10.1016/j.edurev.2013.05.003.

Fullan, M. (2002). The change leader. *Educational Leadership, 59*(8), 16–21.

Gal'perin, P. J. (1969). Stages in the development of mental acts. In M. Cole, & I. Maltzman (Eds.). *A handbook of contemporary Soviet psychology* (pp. 249–273). New York, NY: Basic Books.

Gamoran, A. (1992). Is ability grouping equitable: Synthesis of research. *Educational Leadership, 50*, 11–17.

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology, 47*, 1539–1552. http://dx.doi.org/10.1037/a0025510.

Grimes, K. J., & Stevens, D. D. (2009). Glass, bug, mud. *Phi Delta Kappan, 90*, 677–680. http://dx.doi.org/10.1177/003172170909000914.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55. http://dx.doi.org/10.1080/10705519909540118.

Inspectorate of Education (2015). *Beginnende leraren kijken terug [Beginning teachers look back].* Utrecht, The Netherlands: Inspectorate of Education.

Janssen, J., Scheltens, F., & Kraemer, J. M. (2005a). *Rekenen-wiskunde groep 3–8: Handleidingen [Mathematics test grade 1 through 6: Manuals].* Arnhem, The Netherlands: Cito.

Janssen, J., Scheltens, F., & Kraemer, J. M. (2005b). *Leerling- en onderwijsvolgsysteem rekenen-wiskunde [Student progress monitoring system mathematics].* Arnhem, The Netherlands: Cito.

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS rekenenwiskunde voor groep 3 tot en met 8 [Scientific justification of the mathematics test for grade 1 through grade 6].* Arnhem, The Netherlands: Cito.

Keuning, J., Van Boxtel, H., Lansink, N., Visser, J., Weekers, A., & Engelen, R. (2015). *Actualiteit en kwaliteit van normen: Een werkwijze voor het normeren van een leerlingvolgsysteem [Up-to-dateness and quality of norms: A method to develop norms for a student progress monitoring system].* Arnhem, The Netherlands: Cito.

Klingner, J. K., Ahwee, S., Pilonieta, P., & Menendez, R. (2003). Barriers and facilitators in scaling up research-based practices. *Exceptional Children, 69*, 411–429. http://dx.doi.org/10.1177/001440290306900402.

Koellner, K., & Jacobs, J. (2015). Distinguishing models of professional development: The case of an adaptive model's impact on teachers' knowledge, instruction, and student achievement. *Journal of Teacher Education, 66*, 51–67. http://dx.doi.org/10.1177/0022487114549599.

Kulik, J. A., & Kulik, C. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly, 36*, 73–77. http://dx.doi.org/10.1177/001698629203600204.

Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*, 12–23. http://dx.doi.org/10.1016/j.tate.2008.06.001.

Leonard, J. (2001). How group composition influenced the achievement of sixth-grade mathematics students. *Mathematical Thinking and Learning, 3*, 175–200. http://dx.doi.org/10.1080/10986065.2001.9679972.

Lou, Y., Abrami, P. C., & Spence, J. C. (2000). Effects of within-class grouping on student achievement: An exploratory model. *Journal of Educational Research, 94*, 101–112. http://dx.doi.org/10.1080/00220670009598748.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & D'Appolonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*, 423–458. http://dx.doi.org/10.3102/00346543066004423.

McDonald Connor, C., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., ... Schatschneider, C. (2011a). Testing the impact of child characteristics x instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly, 46*, 189–221. http://dx.doi.org/10.1598/RRQ.46.3.1.

McDonald Connor, C., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science, 315*, 464–465. http://dx.doi.org/10.1126/science.1134513.

McDonald Connor, C., Morrison, F. J., Schatschneider, C., Toste, J. R., Lundblom, E., Crowe, E. C., et al. (2011b). Effective classroom instruction: Implications of child characteristics by reading instruction interactions on first graders' word reading achievement. *Journal of Research on Educational Effectiveness, 4*, 173–207. http://dx.doi.org/10.1080/19345747.2010.510179.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*, 114–140. http://dx.doi.org/10.1037/met0000078.

Murata, A. (2011). Introduction: Conceptual overview of lesson study. In L. C. Hart, A. Alston, & A. Murata (Eds.). *Lesson Study research and practice in mathematics education* (pp. 1–12). London, UK: Springer.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nomi, T. (2010). The effects of within-class ability grouping on academic achievement in early elementary years. *Journal of Research on Educational Effectiveness, 3*, 56–92. http://dx.doi.org/10.1080/19345740903277601.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237–257. http://dx.doi.org/10.3102/01623737026003237.

Prast, E. J., Van de Weijer-Bergsma, E., Kroesbergen, E. H., & Van Luit, J. E. H. (2015). Differentiation in primary school mathematics: Expert recommendations and teacher self-assessment. *Frontline Learning Research, 3*, 90–116. http://dx.doi.org/10.14786/flr.v3i2.163.

Puzio, K., Newcomer, S. N., & Goff, P. (2015). Supporting literacy differentiation: The principal's role in a community of practice. *Literacy Research and Instruction, 54*, 135–162. http://dx.doi.org/10.1080/19388071.2014.997944.

Raven, J. C., Court, J. H., & Raven, J. (1996). *Manual for Raven's standard progressive matrices and vocabulary scales.* Oxford, UK: Oxford Psychologists Press.

Rogers, K. B. (2007). Lessons learned about educating the gifted and talented: A synthesis of the research on educational practice. *Gifted Child Quarterly, 51*, 382–396. http://dx.doi.org/10.1177/0016986207306324.

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

Royal Dutch Academy of the Sciences (2009). *Rekenonderwijs op de basisschool: Analyse en sleutels tot verbetering [Mathematics education in primary school: Analysis and keys to improvement].* Amsterdam, The Netherlands: Royal Dutch Academy of the Sciences.

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*, 1186–1204. http://dx.doi.org/10.1080/13603116.2012.743604.

Schram, E., Van der Meer, F., & Van Os, S. (2013). *Omgaan met verschillen: (G)een kwestie van maatwerk [Responding to differences: (Not) a matter of customisation].* Enschede, The Netherlands: SLO.

Schweizer, K., Goldhammer, F., Rauch, W., & Moosbrugger, H. (2007). On the validity of Raven's matrices test: Does spatial ability contribute to performance? *Personality and Individual Differences, 43*, 1998–2010. http://dx.doi.org/10.1016/j.paid.2007.06.008.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best evidence synthesis. *Review of Educational Research, 57*, 293–336. http://dx.doi.org/10.3102/00346543057003293.

Smeets, E., Ledoux, G., Regtvoort, A., Felix, C., & Mol Lous, A. (2015). *Passende competenties voor passend onderwijs: Onderzoek naar competenties in het basisonderwijs [Adaptive competencies for inclusive education: Research about competencies in primary school].* Nijmegen, The Netherlands: ITS.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms and commentary* (3rd ed.). New York, NY: Oxford University Press.

Tieso, C. L. (2003). Ability grouping is not just tracking anymore. *Roeper Review, 26*, 29–36. http://dx.doi.org/10.1080/02783190309554236.

Tieso, C. L. (2005). The effects of grouping practices and curricular adjustments on achievement. *Journal for the Education of the Gifted, 29*, 60–89.

Van Groenestijn, M., Borghouts, C., & Janssen, C. (2011). *Protocol ernstige rekenwiskunde-problemen en dyscalculie [Protocol severe mathematics difficulties and dyscalculia].* Assen, The Netherlands: Van Gorcum.

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*, 297–312. http://dx.doi.org/10.1177/0016986208321809.

VanTassel-Baska, J., & Stambaugh, T. (2005). Challenges and possibilities for serving gifted learners in the regular classroom. *Theory and Practice, 44*, 211–217. http://dx.doi.org/10.1207/s15430421tip4403_5.

Van de Weijer-Bergsma, E., Kroesbergen, E. H., Jolani, S., & Van Luit, J. E. H. (2016). The monkey game: A computerized verbal working memory task for self-reliant administration in primary school children. *Behavior Research Methods, 48*, 756–771. http://dx.doi.org/10.3758/s13428-015-0607-y.

Van de Weijer-Bergsma, E., Kroesbergen, E. H., Prast, E. J., & Van Luit, J. E. H. (2015). Validity and reliability of an online working memory task for self-reliant administration in school-aged children. *Behavior Research Methods, 47*, 708–719. http://dx.doi.org/10.3758/s13428-014-0469-8.

Vogt, F., & Rogalla, M. (2009). Developing adaptive teaching competency through coaching. *Teaching and Teacher Education, 25*, 1051–1060. http://dx.doi.org/10.1016/j.tate.2009.04.002.

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

Ysseldyke, J., & Bolt, D. M. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36*, 453–467.

Ysseldyke, J., Spicuzza, R., Kosciolek, S., Teelucksingh, E., Boys, C., & Lemkuil, A. (2003). Using a curriculum-based instructional management system to enhance math achievement in urban schools. *Journal of Education for Students Placed at Risk, 8*, 247–265. http://dx.doi.org/10.1207/S15327671ESPR0802_4.