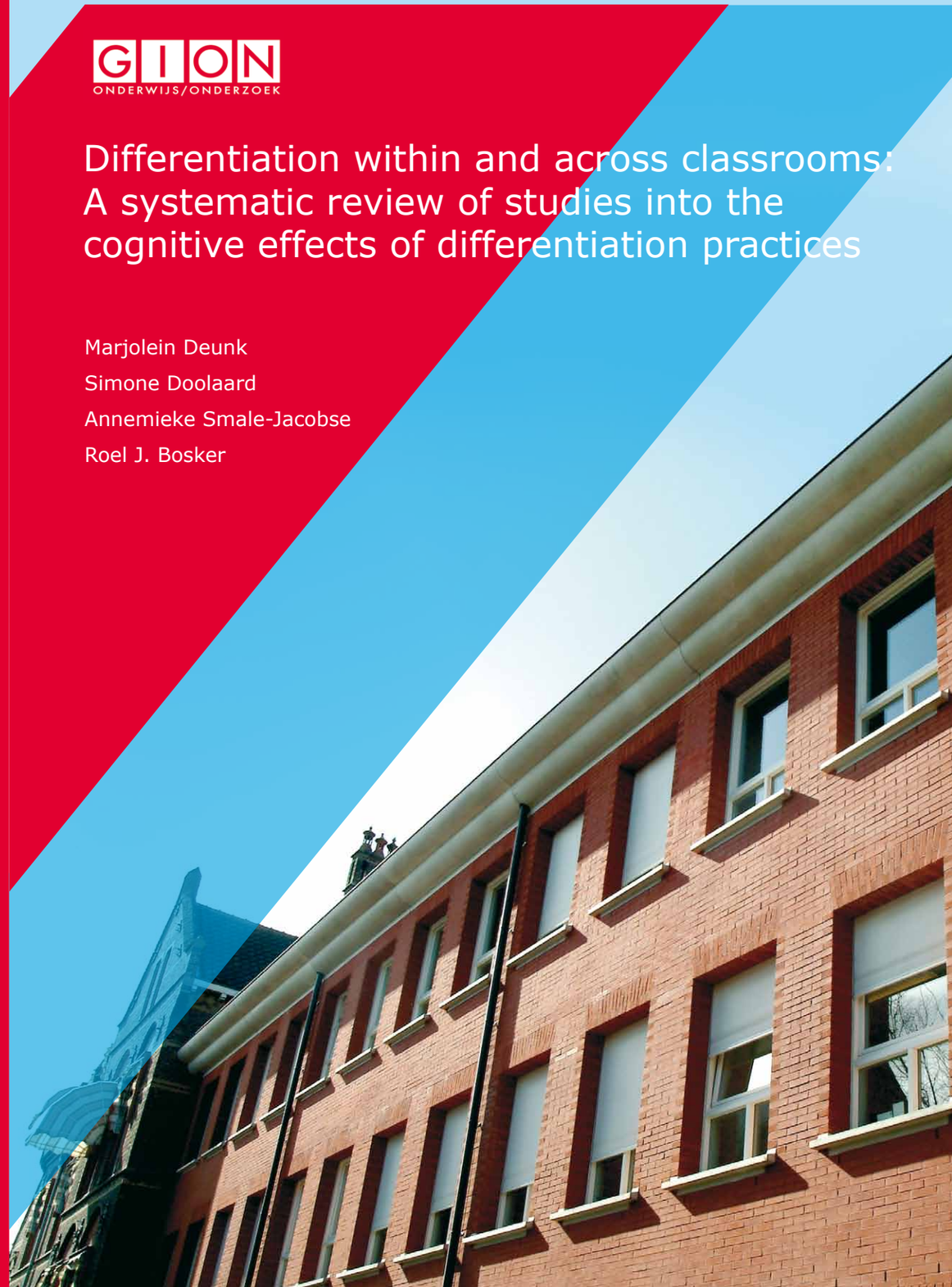
GION
ONDERWIJS/ONDERZOEK

Differentiation within and across classrooms: A systematic review of studies into the cognitive effects of differentiation practices

Marjolein Deunk
Simone Doolaard
Annemieke Smale-Jacobse
Roel J. Bosker



opmaak en foto van cover: www.hansvandijk.nl

Differentiation within and across classrooms: A systematic review of studies into the cognitive effects of differentiation practices

Marjolein Deunk
Simone Doolaard
Annemieke Smale-Jacobse
Roel J. Bosker

ISBN 978-90-6690-586-3

1. Introduction

The quality of schools is for an important part determined by the way teachers deal with cognitive differences between students and adapt their instruction to individual needs. In order to achieve this, teachers need advanced professional skills to deal with these differences, apart from basic skills of classroom management and general didactics. They need to have insight in (differentiated) performance goals, be able to interpret students' current levels based on classwork and test scores, decide what students of different levels need to learn, and they need to know how to teach these students with varying cognitive abilities. Furthermore, teachers need to be aware of school wide decisions about the aim of providing adaptive instruction and the effect of different classroom practices aimed at low, average or high performing students. The combination of these attitudes, knowledge and practices is called differentiation.

There are different teaching strategies that can be used to differentiate in classes and in schools. Schools can create heterogeneous classes or - based on general ability of the students – homogenous classes. Homogeneous classes are generally applied in secondary education (e.g. Ireson, Hallam, & Plewis, 2001), while heterogeneous classes are the standard in early childhood education and primary education. Within heterogeneous classes, teachers can make use of homogeneous grouping (also referred to as ability grouping) or heterogeneous grouping (e.g. Lou et al., 1996; Slavin, 1987a). Furthermore, in heterogeneous classrooms, teachers may provide adapted instruction and offer adapted learning content, in which the lower ability students may receive more time to master the core learning content (e.g. Anderson & Algozzine, 2007; de Koning, 1973; George, 2005; Reezigt, 1993).

Which teaching strategies teachers choose to use seems to relate to the implicit or explicit learning goals they have for their classroom as a whole. From a 'theoretical' point of view teachers can strive for convergence or divergence (Blok, 2004; Bosker, 2005). Teachers aiming at convergence are mainly focusing on reaching a minimum performance level with all of their students, which implies they might have to dedicate additional time and effort to the low achieving children in order for them to reach that minimum performance level, even when this goes at the expense of the high ability children, who by consequence receive less attention. Teachers aiming at divergence mainly focus on helping *all* children to reach their highest potential, equally dividing attention between students with lower and higher ability. Their use of ability-appropriate performance goals for (groups of) students of different ability levels, may lead to a widening of the gap between lower and higher ability students. In practice though, most teachers will combine convergent and divergent goals and will try to reach a minimum performance level with the low ability students, while also offering high ability children the opportunity to extend their knowledge without proceeding (too much) ahead of their peers in the classroom. The achievement distributions resulting from convergent and divergent differentiation are depicted in Figure 1, including the regression lines indicating the relation between post- and pre-test. In the figure on the left hand side, the lines A and B

are initially further apart but approach each other in time, indicating the relative better progress of the initially lower achieving students. In the figure on the right hand side the difference between lines A and C widens over time, indicating the relative better progress of the initially higher achieving students.

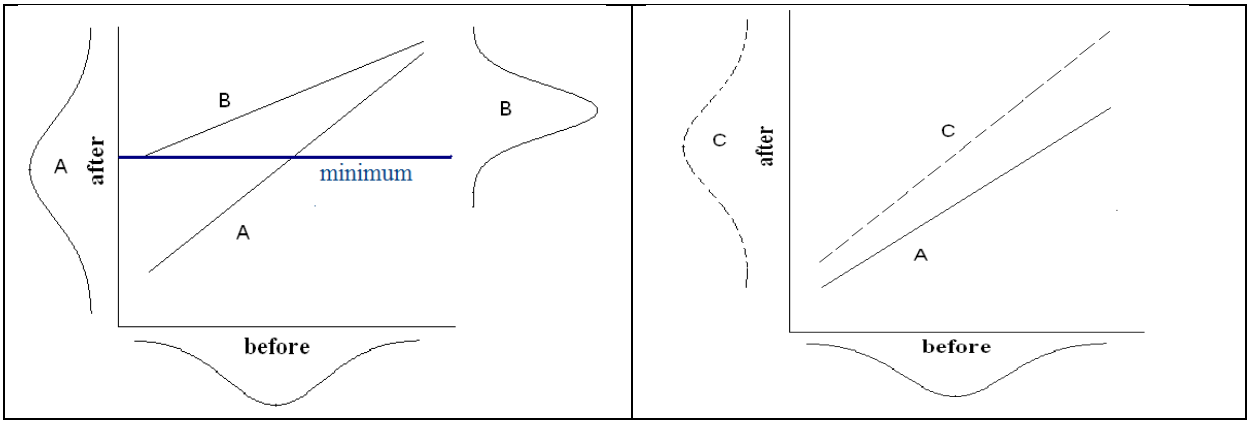


Figure 1: *Convergent (left) and divergent (right) differentiation compared with respect to the effects on the distribution for initially low and high achieving students*

Broadly speaking, there are three problems related to using differentiation in education:

1. teachers are not always fully aware which differentiation goal they (should) strive for (de Koning, 1973),
2. the potential convergent or divergent effects of varying differentiation strategies are not fully clear, as research shows mixed results, and
3. therefore it is difficult for teachers to make explicit decisions on when to use which differentiation strategy, for what goal.

Ability grouping, as a form of differentiation, has been studied extensively. Five key meta-analyses of studies on ability grouping until 1995 are conducted by Kulik and colleagues (1982; 1984), Lou and colleagues (1996) and Slavin (1987a; 1987b; Slavin, 1990). Kulik and colleagues focused on homogeneous ability grouping in primary (1982) and secondary education (1984), Lou and colleagues (1996) focused on homogeneous and heterogeneous grouping in primary and (post)secondary education, and Slavin focused on homogeneous ability grouping in primary (1987a) and secondary education (1990) and on mastery learning in primary and secondary education (1987b). The findings of these key studies will be described in the theoretical framework in chapter 2.

A difficulty in summarizing the effects of studies on ability grouping is that ability grouping is operationalized in different ways and these differences are likely to influence the outcome of the study. Slavin (1987a) pointed to the different ways grouping can be organized, for example temporarily within classes, between classes or between grades (for example Joplin Plan), special classes for high or low achievers or within-class homogeneous ability grouping for specific subjects. This last form of grouping is most common in elementary classrooms. Teachers may assign students to reading or math groups of different achievement

levels or may start with whole-group instruction and offer remediation or enrichment afterwards, while the other students work independently. Many modern learning materials provide content based on ability, with basic content for the whole group, followed by rehearsal or enrichment material, depending on the level of mastery of individual students. This helps teachers in offering differentiated learning content to the students in the classroom.

Partly due to the mixed research results, the use and effects of ability grouping are much debated. Arguments in favor of working with small homogeneous groups are that instruction, learning pace and learning materials can be better adjusted to the needs of the students, which will enhance their learning. Arguments against working with small group homogeneous groups are that students have less interaction with the teacher, who has to divide his/her attention between multiple groups. Most concerns are related to the learning opportunities of low ability students in small homogeneous groups: within these groups, they cannot profit from the input of higher ability peers or from the role models that high ability students can be. Furthermore, teacher expectations of low ability students may be lower, leading students in low ability groups to have less opportunity-to-learn. Finally, students in lower ability groups may experience difficulty in moving upwards to higher ability groups, especially when the gap between lower and higher ability students increases. The variety of research results suggest that children with different ability levels may profit from being part of either homogeneous or heterogeneous groups, but, in general, early selection in which children are placed in low ability homogeneous classes for longer times at a young age will put them at a disadvantage. This is especially relevant for children from impoverished backgrounds and/or minority groups, who might be labeled as being of 'low ability' before they had been able to show their potential. When these children are placed in a low ability class too soon – based on general estimates or even prejudices, rather than on actual performance level - they might encounter low expectations, less demanding teaching and unequal opportunities. Or, according to Slavin: "ability grouping [for a prolonged period, at a young age, SD] goes against our democratic ideals by creating academic elites (…) the use of ability grouping may serve to increase divisions along class, race, and ethnic group lines." (Slavin, 1987a, p.297).

The aim of the current review is to analyze existing research on differentiation from 1995 onwards and add to the insights in how differentiation practices can positively affect the language and math performance of low, average and high ability students. Because of the specific characteristics of different educational age groups, the review will separately focus on early childhood education and kindergarten (2;6 to 6 year olds), primary education (6 to 12 year olds) and early secondary education (12 to 14 year olds)[1]. The review does not focus on grouping only, although many studies may focus on grouping practices without specifying

[1] When interventions were conducted in overlapping age groups, the studies were presented in both sections. In case of follow-up measures, the study is described in the section where the intervention is conducted only. Originally, we intended to include studies from 2;6 to 16 years old, but finally we decided to limit the upper age to 14 years.

whether or not ability grouping creates a context for differentiating in for example learning time, learning content, learning materials, adaptive testing or adaptive instruction. One-to-one tutoring is excluded, since this educational practice is focused on some individuals instead of the performance of the entire class. Studies focusing exclusively on tutoring are excluded as well, although peer tutoring, as such or as an element cooperative learning, can be part of working in differentiated groups. Furthermore, all the different ways in which teachers may take into account performance differences of students are considered in this review.

2. Theoretical framework: Situation up to 1995

2.1. Tracking or whole class ability grouping

Kulik and Kulik (1984) conducted a meta-analysis on ability grouping in primary education. They focused on whole class ability grouping, in which students are assigned to classrooms based on their ability. Overall, students in homogeneously grouped classrooms had better achievement than students in heterogeneous classrooms, although the effect size (ES)[2] is small (ES=+0.19). However, these effects can be explained by studies focusing only on special classes for gifted students. Studies focusing on the entire population of low, average and high achievers show much smaller effects of homogeneous grouping (ES=+0.07). Also Slavin (1987a) described the effects of whole class ability grouping in primary education. He only included programs targeting students from low, average and high ability (thus rejecting whole class grouping for gifted students) and found no overall effect of this type of grouping (effect sizes range from ES=-0.15 to +0.15, with a median of 0.00).

The authors referred to above conducted studies on whole class ability grouping in secondary education as well. Results from the study of Kulik and Kulik (1982) were that performance of students in homogeneous classrooms was higher than performance of students in heterogeneous classrooms. The general effect size was small (ES=+0.10), although the range of effect sizes found in different studies is large, from ES=-1.00 to ES=+1.25. Just like in their study of 1984, effects disappear when only studies are included focusing on the entire population of high, average and low performing students (ES=+0.02). Similar to his study on primary education, Slavin (1990) only included studies that focused on the entire population of low, average and high performing students in his meta-analysis on whole class ability grouping. Overall, he found no effect of grouping, just like in his study on primary education (ES=-0.02).

Regarding differential effects for low, average and high ability students, Slavin (1987a) found inconsistent results for students of different ability levels: some studies included in the review found negative effects for low ability students and positive effects for high ability students, but others found the opposite pattern or no differential effects at all. Effect sizes of individual studies ranged for low achievers from -0.46 to +0.64, for average achievers from -0.11 to +0.22 and for high achievers from -0.24 to +0.54. Kulik and Kulik (1982; 1984) did not report differential effects for whole class ability grouping or tracking in primary or secondary education. They only looked at the effects of grouping programs targeted specifically at gifted or impaired students. Whole class ability grouping for gifted students had positive effects on these gifted students in primary education (ES=+0.49) and in

[2] In this chapter we refer to effect sizes with ES, indicating that these were reported effect sizes. In the chapter where we present the results of our review we will use *d*, since we recalculated all the research results ourselves, and expressed and summarized them as the effect size *d*, being the standardized mean difference between a treated and an untreated group.

secondary education (ES=+0.33), but effects of this 'extraction' of high performing students on the performance of average and low ability students that remain in the regular classrooms were not reported. Slavin (1990) looked at differential effects of ability grouping in secondary education. He found virtually no differential effects for high (ES=+0.01), average (ES=-0.08) and low achievers (ES=-0.02).

2.2. Setting

Setting is between-class ability grouping for specific subjects. It can be organized with parallel classrooms of the same grade level or across grade levels. The regrouping is (in theory) done on the basis of actual performance in the specific subjects, instead of more general intelligence or ability measures.

Slavin (1987a) describes the effect of regrouping for reading and/or mathematics between classrooms, but *within grades*, which is of course only feasible in larger schools. According to Slavin, the studies that qualified for his best evidence synthesis did not provide conclusive evidence on the effectiveness of grouping for specific subjects compared to ordinary heterogeneous classrooms. He considered the quality and quantity of the eligible studies to be insufficient to draw conclusions on the overall effects. The total effect sizes of regrouping for specific subjects compared to heterogeneous classrooms of the individual studies range from -0.28 to +0.43.

Slavin (1987a) also studied the effect of regrouping for specific subjects *across grades*. In this arrangement, students are temporarily regrouped based on performance level, irrespective of grade level, meaning for example that high performing grade 2 students can be placed together with low performing grade 3 students. The studies in Slavin's review show positive effects of between-class grouping across grades (ES=+0.45).

Because Slavin (1987a) considered the studies in his best-evidence synthesis on setting not strong enough to draw firm conclusions of general effects, no overall differential effects are reported either. Individual studies indicate more positive effects for high ability than low ability students though. Effects for high achieving students range from ES=-0.25 to ES=+0.79, for average achieving students from ES=-0.33 to ES=+0.22 and for low achieving students from ES=-0.41 to E.S.=+0.32. Slavin reported no overall significant differential effects for between-class grouping across grades. He stated: "In no case did one subgroup gain at the expense of another; either all ability levels gained more than their control counterparts or (…) none did." (Slavin, 1987a, p.317).

2.3. Within-class ability grouping for specific subjects

Slavin (1987a) described the effect of within-class ability grouping in primary education, a common and relatively easy way of organizing grouping in primary education. According to Slavin, studies regarding this type of grouping are most likely to use random assignment, thus potentially leading to more valid research results in terms of causal attribution. Almost all

eligible studies in Slavin's review, concern within-class ability grouping for mathematics. Generally, the studies show positive effects for homogeneous within-class ability grouping compared to no grouping (randomized studies: ES=+0.32; nonrandomized studies: ES=+0.36). In his study on grouping in secondary education, Slavin (1990) described the few available studies on within-class grouping in secondary education and found no effects (ES=-0.02), contrary to the findings in primary education.

Homogeneous ability grouping is not the only way of handling differences in the classroom. One may also use heterogeneous grouping and let students of different abilities engage in cooperative learning. Lou and colleagues (1996) conducted a meta-analysis of studies on within-class grouping in elementary, secondary and post-secondary education in the period 1965 to 1995 and analyzed the effects of grouping versus whole class activities as well as the effects of homogeneous versus heterogeneous within-class ability grouping. They found a small overall effect of small group instruction, either homogeneous or heterogeneous, over whole class instruction (ES=+0.17). Like in the other reviews, there were substantial differences within individual studies, some favoring small group instruction, some favoring whole class instruction. This was however not caused by the combined analysis of homogeneous and heterogeneous ability grouping, since both had similar positive effects compared to whole class instruction (respectively ES=+0.16 and ES=+0.19). When homogeneous and heterogeneous ability grouping were directly compared, an overall advantage of homogeneous ability grouping was found (ES=+0.12).

Mastery learning can be seen as a special form of within-class ability grouping. Classrooms using mastery learning use regular progress assessment to check whether students reach certain ability levels. The group that does not perform well enough, receives additional instruction inside or outside the classroom. The group that does, may receive advanced materials for enrichment. Every thematic unit starts with whole class instruction; ability groups are created based on students' actual performance. Slavin's (1987b) meta-analysis of studies on mastery learning, in which the control group was provided equal learning time and in which effects were measured using standardized tests, showed a small median effect size (ES=+0.04). Studies which used tests developed by the researchers showed a larger median effect (ES=+0.26). Four other studies compared classrooms with mastery learning with additional instruction time with control classes that did not receive additional time. These studies had a median effect size of +0.31, although Slavin argues that a median effect size is difficult to interpret because the four studies differ too much from each other. Taken together, Slavin concluded that mastery learning is not more effective than traditional instruction, when equal amounts of learning time are provided. But it does seem to help teachers to focus on instructional objectives, as is indicated by the results of studies using researcher developed tests, that resemble the content taught more closely than standardized tests.

Slavin (1987a) cautiously described that within-class homogeneous ability grouping is especially beneficial to low achievers (ES=+0.65), followed by high achievers (ES=+0.41), followed by average achievers (ES=+0.27). Lou and colleagues (1996) found a different

pattern for homogeneous ability grouping within the classroom. They found that only medium ability students benefit from learning in small homogeneous groups (ES=+0.51). Homogeneous within-class grouping had negative effects on low ability students, compared to heterogeneous within-class grouping (ES=-0.60). For high ability students it made no difference whether they were placed in small homogeneous or heterogeneous groups. Lou and colleagues found that grouping in general was beneficial to students of all ability levels, when compared to whole class instruction. They showed that low ability students profited most of small grouping (ES=+0.37), followed by high ability students (ES=+0.28), followed by medium ability students (ES=+0.19).

2.4. Grouping and adaptive teaching

The mixed results of the studies in the meta-analyses indicate that more factors play a role in the effectiveness of ability grouping. Lou and colleagues (1996) and Slavin (1987a) emphasized the important role of adapting instruction to the needs of the group. Lou and colleagues state that "Overall, it appears that the positive effects of within-class [both homogeneous and heterogeneous, MD] grouping are maximized when the physical placement of students into groups for learning is accompanied by modifications to teaching methods and instructional materials. Merely placing students together is not sufficient for promoting substantive gains in achievement." (Lou et al., 1996, p. 448). Also Slavin notices that, for grouping arrangements to have an effect, learning materials and instruction should be adapted: "regrouping for reading and/or mathematics can be effective if instructional pace and materials are adapted to students' needs, whereas simply regrouping without extensively adapting materials or regrouping in all academic subjects is ineffective." (Slavin, 1987a, p.311). Unfortunately, as Slavin notes, many studies do not provide specified information on the instructional practices used in interaction with small ability groups. Lou and colleagues (1996) analyzed the results of studies that did provide (some) information on teacher practices. They found larger effects for within-class grouping when teachers adapted their instruction when teaching to small groups (ES=+0.25) compared to teachers who provided 'whole class instruction' to small groups (ES=+0.02).

From his best evidence synthesis, Slavin (1987a) extracted some criteria that are likely to influence the effect of ability grouping focused on convergent differentiation. The first criterion is that the grouping must lead to homogeneous groups in the skill being taught. Groups based on more general performance may actually not be very homogeneous regarding the skill being taught, leading to poorly formed ability groups. The second criterion is that groups must be flexible. Students assigned to tracked classrooms are likely to remain in the classroom for a long period, while students grouped within or between classrooms only for specific subjects may be reassigned to groups of different levels more easily. The third criterion is that teachers adapt their teaching to the needs of the different ability groups. There appear to be quality differences in the appropriateness of the instruction, learning materials and learning content different ability groups receive. Frequent formative assessment seems to

be necessary to be able to adapt to the students' needs. Another important aspect is the instruction time that students receive. The more ability groups a teacher creates, the less time there is available for each group and the more time students have to spend working independently. The use of three ability groups is most common, but whether this is more effective than for example two or four ability groups remains unclear.

2.5. Evidence from previous meta-analyses

Considering the results from meta-analyses on differentiation up to 1995, several conclusions can be drawn. First of all, whole class ability grouping or tracking seem to have no effects when the entire population of low, average and high performing students is taken into account. Differential effects of tracking are inconclusive. Tracking, or between-class ability grouping may have positive effects, especially when grouping is done across grades. Again, differential effects are inconclusive, although across grade grouping seems to be beneficial for all ability groups. Within-class ability grouping also seems to have positive effects, although effect sizes of this type of grouping are smaller than the effect sizes of between-class grouping. Within-class grouping seems to be beneficial due to the combination of small group instruction and homogeneous grouping. Differential effects however are inconclusive: in the review of Slavin (1987a), within-class ability grouping is most beneficial to low achievers. In contrast, Lou and colleagues (1996) reported that low achievers indeed benefit from grouping, but not from homogeneous grouping. Within-class heterogeneous grouping may be more beneficial for low ability students, according to Lou and colleagues. Slavin as well as Lou and colleagues emphasize the importance of adapted instruction and learning materials in combination with grouping: grouping alone is not enough, it is merely a context for the teacher to apply adequate teaching practices, adapted to the needs of different students. This is confirmed by Slavin (1987b) who suggests that the lack of effects of mastery learning may have to do with insufficient quality and quantity of corrective instruction.

Based on the previous research no general effects are expected for whole class ability grouping or tracking, unless within-class grouping is used within the tracked classrooms or other adaptive high quality teaching methods are used (Slavin, 1990). When differential effects are found, it is expected that whole class homogeneous grouping has negative effects on low ability students, since it is less likely that students are then instructed in smaller groups and since this configuration excludes the possibility to work in heterogeneous ability groups for part of the time. Positive differential effects for streaming and within-class homogeneous grouping are expected, provided that high quality adaptive instruction is offered to the different ability groups. These effects are expected to be positive for low, average and high ability students.

3. Method

The effectiveness of different differentiation practices are studied by applying a best evidence synthesis, which is a meta-analysis extended with additional contextual information on the selected studies, with an emphasis on studies that are particularly relevant to the topic under study (Slavin, Lake, Chambers, Cheung, & Davis, 2009). In an attempt to perform the most comprehensive literature search, both an electronic database search and a cited-references search is conducted. In order to find as many relevant sources as possible, the electronic database search starts broadly and the number of results is narrowed down by manually applying additional selection criteria. Effect sizes are calculated for each eligible study. Content coding is performed in order to create an overview of the different types of studies and the different elements of differentiation studied. This information is used to provide context to the effect size data of the meta-analysis.

3.1. Literature search procedures

An extensive literature search was conducted in the educational databases ERIC, psychINFO and SSCI. The databases were searched by making use of 10 keywords, which were used twice: once in combination with the keyword *achiev** and once in combination with the keyword *effect**. The ten keywords are: *"ability group*", "adapt* instruct*", "adapt* teach*", "aptitude treatment", differentiat*, grouping*, "individuali* instruct*", "individuali* teach*", "mastery learning"* and *streaming*. Papers in which these keywords are mentioned in the abstract were included in the initial selection, provided they were: articles published in peer-reviewed journals, published between 1995 and 2012, written in English and aimed at the age-category 2-16 years (i.e. preschool – secondary education).

In addition to the database search, a 'cited references' search was conducted. Eleven key publications on differentiation were selected, namely Blok (2004), Borman et al. (2005), de Koning (1973), Gamoran and Weinstein (1998), Irseon and Hallam (2001), Kulik and Kulik (1982), Lou et al. (1996), Reezigt (1993), and Slavin (1987a; 1987b; 1990). Three of the key publications (Blok, 2004; de Koning, 1973; Reezigt, 1993) are based on the educational context in the Netherlands. Using the SSCI database, all papers published from 1995 onwards, that made reference to one of these eleven key publications were collected.

These two broad search methods led to a large amount of references, which was narrowed down by manually applying selection criteria. The first broad selection criterion was whether the study was on language or math or not. Language in this case encompasses reading, writing, vocabulary, grammar etc. in the native language of the country under study (i.e. no foreign language studies). The selection was based on title, abstract and keywords. In case of doubt, the paper remained included in the selection. Abstracts which indicated that studies did not focus on students up to 16 years of age, were not linked to education, did not include effects on language- or math performance, were case studies, or did not make use of

empirical research methods, were rejected. Applying these criteria narrowed down the number of references. Of this narrowed down selection, the full text papers were collected.

3.2. Inclusion criteria

A set of 8 final criteria for inclusion was applied to the selection of full text papers. The first criterion focused on the content of the study, the second was practical and the third to eighth focused on the quality of the study. The criteria were based on those used in the best evidence syntheses conducted by Slavin and colleagues (1987a; 2008; 2009).

1. The study addresses effects of differentiation on language or math performance of all students or groups of students in a classroom. The intervention takes place 'inside' the classroom (i.e. no out-of-class tutoring), during the regular school day.
2. The study could have taken place in any country, but the report had to be available in English.
3. The intervention has a minimum duration of 12 weeks, measured from beginning of treatment to posttest.
4. Each treatment group consists of at least 15 students and of at least two teachers that are involved in the study.
5. The study compares children taught in classes using a given intervention to those in control classes using another intervention or standard teaching practice ("business as usual"). Or the study uses secondary data analysis on existing databases in order to compare groups of classes.
6. The study uses random assignments or matching or conditioning with appropriate adjustments for any pretest differences (e.g. ANCOVA). Studies without control groups are excluded.
7. The study provides pretest data, unless the study uses random assignments of at least 30 units (students, classes or schools) and there are no indications of initial inequality. Studies with pretest differences of more than 0.50 of a standard deviation are excluded.
8. The dependent measures include quantitative measures of performance, such as standardized reading measures. Experimenter-made measures were accepted if they were comprehensive measures that would be fair to the control group, but measures inherent to the experimental program were excluded.

From the included papers[3], relevant data was selected to calculate effect sizes. In addition, these studies were coded for content. The content coding included: grade under study, type of differentiation, country (and state) in which the intervention is conducted, sample size, duration of intervention, dependent variables and instrumentation and external variables and covariates (if applicable). In addition, a short summary is made of the study, its effects, drawbacks and strong points, and its relevance for the best evidence synthesis.

[3] A full list of all the references found is available upon request.

3.3. Additional relevant sources

Relevant studies on (aspects of) differentiation could also be found in other sources than papers published in academic, peer-reviewed journals. Therefore, an additional electronic search was performed in the databases ERIC and psychINFO. The search criteria were similar to the search of the journal articles, except for publication type, which could be books, dissertations and theses or reports. The references that were found in this search were checked against the selection criteria applied to the abstracts as described above. Subsequently, the most relevant sources were selected and used for contextual information on differentiation in the different age groups.

3.4. Computation of effect sizes

To be able to compare the effects of the different studies, all results are converted to Cohen's *d*, which is the standardized mean difference between groups. The ways of calculating *d* when using different types of data stemming from various research designs are described in Borenstein et al. (2009). When correlations between pretest and posttest were needed for calculating *d*, but were not provided in the study at hand, a pre-post correlation of 0.70 was assumed. Next to *d* estimates for its 95% confidence interval are presented. If the reader is interested in either more conservative or more liberal intervals, these can be simply derived from the estimates presented.

For every study a general *d* is calculated. When multiple outcome measures are used, they are labeled as either measures of math, vocabulary, reading or reading comprehension, since this is more informative than the names of individual tests, which vary between studies. If possible, differential effect sizes for high, average and low performing students are provided. The effect of differentiation is considered to be *divergent* when the effect size *d* is largest for high ability students and *convergent* when the effect size *d* is largest for low ability students.

3.5. Meta-analysis

In some specific instances it is possible to combine results of different studies into one summary effect size (c.f. Borenstein et al., 2009). These instance are:

1. The studies have the same topic (e.g. within-class ability grouping);
2. The studies are conducted in the same stage of the education system (ECE and kindergarten; primary; early secondary);
3. The studies focus on the same subject domain (either reading or mathematics).

In a statistical meta-analysis the crucial information (an effect size and a standard error suffice) is summarized as a weighted average, with weights being inversely proportional to the magnitude of the standard errors. And the standard error for this summary effect size is derived from the standard errors of the individual studies. A quite surprising result may be that the summary effect size may have a standard error so small that the resulting confidence

interval for the effect size estimate does not contain zero, whereas none of the individual studies had produced a significant effect. The reason of course is that in the summary effect size and its standard error all the samples from the different studies are more or less combined into one very big sample. The meta-analyses were conducted using the CMA-software developed by Borenstein et al. (2009). In meta-analyses in which multiple outcomes from the same study are used, the results are adjusted and the adjustment factor is presented in a note to the table.

4. Results

4.1. General results of the literature search

The broad database search in ERIC, psychINFO and SSCI, using the 10 keywords related to differentiation, led to 2,478 unique references[4]. In addition, a cited reference search was conducted based on the 11 key publications. This led to an additional 262 new references, adding up to 2740 potentially interesting references. Of these, about 500 seemed relevant at first sight, mostly studies regarding primary education[5]. Careful reading of the abstracts led to a selection of approximately 200 papers eligible for further analysis based on their full text versions. The final 8 inclusion criteria (see paragraph 3.2) were applied to the full text papers. A total number of 26[6] journal articles met de inclusion criteria and were used in the meta-analysis.

In addition, potentially interesting books, reports and theses were searched using the same key words used in the general database search. This resulted in 828 publications, of which 97 seemed appropriate, based on the general inclusion criteria. Of this set of books, reports and theses, the 10 most relevant sources were selected manually. They were not included in the meta-analysis, but used for gathering theoretical background information.

4.2. Effects of differentiation in Early Childhood Education and Kindergarten (2;6-6 years)

4.2.1. Overview of differentiation in ECE and Kindergarten

Early childhood education (ECE) is designed to stimulate children in their development, reduce and prevent learning- and language delays and to prepare children for formal education. Preschool and kindergarten teachers have to deal with children from very different (language) backgrounds and different starting levels and aim to help them all to acquire the minimum level needed to enter first grade. The goal of differentiation in early childhood education is thus mainly convergent.

Most studies on differentiation activities in early childhood education focus on (emergent) literacy and early reading. This is not surprising, as language and literacy development is one of the core tasks of ECE, especially when it is aimed at second language learners and/or children from impoverished backgrounds with limited language input at home. The type of differentiation that is typically used is within-class homogeneous ability grouping:

[4] Three of the searches in SSCI resulted in over 1000 hits (*differentiat* & achiev*; differentiat* & effect*; grouping* & effect**). These are narrowed down by selecting the "web of science categories": *education, educational research* and *psychology educational*.

[5] With the distribution ECE and kindergarten : primary education : secondary education being 1 : 2 : 4

[6] The article of Tach and Farkas (2006) is used twice.

the classroom is divided into small groups of children of similar proficiency levels, who receive specific, proficiency level appropriate instruction in literacy or early reading skills.

Ability grouping in preschool and Kindergarten appears to be not as straightforward as depicted above. Ongoing assessment and frequent re-grouping is considered to be important (Slavin, 1987a), but details on how this is applied are often not reported in research. Furthermore, ability groups are not always as homogeneous as they are supposed to be, for example because proficiency is not measured well enough or because other student features are emphasized as well, such as student interest, learning style and gender as a base for grouping. Also secondary goals of grouping play are role, like stimulating self-regulated learning, enhancing student ownership in learning and maintaining a positive classroom atmosphere. These secondary goals are advocated by Tomlinson (2000), a scholar who is specialized in differentiation and writes primarily for an audience of practitioners. Also Howard Gardner's (1984) work on multiple intelligences and variation in learning style is mentioned in this respect. Other factors than ability alone thus seem to play a role in ability grouping.

The problem with this 'broad view' on differentiation is that the more student features are taken into account, the more difficult it becomes to create homogeneous groups. In theory, teachers could first group students based on performance and then make smaller subgroups based on for example learning style, as described by Neel (2008) in her study on reading in first grade. However, this would only be feasible when working with a larger group of students/classrooms in a school. Another problem of grouping based on multiple student features is that it further decreases the transparency of the educational practice of differentiation. In many studies it is not clear on what basis ability groups are formed, how teachers designed their instruction plans focusing on different groups of students and how (well) they implemented them.

A complicating factor in the interpretation of the effects and meaning of grouping in early childhood education is discussed by McCoach (2003), who suggests that grouping for reading in 1st grade reflects a traditional teaching approach, while traditional Kindergarten teachers would probably not use achievement grouping. On the contrary, the Kindergarten teachers who use achievement grouping may be innovative in their teaching and more focused on academic results, according to McCoach. The effects of grouping may thus be confounded by teacher characteristics that are associated with a tendency to use grouping and this may especially be the case in ECE and kindergarten classrooms. This illustrates again that results on grouping are difficult to interpret without detailed information on how teachers create and treat these groups.

4.2.2. Selected studies

In the initial database search, approximately 50 papers focusing on education in preschool or kindergarten were found. Approximately 15 papers were selected for further inspection based on their full text versions. Of these, seven papers met the inclusion criteria, described in the

general method section (paragraph 3.2). These selected papers are alphabetically listed and summarized in appendix 1.

Of the seven selected studies on differentiated instruction in early childhood education, six are based on ECLS-K data. This data originates from the Early Childhood Longitudinal Study (ECLS), in which development, school readiness and school experiences are investigated in three large groups of children. The first group is followed from birth to kindergarten (ECLS-B), the second group is followed from kindergarten (entry in 1998-1999) to 8th grade (ECLS-K) and the third group will be followed from kindergarten (entry in 2010-2011) to 5th grade (ECLS-K: 2011). The study is conducted in the United States by the Institute of Education Sciences and the National Center for Education Statistics. The studies in the current review are based on the first cohort of kindergartners (ECLS-K). The ECLS database is for the most part publically available to researchers. A wide range of child-assessments are used in the ECLS-K: reading, mathematics, general knowledge, social-emotional and physical development. However, most of the studies included in the current review only make use of the reading/literacy measures, and one focuses on math growth.

4.2.3. Literature synthesis

General overview

Ability grouping is measured in different ways in the selected studies, sometimes very broad and sometimes in more detail, ranging from whether grouping is used at all (Adelson & Carpenter, 2011) to how often it is used per week (D. B. McCoach, O'Connell, & Levitt, 2006), to how many time a day is spent on grouping (Chang, 2008; Hong & Hong, 2009; Hong, Corter, Hong, & Pelletier, 2012). In general, ability grouping in early childhood education seems to have a positive effect. Most studies report positive effect sizes for grouping, for students of all levels (*d* ranges from +0.068 to +1.276). Due to too big differences between the studies in terms of operationalization of differentiation, it was not possible to perform meta-analyses on the studied included.

Only two studies look into differential effects for low, average and high performing students (namely Gettinger & Stoiber, 2012; Hong et al., 2012). The effects of the studies seem contradictive and have to do with the amount of instruction time students receive when grouped. Hong and colleagues (2012) conclude that if relatively little time is spent on reading, intensive grouping, compared to whole class instruction, is not beneficial to students of all ability levels. Gettinger and Stoiber (2012) describe an intervention of ability grouping with an emphasis on adaptive education and high quality instruction and found this to be beneficial for all students, including low performing students. Ability grouping under these conditions is most beneficial to average ability students, followed by low ability students, followed by high ability students (Gettinger & Stoiber, 2012). The effect of differentiation in this study is thus neither divergent nor convergent, as the gap between high ability students and their classmates does not enlarge, but the low ability students do not approach their average performing peers either.

Results of the included studies

The only study in the selection using a randomized controlled trial is the one of Gettinger and Stoiber (2012). They studied the effect of an early literacy intervention based on close monitoring and assessment of students' progress and adjusting instruction based on the monitoring results (for key features and estimated effects, see appendix 1). The progress monitoring is used for providing additional small-group instruction to low performing students, adjusting the general whole class instruction and providing additional challenge to the group of high performing students. The way teachers were supposed to monitor performance and adjust their teaching- and lesson plans for different groups of students is described in detail, which is an exception in empirical studies on differentiation in early childhood education and kindergarten. More details on the content of the program are described below in paragraph 4.2.4. A total of 124 3- and 4-year olds in 15 classrooms were included in the study. Eight classrooms (62 preschoolers) were randomly assigned to the intervention condition, which lasted for 4 months. A drawback of the design is that students in the experimental condition received more practice with the content and format of the effect measures due to the monthly progress monitoring and may therefore have been better prepared for the posttests. Overall results were that students from the intervention group scored better on all five literacy measures than matched control students (effect sizes ranging from d=+0.388 to d=+0.911). Positive effects were found for all three achievement levels. On two measures, significant effects are found for all three ability groups: on the reading tasks measuring upper case letter naming and on the reading/reading comprehension task which measured both knowledge of book and print concepts and story comprehension. Average ability students gained most on both measures (respectively d=+1.276 and d=+0.999), followed by low ability students (d=+1.015 and d=+0.876) followed by high ability students (d=+0.675 and d=+0.696).

The other studies described in this section (for key features and estimated effects, see appendix 1) are all based on ECLS-K data, the Early Childhood Longitudinal Study starting in Kindergarten. A drawback is that this database lacks detailed information on the grouping practices of the teachers. Teacher's self-reported use of ability grouping and time spend on language/reading or mathematics is measured with Likert scales. No information is available on the flexibility of groups, the basis on which groups are formed and the way learning content is (differentially) conveyed. This makes interpretation of the results more complex. Nevertheless, the size and the representativeness of the ECLS-K dataset make the studies important for collecting empirical evidence on the effects of grouping for young children.

Hong and colleagues performed two related studies on the relationship between homogeneous grouping, instruction time and reading growth (Hong & Hong, 2009; Hong et al., 2012). They created six categories of educational practice based on instruction time (high or low) and homogeneous grouping (high intensive, low intensive or none). Teachers who reported to spend more than 1 hour a day on literacy instruction were classified as providing 'high' amounts of instruction time. Teachers who reported to spend more than 40% of the

literacy time on instruction to homogeneous groups were classified as using 'high intensive' grouping. "No grouping" means that only whole class instruction was provided. Hong and colleagues used these categorizations in both studies, but in 2009 (10,189 students, 1,858 classrooms, 740 schools) they presented among others the main effects and focused on general reading growth and in 2012 (8,668 students, 1,697 classrooms, 665 schools), they presented differential effects and focused on effects for low, average and high performing students. Results from the 2009 study were that when teachers provide 1 hour or more literacy instruction a day, it is beneficial to use homogeneous grouping compared to whole class instruction. This counted both for high intensity grouping, when students spent 40% or more of the time spend on literacy instruction in homogeneous groups (d=+0.198), and for low intensity grouping, when students spent less time in homogeneous groups (d=+0.164). When teachers provided less than 1 hour a day of literacy instruction, no significant effects of grouping over whole class instruction were found. In this context of low instruction time, high intensity grouping seemed to be less beneficial than whole class instruction, but effects were not significant. In spite of these non-significant results, the authors concluded that the combination of low instruction time with high intensity grouping appeared to have an adverse effect.

Hong and colleagues (2012) therefore studied whether this negative effect of low instruction time in combination with high intensity grouping holds for groups of students of different ability levels. First the effect of grouping was studied for different groups *given that instruction time is low*. Differential effects only reached significance for the low ability group. For these students, whole group instruction was more beneficial than intensive grouping, when instruction time was low (effect sizes for the 5 different literacy measures ranged from d=+0.181 to d=+0.328). The authors also studied whether the effect of intensive grouping was influenced by the amount of time spent on instruction. For all ability groups, intensive grouping was more beneficial when high instruction time was provided than when low instruction time was provided. For high ability students significant effects of high instruction time were found for two of the reading measures (effect sizes d=+0.267 and d=+0.284). For average ability students positive effects were found on all four reading measures (effect sizes range from d=+0.145 to d=+0.174), but not on the measure of reading comprehension. For low ability students positive effects were found on three of the reading measures and the reading comprehension measure (effect sizes range from d=+0.208 to d=+0.268).

The study of Chang (2008) is the only one in the collection of selected papers that focusses on early mathematical development. The longitudinal study of ECLS-K data focuses among others on the effects of grouping on the performance of different groups of minority students, learning English as a second language. Since the current review does not focus on second language learners, only the data of the Caucasian group and the African-American group with English as (only) mother tongue is used here[7] (respectively 5,863 and 1,151

[7] The groups of English only speaking students from the Hispanic and Asian group were small and therefore not used here.

students). Chang studied the relation between the frequency of 4 types of classroom practices and mathematics achievement. The four types of classroom practice were: teacher-directed whole class activity; teacher-directed small-group activity[8]; teacher-directed individual activity; and student-selected individual activity. Teachers indicated the frequency in which they used every type of classroom practice on a 5-point scale, ranging from no use to more than 3 hours a day. Results were that more teacher-directed whole class instruction was significantly related to more math improvement for Caucasian and African-American English-only speakers (*d*=+0.152 and *d*=+0.134 respectively). The other effects were smaller, inconsistent, or not significant: more time spent in teacher-directed small group settings had a negative or no significant effect on math improvement (*d=-0.*045 and *d*=+0.002, 95% CI contains 0); more teacher-directed individual activity had a small positive or negative effect (*d*=+0.008 and *d*=-0.069); more child-selected individual activity had a small positive effect (*d*=+0.012 and *d*=+0.020). In theory, high time can be spent on multiple practices and it is not a case of *either* one classroom practice *or* the other. For example, a combination of intensive whole class instruction and intensive child-initiated individual activity may be effective, but this is not tested here.

McCoach and colleagues (2006) studied, among others, the effects of homogeneous grouping on reading growth based on ECLS-K data. They based their analyses on the data of 10,191 students of 620 schools. The amount of time spent on ability grouping was measured on a 5 point scale, as reported by the teacher. This measure is a rough indication of frequency of grouping: from never to daily. Results were that higher frequencies of ability grouping were related to more reading growth (*d*=+0.127).

Adelson and Carpenter (2011) studied ECLS-K data of over 9,000 students, from almost 1,700 classrooms and 580 schools. They compared, among others, the effect of whole class education with homogeneous grouping on reading growth from fall to spring in Kindergarten K2. The use of ability grouping for reading was measured with a dichotomous question to the teacher (yes/no). Results were that classrooms in which homogeneous grouping took place, students showed more reading growth (d=+0.068). Unfortunately, there was no additional information on the grouping practice, for example on frequency of grouping or time spent in the groups.

Tach and Farkas (2006) used ECLS-K data as well to study the effects of homogeneous grouping. They analyzed among others whether students in Kindergarten classrooms using ability grouping had better reading achievement at the end of the school year[9]. They included almost 12,000 students from over 2,400 classrooms in their analyses and found the use of ability groups in Kindergarten had a positive effect on reading achievement (*d*=+0.346).

[8] Though not explicitly mentioned, this seems to refer to small homogeneous ability groups.
[9] Tach and Farkas also studied the effects of grouping at the end of first grade. These results are described in the section on primary education.

4.2.4. An example of an effective comprehensive program: EMERGE

Because of the importance of implementing high quality, adaptive instruction in order to make differentiation practices like ability grouping effective, an example will be given of a comprehensive program aimed at development in early childhood education which has a clear component of differentiation based on cognitive ability. The EMERGE program, as studied by Gettinger and Stoiber (2012) and which is included in the literature synthesis in paragraph 4.2.3, will be described.

EMERGE is based on the Response-to-Intervention (RTI) approach, which includes screening students, providing differentiated instruction, continuous monitoring and adapting instruction based on the monitoring results. It is in other words a form of differentiation based on actual performance in which ability grouping is used for (part of the) instruction and in which instruction is adapted to the needs of the students. Chambers and colleagues (2010) describe EMERGE in a best evidence synthesis on ECE programs and conclude there is limited evidence of the effectiveness of the program, due to insufficient numbers in the study. However Chambers and colleagues based their conclusion on an older study (Gettinger & Stoiber, 2007) and did not consider their paper from 2012. Due to the strong emphasis on implementation and the connection between grouping and instruction, the program is described here nevertheless.

Gettinger and Stoiber (2012) acknowledge that systematic progress monitoring alone is not sufficient to improve student performance. Teachers should know how to use this monitoring data to adapt their instruction. Therefore, professional development and coaching is part of the intervention. A problem with frequent (monthly) progress monitoring is that it is difficult to find measures sensitive to short-term growth in literacy development in preschool and Kindergarten. The authors therefore aim at developing assessments that are directly linked to the instruction received. Accompanying advantage is that this helps teachers to adapt their instruction to the needs of students, because it is directly clear which elements of the learning content are not well understood. Trained examiners conducted the monthly assessment battery for progress monitoring. The assessments were planned after each thematic unit and measured letter recognition, vocabulary (explicitly taught in the previous thematic unit) and book recognition and book comprehension (of books read in the previous thematic unit). The assessments were administered to all the children in the classroom individually in 10 minutes per child and took place during learning center time. The assessment data was used in instruction, which was divided into two phases: first core literacy instruction and then small group differentiated instruction, based on the progress data.

The core literacy instruction consisted of three elements. The first element is shared book reading, with dialogic reading and a special focus on print. Teachers received detailed cues in order to enhance the quality of the shared book reading and a literacy coach modeled one whole-group reading session a week. Twelve books were used per monthly thematic unit. The second element is explicit vocabulary instruction. Each monthly unit, sixteen words, extracted from the books read in classroom, were discussed. Vocabulary was instructed by

explaining word meanings, as well as providing contexts in which the word is used and stimulating students to provide their own examples. The third element is explicit focus on letters and sounds during book reading and small group instruction. Letters and sounds are not treated in isolation, but embedded in other engaging activities. The literacy coach provides demonstration and feedback on all instructional activities within the core instruction.

In addition to the core instruction, daily 30 minutes small group instruction was provided. Three ability groups were created based on the progress monitoring data. Groups consisted of 4 to 6 children who needed additional instruction and practice though repeated shared book reading and accompanying focus on vocabulary and letter and sound knowledge. High ability students were engaged in additional, more challenging discussions and tasks. A special 5-step plan provided teachers guidance in translating progress data (which they received from the researchers) into differentiated lesson plans. All in all, EMERGE combines ability grouping with frequent progress monitoring and intensive coaching of teachers in how to translate assessment data into differentiated lesson plans and how to provide high quality instruction.

4.3. Effects of differentiation in Primary Education (6-12 years)

4.3.1. Overview of differentiation in Primary Education

In primary education, differentiation is a topic of great concern to teachers. They have to deal with groups of students with a large variation in abilities, which may amount to students within the same class differing four years in didactical age. The desire to fit their instruction to the needs of individual students has led to some widely adopted grouping practices in primary education. One of the most common practices is within-class ability grouping (Kulik & Kulik, 1984; Slavin, 1987a). In this case, teachers form homogeneous groups within the classroom based on students' prior performance and provide instruction in these small homogeneous groups. For instance, in reading instruction, a survey in the United States shows that about two third of the teachers in the first grade of primary education use some type of within-class ability grouping (Chorzempa & Graham, 2006). The within-class ability grouping procedures are typically organized by teachers. Additionally, some articles have addressed using ICT as a tool to facilitate teachers in their within-class ability grouping procedures. ICT programs can be used as a tool to allocate students to groups based on their prior performance and can also be used to facilitate the choice of suitable learning materials for different students.

Another practice used in primary education is setting students in separate homogeneous classes based on their abilities for specific subjects such as reading or mathematics. Setting or regrouping is used frequently in some countries such as the United Kingdom and Australia. This is mostly true in the upper primary school grades. For instance, almost 40 percent of grade 5 and 6 teachers in the United Kingdom use setting for mathematics instruction (Hallam, Ireson, Lister, Chaudhury, & Davies, 2003). The expected

benefit of setting is that teachers can fit whole-group instruction to the needs of the group more easily when the group is quite homogeneous.

4.3.2. Selected studies

Approximately 200 references to studies focusing on differentiation in primary education were found using the database search. Of these, approximately 90 were selected for further inspection based on their full text versions. After applying the 8 final inclusion criteria (see paragraph 3.2), 16 papers remained and were included in the current review.

These 16 articles were divided into four categories: one article described an intervention study on within-class ability grouping; five articles describe natural occurring ability grouping practices; in five articles the effects of computerized testing systems with clues about differentiated instruction for the teacher were described, and in five studies differentiation was part of a broader program. Some of the articles are based on ECLS-K data (see 4.2.2.). In the closing paragraph of this section an exemplary comprehensive program that includes differentiation practices next to all sorts of other educational interventions will be described.

4.3.3. Literature synthesis

Results of an intervention study on within-class ability grouping

Of the included studies in primary education, one study was on an intervention using different types of within-class ability grouping (see appendix 2a). This study of Leonard (2001) comprises two consecutive years. In each year performance on the Maryland Functional Mathematics Test is monitored in a grade six cohort from three classrooms. In the first year of the study, all grade six students in cohort 1 were seated in small heterogeneous groups during mathematics lessons. In the consecutive year, all grade six students in cohort 2 were seated in small homogeneous ability groups. The grouping intervention was executed by clustering students' tables in small groups of three to four based on students' pretest performance and grades. During the year, students of both cohorts collaborated on thematic mathematical activities. The article does not clarify how instruction by the teacher was provided. The effects of homogeneous table grouping compared to heterogeneous table grouping were negative and non-significant (d=-0.250). The intervention does not support the hypothesis that homogeneous grouping has a different effect on students' performance than heterogeneous grouping. Based on qualitative analyses of students group interactions, the author concluded that the way the group collaborated may have been more determinative for achievement than the clustering of students in table groups based on ability level.

Results of studies on naturally occurring ability grouping practices

The second category of studies does not describe intervention programs, but rather analyzes the effects of naturally occurring differentiation practices in education. In these studies, teacher questionnaires or administrative information was used to assess ongoing differentiation practices in classes or schools. In turn, this information was related to student

performance measures for reading and literacy, writing, or math using quantitative analytical procedures. In the studies on the effects of naturally occurring differentiation practices, two types of differentiation were found. The first is within-class ability grouping. The effects of this type of differentiation were assessed in three studies (Condron, 2008; Nomi, 2010; Tach & Farkas, 2006). Another type of differentiation found in the literature on primary education was between-class homogeneous ability grouping (or setting). This type of differentiation was addressed in two studies (Macqueen, 2012; Whitburn, 2001). The key features and findings of these studies are summarized in appendix 2b.

Within-class ability grouping

The articles on the effectiveness of naturally occurring within-class ability grouping are all based on longitudinal data from the ECLS-K cohort, which already was described in paragraph 4.2.2. In the ECLS-K dataset teachers provided information about their grouping procedures. Student performance data is gathered in kindergarten and at the end of first grade. One study also adds third grade performance data to assess the effect of grouping from first to third grade (Condron, 2008). The selected articles in primary education using the ECLS-K data assess the effect of within-class ability grouping on students' reading performance.

In the article of Condron (2008), effects are presented of placing students in reading groups based on their reading performance from kindergarten to first grade and from first to third grade. Using the propensity score matching technique, the author compared the scores of students in a low, middle or high level reading group to scores of non-grouped students with a similar likelihood of being placed in one of these groups. For both first and third grade, placement in a high ability group led to higher gains in reading performance (first grade: d=+0.207; third grade: d=+0.177). Placement in an average level reading group did not have a significant effect on reading performance (first grade: d=-0.043; third grade: d=+0.046), and placement in a low-level group had a significant negative effect on reading performance in both first and third grade (first grade: d=-0.288; third grade: d=-0.245). This shows that within-class ability grouping may lead to divergent differentiation effects.

The articles of Nomi (2010) and Tach and Farkas (2006) both analyze the effect of grouping practices in first grade on first grade spring reading performance. These studies show that within-class ability grouping is frequently used in primary education; in the ECLS-K dataset ability grouping occurs in about 70 percent of the first grade classrooms. Tach and Farkas (2006) used multilevel modeling to estimate effects of grouping on reading performance students in first grade. In this study, the occurrence of ability groups in first grade had a significant negative effect on students' reading performance (d=-0.191). However, additional results show that being in a high ability group positively affected performance. This effect is more profound for African-American or Hispanic students, suggesting that student race interacts with grouping effects. Nomi (2010) used propensity score matching to examine the effects of ability grouping on reading achievement. The reading scores of 8785 students in total were used to analyze the effects of school grouping policies. The author found that on

average, schools using ability grouping served a relatively heterogeneous student population. This confirms the notion that ability grouping is often used as a tool for schools to deal with student diversity. However, in the study of Nomi, no evidence was found of benefits of ability grouping over whole class instruction (*d*=-0.010).

Summarizing the effects found in the ECLS-K studies (Nomi, 2010; Tach & Farkas, 2006), the meta-analyses presented in Table 1 show that overall within-class ability grouping had a small negative effect on students' reading performance (*d*=-0.070). Meta-analyses of the effect of within-class ability grouping for students of differential ability (Condron, 2008; Nomi, 2010) show that in the ECLS-K dataset within-class ability grouping had a small negative effect on the reading performance of low ability students (*d*=-0.232), no effect for students of average ability (*d*=0.000) and a small positive effect on the reading performance of high ability students (*d*=+0.155). Notice furthermore that the confidence intervals for the effect sizes *d* for the three ability types of students do not overlap, indicating significant differential effects in favor of the more able students. Stated otherwise: the results support a divergent pattern. However, caution should be exercised with generalization, since all findings were based on the same dataset.

Table 1: *Meta-analyses: naturally occurring ability grouping practices within classes in primary education; general and differential effects*

Included papers	*School subject*	*Grade*	*Effect sizes (d)*	*95% confidence interval*
Nomi, 2010; Tach & Farkas, 2006	Reading and literacy	K-1	-0.070*	-0.110; -0.029**
Condron, 2008; Nomi, 2010	Reading and literacy	K-3	*Low ability*	
			-0.232*	-0.270; -0.195**
			Average ability	
			0.000	-0.032; +0.031**
			High ability	
			+0.155*	+0.124; +0.186**

* 95% confidence interval of effect size does not contain 0

** The standard errors are multiplied with a factor √2 to account for the fact that the same data is used

Between-class setting

A second type of differentiation we found in studies on naturally occurring practices in primary education is setting students in between-class ability classes for specific topics such as reading or mathematics. Two selected articles discuss the effects of between-class ability grouping on student performance (Macqueen, 2012; Whitburn, 2001). In the article of Macqueen (2012) the gains in performance of students grouped in between-class ability groups were compared to the performance gains of students in heterogeneous classes. Students in both conditions were grouped in heterogeneous home classrooms for most school subjects. However, students in the between-class setting group were allocated to smaller, homogeneous classes for specific school subjects based on their performance on mathematics and literacy. Students in the non-grouping condition remained in their heterogeneous home classrooms throughout the school year. The performance gains between grade three and five for

mathematics, literacy and writing of students in regrouped classes were compared to the gain scores of students in heterogeneous classes. In general, small and non-significant effects of regrouping students based on their literacy abilities on student performance in literacy and writing were found compared to learning in mixed ability classrooms (literacy: *d*=+0.196; writing: *d*=-0.082). Regrouping students based on their mathematical abilities had a small negative and non-significant effect on students' mathematics performance (math: *d*=-0.125). Analysis of differential effects for high, middle and low groups based on mathematical ability or literacy ability also did not show any significant differences between the two conditions (see appendix 2b).

Whitburn (2001) compared mathematics performance between students grouped in homogeneous classes based on their prior mathematics achievement to the performance of students taught in mixed ability classes. Both groups of students were taught using the same interactive, whole class teaching method, which was part of a larger intervention study. Within this intervention, teachers initiated the two different grouping procedures. Mathematical performance in this project was regularly monitored using short written tests about previously taught mathematical topics. These tests were used to analyze grouping effects on student performance in grades three and four. In the article, results are presented of three consecutive cohorts of students. In these three cohorts, approximately 200 students were taught in ability grouped classes and about 1000 students were taught in mixed ability classes. The first cohort had been grouped for 21 months, the second cohort had been grouped for 15 months and the third cohort had been grouped for about 3 months. The analyses in the first cohort show small and non-significant effects of between-class ability grouped students' performance compared to the performance of students in heterogeneous groups (cohort 1 grade 3: *d*=-0.030; cohort 1 grade 4: *d*=-0.270). This finding is replicated in the second cohort (cohort 2 grade 3: *d*=-0.030; cohort 2 grade 4: *d*=-0.130). In the third cohort a significant small negative effect of grouping students in homogeneous classes over heterogeneous classes was found (cohort 3 grade 3: *d*=-0.110; cohort 3 grade 4: *d*=-0.290).

Meta-analyzing the effects of between-class grouping (see Table 2) shows that in the studies of Macqueen (2012) and Whitburn (2001), between-class setting based on mathematical ability has a significant negative effect on students' mathematics performance (*d*=-0.142*). The effect of setting is negative and significant for both low ability students (*d*=-0.224*), average ability students (*d*=-0.437*), and high ability students (*d*=-0.162*). Furthermore, the confidence intervals for the effect sizes *d* for the various ability groups do show quite some overlap, which indicates the absence of differential effects, be it divergent or convergent.

Table 2: *Meta-analyses: naturally occurring ability grouping practices between classes in primary education; general and differential effects (compared to heterogeneous classes)*

Included papers	*School subject*	*Grade*	*Effect sizes (d)*	*95% confidence interval*
Macqueen, 2012;	Mathematics	3-6	*Overall*	
Whitburn, 2001			-0.142*	-0.245; -0.038
			Low ability	
			-0.224*	-0.382; -0.065
			Average ability	
			-0.437*	-0.593; -0.282
			High ability	
			-0.162*	-0.314; -0.009

* 95% confidence interval of effect size does not contain 0

Summarizing, in the studies on naturally occurring differentiation practices, we found that the effects of within-class grouping vary depending on students' ability. Small positive effects of grouping for high ability students were found, but overall within-class ability grouping had a negative effect on early elementary students' reading performance. For setting, or regrouping, a meta-analysis of two studies shows a negative effect on students' mathematics performance. However, there are some concerns about the generalizability of these findings. One methodological concern for the within-class grouping analyses is that they were all based on the same dataset. And for the analyses of the effects of setting only two studies met the inclusion-criteria. Moreover, a major drawback of the articles about naturally occurring practices is that they often do not give insight in the instruction teachers provide. Thus, it is unclear whether and how instruction within these ability groups was tailored to the needs of students.

Results of studies on differentiation based on computerized systems

The third category of studies concerns differentiation guided by computer systems. In most educational settings, ability grouping practices are based on teacher-directed allocation of students based on students' prior performance. However, recent developments show that computer technology can also be used as a tool to support differentiation in primary education. Computer algorithms may be used to give suggestions on homogeneous grouping procedures based on students' prior performance. They can also be used to determine which type of instruction is most suitable for students' needs based on analyses of their prior performance. Using computer technology to support differentiation in such a manner is described in the articles of Connor and colleagues (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Connor, Morrison et al., 2011a; Connor, Morrison et al., 2011b) and Ysseldyke and colleagues (Ysseldyke et al., 2003; Ysseldyke & Bolt, 2007). An overview of these studies can be found in appendix 2c.

Connor and colleagues published several articles on the effects of individualizing student instruction (ISI) using A2i software (Assessment-to-Instruction). The ISI intervention is designed to support teachers in their efforts to provide optimally effective reading instruction for all students. The computerized system advices the teacher about the amount of

teacher- or student-managed instruction suitable for the specific child based on students' prior performance. Additionally, the program provides teachers with suggestions about the content of the instruction regarding whether the reading instruction should be more code focused or meaning focused. Based on the suggestions made by the computer program, teachers can provide reading instruction to small homogeneous groups of students . In the review, three articles of Connor and colleagues were included which used a student-level cognitive output measure (Connor et al., 2007; 2011a; 2011b).

In the article of 2007, the authors report on the effectiveness of the ISI treatment on student language and literacy outcomes. The growth of first grade students from schools in which teachers used the ISI program to differentiate their reading instruction was compared to students' growth in reading performance in matched control schools. Teachers using the ISI intervention received the program and a professional development course in the use of differentiated reading instruction. Control group teachers did not receive any professional development nor did they use the computer program. Results show that the individualized instruction had a small but significant positive effect on students' reading achievement on a standardized test (d=+0.183). Although these results were presumably affected by teachers' professional development in the experimental group, the authors show that students' growth in the experimental group was related to the amount of time spent on the intervention in the classroom, suggesting that the intervention in itself was also related to students' reading outcomes.

Connor et al. (2011a) replicated the first grade results in their study. They analyzed the effectiveness of the ISI-intervention on students' word reading skills in comparison to a business as usual control group. Teachers in the experimental group used the suggestions by the computer program to form ability groups and to choose the content of their instruction based on students' needs. They were supported in the use of the ISI intervention by professional development instruction and coaching. In the control group, teachers spent an equal amount of time on small group reading instruction, but did not have access to the computer program. Classroom observations showed that teachers in the ISI-condition were better able to fit the content instruction to student-needs based on prior performance than teachers in the control condition. Matching the instruction to recommendations of the computerized algorithm strongly predicted students' reading outcomes. Multilevel analyses show that the ISI-intervention had a significant positive effect (d=+0.249) on students' word readings scores on a standardized test collected in spring of the school year. The authors argue that the effectiveness of the treatment had increased compared to the study in 2007 since they made the computer program more user-friendly and the professional development program for teachers was improved.

Another study on the effectiveness of the ISI-treatment reports treatment effects on student results in third grade (Connor et al., 2011b). In this study, effects on students' reading performance of the intervention were compared to an alternative intervention based on vocabulary instruction. In the ISI-treatment condition, teachers assessed students'

performance three times a year, used the computerized instructions to determine the focus and content of their instruction. Teachers also received a professional development training on implementation of the treatment. In the vocabulary treatment condition, teachers received a professional development training in which they read and discussed instruction principles from a vocabulary handbook and designed and evaluated their lessons collaboratively with a focus group of other teachers. Classroom observations during the school year showed that teachers in both conditions did not differ in the amount of individualized instruction, in their organization and planning activities, in the use of strategies and in classroom-management styles. However, teachers from the ISI group did match their instruction more closely to the content suggested by the computer algorithm. Multilevel analyses of student results show that the ISI-training had a small significant positive effect on students' reading comprehension *(d* =+0.191) and on vocabulary performance (*d*=+0.033) in comparison to the vocabulary intervention.

Ysseldyke and colleagues (Ysseldyke et al., 2003; Ysseldyke & Bolt, 2007) used a computer program to support differentiated mathematics instruction. The program they used is called Accelerated Math. In the article of 2003, the effectiveness of the program on student results in third, fourth, and fifth grade was assessed. Accelerated Math generates mathematics exercises for students of different levels of proficiency. After completing the exercises, students scan their work and the computer provides them with immediate feedback. Also, the program provides teachers with suggestions about content and grouping practices based on each student's individual performance. In this study, teachers from four schools volunteered to use the computer program during mathematics instruction. Of all classrooms, teachers in ten classrooms fully implemented the program. Scores of students from classrooms in which teachers used Accelerated Math were compared to students from other classrooms in these schools and a random group of students from the district's testing database. Within schools, significant small to medium positive effects were found of using the program on a standardized math test (*d*=+0.189) and on a computerized adaptive math test (*d* =+0.268). In the study published in 2007, Ysseldyke and Bolt investigated the effect of the same system in both primary and secondary schools. Classrooms were randomly assigned to within-school experimental and control groups. Again it turned out that when teachers implemented the continuous progress monitoring system as intended, their students gained significantly more than (Terra Nova test: *d*=+0.469; STAR Math test: *d*=+0.458).

A meta-analysis on the effect estimates from the studies on the computer-based differentiation interventions shows that both in math and in reading, computer algorithms fostering differentiation can positively affect student performance (see Table 3). The meta-analysis of the articles of Connor and colleagues (2007; 2011ab; 2011ba) shows a significant small positive effect of the computer intervention on students' reading performance (*d*=+0.204). A meta-analysis of the two articles of Ysseldyke (2003; 2007) shows a significant medium positive effect of the computerized differentiation intervention on students' math performance (*d*=+0.345). Although the number of articles included in this meta-analysis is

small, the cumulated effects show that a computer supported approach to differentiation in which both grouping and instructional content is addressed can be beneficial for students' performance in primary education.

Table 3: *Meta-analyses; differentiation based on computerized systems in primary education; effects for reading and mathematics*

Included papers	*School subject*	*Grade*	*Effect sizes (d)*	*95% confidence interval*
Connor et al., 2007; 2011a; 2011b	reading	1-3	+0.204*	+0.104; +0.303
Ysseldyke et al.2003; Ysseldyke & Bolt, 2007	mathematics	2-6	+0.345*	+0.232; +0.458

* 95% confidence interval of effect size does not contain 0

Results of studies on differentiation as part of a broader school reform program

The fourth category of articles describes differentiation in the context of a broader program. Implementing differentiation practices cannot be done in isolation, and moreover synergetic effects can be expected when differentiation is one of the many elements of a well-designed comprehensive program. This paragraph looks into studies on the effects of such programs, although one has to bear in mind that effects (or absence of effects) cannot – by definition – be solely attributed to the differentiation component of such a program. The key features and summary estimated effects for the various studies are presented in appendix 2d .

The most well known and most researched program is Success for All. Success for All aims at comprehensive school reform to ensure that all children can read. For reading instruction pupils are regrouped across grades according to specific performance levels (i.e. setting). Every nine weeks pupils are assessed and regrouped when necessary. Pupils that need additional help receive one-to-one-tutoring to get them back on track so as to achieve convergent differentiation. The article of Borman, Slavin, Cheung, Chamberlain, Madden and Chambers (2007) reports final literacy outcomes for a 3-year longitudinal sample of pupils from 35 schools who participated in an effect study of Success for All (cluster randomized controlled design) from kindergarten to second grade. The significant effects of the treatment were as large as one third of a standard deviation on all three outcome measures (Word Identification: d=+0.220, Word Attack: d=+0.330, Passage Comprehension: d=+0.210).

The second article that matches the criteria for inclusion is an article of Stevens and Slavin (1995) in which achievement (among other measures) of grade two to six students of two cooperative elementary schools were compared to the achievement of comparable students in three control schools. Being a cooperative school implied several elements: using cooperative learning across a variety of content areas, full-scale mainstreaming of academically handicapped students, teachers using peer coaching, teachers planning cooperatively, and parent involvement in school. For the present study, teachers were trained to work with two comprehensive programs designed to accommodate student diversity: CIRC (Cooperative Integrated Reading and Composition) and TAI (Team Assisted

Individualization-Mathematics). In both programs students worked in heterogeneous learning teams but received instruction in relatively homogeneous teaching groups, elements that are also included in Success for All. Students' achievement was tested in reading, language and mathematics after one and after two years. During the first years the two schools were implementing the program and students' achievement only differed – in favor of the cooperative schools – on reading vocabulary (*d*=+0.170). After two years, students of the cooperative schools also performed better at reading comprehension (*d*=+0.280), language expression (*d*=+0.210), and math computation (*d*=+0.290). In language mechanics and math application treatment and control schools do not differ.

Because the programs (cooperative school, CIRC and TAI) had so many components it is difficult to ascribe the outcomes to any single element. However, according to the authors, the results of the study support the hypothesis that cooperative learning can be effective in producing higher student achievement. In terms of differentiation, this finding supports the effectiveness of working in heterogeneous learning teams - which involves group goals based on group members' individual learning performance - and homogeneous teaching groups.

Reis, McCoach, Coyne, Schreiber, Eckert and Gubbins (2007) combined their School-wide Enrichment Model in Reading Framework (SEM-R) with Success for All. This article discusses an experiment executed in two primary schools serving a primarily culturally diverse, high poverty group of students. The schools participating in the study were required to give reading instruction each afternoon in addition to the Success for All program which they used in the morning. In the experiment, effectiveness of two types of reading instruction is evaluated by randomly assigning teachers and students to two conditions. Teachers were frequently coached and observed during the experiment. Students in the control condition received twelve weeks of literacy instruction based on whole group instruction with workbook-materials and test-preparation assignments. Students in the experimental condition used the School-wide Enrichment Model in Reading Framework (SEM-R) for twelve weeks. In SEM-R teachers first read aloud and use higher order questioning and thinking-skills instruction. Then, students were encouraged to select books suitable for their ability level. During this phase, teachers gave individualized support and differentiated instruction about reading strategies. In the third phase, students chose between different literacy-related activities with varying complexity. Posttest results showed a positive effect of SEM-R on students reading fluency (*d*=+0.299), but no significant effects on students reading comprehension (*d*=+0.220).

After this experiment Reis, McCoach, Little, Muller and Kaniskan (2011) implemented SEM-R without Success for All in five primary schools serving a primarily culturally diverse, high poverty group of students. This article discusses a cluster-randomized experiment in which teachers were randomly assigned to a control or treatment condition. In both conditions teachers had a two-hour block of reading and arts instruction every day for five months. In the control condition, the full two hours were devoted to the regular reading and language arts program. This program was mostly teacher led and consisted of silent

reading activities, test preparation activities, workbook exercises and some small group or individual instruction (21% of the time). The teachers assigned to the experimental condition used the same program for the first hour and used SEM-R during the second hour. Matching students on individual performance, teachers provided students with feedback and individual instructions. In the third phase, students chose between different literacy-related activities with varying complexity. Posttest results for reading fluency and reading comprehension were mixed. Students in the control and the experimental group both increased their performance after the intervention. In two schools students receiving SEM-R outperformed control students, but in the other three schools no apparent differences were found. The authors suggest that the SEM-R approach may be especially suitable for (sub)urban schools. Nevertheless, the overall effects were non-significant (Fluency: d=+0.254, Comprehension: d=+0.145).

In the Netherlands, Houtveen and van de Grift (2012) reported on the effects of the Reading Acceleration Programme (RAP). The program aims at reducing the percentage of struggling readers in the first year of formal schooling. A quasi-experimental study was carried out. The teachers in the experimental group had been trained to improve their core instruction (Tier 1), to broaden their instruction for struggling readers (Tier 2) and to implement special measures for pupils who did not respond sufficiently to the interventions (Tier 3). The aim of Tier 2 and 3 is to make it possible for the students to attend the whole group instruction successfully (convergent differentiation). After correcting for pre-test, age, intelligence, socioeconomic status and ethnic minority a significant difference on reading was found in favor of the pupils in the experimental group (Word Decoding: d=+0.280, Fluency: d=+0.620).

A meta-analysis on the effects presented in the articles of Stevens and Slavin, Borman et al. and Reis shows a small significant positive effect of the programs on reading comprehension (d=+0.231); see Table 4. The meta-analysis of the effects from the studies of Borman et al., Houtveen and van de Grift and Reis shows a significant medium positive effect of the programs on basic reading (d=+0.375). Mathematics and language were only covered by the study of Stevens and Slavin. These effects are non-significant or very small.

The main drawback of these programs in terms of this best-evidence review on differentiation is the fact that it is unclear which part of the program causes the effect. Probably all aspects ‘work’ together, which leads to higher achievement of students.

Table 4: *Meta-analyses: differentiation as part of comprehensive programs in primary education; effects on basic reading and reading comprehension*

Included papers	*School subject*	*Grade*	*Effect sizes (d)*	*95% confidence interval*
Borman et al., 2007; Reis et al, 2007; Reis et al., 2011; Stevens & Slavin, 1995	reading comprehension	Grades 2 to 6	+0.231*	+0.128; +0.333
Borman et al., 2007; Houtveen et al., 2012; Reis et al, 2007; Reis et al., 2011	basic reading	Grades 2 to 6	+0.375*	+0.279; +0.471

* 95% confidence interval of effect size does not contain 0

4.3.4. An example of an effective comprehensive program: Success for All

SfA - its effects were presented in the previous paragraph - is a school wide program for students in grades pre-K to 6 which organizes resources to ensure that virtually every student will reach the third grade on time with adequate basic skills and build on this basis throughout the elementary grades. The main element is the reading program. In grades K-1 (in Kindergarten: Stepping Stones and KinderRoots incorporated in KinderCorner, in grade 1: Reading Roots containing FastTrack Phonics, Shared Stories, Story Telling and Retelling (STAR) and Language Links) it emphasizes language and comprehension skills, phonics, sound blending and use of shared stories that students read to one another in pairs. The stories combine teacher-read material with phonetically regular student material to teach decoding and comprehension in the context of meaningful, engaging stories. In grades two to six (Reading Wings, an adaptation of Cooperative Integrated Reading and Composition - CIRC) students use "real" novels and books but not workbooks. The program emphasizes cooperative learning and partner reading activities, comprehension strategies such as summarization and clarification built around narrative and expository texts, writing and direct instruction in reading comprehension skills.

During daily 90-minute reading periods, students from all heterogeneous 'home room' classes (grade 1 to 6) are regrouped across age lines so that each reading class contains students all at one reading level. Use of tutors as reading teachers during reading time reduces the size of most reading classes to about twenty students. Students in first to sixth grade are assessed every trimester to determine whether they are making adequate progress in reading. This information is used to suggest alternate teaching strategies in the regular classroom, changes in reading group placement and provision of tutoring services. Specially trained teachers and paraprofessionals offer tutorial services in grade one to three to students who are failing to keep up with their classmates in reading. Tutorial instruction is closely coordinated with regular classroom instruction. It takes place in one-to-one settings, twenty minutes daily during times other than reading periods.

The instruction process is based on research-proven practices combined in the model of instructional effectiveness called QAIT, quality, adaptation (to the level and pace of each student), incentive (strategies to increase students' motivation to learn) and time. Cooperative

learning is a central feature in SfA: groups can earn recognition only if all team members have learned, so they encourage and help each other to master academic content.

SfA further consists of comprehensive, theme-based preschool (Curiosity Corner) and Kindergarten (KinderCorner) programs, a professional development program, a school facilitator and a solutions team in each school to plan school wide strategies for parental and community involvement, attendance and school climate.

4.4. Effects of differentiation in Early Secondary Education (12-14 years)

4.4.1. Overview of differentiation in Early Secondary Education

While primary education is generally a heterogeneous environment, secondary education tends to be more homogeneous, due to external differentiation or tracking. Students in secondary education are generally assigned to educational tracks or grouped for specific subjects, mostly language and math (setting). Tracking and setting are based on student's cognitive abilities, leading to homogeneous classes or courses. In the first one or two years of secondary education, a mitigated form of external differentiation may be used, with students with adjacent educational levels grouped together. Students are provided with differentiated assignments and tests, with additional work or test items for the more able students. This way, the most appropriate level for every student should emerge during the early secondary school years. After the first basic years of secondary education, students choose vocational tracks or curricular profiles based on their own interests.

Grouping in secondary education leads to divergent differentiation in the student population as a whole, although within classrooms or curricular subjects convergent differentiation is pursued. Within tracked classrooms, although the groups are homogeneous based on general levels of ability, large individual differences between students may still exist, which requires within-class differentiation. However, differentiation is not an educational practice that teachers in secondary schools tend to apply, especially in the higher pre-academic tracks (Inspectorate of Education, 2013).

Countries differ in the way secondary education is organized: the degree to which external differentiation is implemented and the age at which students are tracked differs. This international variation in educational systems makes it difficult to study the effects of external differentiation. Most studies make use of cross-sectional international assessments of IEA-TIMSS or OECD-PISA, and thus are suffering from all sorts of methodological flaws that hinder causal conclusions to be made about the relation between differentiation and student achievement. The most obvious problem is that students are selected into tracks at an early age, so one never knows whether the student achievement differences between integrated and differentiated educational systems – say at the age of 15 - are the result of the system differences or differences already present at an earlier age – say the age of 12. Clever solutions have been tried to circumvent this problem, like naturally occurring experiments in Great-Britain and Sweden where integrated and differentiated systems co-existed for a while

(c.f. Luyten, 2008), Difference-in-Difference models in which many countries with and without early tracking were compared with respect to the within-country differences between secondary and primary school performance (Hanushek & Woessmann, 2006), or propensity score matching techniques in which students from the integrated Polish System were matched to similar students from the tracked system before the education reform (Jakubowski, Patrinos, Porta, & Wisniewski, 2010). The results are not very clear-cut, but at least seem to indicate that integrated systems in general do not perform worse than differentiated systems. And moreover, as was described in the theoretical framework, no effects of tracking or setting are found when the results of students of lower, average and higher ability are taken into account simultaneously. Below we will concentrate on reviewing systematically studies that where conducted within one country with a direct comparison of differently differentiated groups of students.

4.4.2. Selected studies

In the initial database search, approximately 100 papers focusing on early secondary education (12-16 years) were found. Of these, approximately 40 were selected for further inspection based on their full text versions. In order to maintain the focus on *early* secondary education and/or middle school, the general age criteria were sharpened and restricted to the first two years of secondary education (grades 7 and 8; approximately 12-14 years of age). Four of the obtained papers met these new age criteria and the 8 final inclusion criteria (paragraph 3.2). These selected papers are alphabetically listed and summarized in appendix 3.

4.4.3. Literature synthesis

General overview

The studies selected for this review all focused on differentiation practices for mathematics only. Two studies from the same authors (Burris et al., 2006; 2008) are on the effects of an accelerated math curriculum in heterogeneous classrooms. The study by Barrow c.s. (2009) focuses on computer assisted mathematics instruction according to general principles of mastery learning. And a study by Linchevski and Kutscher (1998) focuses on the question whether small heterogeneous groups have different effects on mathematics achievement that homogeneous groups. Key features and summary of estimated effects for each of the studies are presented in appendix 3. Due to large differences between the studies in terms of operationalization of differentiation and/or the criterion variables used, it was not possible to perform meta-analyses on the studied included.

Results of the included studies

Barrow, Markman and Rouse (2009) conducted a randomized controlled trial on the use of individualized computerized (pre-)algebra instruction. Within schools, grade 8 classrooms were randomly assigned to the experimental condition using computerized instruction, or to the control condition using traditional forms of instruction. Each computerized mathematics lesson consisted of a pretest, a review of prerequisite knowledge, the subject content, a review

and a comprehensive test. Students repeat the lesson until they reach sufficient mastery. The teacher receives progress reports and provides individualized instruction to students who need it. Use of the computerized instruction positively influenced algebra achievement of the students (*d*=+0.416).

Burris, Heubert, and Levin (2006) studied the effect of offering an accelerated math curriculum in heterogeneous classrooms in middle school on students' math achievement and completion of advanced courses. They studied whether more students would take and pass advanced math classes in high school when heterogeneous, advanced math classes were offered to all students in middle school and whether providing heterogeneous math classes to students of all ability levels would influence the performance of initial high achievers. The study focusses on cohorts of students before and after a curriculum change, in which accelerated mathematics was implemented in middle school. The accelerated mathematics included offering the regular 3-year math curriculum for grades 6, 7 and 8 of middle school in 2 years, creating time to offer a more advanced algebra course in 8th grade. Originally, only selected students took part in the accelerated program, but after a while schools were mandated to offer accelerated mathematics for all students, in heterogeneous classrooms. Additional math support was available for students struggling with the advanced curriculum. Results showed that opening up the curriculum for all students in heterogeneous classrooms led to more students successfully completing two of the three advanced mathematics courses that increase in difficulty (*d*=+1.450 and *d*=+1.511).

In a later study, Burris and colleagues (2008) again studied the effect of offering an accelerated math curriculum to all students, making use of the system change in a New York state school district. This time, instead of studying the relationship between detracking and completing mathematics courses, they looked at the relationship between detracking and receiving diplomas tied to state-wide or international standards. These diplomas are additional to local school diplomas and reflect rigorous achievement requirements. Results show that detracked students had a greater chance of receiving a state diploma than tracked students (*d*=+3.187)[10] No significant differences between detracked and tracked students were found for receiving the prestigious international baccalaureate diploma.

Linchevski and Kutscher (1998) studied the effect of teaching mathematics in heterogeneous groups. The schools participating in the study had heterogeneous classrooms in which students worked sometimes in whole class settings, small heterogeneous groups, small homogeneous groups and large homogeneous groups. Large (whole) group learning was mainly teacher driven, while small group learning was fostered by cooperative learning. After one school year, heterogeneous classrooms (with cooperative learning and instruction in homogeneous groups when needed) had a significant small positive effect on math performance compared to the performance that was expected when students would have been

[10] This somewhat unusual large effect size is calculated by transforming the LogOdds-Ratio of 5.78 into the effect size *d* applying the equality d=LogOddsRatio x ($\sqrt{3}$)/ π (Borenstein et al., 2009, p. 47).

homogeneously grouped throughout the year (*d*=+0.112). Retention effects for a small group of schools at the end of 8th grade were not significant.

4.4.4 An example of an effective comprehensive program: IMPROVE

Comprehensive programs of which differentiation is an integral part do exist, but solid proof that such programs are effective only exists in the domain of mathematics (Slavin, Lake, & Groff, 2009) and not for reading and/or science. The Best Evidence Encyclopedia[11] only mentions two, namely STAD and IMPROVE. The reason why these were not initially included in the meta-analysis were that no references were found in our search to STAD, due to the fact that the key element of this program is cooperative learning, rather than differentiation. The literature search did result in references to IMPROVE , but these were rejected as the quintessential element of the program is metacognitive instruction rather than differentiation. However, IMPROVE and STAD do contain differentiation as an element, albeit less pronounced than other elements. Therefore, therefore IMPROVE be described here as an example of a successful comprehensive program for early secondary education. IMPROVE (Mevarech & Kramarski, 1997) is developed as an alternative to streaming or setting, and was evaluated in Israeli schools. The acronym stands for: Introducing new (mathematical) concepts, Metacognitive questioning, Practicing, Reviewing and reducing difficulties, Obtaining mastery, Verification, and Enrichment. Important elements are that within the heterogeneous groups students question each other metacognitively (which implies cooperative learning based on peer interaction), continue learning for mastery up till 80% correct, and based on this criterion students either continue for enrichment or individualized corrective instruction. The evaluation studies are relevant because IMPROVE is compared to business as usual in ability tracked classrooms. All in all, students in the IMPROVE condition outperform the control students, but the results are somewhat mixed. In a first study the main effect of IMPROVE for algebra is *d*=+0.301, and there are some indications for treatment x aptitude interactions, meaning that IMPROVE is effective for low, middle, and high ability students, but especially for the latter two groups. A second study produced similar main effects, and also the treatment x aptitude interactions seemed to indicate that IMPROVE was somewhat more effective for middle and high ability students than that it was for low ability students. Stated somewhat conservative: IMPROVE is effective, but there are no indications that it leads to convergent differentiation. The authors indicate that "It is possible that lower achieving students need additional support in order to further enhance their achievement" (Mevarech & Kramarski, 1997, p. 385). Although the effects are positive, once again one has to bear in mind, that it is the synergetic effect of various elements (a.o. metacognitive strategies, cooperative learning, regular assessments, learning for mastery, corrective instruction) that is probably generating the effects and not differentiation as such.

[11] Retrieved from http://www.bestevidence.org/math/mhs/top.htm at November 14, 2014.

4.5. Reflection on the included studies

Having presented and discussed the many findings from the 26 studies, we have to consider the possibility that the results may suffer from selection problems. Although our literature search initially resulted in almost 2,500 references, our very strict substantive and rigorous methodological inclusion criteria ruled out the vast majority of these references. Valuable as many of these references may have been from a conceptual, theoretical, and/or practical point of view, or as a rich qualitative description of occurring differentiation practices, for this review we were solely interested in studies that could shed light on the association between differentiation practices and students' cognitive outcomes. This type of selection was thus intended. Another kind of selection, however, could not be controlled by us, and that is that valuable studies may not have found their way to scientific journals since the results were viewed as disappointing or not ground breaking enough. Such selection often starts with researchers who themselves may not find it worthwhile to put effort in trying to get non-significant effects published. And, in second instance, journal editors and reviewers may be biased towards accepting manuscripts that contain statistically significant effects. To gain insight in the prevalence of this second type of selection within our dataset we assume the following model underlying publication bias Studies that do not have much statistical power as a result of small samples, only get published if they produce large effects that counterbalance the large standard errors. Studies that produce smaller effects find their way only to journals if they have (considerably) more statistical power (resulting from a big sample with consequently small standard errors). If this model is true, then the distribution of reported effect sizes is strongly biased (normally positively biased, but that of course depends on the phenomenon of interest and the scaling of the variables) as a function of an increasing standard error. A visual inspection of the relation between effect size and confidence interval may help us to sort this out. For that purpose we selected one finding for mathematics and language respectively per study (in case there were multiple cohorts we treated each cohort as a separate study), discarding the studies of Burris that focused on other outcomes (taking an advanced course or getting a diploma). See Figure 2.

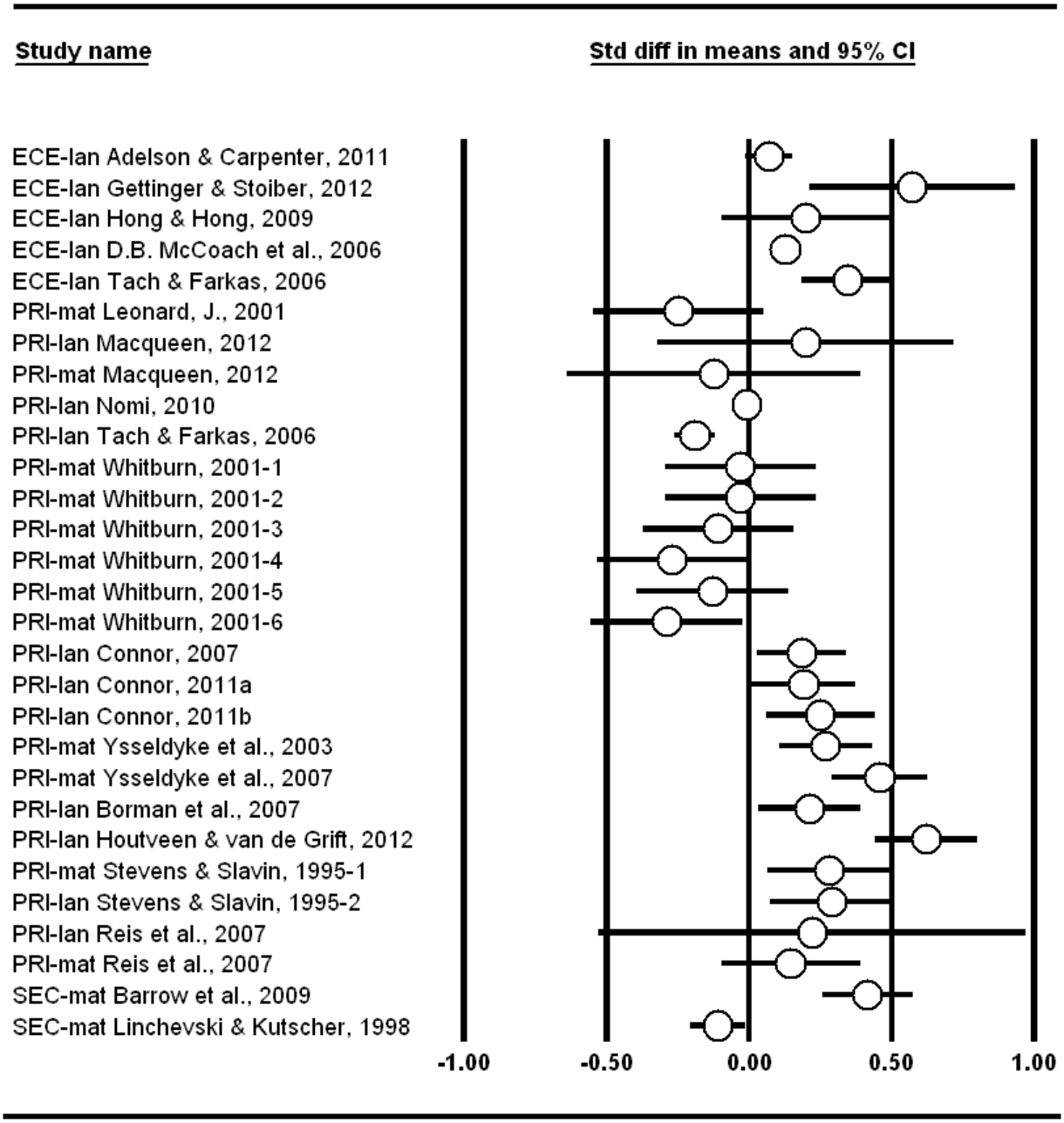


Figure 2: *Forest plot for the studies (one finding per subject per study selected) in the review*

There is a slight tendency that the studies with the smaller effect sizes also have the smaller confidence intervals, and at least for language in early childhood education, kindergarten and primary education the larger effect sizes are accompanied with wider confidence intervals. A second aid to detect potential publication bias may be of further help, and this is to be found in Figure 3.

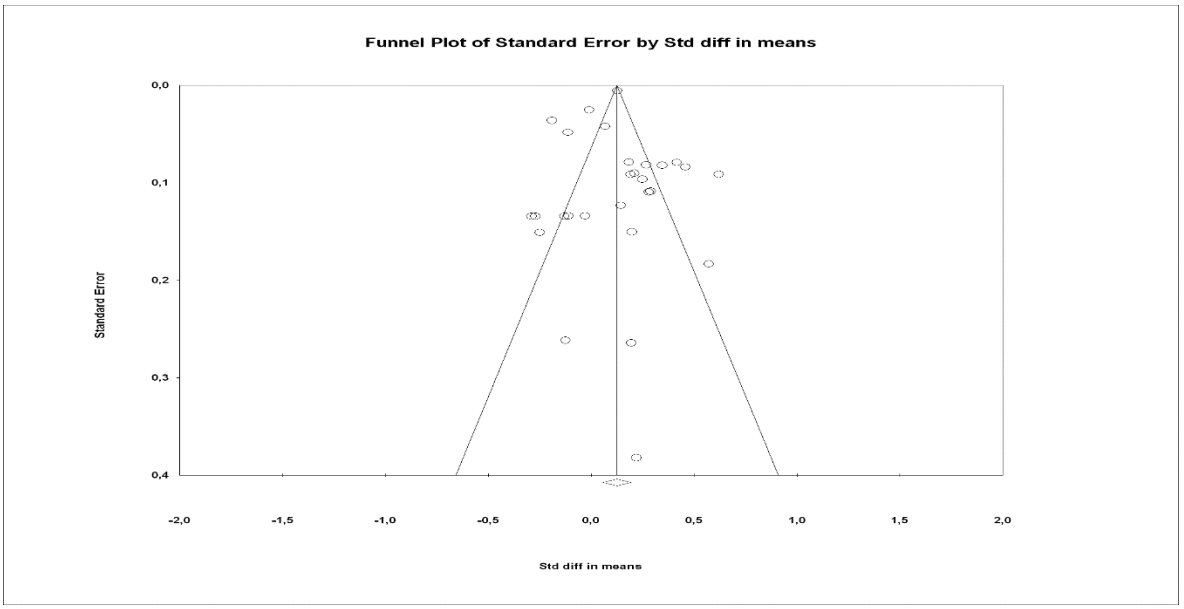


Figure 3: *Funnel plot to inspect publication bias from the studies reviewed*

The vertical line in the middle represents the average effect in a meta-analysis using a random effects model[12]. The picture shows that all the effect sizes are evenly distributed to the left and the right of the line. Assuming the correctness of our model of publication bias, our results thus do not seem to be overwhelmingly plagued with this phenomenon.

Finally, we can analyze whether differences in effect sizes found are related to the sector studied (Early Childhood, primary or secondary education), the type of differentiation (ability grouping (either within or across classes) or otherwise), whether it is computer supported or not and if differentiation was studied as being an element of a broader program. Table 5 contains the regression coefficients from a meta-regression model in which the effect sizes were regressed on these study characteristics. The meta-regression analyses were conducted using HLM software (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011).

Table 5: *Meta-regression results (standard errors in brackets) from regressing effect sizes on study characteristics*

	Regression coefficient	*95% confidence interval*	*t-ratio*	*p-value*
Intercept	+0.176 (0.054)	+0.070; +0.282	+3.354	.004
primary vs ECE	-0.293 (0.083)	-0.456; -0.130	-3.548	.002
secondary vs ECE	-0.227 (0.127)	-0.476; +0.022	-1.793	.086
ability grouping vs otherwise	-0.011 (0.089)	-0.185; +0.163	-0.122	.905
computer supported or not	+0.401 (0.088)	+0.229; +0.573	+4.566	<.001
part of broader program or not	+0.428 (0.085)	+0.261; +0.595	+5.024	<.001

[12] The difference between a fixed and random effect model is, that in the first we assume that in all the studies the true effect size is the same, whereas in the latter we do not.

The meta-regression results for the selected findings indicate that differentiation practices in primary are less effective than those in Early Childhood Education; that differentiation practices in secondary education are almost even effective as those in primary education; that using computer supported differentiation is more effective than other differentiation practices; and that broader programs of which differentiation is one of many key elements are the most effective. Ability grouping, either within or across classes, is not less effective than other differentiation practices given the other study characteristics[13]. The meta-regression results may be of help in finding some structure amidst all the associations reported.

[13] Not reported here are the results of an additional meta-regression analysis, in which also a dummy for subject domain (mathematics versus language) was included. This model produced similar results and there appear to be no differences between the two subject domains.

5. Conclusion and discussion

Students differ, and they may differ quite a lot even if they are in the same classroom. Didactical age differences between children in the same class may amount to 4 years, implying that, for instance in a grade 4 class of a primary school, some students perform at the average level of grade 2, whereas others have already advanced up till the average level of grade 6. Differentiation and adaptive instruction together are seen as a way to address these differences, but how these practices can be implemented well in the classroom is less clear. Differentiation is essential, but there are many forms. Grouping may be one, allowing time differences for mastering curricular subjects another. What are proven effective practices?

In this systematic review we summarized the results of studies into the effects of differentiation practices along three stages in the education system: early childhood education and kindergarten (2;6 to 6 years), primary education (6 to 12 years), and early secondary education (12-14 years). We also described exemplary effective comprehensive programs, in which differentiation was one of many elements, for each stage. From the almost 2,500 references related to differentiation found in the literature search around 1% met the inclusion criteria set for this review.

5.1. Early Childhood Education and Kindergarten

Early Childhood Education and Kindergarten was not part of the reviews on studies on differentiation up to 1995 (Kulik & Kulik, 1984; Lou et al., 1996; Slavin, 1987a; Slavin, 1987b). These reviews include studies with grade 1 as the youngest age group and therefore do not provide information on differentiation at earlier ages. Only the study of Lou and colleagues might be informative in this respect. They compared the effects of within-class homogeneous grouping between early and late elementary grades (respectively grades 1-3 and grades 4-6) and found that the effects in the earlier grades were much smaller (d=+0.08, 95% CI [+0.02;+0.14]) than the effects in the later grades (d=+0.29, 95% CI [+0.24;+0.35]). One may infer from this finding that homogeneous ability grouping is less effective at lower grades, and therefore as well in pre-K and K. On the other hand, since language and literacy development is a main goal of Early Childhood Education, especially for second language learners and children with limited language input at home, (convergent) differentiation practices are probably applied. In order to gather empirical evidence on this matter, in the current review studies on differentiation practices in Early Childhood Education and Kindergarten are taken into account.

The general result from the systematic review is that within-class homogeneous ability grouping has a moderate positive effect on the language performance of the classroom, with effect sizes for undifferentiated effects ranging from d=+0.068 to d=+0.911. The existence and direction of differential effects for differentiation on language growth are studied less and are inconclusive. It is therefore difficult to draw conclusions about the convergent or divergent

effect of differentiation practices in Early Childhood Education. Mathematical performance was only addressed in one study (Chang, 2008), in which spending relatively large amounts of time in small groups had no or negative effects. There are several factors that should be taken into account when interpreting these findings.

Important to note is that only seven studies on differentiation in ECE and Kindergarten met de inclusion criteria, of which six were based on data from the same longitudinal study, ECLS-K. This means only a fraction of the studies on teaching practices and child development in ECE and Kindergarten was selected for the current review and results may therefore be hard to generalize. Perhaps studies in this field generally do not explicitly focus on achievement in relation to grouping or other differentiation practices and/or do not describe these practices in terms of 'differentiation'. In order to get a better view on differentiation at these younger ages, in a future study, it may be worthwhile to look in more detail at the jargon used for describing differentiation practices in ECE and Kindergarten and to include studies using other, more descriptive, research methods as well.

The differentiation practice used in the selected studies is 'within-class homogeneous ability grouping'. Due to different combinations of variables from the ECLS-K database, these studies vary in their operationalization of 'homogeneous grouping', from broad dichotomous grouping/no grouping to combinations of intensity of grouping and intensity of instruction. The studies based on ECLS-K data do not specify how the ability groups are formed and on what information they are based. Furthermore, they do not specify the type and quality of the instruction and materials provided to these ability groups. The importance of this information is illustrated with the study of Hong and Hong (2009), who found that homogeneous ability grouping, of either high or low intensity, had positive effects on reading growth *only* if students receive at least one hour of reading instruction a day. When students received less instruction, grouping did not make a difference compared to whole class activities. This emphasizes that the effects of grouping as such are difficult to interpret as long as it is unknown what the teacher does with these groups. This is in line with the conclusion Lou and colleagues drew from their review: "It appears that the positive effects of within-class grouping are maximized when the physical placement of students into groups for learning is accompanied by modifications to teaching methods and instructional materials. Merely placing students together is not sufficient for promoting substantive gains in achievement." (Lou et al., 1996, p.448). Making use of existing databases, like ECLS-K, implies having to work with available data and therefore not being able to gather additional information on differentiation practices, unfortunately.

One study included in the current literature review does provide more information of the implementation of differentiation, namely the study on the effect of the comprehensive literacy program EMERGE (Gettinger & Stoiber, 2012). Within EMERGE, within-class ability grouping is part of a broader package of frequent process monitoring, enriched literacy content, and intensive teacher coaching. What is relevant is not the amount of time students spend in homogeneous ability groups (which is 30 minutes daily), but the fact that groups are

created based on recent performance data and that teachers are guided towards offering students of different performance levels appropriate, differentiated instruction and activities. This approach is fundamentally different compared to the ECLS-K studies, which only look at intensity or frequency of grouping.

5.2. Primary Education

Overall, based on reviews summarizing studies on differentiation up to 1995, previous studies did not report clear effects of between-class homogeneous ability grouping in primary education, but they did report some positive effects of providing students with instruction in small (homogeneous) ability groups within the classroom. Furthermore, both Slavin (1978a) and Lou and colleagues (1996) argue that the key of successful differentiation may not be merely placing students in groups, but actually adapting the teaching to the needs of different ability groups. Aim of this review was to replicate and extend the knowledge on the effects of differentiation practices. In the current systematic review, we included sixteen articles dealing with differentiation practices in primary education. Within these articles, we discerned four types of studies: studies of an intervention using ability grouping , studies analyzing the effects of naturally occurring grouping practices, studies on differentiation practices supported by computer systems, and studies in which differentiation was part of a broader school reform.

In the studies on naturally occurring practices, we found two types of differentiation practices which were also described in previous studies: within-class homogeneous ability grouping and between-class homogeneous grouping (also called setting). The two between-class ability grouping studies in our sample were on the effects of regrouping students for specific subjects or tracking students in homogeneous classes. Summarizing the effects of the two studies, a small negative effect was found of streaming or tracking on students' mathematics performance in homogeneous ability grouped classes compared to heterogeneous classes, especially for average ability students. This in contrast to previous reviews (Lou et al., 1996; Slavin, 1987a), in which no clear differential effects were found.

Another two studies of naturally occurring practices in primary education compared within-class ability grouping to not grouping students. Here, effects of near zero were found. However, the two studies providing insight in differential effects, show that homogeneous ability grouping overall had a small positive effect on high ability students' reading performance and a small negative effect on low ability students' performance. In this respect, within-class ability grouping could have a divergent effect, widening the gap between high and low ability students' performance. Only one study in our sample evaluated the effectiveness of an intervention which was specifically aimed at grouping students in either homogeneous versus heterogeneous ability groups within the classroom. In this study, a negative and non-significant effect of homogenous grouping was found compared to heterogeneous grouping. The finding from the meta-analysis of Lou et al. (1996) in which heterogeneous grouping was more beneficial for low ability students could not be replicated.

One reason why our findings on the effects of within-class ability grouping were not in line with previous positive findings on within-class ability grouping (Lou et al., 1996; Slavin, 1987a) may be that the studies on natural occurring grouping practices only gave insight in whether teachers used grouping or not, but not in how the grouping was actually used to provide adapted instruction. As noted previously, grouping may only be effective in cases in which instruction is also adapted to students' specific academic needs. The fact that ability grouping should be combined with instructional practices is illustrated by our review of the effectiveness of the use of adaptive computer systems for students' performance in reading and mathematics. In these studies, the computer adaptive system evaluated students' prior performance and used this to provide suggestions on the instructional content that students needed, which in turn influenced the grouping practices. Our meta-analyses of the findings of the studies using such a combination of adaptive testing, feedback and differentiated instruction show that this type of within-class differentiation can positively affect students' performance. Such computerized aids for supporting differentiation practices seem to be an interesting addition to the literature on differentiation from 1995 onwards.

Lastly, the effects of school reform programs in which differentiation was a prominent part of the program were evaluated. These comprehensive school reform programs such as Success for All, SEM-R and the Reading Acceleration Program overall had small to medium positive effects on students' reading performance. Again, it seems that the positive effect is magnified by combining different grouping practices with a varied offer of instructional content and school wide reform. For instance, in the Success for All program, students are regrouped across classes for daily reading periods. In the small reading classes, students' progress is frequently monitored and powerful instructional strategies aimed at increasing achievement and motivation are applied by well-trained tutors. Also, in the program, students work in cooperative groups frequently. This is another way to flexibly group students according to their instructional needs.

5.3. Early Secondary Education

The big differentiation question in secondary education can be framed as: "To track or not to track?" International debates about comprehensive or differentiated systems are heated, but the problem is that decisive scientific information can hardly be found since comparing the performance of national education systems mostly is based on international cross-sectional assessment studies like OECD-PISA or IEA-TIMSS. Another problem is that it is hard to ascribe differences between students to the effects tracking, since these may be due to existing differences that led them to be placed in a certain track in the first place. The results of studies on this topic are not very clear-cut, but at least seem to indicate that integrated systems in general do not perform worse than differentiated systems. And moreover, as was described in the theoretical framework, no effects of tracking or setting are found when the results of students of lower, average and higher ability are taken into account simultaneously.

The early review studies of Kulik and Kulik (1982) and Slavin (1990) on differentiation in secondary education concern the effects of ability grouping practices. A rigorous approach to assessing effects of ability grouping practices is to consider the whole population of students and not a selected subpopulation (e.g. gifted students or low ability students). Unfortunately, many studies do not address the effects ability grouping practices may have for the students *not* included. Studies on ability grouping practices for high ability students, for example, often fail to study the effects that separating high from average and low ability students may have on the performance of these latter two groups. In the end we only found four studies that both met are substantive and methodological inclusion criteria and studied the whole range of students varying in abilities.

The studies differ quite a bit. One study focused on computer aided mastery learning in the domain of mathematics, provided individualized instruction to students. Moreover, using progress reports from the computer system teachers provided additional individual support to students who need this. The effects of this approach were near medium (d=+0.416), and in line with findings reported for similar differentiation practices in primary education.

Two studies by Burris and colleagues (2006, 2008) looked into the effects of an accelerated math curriculum - the same curricular content was offered in two rather than the usual three years - that was taught in heterogeneous ability classes (rather than in the usual homogeneous ability classes), with additional instructional help for struggling learners. Unlike the other studies in our review the effects studied where not the cognitive math effects, but whether or not students opted for advanced math subjects after those two years and/or received a prestigious diploma afterwards. The results of these studies indicated that this was indeed the case, leading the authors to the conclusion that detracking can be done successfully.

Linchevski and Kutscher (1998) also looked for the effects of detracking grouping strategies in the mathematics domain. They studied an intervention that consisted of a mix of either heterogeneous or homogenous grouping after whole class instruction, with small group learning being fostered by cooperative learning. Heterogeneous grouping had a slight advantage over homogeneous grouping (d=+0.112), but retention effects could not be established.

Integrating differentiation practices in comprehensive programs that includes many more elements seems very promising. Once again, however, successful studies only have been conducted in the domain of mathematics. Similar studies on comprehensive programs for language were either designed with less rigor or produced less promising findings. We discussed the IMPROVE program, as an example of a proven effective broad program (d=+0.301). Important elements are that within the heterogeneous groups students question each other metacognitively (which implies cooperative learning based on peer interaction), continue learning for mastery up till 80% correct, and based on this criterion students either continue for enrichment or individualized corrective instruction.

5.4. Recommendations for research and practice

When trying to understand the effects of differentiation, it is important to use an ecologically valid operationalization of differentiation. Differentiation is more than within-class homogeneous ability grouping, and within-class homogeneous ability grouping is more than placing students together at a table for a certain amount of time. The real question is how teachers take into account differences between students in daily classroom practice and how they can be supported in doing so. Sensible ability grouping (both homogeneous and heterogeneous) and sensible application of other differentiation practices, like adaptive questioning during whole class activities, assume two things: teachers need to have an accurate view of students' level of understanding and teachers need to know which instruction and learning activity is appropriate for children at different levels, given the goals they strive for. Therefore, differentiation might be best applied within the context of comprehensive programs aimed at supporting teachers to adapt their teaching towards the needs of students. Most research on comprehensive programs we found focuses on reading and literacy. Differentiation in the domain of mathematics is often approached by using computer software. Software, either aiming at the domain of mathematics or language, can take part of the assessment and diagnosing out of the hand of the teacher and may provide instructional suggestions. Computer supported differentiation practices open the gates for completely individualized learning and instruction routes. Although computerized programs can be a helpful tool, it is the teacher who implements the differentiation practices and using differentiation software is not a guarantee for actual differentiation in the classroom.

For future research into differentiation practices our recommendations are the following:

1. Differentiation is not a concept that is used much in studies in Early Childhood Education. However, it is likely to be part of ECE classrooms with their child-following perspective of ECE, emphasis on play and on "naturally occurring" learning and instruction. It is therefore worthwhile to study the differentiation practices and their potential beneficial effects within the context of rich educational programs in more detail.
2. Computer supported differentiation practices seem promising. In our description of these practices we encountered elements such as assessment, using data for diagnosis, suggesting individual learning routes and indicating the need of supplementary support, etc. Comprehensive computerized programs may thus support teachers in implementing differentiation. Further research on how these programs influence teaching practices will help to understand how to use software as an effective teaching tool.
3. The most promising route for differentiation seems to be to embed it in a broader structure, either within a computerized system or a comprehensive educational program, which includes, for instance, meta-cognitive learning strategies, cooperative learning, regular assessment, remedial instruction, and flexible grouping. Studying the

effect of differentiation within such a broader structure is complicated, since all elements intertwine. Nevertheless, it seems important to further study the effects of differentiation when it is combined with other support systems, in order to determine how differentiation practices can be embedded within the classroom and the school.

References

* Adelson, J. L., & Carpenter, B. D. (2011). Grouping for achievement gains: for whom does achievement grouping increase kindergarten reading growth? *Gifted Child Quarterly, 55*(4), 265-278.

Anderson, K. M., & Algozzine, B. (2007). Tips for teaching: Differentiating instruction to include all students. *Preventing School Failure, 51*(3), 49-54.

* Barrow, L., Markman, L., & Rouse, C. E. (2009). Technology's edge: The educational benefits of computer-aided instruction. *American Economic Journal-Economic Policy, 1*(1), 52-74.

Blok, H. (2004). Adaptief onderwijs: betekenis en effectiviteit. *Pedagogische Studiën, 81*(1), 5-27.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Cha, b., B. (2005). The national randomized field trial of Success for All: second-year outcomes. *American Educational Research Journal, 42*(4), 673-696.

* 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*(3), 701-731.

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

* Burris, C. C., Wiley, E., Welner, K. G., & Murphy, J. (2008). Accountability, rigor, and detracking: Achievement effects of embracing a challenging curriculum as a universal good for all students. *Teachers College Record, 110*(3), 571-607.

* Burris, C. C., Heubert, J. P., & Levin, H. M. (2006). Accelerating mathematics achievement using heterogeneous grouping. *American Educational Research Journal, 43*(1), 105-136.

* References with an * are included in the meta-analysis.

Chambers, B., Cheung, A., Slavin, R. E., Smith, D., & Laurenzano, M. (2010). *Effective early childhood education programs: a systematic review.* Downloaded from: www.bestevidence.org: Best Evidence Encyclopedia.

* Chang, M. (2008). Teacher instructional practices and language minority students: A longitudinal model. *Journal of Educational Research, 102*(2-), 83-97.

Chorzempa, B. F., & Graham, S. (2006). Primary-grade teachers' use of within-class ability grouping in reading. *Journal of Educational Psychology, 98*(3), 529-541.

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

* Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science, 315*(5811), 464-465.

* Connor, C. M., Morrison, F. J., Schatschneider, C., Toste, J. R., Lundblom, E., Crowe, E. C., & Fishman, B. (2011a). Effective classroom instruction: Implications of child characteristics by reading instruction interactions on first graders' word reading achievement. *Journal of Research on Educational Effectiveness, 4*(3), 173-207.

* Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., . . . Schatschneider, C. (2011b). Testing the impact of child characteristics x instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly, 46*(3), 189-221.

De Koning, P. (1973). *Interne differentiatie*. Amsterdam: APS/RITP.

Gamoran, A., & Weinstein, M. (1998). Differentiation and opportunity in restructured schools. *American Journal of Education, 106*(3), 385-415.

Gardner, H. (1984). *Frames of mind: the theory of multiple intelligences*. London: Heinemann.

George, P. S. (2005). A rationale for differentiating instruction in the regular classroom. *Theory into Practice, 44*(3), 185-193.

Gettinger, M., & Stoiber, K. (2007). Applying a response-to-intervention model for early literacy development in low-income children. *Topics in Early Childhood Special Education, 27*(4), 198-213.

* Gettinger, M., & Stoiber, K. C. (2012). Curriculum-based early literacy assessment and differentiated instruction with high-risk preschoolers. *Reading Psychology, 33*(1-2), 11-46.

Hallam, S., Ireson, J., Lister, V., Chaudhury, I. A., & Davies, J. (2003). Ability grouping practices in the primary school: A survey. *Educational Studies, 29*(1), 69-83.

Hanushek, E. A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116*(510), 63-76.

* Hong, G., Corter, C., Hong, Y., & Pelletier, J. (2012). Differential effects of literacy instruction time and homogeneous ability grouping in kindergarten classrooms: Who will benefit? Who will suffer? *Educational Evaluation and Policy Analysis, 34*(1), 69-88.

* Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: an application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis, 31*(1), 54-81.

Houtveen, T., & van de Grift, W. (2012). Improving reading achievements of struggling learners. *School Effectiveness and School Improvement, 23*(1), 71-93.

Inspectorate of Education. (2013). *De staat van het onderwijs. Onderwijsverslag 2011/2012*. Utrecht: Inspectie van het Onderwijs.

Ireson, J., & Hallam, S. (2001). *Ability grouping in education*. London: Paul Chapman Publishing.

Ireson, J., Hallam, S., & Plewis, I. (2001). Ability grouping in secondary schools: Effects on pupils' self-concepts. *British Journal of Educational Psychology, 71*(2), 315-326.

Jakubowski, M., Patrinos, H. A., Porta, E. E., & Wisniewski, J. (2010). *The impact of the 1999 education reform in Poland*. World Bank: Policy Research Working paper 5263.

Kulik, C. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: a meta-analysis of evaluation findings. *American Educational Research Journal, 19*(3), 415-428.

Kulik, C. C., & Kulik, J. A. (1984). *Effects of ability grouping on elementary school pupils: a meta-analysis* (Paper presented at the annual meeting of the American Psychological Association ed.)

* Leonard, J. (2001). How group composition influenced the achievement of sixth-grade mathematics students. *Mathematical Thinking and Learning, 3*(2-3), 175-200.

* Linchevski, L., & Kutscher, B. (1998). Tell me with whom you're learning, and I'll tell you how much you've learned: Mixed-ability versus same-ability grouping in mathematics. *Journal for Research in Mathematics Education, 29*(5), 533-554.

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*(4), 423-458.

Luyten, H. (2008). *Empirische evidentie voor effecten van vroegtijdige selectie in het onderwijs, literatuurstudie in opdracht van het Ministerie van OCW*. Enschede: Universiteit Twente.

* Macqueen, S. (2012). Academic outcomes from between-class achievement grouping: the Australian primary context. *Australian Educational Researcher, 39*(1), 59-73.

* McCoach, D. B., O'Connell, A. A., & Levitt, H. (2006). Ability grouping across kindergarten using an early childhood longitudinal study. *Journal of Educational Research, 99*(6), 339-346.

McCoach, D. E. (2003). Does grouping matter? A cross-classified random effects model of children's reading growth during the first two years of school. *ProQuest, Dissertation Abstracts International Section A: Humanities and Social Sciences, 64*(5). (2003-95021-019).

Mevarech, Z. R., & Kramarski, B. (1997). IMPROVE: A multidimensional method for teaching mathematics in heterogeneous classrooms. *American Educational Research Journal, 34*, 365-394.

Neel, J. L. (2008). The effects of differentiated developmentally appropriate instruction of first grade learners. *ProQuest, Dissertation Abstracts International Section A: Humanities and Social Sciences, 68*(7). (2008-99011-064).

* Nomi, T. (2010). The effects of within-class ability grouping on academic achievement in early elementary years. *Journal of Research on Educational Effectiveness, 3*(1), 56-92.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM7 - Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International.

Reezigt, G. J. (1993). *Effecten van differentiatie op de basisschool*. Groningen: RION.

* Reis, S. M., McCoach, D. B., Coyne, M., Schreiber, F. J., Eckert, R. D., & Gubbins, E. J. (2007). Using planned enrichment strategies with direct instruction to improve reading fluency, comprehension, and attitude toward reading: An evidence-based study. *Elementary School Journal, 108*(1), 3-24.

* Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal, 48*(2), 462-501.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). *Effective beginning reading programs: A best-evidence synthesis*. Baltimore: Best Evidence Encyclopedia.

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research, 78*, 427-515.

Slavin, R. E. (1987a). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57*(3), 293-336.

Slavin, R. E. (1987b). Mastery learning reconsidered. *Review of Educational Research, 57*(2), 175-213.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60*(3), 471-499.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*(3), 427-515.

* Stevens, R. J., & Slavin, R. E. (1995). The cooperative elementary-school - effects on students achievement, attitudes, and social-relations. *American Educational Research Journal, 32*(2), 321-351.

* Tach, L. M., & Farkas, G. (2006). Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. *Social Science Research, 35*(4), 1048-1079.

Tomlinson, C. A. (2000). *Differentiation of instruction in the elementary grades*. Champaign, IL: ERIC Digest.

* Whitburn, J. (2001). Effective classroom organisation in primary schools: mathematics. *Oxford Review of Education, 27*(3), 411-428.

* Ysseldyke, J., & Bolt, D. M. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36*(3), 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*(2), 247-265.

Appendix 1: Included studies ECE and Kindergarten

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% confidence interval*
Adelson & Carpenter, 2011	homogeneous ability grouping for reading	USA (ECLS-K)	580 schools, 1690 classrooms, 9340 students	fall-spring K2	achievement	Relationship achievement grouping for reading (yes/no, as indicated by teacher) and reading growth	+0.068*	+0.028; +0.109
Chang, 2008	grouping*activity for math	USA (ECLS-K)	5863 Caucasian English only speaking students; 1151 African-American English only speaking students	spring K2, with follow ups to spring grade 5	achievement and interest	Relationship time spent on different classroom practices (as indicated by teacher on a 5-point scale ranging from 0 to 3+ hours a day) and math growth	*Caucasian* whole cl. +0.152* small gr. -0.045* indiv. +0.008* child sel. +0.012* *Afr.-Am.* whole cl. +0.134* small gr. +0.002 indiv. -0.069* child sel. +0.020*	 +0.151; +0.153 -0.047; -0.044 +0.007; +0.009 +0.011; +0.013 +0.128; +0.141 -0.005; +0.008 -0.076; -0.063 +0.013; +0.027
Gettinger & Stoiber, 2012	progress monitoring and adjusted instruction for reading	USA, large urban metropolis in the Midwest	15 classrooms, 124 students	4 months	achievement	classrooms randomly assigned to intervention condition with close monitoring/ formative assessment and adapted instruction for low, general, and high performing students.	*overall* V +0.837* R1 +0.388* R2 +0.574* R3 +0.911* R/RC +0.572* *High ability* V +0.243 R1 +0.474 R2 +0.468	 +0.470; +1.204 +0.033; +0.743 +0.215; +0.933 +0.542; +1,281 +0.213; +0.931 -0.357; +0.843 -0.133; +1.080 -0.138; +1.074

						V=vocabulary R1=rhyme awareness and alphabet knowledge R2=print knowledge and phonological awareness R3=upper case letter naming R=reading RC=reading comprehension	R3 +0.675* R/RC +0.696* *Average ability* V +0.465 R1 +0.500 R2 +0.845* R3 +1.276* R/RC +0.999* *Low ability* V +0.337 R1 +0.594 R2 +0.500 R3 +1.015* R/RC +0.876*	+0.060; +1.289 +0.081; +1.312 -0.121; +1.050 -0.087; +1.087 +0.241; +1.448 +0.642; +1.910 +0.386; +1.612 -0.331; +1.004 -0.083; +1.271 -0.173; +1.173 +0.311; +1.719 +0.182; +1.570
Hong & Hong, 2009	homogeneous ability grouping for reading	USA (ECLS-K)	740 schools, 1858 classrooms, 10189 students	fall-spring	achievement	Relationship between instruction time (high or low) * grouping (G - high, low or no) and reading growth. nb no grouping = whole class	*Grouping under low instr. time* low G +0.036 high G -0.040 *Grouping under high instr. time* low G +0.164* high G +0.198*	 -0.094; +0.165 -0.173; +0.094 +0.047; +0.281 +0.051; +0.346
Hong et al., 2012	homogeneous ability grouping for reading	USA (ECLS-K)	665 schools, 1697 classrooms, 8668 students	fall-spring	achievement	Relationship between instruction time (high or low) * grouping (G high, low or no) and reading growth for 3 groups of students (high, medium, low ability) R1=letter recognition R2= beginning sounds R3=ending sounds	Whole class vs intensive grouping under low instr. time *high ability* R1 -0.064 R2 +0.083 R3 +0.088 R4 +0.184 RC +0.142	 -0.281; +0.152 -0.134; +0.299 -0.128; +0.305 -0.033; +0.401 -0.074; +0.359

	R4=sight words	*Average ability*	
	RC=reading comprehension	R1 +0.031	-0.070; +0.131
		R2 +0.048	-0.052; +0.148
		R3 +0.038	-0.062; +0.139
		R4 +0.072	-0.029; +0.127
		RC +0.023	-0.077; +0.124
		Low ability	
		R1 +0.236*	+0.070; +0.402
		R2 +0.181*	+0.015; +0.346
		R3 +0.220*	+0.054; +0.386
		R4 +0.325*	+0.159; +0.491
		RC +0.328*	+0.162; +0.494
		High vs low instruction time under intensive grouping	
		High ability	
		R1 +0.073	-0.181; +0.327
		R2 +0.267*	+0.011; +0.522
		R3 +0.175	-0.079; +0.430
		R4 +0.284*	+0.029; +0.539
		RC +0.255	0.000; +0.510
		Average ability	
		R1 +0.158*	+0.040; +0.277
		R2 +0.145*	+0.027; +0.263
		R3 +0.152*	+0.034; +0.270
		R4 +0.174*	+0.055 - +0.292
		RC +0.118	0.000; +0.236
		Low ability	
		R1 +0.236*	+0.045; +0.427
		R2 +0.170	-0.020; +0.360
		R3 +0.234*	+0.043; +0.424
		R4 +0.268*	+0.077; +0.459
		RC +0.208*	+0.018; +0.398

D.B. McCoach et al., 2006	homogeneous ability grouping for reading	USA (ECLS-K)	620 schools, 10191 students	fall-spring	achievement	Relationship between frequency of ability grouping per week (as indicated by teacher on a 5 point scale ranging from never to daily) and reading growth	+0.127*	+0.068; +0.186
Tach & Farkas, 2006	Homogeneous ability grouping for reading	USA (ECLS-K)	Kindergarten sample: 2420 classrooms, 11769 students	fall-spring K	achievement	Multi-level analysis studying the relationship between ability grouping in Kindergarten and reading achievement at the end of the school year	+0.346*	+0.265; +0.427

* 95% confidence interval of effect size does not contain 0

Appendix 2: Included studies Primary Education

Appendix 2a: An intervention study on ability grouping

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% confidence interval*
Leonard, J., 2001	Within-class heterogeneous small groups versus within-class homogeneous small groups	USA	177 students from 3 classes: 88 students heterogeneous cohort (16 low, 34 average, 43 high); 89 students homogeneous cohort (37 low, 29 average, 28 high)	One school year (fall – spring)	Achievement	Comparison of students' mathematics achievement in the homogeneously grouped cohort versus the heterogeneously grouped cohort	*Overall* -0.250 *Low ability* -0.397 *Average ability* -0.133 *High ability* -0.185	 -0.546; +0.046 -1.006; +0.213 -0.644; +0.379 -0.675; +0.305

* 95% confidence interval of effect size does not contain 0

Appendix 2b: Ability grouping studies

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% confidence interval*
Condron, 2008	Within-class ability grouping	USA	K – 1: 13,625 students (ungrouped: 4718 students, low group: 2219, average group: 3380, high group: 3308) Grade 1 – 3: 13.010 students (ungrouped: 6873, low group: 1436, middle group: 2067, high group: 2634)	Growth from kindergarten to the end of grade one and from grade one to the end of grade three	Achievement	Propensity score matching is used to estimate the effect of placement in a high, average or low ability group in comparison to non-grouped instruction. We report the general effects cumulated over the various strata	K – grade 1 *Low ability* -0.288* *Average ability* -0.043 *High ability* +0.207* Grade 1 - 3 *Low ability* -0.245* *Average ability* +0.046 *High ability* +0.177*	 -0.343; -0.233 -0.088; +0.002 +0.158; +0.256 -0.305; -0.185 -0.005; +0.097 +0.129; +0.225
Macqueen, 2012	Between-class ability grouping	Australia	8 schools. Literacy: regrouping	Growth from grade	Achievement	Comparison of growth scores of students in between-class ability	*Overall Literacy* +0.196	 -0.170; +0.561

(setting)	50 students, heterogeneous 68 students Writing: regrouping 29 students, heterogeneous 47 students Math: regrouping 51 students, heterogeneous 69 students	three and five	grouped classes versus students in heterogeneous classes in the areas of literacy, writing and mathematics. *Low lit group:* Low level literacy group versus heterogeneous *Average lit group:* Average level literacy group versus heterogeneous *High lit group:* High level literacy group versus heterogeneous *Low math group:* Low level math group versus heterogeneous *Average math group:* Average level math group versus heterogeneous *High math group:* High level math group versus heterogeneous	*Overall Writing* -0.082 *Overall Math* -0.125 *Low lit group: Literacy* -0.379 *Average lit group: Literacy* +0.275 *High lit group: Literacy* +0.218 *Low lit group: Writing* +0.038 *Average lit group: Writing* -0.023 *High lit group: Writing* +0.196 *Low math group: Math* -0.776 *Average math group: Math* -0.061 *High math group: Math* +0.171	 -0.545; +0.381 -0.488; +0.237 -1.290; +0.532 -0.286; +0.836 -0.243; +0.678 -1.130; +1.206 -0.738; +0.691 -0.463; +0.855 -1.620; +0.067 -0.605; +0.483 -0.294; +0.636

Nomi, 2010	Within-class ability grouping	USA	13512 schools with 13512 students: 3922 schools with 2043 students ungrouped, 9590 schools with 6742 students ability-grouped ;	Achievement from kindergarten to the end of first grade	Achievement	Propensity score matching is used to estimate the effect on reading scores of placement in a high, average or low ability group in comparison to a non-grouped classroom.	*Overall* -0.010 *Low ability* -0.030 *Average ability* 0.021 *High ability* -0.059	 -0.060; +0.039 -0.126; +0.066 -0.063; +0.105 -0.141; +0.023
Tach & Farkas, 2006	Within-class ability grouping	USA	First grade sample: 3133 classes with 13.010 students (ability grouped classes: 2256)	Achievement from kindergarten to the end of first grade	The authors analyze which variables affect teachers' grouping practices. Students' prior achievement has the strongest effect on grouping. Also, grouping effects were found for students' learning	Multilevel analyses are used to determine the effect of having ability groups present in the classroom on students' reading performance	-0.191*	-0.261; -0.120

					behavior, SES, age, and classroom-level variables related to average performance, ethnicity, SES and age.			
Whitburn, 2001	Between-class ability grouping (setting)	United Kingdom	1200 students (200 students in homogeneous classrooms and 1000 in heterogeneous classrooms)	Cohort 1: 21 months Cohort 2: 15 months Cohort 3: 3 months	Achievement	Comparison of mathematics performance between students taught in homogeneous (set) classes and students in mixed ability classes.	*First Cohort* *Total grade 3* -0.030 *Total grade 4* -0.270* *Low ability grade 3* +0.040 *Low ability grade 4* -0.340 *Average ability grade 3* -0.670* *Average ability grade 4* -0.690* *High ability grade 3* +0.080 *High ability grade 4* -0.090	 -0.292; +0.232 -0.533; -0.007 -0.353; +0.433 -0.735; +0.055 -1.071; -0.269 -1.091; -0.289 -0.313; +0.473 -0.483; +0.303

	Second Cohort	
	Total grade 3	-0.292; +0.232 -0.030
	Total grade 4	-0.393; +0.133 -0.130
	Low ability grade 3	-0.333; +0.453 +0.060
	Low ability grade 4	-0.453; +0.333 -0.060
	Average ability grade 3	-0.604; +0.184 -0.210
	Average ability grade 4	-0.735; +0.055 -0.340
	High ability grade 3	-0.463; +0.323 -0.070
	High ability grade 4	-1.030; -0.230 -0.630*
	Third Cohort	
	Total grade 3	-0.373; +0.153 -0.110
	Total grade 4	-0.553; -0.027 -0.290*
	Low ability grade 3	-0.847; -0.053 -0.450*
	Low ability	

	grade 4	-0.877; -0.083
	-0.480*	
	Average ability	
	grade 3	-0.847; -0.053
	-0.450*	
	Average ability	
	grade 4	-0.877; -0.083
	-0.480*	
	High ability	
	grade 3	-0.433; +0.353
	-0.040	
	High ability	
	grade 4	-0.877; -0.083
	-0.480*	

* 95% confidence interval of effect size does not contain 0

Appendix 2c: Studies on computerized systems

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% confidence interval*
Connor et al., 2007	Within-class differentiated instruction	USA	10 schools, 47 classes (treatment 22 classes, control 25 classes), 616 students	fall-spring	performance	A cluster-randomized field trail is used in which effects of differentiated instruction using the computer program are compared to students results in matched control schools on a language and literacy outcome measure.	+0.183*	+0.025; +0.342
Connor et al., 2011a	Within-class differentiated instruction	USA	7 schools, 33 classes (16 treatment, 17 control), 464 students (experimental group: 219 students, control group: 229 students)	fall-spring	performance	Multilevel modeling is used to analyze the effects of differentiated instruction using the computer program in comparison to a vocabulary instruction intervention on reading comprehension an vocabulary outcome measures.	*Reading comprehension* +0.191* *Vocabulary* +0.033	+0.005; +0.377 -0.153; +0.219
Connor et al., 2011b	Within-class differentiated instruction	USA	7 schools, 25 classes, 396 students ((16 treatment, 17 control), 464 students (experimental group: 3	fall-spring	performance	Multilevel modeling is used to analyze the effects of differentiated instruction using the computer program in comparison to a control group on a language and literacy outcome measure.	+0.249*	+0.050; +0.448

			schools, 14 classes, 222 students; control group: 4 schools, 11 classes, 174 students))					
Ysseldyke et al., 2003	Within-class differentiated instruction	USA	Experimental group: 18 classes, 397 students. Within-school control group: 484 students	September - June	performance	An analysis of variance of the mean scores on two mathematics tests (NALT and STAR Math) of the experimental and the control group	*NALT* +0.189* *STAR Math* +0.268*	+0.030; +0.349 +0.109; +0.428
Ysseldyke et al., 2007	Within-class differentiated instruction	USA	Experimental condition: 8 schools, 41 classrooms; Control condition: 8 schools, 39 classrooms	October - May	performance	An analysis of variance of the mean scores on two mathematics tests (NALT and STAR Math) of the experimental and the control group in primary education	*Terra Nova* +0.469* *STAR Math* +0.458*	+0.312; +0.626 +0.294; +0.622

* 95% confidence interval of effect size does not contain 0

Appendix 2d: Studies on differentiation as part of a broader program

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% confidence interval*
Borman, et al., 2007	Ability grouping across grades for reading, as a part of a whole school comprehensive reform	USA	35 schools: 1445 students in Grade 2 (longitudinal sample of students that started in K)	3 years, from kindergarten to second grade	Achievement, measured every 9 weeks	Cluster randomized design	Word Identification: +0.220*[2] Word Attack: +0.330*[2] Passage Comprehension: +0.210*[1]	+0.024; +0.416 +0.114; +0.546 +0.034; +0.386
Houtveen & van de Grift, 2012	Direct instruction in heterogeneous group, and intensive small group instruction, aimed at convergent differentiation	The Netherlands	37 schools; 21 treatment schools, 16 control schools, 1021 students	December-May	Achievement	Quasi-experimental	Word Decoding: +0.280*[2] Fluency: +0.620*[2]	+0.156; +0.404 +0.494; +0.746
Stevens & Slavin, 1995	Students work in heterogeneous learning teams but receive instruction in relatively homogeneous teaching groups, as part of a whole school reform program	USA	5 schools: 2 treatment schools, 3 control schools, grade 2 – 6	After 1 and 2 years	Achievement	Quasi-experimental	After 1 year: Read voc: +0.170* Read comp: +0.130 Lang mech: -0.010 Lang expr: +0.080 Math comp: +0.120 Math appl: -0.050 After 2 years: Read voc: +0.210* Read comp: +0.280*[1] Lang mech: +0.100 Lang Expr: +0.210* Math comp: +0.290 Math appl: +0.100	 +0.014; +0.326 -0.026; +0.286 -0.164; +0.144 -0.074; +0.234 -0.056; +0.296 -0.204; +0.104 +0.075; +0.345 +0.128; +0.432 -0.069; +0.269 +0.069; +0.351 +0.139; +0.441 -0.058; +0.258
Reis et al., 2007	SEM-R (School-wide Enrichment Model in Reading Framework): differentiated,	USA	2 schools, 14 (7 treatment, 7 control) teachers, 226 students,	12 weeks	Teacher's judgment	Randomized design	Fluency: +0.299*[2] Comprehension: +0.220[1]	+0.005; +0.594 -0.529; +0.970

	individual reading instruction among other things (all students participate in SfA in the morning)		grade 3 – 6, teachers and students were randomly assigned to treatment or control group					
Reis et al., 2011	SEM-R (School-wide Enrichment Model in Reading Framework): differentiated, individual reading instruction among other things	USA	5 schools, 63 teachers, 1192 students (grade 2, 3, 4, 5), teachers/classes were randomly assigned to control or treatment groups	24 weeks	Teacher's judgment	Cluster-randomized design	Fluency: +0.254[2] Comprehension: +0.145[1]	-0.063; +0.571 -0.096; +0.386

1) Effects included in the meta-analysis of comprehensive reading

2) Effects included in the meta-analysis of basic reading, correction is made for including two measures from 1 study by multiplying the standard error with √2.

* 95% confidence interval of effect size does not contain 0

Appendix 3: Included studies Early Secondary Education

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% confidence interval*
Barrow et al., 2009	computer aided individualized instruction	USA, 3 large urban districts in Northeast, Midwest and South	1605 students, 59 teachers, 146 classrooms, 17 schools	1 school-year	n/a random assignment	Within school random assignment of classrooms	+0.416*	+0.261 - +0.571
Burris et al., 2006	heterogeneous ability grouping with advanced courses for all and remediation if necessary	USA, suburban area	6 cohorts, 3 before (477 students) and 3 after (508 students) accelerated math curriculum for all	2 years	Heterogeneous grouping; all students included (achievement)	Quasi-experimental longitudinal cohort study M1=sequential maths M2=calculus M3=advanced place calc	M1: +1.450* M2: +1.511* M3: +1.097	+0.062 - +2.838 +0.204 - +2.817 -0.171 - +0.2365
Burris et al., 2008	heterogeneous ability grouping with advanced courses for all and remediation if necessary	USA	6 cohorts, 1300 students	2 years	Heterogeneous grouping; all students included (achievement)	Quasi-experimental cohort study state=diploma tied to state-wide standards int.= diploma tied to international standards	state: +3.187* int.: +0.965	+1.700; +4.673 -0.302; +2.232
Linchevski & Kutscher,	Effect of mixed ability grouping	Israel	1629 students,	1 school-year (7th	Achievement	regression-discontinuity design. Study 1: analysis	gr 7: +0.112* gr.8: +0.164	+0.018; +0.207 -0.037; +0.364

1998	on math	12 schools, 40 classrooms /groups	grade)	on school level. For 4 schools (12 groups) retention effects at the end of 8th grade.

* 95% confidence interval of effect size does not contain 0