Effective differentiation Practices

Deunk, Marjolein I.; Jacobse, Annemieke E.; de Boer, Hester; Doolaard, Simone; Bosker, Roel J.

Published in:
Educational Research Review

DOI:
10.1016/j.edurev.2018.02.002

Publication date:
2018

Effective Differentiation Practices:

A Systematic Review and Meta-Analysis of Studies on the Cognitive Effects of Differentiation Practices in Primary Education

Marjolein I. Deunk, Annemieke E. Smale-Jacobse, Hester de Boer, Simone Doolaard and Roel J. Bosker

GION Education/Research, Faculty of Behavioral and Social Sciences of the University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, the Netherlands

Email addresses authors:

Marjolein Deunk (corresponding author): m.i.deunk@rug.nl

Annemieke Smale[1]: a.e.smale-jacobse@rug.nl

Hester de Boer: hester.de.boer@rug.nl

Simone Doolaard: s.doolaard@rug.nl

Roel Bosker: r.j.bosker@rug.nl

[1] Present address: Department of Teacher Education, Faculty of Behavioral and Social Sciences, Nieuwe Kijk in 't Jatstraat 70, 9712 SK University of Groningen, the Netherlands

Abstract

This systematic review gives an overview of the effects of differentiation practices on language and math performance in primary education, synthesizing the results of empirical studies (*n* = 21) on this topic since 1995. We extracted 78 effect sizes from the included studies. We found that using computerized systems as a differentiation tool and using differentiation as part of a broader program or reform had small to moderate positive effects on students' performance. Between- or within-class homogeneous ability grouping had a small negative effect on low-ability students, but no effect on others. The finding that computer technology can be a useful tool to facilitate differentiated instruction is not covered in earlier reviews. Moreover, our findings emphasize that homogeneous ability grouping alone is not enough to guarantee differentiated instruction. This stresses the importance of embedding differentiation practices in a broader educational context. *(141 words)*

Keywords: Differentiation practices; ability grouping; primary education; systematic review; meta-analysis

Effective Differentiation Practices: A Systematic Review and Meta-Analysis of Studies on the Cognitive Effects of Differentiation Practices in Primary Education

1. Introduction: differentiation in primary education

Student ability in untracked primary classrooms may vary widely, which poses challenges for teachers. This variability does not only occur in schools with a policy of full inclusion, but in all classrooms that are created based on student age (Tomlinson et al., 2003). The quality of schools is largely determined by how teachers deal with these (cognitive) differences between students and by how they adapt their instruction to individual needs (e.g., Hamre & Pianta, 2005). This requires teachers to develop advanced professional skills in addition to their basic skills of classroom management and general didactics. Note that this hierarchy of professional skills stems from practice, not from principle: although taking into account individual student needs is fundamental of good teaching and therefore should be a basic skill, research shows that novice teachers first need to master other skills before they can start attending to differences between students well (Maulana, Helms-Lorenz, & van de Grift, 2014; Van de Grift, 2007; Van de Grift, Van der Wal, & Torenbeek, 2011). These advanced professional skills are summarized in the concept ‘differentiation’. Differentiation is a combination of careful progress monitoring and adapting instruction in response (Heitink, Van der Kleij, Veldkamp, Schildkamp, & Kippers, 2016; Prast, van de Weijer-Bergsma, Kroesbergen & van Luit, 2015; Roy, Guay & Valois, 2013). It is “an approach to teaching in which teachers proactively modify curricula, teaching methods, resources, learning activities, and student products to address the diverse needs of individual students and small groups of students to maximize the learning opportunity for each student in a classroom” (Tomlinson et al, 2003, p. 120). It is related to the concept of aptitude-

treatment interaction, which emphasizes that education is most effective when instruction is closely matched to the student's own capacities and talents, and also acknowledges the complex interplay between characteristics of the student, task and instruction (Snow, 1989). Differentiation is an overall approach to teaching and can include combinations of many practices, like flexible (heterogeneous or homogeneous) grouping, detailed progress monitoring, using adaptive computer programs or learning materials, modifying learning content, adapting instruction for weaker students, and providing opportunities for acceleration for stronger students. Differentiation practices can be applied to areas of learning content, learning process, learning product (Roy, Guay, &Valois, 2013). Tomlinson (2014) extends this list with affect or environment. Furthermore, teachers may not only take into account differences in students' cognitive abilities, but also other differences such as in students' motivation or interest for example. This broad array of differentiation options is appealing, but does pose some challenges in a theoretical sense because of the many practices and understandings that it may entail. To assure a clear focus, and therefore aim at larger practical and theoretical relevance, the current review study is limited to differentiation in which student differences in ability or performance are taken into account. The potential relevance of this type of differentiation is clear from theoretical underpinnings in theories such as Vygotsky's Zone of Proximal Development (1978), which describes how learning could be advanced by providing students tasks that are just outside their current level of mastery. Therefore, the definition of differentiation we will use is in this study is: teaching modified to address the diverse cognitive needs of *all* students [2].

[2] This means singling out students by individual out-of-class tutoring or by creating separate classrooms for the gifted is beyond the scope of this study.

How teachers choose to apply differentiation seems to be related 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; [Author], 2005). Teachers aiming at convergence mainly focus on helping all their students to reach a basic performance level. This implies that they may dedicate additional time and effort to low-achieving students in order to help them reach a minimum performance level, even when this is at the expense of time they had reserved for high-ability students. Teachers aiming at divergence, on the contrary, mainly focus on helping all students to reach their highest potential, dividing attention equally between students with lower, average, and higher ability. Their use of ability-appropriate performance goals for (groups of) students at different ability levels may lead to a widening of the gap between lower- and higher-ability students. Convergent and divergent goals thus lead to different pedagogical-didactical decisions. In practice, though, most teachers are likely to combine convergent and divergent goals, and will aim to reach a minimum performance level with low-ability students while also offering high-ability students the opportunity to extend their knowledge without proceeding (too much) ahead of their classmates (Denessen, 2017).

Differentiation in education is a highly debated topic, especially when it is applied in the form of homogeneous grouping. Teachers appear less accurate in estimating students' cognitive abilities when they are placed in homogeneous classrooms (Machts, Kaiser, Schmidt & Möller, 2016). Most concerns regarding homogeneous grouping are related to the reduced learning opportunities for low-ability students: within these groups, students cannot profit from the input of higher-ability peers or from the role models that high-ability students can be (e.g., Burris, Heubert, & Levin, 2006). Furthermore, teachers may have lower expectations of low-ability students and, therefore, unconsciously limit their opportunity to learn. This is especially relevant

for students from impoverished backgrounds or minority groups, who might be labeled as being of "low ability" even before they have had the opportunity to show their potential (Denessen, 2017). Teacher expectations and beliefs are found to correlate with the SES of students (e.g. Lee & Ginsburg, 2007; Ready & Wright, 2011). When students from low SES families are placed in a low-ability group too soon – based on general estimates or prejudices, rather than on actual performance level – they might encounter lower expectations and, as a result, less demanding teaching and unequal learning opportunities. The debate on how to implement differentiation in such a way that students of *all* ability levels profit from it should be informed by empirical research data. Review studies of the effects of differentiation practices are, therefore, important.

1.1 Evidence on the Effects of Differentiation: Situation up to 1995

One of the most common differentiation practices in primary education is within-class homogeneous ability grouping (e.g. Anderson & Algozzine, 2007; Chorzempa & Graham, 2006; de Koning, 1973; George, 2005; Kulik & Kulik, 1984; ; Reezigt, 1993; Slavin, 1987a). This organizational tool can be used as a context for fitting instruction to the needs of individual students in academically diverse classrooms. Five key systematic reviews and meta-analyses on differentiation in primary education until 1995 were conducted by Kulik and Kulik (1984), Kulik (1992), Lou and colleagues (1996), and Slavin (1987a; 1987b). Slavin's latter review has been part of a public academic dispute (see Kulik, Kulik & Bangert-Drowns, 1990; Slavin, 1990), which illustrates the relation between decisions of the researcher and outcomes of a (review) study, especially when fuzzy constructs like 'differentiation' are the topic of concern. We consider Slavin's review relevant for the current study, though. In addition, Steenbergen-Hu, Makel and Olszewski-Kubilius (2016) conducted a meta-meta-analysis on reviews conducted up

to 1995, which included three of the five reviews (Kulik & Kulik, 1984; Lou et al. 1996; Slavin, 1987a). However, in the meta-meta-analysis no distinction is made between primary and secondary education, which makes the results not fully comparable with the described systematic reviews and more difficult to interpret for the purpose of the current study. Four of the five reviews on differentiation as well as the meta-meta-analysis focus on different forms of grouping based on academic performance or ability: general whole-class homogeneous ability grouping; temporary whole-class homogeneous ability grouping for specific subjects (setting); temporary within-class homogeneous grouping for specific subjects; and small-group formation in general, whether homogeneous or heterogeneous. The fifth review is about mastery learning, a form of convergent differentiation.

The review studies do not lead to a clear conclusion about the effects of differentiation. Different forms of grouping seem to create different opportunities for effectively adapting teaching to students' needs. In general, homogeneous whole-class ability grouping does not seem to be very effective for students in primary education, nor does it seem to positively influence the well-being of students of all ability levels (in secondary education, Belfi, Goos, De Fraine & Van Damme, 2012). Kulik and Kulik (1984) summarized the effects of 19 studies and report an overall effect size of +0.07. They found a higher effect size for homogeneous grouping of gifted and high performing students, but without information on the effects of the extraction of gifted and high performing students out of the classroom on other students, this finding biases the effect of homogeneous whole class grouping. Kulik (1992) reviewed 51 studies, of which 26 took (partly) place in primary education. The individual effect sizes of these 26 studies range from -0.95 to +0.46. Slavin (1987a) summarized 17 studies and reports an overall effect size of 0.00. The findings on the differential effects of this type of grouping are inconclusive, although there

are some indications that this practice is more profitable for high performing students and less profitable for low performing students. The results of the meta-meta-analysis of Steenbergen-Hu and colleagues (2016) are in line the results of the reviews described above (effect sizes: overall -0.03; low ability +0.03, average ability -0.04, high ability +0.06; all effect sizes are non-significant).

Homogeneous whole-class ability grouping for specific subjects (setting) seems more promising than full time whole-class homogeneous ability grouping. When students are temporarily regrouped across grades, high performing grade 2 students could for example be placed together with low performing grade 3 students for a specific subject. Slavin (1987a) reviewed 14 studies with this kind of arrangement and reported an overall effect size of +0.45. Kulik (1992) reviewed as well 14 studies on across grade grouping and reported an overall effect size of +0.33. Neither review study contained enough information on the performance of students of low, average and high ability to draw conclusions on differential effects. In the meta-meta-analysis (Steenbergen-Hu et al., 2016) a slightly lower overall effect size of +0.26 is reported, and no differential effects.

Another, probably more feasible, form of grouping is within-class homogeneous ability grouping for specific subjects. This type of grouping has small positive overall effects, especially when it is compared with whole-class teaching. Slavin (1987a) reviewed 8 studies and reported an effect size of +0.32 (based on 5 of the 8 studies which used a randomized design). Kulik (1992) reviewed 11 studies on within class grouping, of which eight focused on primary education , and reported an overall effect size of +0.25. The positive effects of this type of grouping are smaller, however, when a comparison with within-class heterogeneous grouping is made. Lou and colleagues (1996) reviewed 20 studies on primary, secondary and post-secondary

level which compared homogeneous with heterogeneous grouping and reported and overall effect size of +0.12. These findings indicate that the positive effects of within-class homogeneous grouping may be the result of forming small groups, rather than the result of a specific configuration of the groups. This suggestion is supported by the finding of Lou and colleagues (1996) that both homogeneous and heterogeneous within-class grouping are more effective than whole-class teaching (grades 1-3, ES=+0.08; grades 4-6, ES=+0.29). Again, differential effects are inconclusive. Kulik (1992) reports positive overall effects for students of low (ES=+0.16), average (ES=+0.18) and high (ES=+0.30) ability. Slavin (1987a) as well reported positive differential effects for students of all ability levels, although he did not calculate overall effect sizes. However, the review of Lou and colleagues (1996) in which homogeneous within-class grouping was compared with within-class heterogeneous grouping in primary to (post)secondary education, reported negative effects for low-ability students (ES=-0.60), positive effects for average-ability students (ES=+0.51), and small positive effects for high-ability students (ES=+0.09). The results of the meta-meta-analysis of Steenbergen-Hu and colleagues (2016) partly confirm the findings from the four systematic review studies described above: in line with the other studies, an overall positive effect for within-class homogeneous grouping is reported (+0.25), but no evidence for a negative effect of this type of grouping for subgroups of students are reported (low ability: +0.30, average ability: +0.19, high ability +0.29).

The studies described above focused on different types of grouping as a context for differentiation. The fifth systematic review study focused on mastery learning as a differentiation strategy (Slavin, 1987b). mastery learning entails that regular progress assessments are used to check whether students have reached certain ability levels. The group of students that does not perform well enough receives additional instruction inside or outside the classroom. The group

that meets the standards may receive advanced materials for enrichment. Key to mastery learning is allowing students enough time for learning, which implies some students will need more instruction and practice than others (Bloom, 1971). Five of the studies reviewd by Slavin were conducted in elementary classrooms, and included control classrooms which spent the same amount of time on the subject matter as the experimental classrooms and used standardized tests. The overall effects of mastery learning in this selection of studies ranged from 0.00 to +0.25. When studies in which experimenter-made tests were used instead of standardized tests were considered (n = 5), the range in effect sizes widened. No differential effects were reported.

Overall, the conclusion that can be drawn from the review studies is that (homogeneous) ability grouping may have positive effects, especially when students are regrouped for specific subjects and when the resulting ability groups are small. Differential effects for low-, average-, and high-ability students are inconclusive, however. These mixed findings may be the result of the way grouping is used as a context for taking into account students' needs. Clearly, just grouping students and placing them together physically does not ensure differentiated teaching. Referring to both homogeneous and heterogeneous grouping, Lou and colleagues state the obvious that "Overall, 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). Lou and colleagues (1996) analyzed the results of a sub-selection of studies (conducted in primary, secondary, and postsecondary education) which gave (some) information on what teachers actually did after they created groups. As expected, they found larger effects for within-class grouping when teachers adapted their instruction (ES = +0.25) than when teachers provided their

regular whole-class instruction to the small groups. Unfortunately, as Slavin (1987a) already noted, many researchers do not provide specified information on the instructional practices used in interaction with ability groups and therefore it is often hard to reconstruct the operationalization of differentiation in the different studies.

1.2 Research Question and Hypotheses

Differentiation practices seem promising, but due to the fuzziness of the concept under which conditions and in which form differentiation is effective for students of all ability levels remains unclear. The aim of the current review was to analyze recent evidence on the effects of differentiation and add to the understanding if and how differentiation in primary education can positively affect the language and math performance of low-, average-, and high-ability students. Our research question was as follows: *What are the cognitive effects of differentiation practices on students in primary education?* In answering this question, we also considered a related question on the operationalization of differentiation practices in different studies. The review builds on previous research and includes recent empirical studies, published since 1995.

We expected differentiation in all its forms to have positive effects on students of all ability levels, as long as the teachers actually adapted their instructions to the needs of students. We expected grouping to be potentially effective, because it can serve as a good context for applying other differentiation practices specifically aimed at students' needs, like explaining content again in another way to weaker students, providing additional worksheets for stronger students, or designing different assignments for small mixed-ability groups. Based on the findings of previous reviews described above, we did not expect overall effects of general whole-class homogeneous ability grouping. We expected positive effects of within-class homogeneous

and heterogeneous ability grouping for specific subjects on the performance of students of all ability levels.

2. Method

We investigated the effectiveness of different differentiation practices in the form of a systematic review, conducting a meta-analysis where possible. We extended the review with additional contextual information on the selected studies, emphasizing studies that are particularly relevant to the topic of interest (Slavin, Lake, Chambers, Cheung, & Davis, 2009). To ensure the most comprehensive literature search, we conducted both an electronic database search and a cited-references search. In order to find as many relevant sources as possible, we started the literature search with a broad electronic database search. We then narrowed down the number of results by manually applying additional selection criteria. We calculated effect sizes for each eligible study, and performed content coding in order to create an overview of the different types of studies and the different elements of differentiation investigated. We used this information to provide context to the effect size data of the meta-analysis.

2.1 Literature Search Procedures

We conducted an extensive literature search in the educational databases ERIC, psycINFO, and SSCI. We used each of 10 keywords twice: once in combination with the keyword *achiev** and once in combination with the keyword *effect**. The set of 10 keywords consists of 5 general terms related to differentiation (*"adapt* instruct*", "adapt* teach*", differentiat*, "individuali* instruct*", "individuali* teach*"*) and 5 more specific terms (*"ability group*", "aptitude treatment", grouping*, "mastery learning"*, *streaming*). We added

the specific terms in an attempt to reduce the effects of the fuzziness of the concept differentiation. Papers in which these keywords were 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 6–12 years (i.e., primary education; grades 1 to 6 in the US system).[3]

In addition to the database search, we conducted a cited references search using the SSCI database. We selected 11 key publications on differentiation, namely, Blok (2004), Borman et al. (2005), de Koning (1973), Gamoran and Weinstein (1998), Ireson and Hallam (2001), Kulik and Kulik (1984), Lou et al. (1996), Reezigt (1993), and Slavin (1987a; 1987b; 1990). All peer-reviewed papers published since 1995 that made reference to one of these 11 key publications were collected. The searches were conducted the end of April, 2012.

These two broad search methods led to a collection of around 1,430 references, which we narrowed down by manually applying further selection criteria. The first broad selection criterion was whether the study was on language or math, or not. Language in this case encompassed reading, writing, vocabulary, grammar, etcetera, in the native language of the country under investigation (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. We rejected abstracts which indicated that studies were not focused on students of 6 to 12 years of age (even though this had been one of the original search criteria), were not linked to education, did not include

[3] The current review is an adaptation of a report on the effects of differentiation practices in Early Childhood Education, Primary Education, and early Secondary Education ([Authors], 2015). The original research report had a wider scope than the current review and included studies focusing on students within the age range 2-16 years (i.e. early childhood education to first years of secondary education).

effects on language or math performance, were case studies, or did not use quantitative research methods. In general, all the different ways in which elementary school teachers may take into account student performance differences were considered eligible for this review, but studies on the effects of one-to-one tutoring were excluded, because this educational practice is focused on selected individuals, instead of the entire class. We also excluded studies focusing exclusively on tutoring, although peer tutoring could be part of working in small groups. Applying all these criteria narrowed down the number of references to approximately 90. We collected the full-text papers of this narrowed-down selection.

2.2 Inclusion Criteria

We applied a set of seven final inclusion criteria to the selection of full-text papers. The first criterion focused on the content of the study. This was necessary because we had applied the previous broad selection criteria leniently. Therefore, some irrelevant studies were possibly still in the collection of full papers. The second to seventh criteria focused on the quality of the study. These seven final inclusion criteria were based on those used in the best evidence syntheses conducted by Slavin and colleagues (Slavin, 1987a; Slavin, & Lake, 2008; Slavin et al., 2009).

1. The study addresses effects of cognitive differentiation on language or math performance of all students or groups of students in a classroom (i.e., no studies focusing solely on classrooms for gifted students). The intervention takes place inside the classroom (i.e., no out-of-class tutoring), during the regular school day.
2. The intervention has a minimum duration of 12 weeks. If the duration is not mentioned in the paper, it is measured from beginning of treatment to posttest, or from pretest to posttest.

3. Each treatment group consists of at least 15 students.
4. The study compares students taught in classrooms using an intervention to those in control classrooms using another intervention or standard teaching practice ("business as usual"). Or the study uses secondary data analysis on existing data of large scale survey studies in order to compare groups of classrooms.
5. The study uses random assignment, matching, or uses with appropriate adjustments for any pretest differences (e.g., ANCOVA). Studies without comparison groups are excluded.
6. The study provides pretest data, unless the study uses random assignment of at least 30 units (students, classrooms or schools) and there are no indications of initial inequality.
7. 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. There is sufficient statistical data available in order to calculate effect sizes.

The criteria were applied consecutively: 54 studies did not meet criterion 1 and were disregarded from that point onwards. Over 20 of the remaining studies were rejected on the base of one of the other 6 criteria, or had in hindsight failed to meet the criteria of the first round of selection. Applying all these criteria led to the final selection of 21 studies, from which we selected relevant data to calculate effect sizes. In addition, we coded the studies for content in order to write a short summary of every study. The content coding included: grade, country (and if applicable: state) in which the intervention was conducted, sample size, duration of intervention, dependent variables and instrumentation, and external variables and covariates.

2.3 Computation of Effect Sizes

To be able to compare the effects of the different studies, we converted all research results to Cohen's *d*, which is the standardized mean difference between groups. We recalculated effect sizes for all studies, even when a study already reported effect sizes. In the case of a difference between reported and recalculated *d*, we used the recalculated measure. Methods of calculating *d* using different types of data stemming from various research designs are described in Borenstein, Hedges, Higgins, and Rothstein (2009).

For every study we calculated a general *d*. When multiple outcome measures were used, we labeled these as measures of "math", "vocabulary", "reading", or "reading comprehension", because these labels are more informative than the names of individual tests, which vary between studies. In the appendices, these labels were used in combination with the specific test names. Some studies provide multiple outcome measures of the same cognitive (sub) domain. In these cases, we took all measures together to compute one mean effect size. If possible, we provided differential effect sizes for high-, average-, and low-performing students, using the categorization of the authors of the individual papers.

2.4 Meta-Analysis

Where possible, we combined the results of different studies into one summary effect size (c.f. Borenstein et al., 2009). This was done for studies with the same type of differentiation practice. We conducted the meta-analyses using the CMA software developed by Borenstein et al. (2009). We used a random effects model for the computation of weighted summary effects, and a mixed effects model for moderator analyses for analyzing whether context variables

influenced the effects. For meta-regression analyses, we used the statistical program HLM (Raudenbush, Bryk, Cheong, Congdon, & Du Toit, 2011).

3. Results

3.1 General Results of the Literature Search

We divided the 21 articles thematically into four categories: studies on between-class homogeneous ability grouping (n = 3), studies on within-class homogeneous ability grouping (n = 6), studies using computerized systems as a differentiation tool (n = 6), and studies in which differentiation was part of a broader program of school reform (n = 6). In total 78 effect sizes were extracted from these studies.

3.2 Literature Synthesis

3.2.1 Between-class homogeneous ability grouping.

Three of the studies included in the current review focused on between-class homogeneous ability grouping in primary education (see appendix A). One of these studies considered whole-class homogeneous grouping based on general abilities (tracking; Lefgren, 2004). The other two considered setting: the formation of homogeneous classrooms for specific subjects, in these cases by regrouping students from parallel classrooms (Macqueen, 2012; Whitburn, 2001).

Lefgren's (2004) study on tracking explored the differences between tracked and untracked schools in the reading and mathematics performances of students in grade 3 and 6. The author recognized that the students were probably non-randomly placed within the schools. He therefore investigated the interaction between the tracking policy of the school and the students'

observed initial achievement on reading and math. The overall effects on reading and math performance in both grades were zero. No differential effects were reported.

The two studies on setting compared the performances of students in temporarily regrouped homogeneous classrooms for specific subjects to the performance of students that remained in their regular heterogeneous classroom all the time. Macqueen (2012) focused on setting for literacy and mathematics. Between-class homogeneous ability grouping was done by reassigning students from parallel classrooms to homogeneous classrooms. Schools which regrouped made sure that the homogeneous classrooms with low achievers were smaller than the homogeneous classrooms with average- and high-achieving students, indicating a convergent aim of differentiation. The performance gains between grades 3 and 5 for mathematics, literacy, and writing of students in temporarily regrouped homogeneous classrooms were compared with the gain scores of students in regular heterogeneous classrooms. The author reported small but non-significant overall effects of between-class homogeneous ability grouping on literacy, writing, and math performance (literacy: d = +0.196; writing: d = -0.082, math: d = -0.125). Analysis of differential effects for high-, average-, and low-performing students did not show any significant effects either.

Whitburn (2001) investigated the effects of between-class homogeneous ability grouping for mathematics, compared with mathematics instruction in students' regular heterogeneous classrooms. Between-class grouping was done by reassigning students from parallel classrooms based on their mathematics level to homogeneous classrooms for mathematics lessons. Students in both conditions were taught using the same interactive, whole-class teaching method, which was part of a larger intervention study. Mathematical performance in this project was monitored regularly using short written tests of previously taught mathematical topics. These tests were

used to analyze grouping effects on student performance in grades 3 and 4. The article presents the results of three consecutive cohorts of students. In these three cohorts, approximately 200 students were taught mathematics in homogeneously regrouped classrooms, and about 1,000 students were taught mathematics in their regular heterogeneous classrooms. Analyses of the performance of the three cohorts showed small, negative, but non-significant overall effects of between-class homogeneous ability grouping for mathematics (effect sizes ranged between d = -0.248 and d = -0.101). Similar small, negative, and non-significant results were found for students of different ability levels (effect sizes ranged from d = -0.350 to d = -0.050).

Meta-analysis of the effects of between-class homogeneous grouping showed no overall effect on students' academic performance. Subgroup analysis revealed a significant negative effect for low-ability students (Table 1). However, the confidence intervals for the effect sizes d for the three ability groups overlapped, indicating an absence of significant divergent or convergent differential effects ($Q_{between}$ = 1.189; df = 2; p = 0.552).

Table 1
Meta-analyses. General and Differential Effects of Between-class Homogeneous Ability Grouping

Included papers	*Effect sizes (d)*	*95% CI*
Lefgren, 2004;	*Overall*	
Macqueen,[4] 2012;	-0.065	-0.169; +0.038
Whitburn, 2001	*Low ability*	
	-0.300*	-0.554; -0.046
	Average ability	
	-0.161	-0.402; +0.080
	High ability	
	-0.112	-0.348; +0.123

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

[4] Macqueen compared three different homogeneous ability groups with one regular heterogeneous control group. The variances for using the same comparison group multiple times were corrected. This was done by dividing the number of students in the comparison group by three and then re-computing the variances using the statistical package CMA.

3.2.2 Within-class homogeneous ability grouping.

Six studies evaluated the effects of within-class homogeneous ability grouping (see appendix B). Three of these reported on an intervention (Crijnen, Feehan, & Kellam, 1998; Hunt, 1996; Leonard, 2001): two compared homogeneous grouping with heterogeneous grouping and one made the comparison with whole-class teaching. The other three studies re-analyzed existing data in order to investigate the effects of ability grouping compared with regular classroom teaching (Condron, 2008; Nomi, 2010; Tach & Farkas, 2006).

Leonard (2001) investigated the effects of homogeneous small groups compared with those of heterogeneous small groups on mathematics achievement. The study was conducted over two consecutive years. In the first year, all grade 6 students (cohort 1) were placed in small heterogeneous groups during mathematics instruction. In the following year, all grade 6 students (cohort 2) were placed in small homogeneous ability groups during mathematics instruction. During the school year, students collaborated on thematic mathematical activities. The article did not provide details of the content and form of instruction provided by the teacher. The effects of homogeneous grouping compared with heterogeneous grouping were negative, but non-significant (overall: d = -0.250, low ability: d = -0.397, average ability: d = -0.133, high ability: d = -0.185). Based on qualitative analyses of students' group interactions, the author of the study concluded that how the group collaborated may have been more important for determining achievement than grouping based on ability level.

Hunt (1996) also investigated the effects of using homogeneous small groups on mathematics achievement, which she compared with the use of heterogeneous small groups. Although the main focus of the study was the effect of grouping on gifted students, the effects on

average and low-ability students were taken into account as well. More than 200 6th graders were randomly assigned to classrooms in which either homogeneous or heterogeneous grouping was used. The group of gifted students consisted of both students who had been identified as such by the state (n = 15) and students who had scored high on a pretest (n = 17). The study revealed positive but non-significant effects on math achievement for homogeneous grouping (gifted students identified by the state: d = +1.061; other gifted students: d = +0.183; students with average ability: d = +0.137; students with low ability: d = +0.013).

The third intervention study examined the effects of within-class homogeneous ability grouping through comparison with regular whole-class teaching. Crijnen and colleagues (1998) evaluated the effects of a mastery learning intervention for reading in grade 1, and its effects throughout elementary school. The study was conducted in schools in which at least one classroom received the intervention and one classroom did not. Differentiation was applied by providing extra learning time and individual help to (groups of) students who needed it. In addition, the classroom as a whole would only continue to the next learning unit when 80% of the students had mastered 80–85% of the learning goals, implying a convergent goal of differentiation. It was found that students in the intervention condition more often showed average expected (or even greater) growth in test scores over the course of a year than students in the control classrooms (d = +0.138), but this effect was not significant. No long term effects (up to grade 5) were found.

The next three studies (Condron, 2008; Nomi, 2010; Tach & Farkas, 2006) analyzed the effects of within-class homogeneous ability grouping using the publicly available ECLS-K database. The ECLS-K database is part of the Early Childhood Longitudinal Study (ECLS) conducted in the United States by the Institute of Education Sciences and the National Center for

Education Statistics. Its aim is to investigate the development, school readiness, and school experiences of three large cohorts of children. The ECLS-K database consist of data from a cohort of children followed from kindergarten (entry in 1998–1999) to grade 8. A wide range of child-assessments was used in the ECLS-K: reading, mathematics, general knowledge, social-emotional, and physical development. In the ECLS-K dataset, teachers provided some information about their grouping procedures: for example, whether and how frequently they used homogeneous ability grouping. The three studies selected for this review all assessed the effects of within-class homogeneous ability grouping on students' reading performance.

Condron (2008) followed student reading performance from kindergarten to grade 1 and from grade 1 to 3. Using a propensity score matching technique, the author compared the scores of students in low-, average-, and high-level reading groups with the scores of non-grouped students with a similar likelihood of being placed in one of these groups. Placement in a high-ability group led to significantly higher gains in reading performance (grade 1: d = +0.207; grade 3: d = +0.177). Placement in a low-ability group had a significant negative effect on reading performance (grade 1: d = -0.288; grade 3: d = -0.245). Placement in an average-level reading group did not have significant effects on reading performance (grade 1: d = -0.043; grade 3: d = +0.046).

Nomi (2010) used propensity score matching to analyze the effects of school grouping policy on the reading scores of almost 9,000 students. The author noticed that schools using within-class homogeneous ability grouping generally served a relatively heterogeneous student population. The study rendered no evidence for advantages of within-class homogeneous ability grouping over whole-class instruction: a negative, very small and non-significant effect was found (d = -0.010). The effects for the various ability groups were also examined; all effects

were very small and non-significant (low ability: d = -0.030, average ability: d = +0.021, high ability: d = -0.059).

Tach and Farkas (2006) used multilevel modeling to estimate the effects of teaching homogeneous small groups. Prior reading performance and other student characteristics (math performance, sex, ethnicity, and SES) where taken into account as background variables in the models. They found that the use of homogeneous ability groups in the classroom had a significant overall negative effect on students' reading performance (d = -0.191). No differential effects were reported.

Because Condron (2008), Nomi (2010), and Tach and Farkas (2006) used the same ECLS-K dataset, we treated the three studies as one study with multiple outcome measures in the meta-analysis. When we summarized the effects over all six studies (Table 2), within-class homogeneous ability grouping appeared to have no overall effect on students' performance. Subgroups analysis revealed significant differential effects between students with different ability levels: within-class homogeneous ability grouping had a significant negative effect on the performance of low-ability students, and small but non-significant effects on the performance of students with average or high ability levels. The effect sizes for the three ability groups differed significantly from each other ($Q_{between}$ = 12.511; df = 2; p = 0.002), which indicates a divergent effect of using small homogeneous groups within the classroom.

Table 2
Meta-analyses. General and Differential Effects of Within-class Homogeneous Ability Grouping

Included papers	*Effect sizes (d)*	*95% CI*
Crijnen et al., 1998;	*Overall*	
ECLS-K studies (Condron,	-0.007	-0.146; +0.132
2008; Nomi, 2010; Tach &	*Low ability*	
Farkas, 2006);	-0.192*	-0.310; -0.074
Hunt, 1996;	*Average ability*	
Leonard, 2001	+0.006	-0.049; +0.061
	High ability	
	+0.103	-0.023; +0.229

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

3.2.3 Computerized systems as a differentiation tool.

The third category of studies concerned differentiation practices supported by computer systems. Computer programs may be used to collect information about students' performance level, which teachers can use for making grouping decisions. Computer programs may also provide teachers with suggestions about which type of instruction or content is most suitable for students with different needs. Connor and colleagues and Ysseldyke and colleagues investigated the use of such computer technology for supporting differentiation practices. An overview of these studies can be found in appendix C.

Connor and colleagues (Connor et al., 2011a; Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Connor et al., 2011b) published several articles on the effects of individualizing student instruction (ISI) using a special type of software (A2i, Assessment-to-Instruction). The ISI intervention was designed to support teachers in their efforts to provide optimal reading instruction for students of all levels. The computerized system advised the teacher about the amount of teacher- and/or student-managed instruction suitable for a specific student, based on prior performance. Low-ability students received more attention than high-ability students, suggesting a convergent aim of the intervention. Additionally, the program

provided teachers with suggestions about the content of the instruction, helping teachers to offer more code- or meaning-oriented instruction and tasks to small homogeneous groups of students.

Connor and colleagues (2007) investigated the effects of the ISI intervention on reading performance in grade 1. Teachers in the ISI condition received a professional development course on the use of differentiated reading instruction. Teachers in the matched control group did not receive any professional development course, nor did they use the computer program A2i. The intervention was found to have a small but significant positive effect on students' reading achievement (d = +0.183). Although this result is likely to have been affected by the professional development course, the authors reported that the students' improvement in reading was related to the amount of time teachers spent using the A2i software in the classroom. In their view, this suggested that implementation of the computer program in itself was at least partly related to the students' reading outcomes.

A few years later, Connor and colleagues replicated their study (Connor et al., 2011b) and again investigated the effectiveness of the ISI intervention on first-grade students' word-reading skills compared with a "business as usual" control group. The teachers in the experimental group used the suggestions of the computer program A2i to form ability groups and to select the appropriate content of their instruction. They were supported by professional development courses 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, nor did they receive any professional development on differentiated instruction. Classroom observations showed that teachers in the ISI condition were better able to fit the content of instruction to the needs of the students than teachers in the control condition, and that matching the instruction to the recommendations of the computerized algorithm strongly predicted students' reading

outcomes. Multilevel analyses showed that the ISI intervention had a significant positive effect (d = +0.249) on students' word-reading scores. The authors argued that the effectiveness of the intervention had increased since 2007 due to improvements in the computer program, which was now more user-friendly, and due to the improvement of the professional development program for teachers.

The third study on the effectiveness of the ISI intervention focused on its effects on student performance in grade 3 (Connor et al., 2011a). The effects of ISI were compared with those of an alternative vocabulary intervention. In the ISI condition, teachers again used the A2i software and received professional training. In the control condition, teachers received more general training in how to provide better vocabulary instruction. Classroom observations during the school year showed that teachers in both conditions were similar in the amount of individualized instruction they provided, in their organization and planning activities, in their use of strategies, and in their classroom-management styles. Multilevel analyses of student results showed that the ISI intervention had a small significant positive effect on reading comprehension (d = +0.191) compared with the general vocabulary intervention.

Ysseldyke and colleagues (Ysseldyke, & Bolt, 2007; Ysseldyke et al., 2003; Ysseldyke, Tardrew, Betts, Thill, & Hannigan, 2004) used a computer program called "Accelerated Math" (AM) to support differentiated mathematics instruction[5]. In the AM program, students were provided with computer-adaptive math tests. Based on test performance, the computer program generated individual level-appropriate mathematics exercises. After completing their exercises, students scanned their work and the computer provided them with immediate feedback. Then the

[5] Studies on the related program "Accelerated Reading" (e.g. Nunnery, Ross & McDonald, 2006) were not found by applying the search criteria in the current systematic review.

computer offered students new exercises based on their performance, indicating a divergent goal. The program provided teachers with information about students' progress, which teachers could use to adapt their instruction to students' needs.

The effects of AM on students' performance were evaluated in the study by Ysseldyke and colleagues (2003). They investigated the effects of using the program in math lessons on grade 3, 4, and 5 student test results. Teachers from 18 classrooms in four schools (almost 400 students) volunteered to use the computer program during mathematics instruction; of these, teachers from 10 classrooms fully implemented the program. Scores of students from the classrooms in which teachers fully implemented AM were compared with scores of a control group of students from other classrooms within these schools.[6] Within schools, significant small to medium positive effects of fully implementing the AM program were found, compared with the control group (d = +0.189 and d = +0.268).

In a following study, Ysseldyke and Bolt (2007) investigated the effect of AM on students' math performance in elementary and secondary schools. After volunteering to participate in the study, teachers from seven elementary schools were randomly assigned to three groups: an experimental group using the AM program throughout the year (41 classrooms), an experimental group using the AM program from midway through the school year and onwards (20 classrooms), and a control group not using the program (39 classrooms). Students in the experimental classrooms in which AM was fully implemented scored significantly higher than students in control classrooms (AM full year: d = +0.491; AM half year: d = +0.324).

[6] The performance of students in classrooms where AM was fully implemented were also compared with those of a random group of students from the district's testing database, but because this is a less optimal way of forming a control group, these results were not used in the current systematic review.

Ysseldyke and colleagues (2004) also looked into the usefulness of the AM computer program for differentiation aimed at gifted students in regular classrooms in grades 3 to 6. The teachers in this study used the AM program in their classrooms for about four months. In the experimental classrooms, gifted as well as non-gifted students worked on the exercises from the AM program regularly. In the control classrooms, neither gifted nor non-gifted students had access to the program. Gifted students inthe experimental classrooms scored significantly higher than gifted students from control classrooms (d = +0.456). Similar results were found for the other students in the classroom: the non-gifted students from AM classrooms scored significantly higher than non-gifted students in control classrooms (d = +0.369).

A meta-analysis of the effects of the two computer-based differentiation interventions showed that they positively affect student performance. There was a significant small to medium overall effect of the six studies on computer-based interventions (d = +0.290; 95% CI [0.206, 0.373]). This result indicates that a blended learning approach to differentiation in which both analyzing students' progress and selecting appropriate instruction practices and content are addressed, is beneficial to students' performance. It was not possible to perform a subgroup analysis of the differential effects for students of various ability levels, because, except for Ysseldyke and colleagues (2004), none of the studies contained data for subgroups of students.

3.2.4 Differentiation as part of a broader program or school reform.

The fourth category of articles focused on differentiation in the context of a broader program or reform. Embedding differentiation in a supportive context can be a good way of helping teachers applying differentiation and thereby ensuring implementation fidelity. Six

studies on differentiation as part of a broader program were included in the current review (see appendix D).

The first article (Borman et al., 2007) focused on differentiation for reading as part of the program "Success for All" (SfA). During reading instruction, students were regrouped between classrooms and across grades, based on their performance level. Student performance was assessed every nine weeks and students were regrouped if necessary. One-to-one-tutoring was available for students who needed additional help. The combination of across grade ability grouping and optional tutoring indicates that SfA had both a divergent and convergent aim. The study, in which students from 35 schools were monitored from kindergarten to grade 2, used a cluster randomized controlled design. The final literacy outcomes of the students in schools using SfA were compared with the outcomes of students in control schools. Results showed that students in intervention schools scored significantly higher on the three literacy measures than students in control schools (d = +0.220, d = +0.330, d = +0.210).

Success for All was also part of the study by Reis and colleagues (2007). They evaluated the effects of a comprehensive reading intervention (School-wide Enrichment Model in Reading Framework, SEM-R) combined with SfA. The article discussed the effects of SEM-R in two elementary schools serving a culturally diverse, high-poverty population. Both schools used SfA in the morning and implemented a one-hour reading program every afternoon. Half of the teachers were randomly assigned to the experimental group, in which SEM-R was used as the afternoon reading program. The other half of the teachers formed the control group, in which the state-mandated reading program based on whole-group instruction was used in the afternoons. In the SEM-R condition, teachers first read aloud and used higher order questioning and thinking-skills instruction. Afterwards, students were encouraged to select challenging books, somewhat

above their current reading level, for individual reading. During this phase, teachers gave individualized support and differentiated instruction about reading strategies, from vocabulary use with lower level readers to information synthesis with advanced readers. In the third phase, students could choose different literacy-related activities of varying complexity. Due to the phase of differentiated instruction, and the offering of books and activities suitable for students with different performance levels, we consider SEM-R as a program that focusses on cognitive differentiation. Teachers in the experimental group received a one-day training in SEM-R. Coaching and support were available to all teachers, both in the experimental and the control condition, during the 12-week intervention period. The results showed a significant positive effect of SEM-R on reading fluency (d = +0.299), but no significant effects on reading comprehension (d = +0.220).

Reis, McCoach, Little, Muller, and Kaniskan (2011) continued the investigation the effect of SEM-R, this time in schools that did not use Success for All. Their study was set up as a cluster randomized experiment, in which teachers were randomly assigned to a control or treatment condition. In both conditions, teachers gave 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. The teachers assigned to the experimental condition used the same program for the first hour and SEM-R during the second hour. The results showed that students in both the control and the experimental group improved their performance. The overall effect of SEM-R compared with the regular program was positive, but non-significant (reading fluency: d = +0.254, reading comprehension: d = +0.145).

Stevens and Slavin (1995) investigated differentiation as part of a program focusing on cooperative learning. The achievements of students in grades 2 to 6 in two elementary schools using cooperative learning were compared with those of comparable students in three control schools. The experimental schools had the following features: they used cooperative learning and peer coaching across a variety of content areas, teachers planned cooperatively, academically handicapped students were mainstreamed full-scale, and parent involvement in school was stimulated. In addition, teachers in these schools were trained to use two comprehensive programs designed to accommodate student diversity: CIRC (Cooperative Integrated Reading and Composition) and TAI (Team Assisted Individualization-Mathematics). Students worked in heterogeneous learning teams in both programs, but received instruction in relatively homogeneous teaching groups. Students lagging behind received additional instruction, indicating a convergent aim of differentiation. In sum, the experimental schools implemented a very broad reform in which working in heterogeneous and homogeneous groups was an important part of the day-to-day program. To investigate the effects of the reform, student achievement in reading, language, and mathematics was assessed. After two years, students in the cooperative schools scored significantly higher on measures of vocabulary (d = +0.210), reading comprehension (d = +0.280), language expression (d = +0.210), and math computation (d = +0.290).

Another intervention in which differentiation was part of a broader reading program was described by Houtveen and van de Grift (2012). They conducted a quasi-experimental study on the effects of the “Reading Acceleration Programme” (RAP), which aimed at reducing the percentage of struggling readers in grade 1. 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 provide special help to students who did not respond sufficiently to the intervention (tier 3). The aim of tiers 2 and 3 was to allow struggling readers to participate successfully in whole-group instruction, which implies that RAP was aimed at convergent differentiation. Students in the control group received instruction in the same way as they always had. After the pre-test data (age, intelligence, socioeconomic status, and ethnic minority status) were corrected for, a significant difference in reading performance was found in favor of students in the experimental schools (Decoding skills: d = +0.280, reading fluency: d = +0.620).

The last study on differentiation as part of a broader reform was conducted by Sterbinsky, Ross, and Redfield (2006). They investigated the effects of four types of school reform on reading performance. Although differentiation (in the form of within-class homogeneous grouping) was only explicitly part of two of the four reforms (namely, Success for All and Direct Instruction), the observations made by the researchers showed that differentiated instruction was applied in all intervention conditions. Furthermore, ability grouping appeared to be used more often by the experimental schools than by the control schools. The results show that after three years students in schools applying one of the reforms scored significantly higher on various reading measures (d ranged from +0.286 to +0.429) than students in control schools. The four types of reform were not compared due to the small numbers of schools in each program.

A meta-analysis of the included studies of differentiation as part of a broader school reform showed a significant positive effect on students' academic performance. The summary effect was d = +0.296 (95% CI [0.197, 0.395]). Because none of the studies in this category published results for students of different ability levels, differential effects could not be calculated.

3.3 Overall Results

The 21 studies selected for this review were categorized by the type of context which can facilitate the implementation of differentiated instruction. The meta-analyses showed that some types of contexts had larger summary effects than others (Table 3). Studies on differentiation aided by computerized systems and differentiation which was part of a broader school reform program had on average significant small to moderate positive effects on students' cognitive outcomes. In contrast, studies on differentiation which was comprised solely of between-class or within-class homogeneous ability grouping did not show any significant effects. Moderator analysis, which is used to see whether the different contexts lead to different effects on student performance, showed that the differences between the effects of the four types of contexts were significant ($Q_{between}$ = 40.068; df = 3; $p < 0.001$).

Table 3
Meta-analyses. General Effects of Contexts for Differentiation Practices

Category	*Effect sizes (d)*	*95% CI*
Between-class grouping	-0.065	-0.169; +0.038
Within-class grouping	-0.007	-0.146; +0.132
Computer system	+0.290*	+0.206; +0.373
Broader Program	+0.296*	+0.197; +0.395

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

Figure 1 provides a forest plot with an overview of the average effect size of each individual study (depicted with squares). The summary effect is also reported (depicted with a diamond). The summary effect shows that, overall, differentiation practices in primary education have a small significant positive effect on students' academic performance (d = +0.146; 95% CI [0.066, 0.226]). Subgroup analysis could only be conducted on the six studies that reported subgroup data, which all concerned between-class or within-class grouping. The findings reveal a small significant negative effect of differentiation for low-ability students (d = -0.195, 95% CI

[-0.264, -0.126]), but no significant effects for the other ability groups (average ability: d = -0.001, 95% CI [-0.060, 0.058]; high ability: d = +0.018, 95% CI [-0.131, 0.168]). The differences between the ability groups are significant ($Q_{between}$ = 19.129; df = 2, p < 0.001).

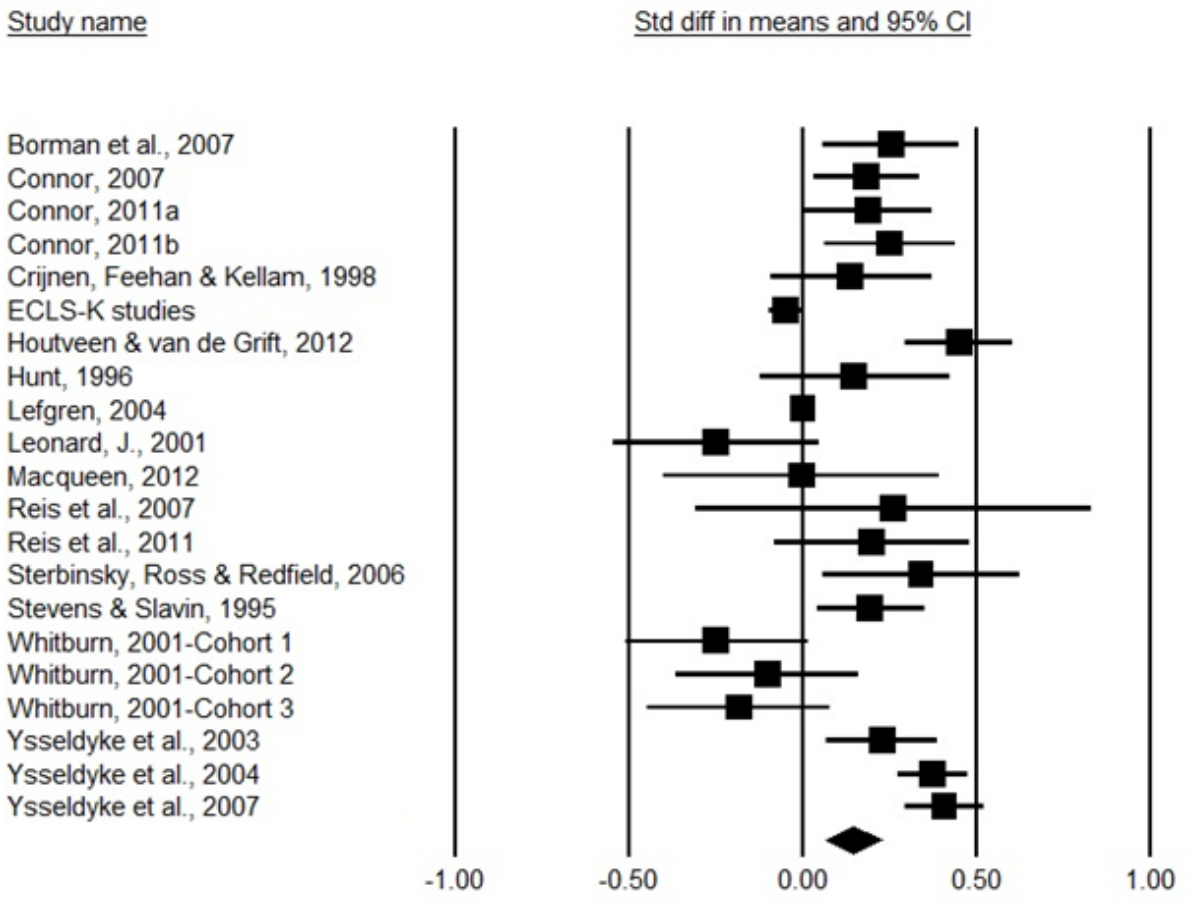

Figure 1. Forest plot for the included studies. The squares represent the average effects of the individual studies and the diamond the summary effect. The lines around the squares and the diamond represent the confidence interval.

3.4 Reflection on the Included Studies

There is a possibility that our findings are influenced by bias. Although the initial literature search resulted in around 1,430 references, the rigorous methodological inclusion criteria ruled out the majority of these. We acknowledge that many of the excluded references may have been valuable from a conceptual, theoretical, or practical point of view, providing, for

example, rich qualitative descriptions of differentiation practices and their outcomes. However, the strict inclusion criteria fitted the aim of this review: to investigate the effects of differentiation practices on students' cognitive outcomes. This type of bias was thus intentionally applied.

There may be an unintended second source of bias: hypothetically eligible studies with non-significant ('disappointing') results may not have been published at all. The possible effects of this type of publication bias are that (a) studies lacking statistical power as a result of a small sample size are only published if they produce large effects that counterbalance the large standard errors, and (b) smaller effects are only reveiled by studies witj considerable statistical power, resulting from large sample sizes with consequently small standard errors. These two mechanisms lead to a bias in the distribution of reported effect sizes, as a function of an increasing standard error. To explore the prevalence of this bias, we created a funnel plot (Figure 2). The vertical line in the middle represents the average effect in a meta-analysis using a random effects model. We used Duval and Tweedie's trim and fill method for a random effects model (Borenstein et al., 2009; Peters, Sutton, Jones, Abrams, & Rushton, 2008) to check whether studies were missing due to publication bias. The results show that the effect sizes in individual studies are evenly distributed to the left and the right of the vertical line, indicating that there are no missing studies. The white diamond at the bottom shows the general summary effect, and the black diamond shows the summary effect after correction for publication bias. Because no publication bias was detected, both effects are the same.

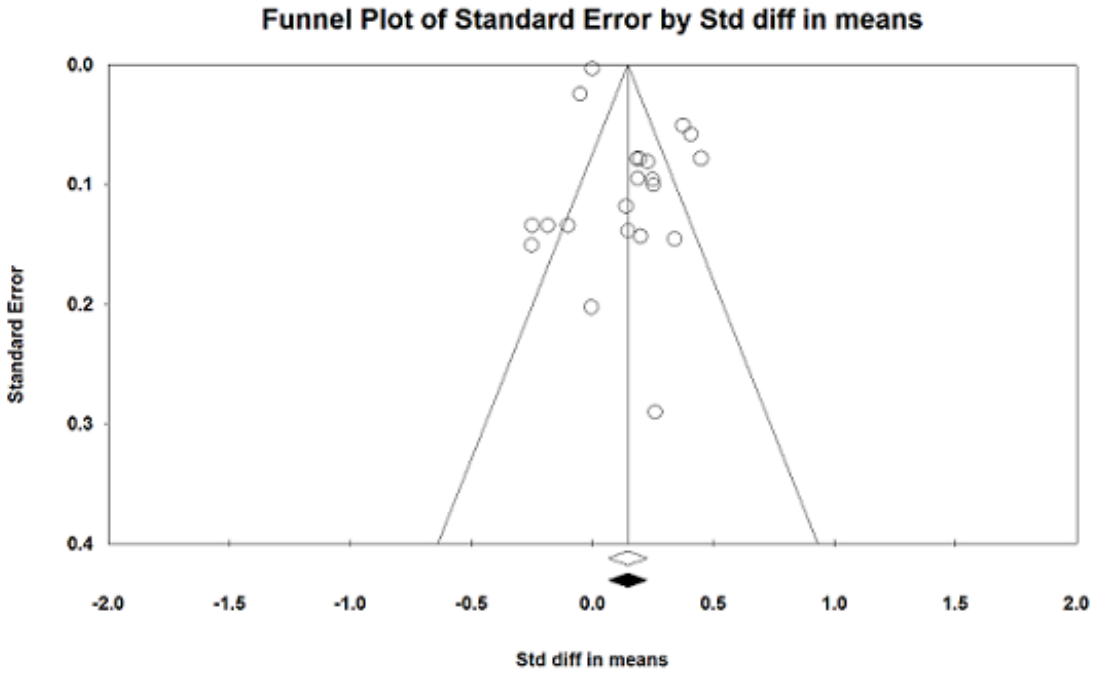

Figure 2. Funnel plot to check for publication bias in the included studies.

4. Conclusion and discussion

The importance of dealing with cognitive differences of students by applying differentiation practices which are knowledge- and learner centered (Tomlinson et al., 2003), is currently greatly emphasized by educationalists. Partly due to the fuzziness of the construct, the effectiveness of differentiation is unclear. Previous (meta-)meta-analyses on differentiation practices were mainly focused on different forms of grouping: between-class or within-class, full-time or only for specific subjects, whole group or small group, homogeneous or heterogeneous. The overall conclusion that can be drawn from these previous studies is that grouping can create a context for differentiated instruction, but that it should be ensured that this differentiated instruction is indeed offered. Although this precondition has been emphasized by previous (e.g. Kulik, 1992; Lou et al., 1996; Slavin, 1987a) and current researchers (e.g. Roy, Guay & Valois, 2013), apparently it is still a relevant point to make. A second important

conclusion that can be drawn from the precious studies is that the differential effects of differentiation are still inconclusive. The aim of the current review was to extend knowledge of the effects of differentiation practices in primary education.

The 21 studies included in this review can be divided into four types: (a) studies on the effects of between-class homogeneous ability grouping, (b) studies on the effects of within-class ability grouping, (c) studies on differentiation practices supported by computer systems, and (d) studies in which differentiation was part of a broader program or school reform.

In general, we found that differentiation had a small overall positive effect on students' academic performance (d = +0.146), especially when the practice was embedded in a supportive context: either a computer-assisted environment (d = +0.290) or a broader school reform (d = +0.296). We did not find a significant overall effect for between- or within-class homogeneous grouping. This supports the conclusion of the prior reviews that grouping alone is not enough and should be accompanied by differentiated teaching practices. However, the overall positive effect does not necessarily mean that students of all ability levels benefit from differentiation practices. Differential effects could only be calculated for between- and within-class homogeneous grouping. These types of differentiation practices appeared to have a small negative effect for low-achieving students (d = -0.195) and no significant effects for average- and high-ability students. This discouraging result is not in line with the meta-meta-analysis of Steenbergen-Hu and colleagues (2016), although comparability is limited, because this study takes into account secondary education as well.

A possible reason for the absence of significant effects of between- and within-class homogeneous ability grouping is that although the teachers in these studies reported to use grouping, they may not have used grouping to provide differentiated instruction. Because

detailed procedural information was not given, how the instruction was tailored to students' needs remained unclear in most of the studies. This may also indicate that teachers were not supported in effectively using their grouping to improve differentiated instruction. The findings that differentiation was more effective when it was embedded in a broader context, like a computerized environment or a school reform, supports this suggestion. These computerized environments or more general reforms are more likely to include teacher professional development, which help to ensure implementation and to improve quality of teaching (Timperley. Wilson, Barrar, & Fung, 2007).

The contribution of the current review to existing knowledge of the effects of differentiation in primary education on students' performance is twofold. It gives an updated overview of the overall effects of all experimental and correlational studies conducted in this area since 1995, including information on the possibilities of computer technology as a tool for differentiation, which is an interesting addition to the previous literature syntheses. Furthermore, in the current review we examined the characteristics of effective differentiation practices by conducting a moderator analysis, in order to see how different contexts for differentiation render different effects.

4.1 Limitations

Systematic reviewing is a technique to thoroughly examine all empirical evidence on a certain topic. The operationalization of the topic of interest in a set of search terms is therefore essential. We decided to define two sets of search terms. The first set comprised general ways of describing differentiation. In order to capture studies that described differentiation practices under a different name, we selected an additional set of terms with more specific terms for

differentiation practices or contexts for differentiated instruction. These terms were selected on the basis of previous research, but the list is not exhaustive. Our decision to add this second set of search terms enriched the search, but also directed it in a specific way, leading to the inclusion of studies on for example grouping, while studies in which differentiation contexts or practices went under yet a different name possibly remained undetected.

One of the main critiques of review studies and meta-analyses is that they try to compare incomparable elements. A solution to this problem is to try to capture differences between studies as variables and to control for them, but this requires a large number of individual studies: approximately ten studies per variable (Borenstein et al., 2009). Even when some variables are controlled for, studies need to be combined in order to perform a meta-analysis, and this inevitably leads to some loss of detail in individual studies and forces researchers to make relatively rough categorizations. This happened, for example, in the current review study when we described the category of studies on between-class homogeneous grouping: studies on tracking and setting were taken together, even though these are very different types of grouping. Similarly, in our analysis of the studies on within-class homogeneous ability grouping, studies comparing homogeneous grouping with heterogeneous grouping and studies comparing homogeneous grouping with whole-class teaching were combined into one category and taken together in one meta-analysis, even though it was known from earlier studies that this type of comparison is likely to influence the findings (Lou et al., 1996). Meta-analyses are thus inherently less fine-grained than is desirable. Combining a meta-analysis with relatively extended descriptions of the included studies is a way to mitigate this drawback. For this reason, we have provided summaries of all 21 studies included in this review in the results section.

Another drawback of the current review is the limited number of studies included. The very strict inclusion criteria meant that only very few were considered eligible: this also limited the statistical possibilities. Approximately 1,430 studies on differentiation were found in the initial literature search, but only1.5% of these met the final inclusion criteria used. The rigorous inclusion criteria follow directly from the decision to focus on the effects of differentiation practices on the cognitive outcomes of students, and not on other student or teacher outcomes. Studies on differentiation that were focused on different types of outcomes and used different research designs may have offered interesting insights into differentiation as an educational practice. Future reviews which include these types of studies would, therefore, be valuable.

Furthermore, the inclusion criteria ensured that studies met certain methodological requirements. This enabled us to avoid the "garbage-in-garbage-out" effect and to only synthesize the best evidence available (Slavin, 1995). Because of our methodological requirements we decided to only search for research described in peer-reviewed journals, although this meant that possibly relevant research from dissertations, technical reports, conference proceedings and unpublished monographs remained undetected. These decisions may have led to bias, but the relation between the effect sizes and the statistical power of the included studies did not show any indication of this.

Another limitation is related to the category of studies in which differentiation practices were part of a broader reform. Analysis of the results of this type of study complex because effects cannot be attributed solely to one specific element. However, excluding such studies might lead to an underestimation of possible effects of differentiation practices. Furthermore, by excluding these studies an important message for educational practice would be lost, namely, the importance of creating an extensive supportive context when implementing an educational

innovation. Therefore, despite the difficulties of assigning possible effects of a broader program specifically to differentiation practices as one of its components, we considered it important to include this type of study in the review.

It would also be useful to investigate the influence of the theoretical goal of differentiation. Differentiation practices can have different goals ([Author], 2005; Denessen, 2017). At least theoretically, differentiation is focused either on keeping the lower-performing students on track and reducing the variation in performance within the classroom (a convergent aim), or on helping all students to proceed as well and as fast as they can, thereby increasing the differences between low- and high-performing students (a divergent aim). Taking into account the aim of differentiation in every study when analyzing its effects would have been interesting. However, many studies do not mention whether the differentiation practice described had a convergent or a divergent aim, and inferring this is often hard because of a lack of detail in the description of the intervention. Furthermore, the theoretical distinction between convergent and divergent aims might be diffuse in practice (Denessen, 2017). Another solution for taking into account the fundamental difference between convergent and divergent differentiation is to analyze the effect sizes of the differential results. When *d* is largest for the high-performing students, the differentiation practice resulted in divergence. When *d* is largest for the low-performing students, the differentiation practice resulted in convergence. However, differential effects could be calculated for only two of the four contexts for differentiation. Therefore, the question how the aim of differentiation influences student performance cannot be answered based on the current systematic review study.

Yet another important aspect of differentiation which was not covered systematically in the current review is the amount of instruction that students receive. The more ability groups a

teacher creates, the less time there is for each group, and the more time students have to spend working independently. The formation of small groups in combination with adapted instruction may thus be effective, but it is likely that there is an optimum number of groups or amount of time spent in groups, and an optimum amount of instruction time that should be provided. Hong and colleagues support this suggestion. They report that "intensive grouping" in combination with low instruction time has negative effects, especially for low-performing students (Hong & Hong, 2009; Hong, Corter, Hong, & Pelletier, 2012; also see van de Pol, Volman, Oort & Beishuizen. 2015). The use of three ability groups in combination with some whole-class instruction seems to be most common in everyday practice in primary education, but whether this is more effective than, for example, having two or four ability groups remains unclear. Although this is a very relevant question for practitioners, the current review does not shed light on this matter.

4.2 Recommendations for Future Research and Practice

To understand the effects of differentiation, it is important to use an ecologically valid operationalization, which is inherently somewhat fuzzy, because differentiation is an educational approach in which multiple practices are combined. As has been made clear in prior reviews and is confirmed by the current one, differentiation is more than ability grouping, and ability grouping should be more than physically 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 (Tomlinson, 2014). Successful application of differentiation practices assumes two things: (a) teachers need to have an accurate view of students' level of understanding, informed by data, and (b) teachers need to know which

instruction and learning activities are appropriate for students of different ability levels, given their goals. However, the decision on how to adapt instruction might also be influenced by external factors like the amount of preparation it requires (Roy, Guay, & Valois, 2012; Tomlinson et al., 2003). Therefore, differentiation is best applied within a supportive environment for teachers. In general, such a supportive environment could be created by organizing teacher collaboration in which experiences and expertise can be shared (Vangrieken, Dochy, Raes, & Kyndt, 2015), facilitated by a school leader with a strong focus on educational leadership (Hubbard, Datnow, & Pruyn, 2014; Leithwood, Harris, & Hopkins, 2008). Furthermore, the current findings suggest that the formation of such supportive environments can be stimulated by implementing adaptive computer programs or broader comprehensive programs.

Software can be used to take care of some of the assessment and diagnosing, and may provide suggestions for tailored instruction, content, or materials. However, it is still the teacher who implements the differentiation practices. As with homogeneous ability grouping, using differentiation software is not a guarantee for actual differentiation in the classroom. Coaching and support is needed to help teachers implement differentiation and to ensure differentiation goes beyond grouping. Research on how computerized systems influence teaching practices may further our understanding of how to use software as an effective teaching tool.

Another promising route for differentiation is to embed it in a broader structure, in which educational practices like cooperative learning, regular assessment, remedial instruction, and flexible grouping are combined. As noted above, investigating the effects of differentiation within such a broader structure is complicated, because the components intertwine. Nevertheless, it is important to further investigate the effects of differentiation practices when they are

combined with other support systems in order to determine how differentiation practices can be embedded within the classroom and the school.

Finally, referring back to the claim that it is important to use an ecologically valid operationalization of differentiation, future researchers who conduct effect studies on differentiation should make sure to include enough information on the actual differentiation practices used (Janssen, Westbroek & Doyle, 2015). Information on implementation fidelity is crucial, as well as information on the intervention itself: it should not only be clear whether teachers implemented differentiation as intended, but also what the (intended) differentiation practices exactly entailed. When it is unclear how differentiation took form within the classroom and what teachers actually did in order to differentiate, its effects in terms of higher test performance are difficult to interpret and of less theoretical and practical value. The findings of such studies are therefore of less relevance to the applied goal of the educational sciences of making sure that all students receive a suitable education in order to fulfill their full potential.

References

Anderson, K. M., & Algozzine, B. (2007). Tips for teaching: Differentiating instruction to include all students. *Preventing School Failure, 51*(3), 49–54. doi:10.3200/PSFL.51.3.49-54

Belfi, B., Goos, M., De Fraine, B. & Van Damme, J. (2012). The effect of class composition by gender and ability on secondary school students' school well-being and academic self-concept: A literature review. *Educational Research Review, 7*, 62–74. doi:10.1016/j.edurev.2011.09.002

Blok, H. (2004). Adaptief onderwijs: Betekenis en effectiviteit [Adaptive education: Meaning and effectivity]. *Pedagogische Studiën, 81*, 5–27.

Bloom, B. S. (1971). Learning for mastery. In: B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. doi:10.1002/9780470743386

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal, 42*, 673–696. doi:10.3102/00028312042004673

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

Author. (2005). Blinded for review.

Burris, C. C., Heubert, J. P., & Levin, H. M. (2006). Accelerating mathematics achievement using heterogeneous grouping. *American Educational Research Journal, 43,* 105–136. doi:10.3102/00028312043001105

Chorzempa, B. F., & Graham, S. (2006). Primary-grade teachers' use of within-class ability grouping in reading. *Journal of Educational Psychology, 98,* 529–541. doi:10.1037/0022-0663.98.3.529

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

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

*Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S.,... Schatschneider, C. (2011a). Testing the impact of child characteristics x instruction

interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly, 46*, 189–221. doi:10.1598/RRQ.46.3.1

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

*Crijnen, A. A. M., Feehan, M., & Kellam, S. G. (1998). The course and malleability of reading achievement in elementary school: The application of growth curve modeling in the evaluation of a mastery learning intervention. *Learning and Individual Differences, 10*, 137–157. doi:10.1016/S1041-6080(99)80138-1

de Koning, P. (1973). *Interne differentiatie* [Internal differentiation]. Amsterdam, the Netherlands: APS/RITP.

Denesse, E. (2017). *Verantwoord omgaan met verschillen:sociale-culturele achtergronden en differentiatie in het onderwijs* [Dealing with differences soundly: social-cultural background and differentiation in education]. Inaugural speech. Retrieved from https://openaccess.leidenuniv.nl/handle/1887/51574

Authors. (2015). Blinded for review.

Gamoran, A., & Weinstein, M. (1998). Differentiation and opportunity in restructured schools. *American Journal of Education, 106*, 385–415. doi:10.1086/444189

George, P. S. (2005). A rationale for differentiating instruction in the regular classroom. *Theory into Practice, 44*, 185–193. doi:10.1207/s15430421tip4403_2

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*, 949–967. doi:10.1111/j.1467-8624.2005.00889.x

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K. & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review, 17*, 50–62. doi: 10.1016/j.edurev.2015.12.002.

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*, 69–88. doi:10.3102/0162373711424206

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*, 54–81. doi:10.3102/0162373708328259

*Houtveen, T., & van de Grift, W. (2012). Improving reading achievements of struggling learners. *School Effectiveness and School Improvement, 23*, 71–93. doi:10.1080/09243453.2011.600534

Hubbard, L., Datnow, A., & Pruyn, L. (2014). Multiple initiatives, multiple challenges: The promise and pitfalls of implementing data. *Studies in Educational Evaluation, 42,* 54-62. doi.org/10.1016/j.stueduc.2013.10.003

*Hunt, B. (1996). The effect on mathematics achievement and attitude of homogeneous and heterogeneous grouping of gifted sixth-grade students. *Journal of Secondary Gifted Education, 8,* 65–73.

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

Janssen, F., Westbroek, H. & Doyle, W. (2015). Practicality Studies: How to Move From What Works in Principle to What Works in Practice. *Journal of the Learning Sciences, 24(1),* 176-186. doi.org/10.1080/10508406.2014.954751

Kulik, J. A. (1992). *An Analysis of the research on ability grouping: historical and contemporary perspectives*. Storrs, CT: National research center on the gifted and talented, University of Connecticut.

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

Kulik, J. A., Kulik, C. C. & Bangert-Drowns, R. L. (1990). Is there better evidence on Mastery Learning? A response to Slavin. *Review of Educational Research, 60(2), 303-307.*

Lee, J. S., & Ginsburg, H. P. (2007). Preschool teachers' beliefs about appropriate early literacy and mathematics education for low-and middle-socioeconomic status children. *Early Education and Development, 18*, 111−143. doi:10.1080/10409280701274758

*Lefgren, L. (2004). Educational peer effects and the Chicago public schools. *Journal of Urban Economics, 56*, 169–191. doi:10.1016/j.jue.2004.03.010

Leithwood, K., Harris, A., & Hopkins, D. (2008). Seven strong claims about successful school leadership. *School Leadership and Management, 28(1),* 27-42. doi.org/10.1080/13632430701800060

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

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

*Macqueen, S. (2012). Academic outcomes from between-class achievement grouping: The Australian primary context. *Australian Educational Researcher, 39*, 59–73. doi:10.1007/s13384-011-0047-3

Machts, N., Kaiser, J., Schmidt, F. T. C. & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review, 19*, 85–103. doi:10.1016/j.edurev.2016.06.003

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*, 169–194. doi:10.1080/09243453.2014.939198

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

Nunnery, J. A., Ross, S. M., & McDonald, A. (2006). A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6. *Journal of Education for Students Placed at Risk, 11(1)*: 1-18. doi: 10.1207/s15327671espr1101_1

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology, 61*, 991–996. doi:10.1016/j.jclinepi.2007.11.010

Prast, E. J., van de Weijer-Bergsma, E., Kroesbergen, E. H., & van Luit, J. E. H. (2015). Readiness-based differentiation in primary school mathematics: Expert recommendations and teacher self-assessment. *Frontline Learning Research, 3 (2),* 90-116. doi:10.14786/flr.v3i2.163

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & Du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48,* 335–360. doi:10.3102/0002831210374874

Reezigt, G. J. (1993). *Effecten van differentiatie op de basisschool* [Effects of differentiation in primary school]. Groningen, the Netherlands: 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,* 3–24. doi:10.1086/522383

*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,* 462–501. doi:10.3102/0002831210382891

Roy, A., Guay, F., & Valois, P. (2012). Teaching to address diverse learning needs: development and validation of a Differentiated Instruction Scale. *International Journal of Inclusive Education, 17:11,* 1186-1204. doi:10.1080/13603116.2012.743604

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

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

Slavin, R. E. (1987b). Mastery learning reconsidered. *Review of Educational Research, 57*, 175–213. doi:10.3102/00346543057002175

Slavin, R. E. (1990). Mastery learning re-reconsidered. *Review of Educational research, 60(2), 300-302*. doi:10.3102/00346543060002300

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60*, 471–499. doi:10.3102/00346543060003471

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*, 427–515. doi:10.3102/0034654308317473

Snow, R. E. (1989). Aptitude-Treatment Interaction as a framework for research on individual differences in learning. In: P. L. Ackerman, R. J. Sternberg, & R. Glaser (eds.), *Learning and Individual Differences: Advances in theory and research* (pp. 13-59). New York: W. H. Freeman and Company.

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K-12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research, 86 (4),* 849-899. doi:10.3102/0034654316675417

*Sterbinsky, A., Ross, S. M., & Redfield, D. (2006). Effects of comprehensive school reform on student achievement and school change: A longitudinal multi-site study. *School Effectiveness and School Improvement, 17,* 367–397. doi:10.1080/09243450600797661

*Stevens, R. J., & Slavin, R. E. (1995). The cooperative elementary-school - effects on students achievement, attitudes, and social-relations. *American Educational Research Journal, 32,* 321–351. doi:10.3102/00028312032002321

*Tach, L. M., & Farkas, G. (2006). Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. *Social Science Research, 35,* 1048–1079. doi:10.1016/j.ssresearch.2005.08.001

Timperley, H. S., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development. Best evidence synthesis [BES]*. Wellington, New Zealand: Ministry of Education.

Tomlinson, C. A. (2014). *The differentiated classroom. responding to the needs of all learners* (2nd ed.). Retrieved from https://ebookcentral.proquest.com

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K.,… Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest

and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted, 27*, 119-145. doi:10.1177/016235320302700203

Tomlinson, C. A. (2005). Grading and differentiation: Paradox or good practice? *Theory into Practice, 44*, 262–269. doi:10.1207/s15430421tip4403_11

Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research, 49*, 127–152. doi:10.1080/00131880701369651

Van de Grift, W. J. C. M., Van der Wal, M. & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs [Development of the pedagogical didactical competences of teachers in primary education]. *Pedagogische Studiën, 88*, 416–432.

Van de Pol, J., Volman, M., Oort, F. & Beishuizen, J. (2015). The effects of scaffolding in the classroom: support contingency and student independent working time in relation to student achievement, task effort and appreciation of support. *Instructional Science, 44(5),* 615-641. doi.org/10.1007/s11251-015-9351-z

Vangrieken, K., Dochy, F., Raes, E., & Kyndt, E. (2015). Teacher collaboration: A systematic review. *Educational Research Review, 15*, 17-40. doi.org/10.1016/j.edurev.2015.04.002

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

*Whitburn, J. (2001). Effective classroom organisation in primary schools: Mathematics. *Oxford Review of Education, 27*, 411–428. doi:10.1080/3054980120067438

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

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

*Ysseldyke, J., Tardrew, S. P., Betts, J., Thill, T. L., & Hannigan, E. (2004). Use of an instructional management system to enhance math instruction of gifted and talented students. *Journal for the Education of the Gifted, 27*, 293–310. doi:10.4219/jeg-2004-319

Appendix A: Studies on between-class homogeneous ability grouping

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% CI*
Lefgren, 2004	Between-class ability grouping (tracking)	USA, Chicago	More than 170,000 students (80,003 grade 3, 94,230 grade 6)	One school year	Achievement	Comparison of growth scores of students in tracked vs untracked classrooms, using the Iowa test of Basic Skills, reading and math scores.	Reading *Grade 3* 0.000 *Grade 6* 0.000 Math *grade 3* 0.000 *grade 6* 0.000	 -0.007; +0.007 -0.006; +0.007 -0.007;+0.007 -0.006; +0.007
Macqueen, 2012	Between-class ability grouping (setting)	Australia	8 schools. Literacy: regrouped 50 students, heterogeneous 68 students Writing: regrouped 29 students, heterogeneous 47 students Math: regrouped 51 students, heterogeneous 69 students	Growth from grade 3-5	Achievement	Comparison of growth scores of students in between-class ability grouped classrooms vs students in heterogeneous classrooms in the areas of literacy, writing and mathematics (Basic Skills Test). *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 Literacy* +0.196 *Overall Writing* -0.082 *Overall Math* -0.125 Literacy *Low lit group:* -0.379 *Aver. lit group:* +0.275 *High lit group:* +0.218 Writing *Low lit group:* +0.038 *Aver. lit group:* -0.023 *High lit group:* +0.196 Math *Low math grp:* -0.776 *Aver. math grp:* -0.061 *High math grp:* +0.171	 -0.170; +0.561 -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

Whitburn, 2001	Between-class ability grouping (setting)	United Kingdom	±200 students in homogeneous class-rooms and ±1,000 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) classrooms and students in mixed ability classrooms. Performance measured using researcher/teacher-designed tests.	Cohort 1 *Overall* -0.248 *Low ability* -0.389 *Average ability* -0.350 *High ability* -0.050 Cohort 2 *Overall* -0.101 *Low ability* -0.281 *Average ability* -0.166 *High ability* -0.228 Cohort 3 *Overall* -0.184 *Low ability* -0.213 *Average ability* -0.098 *High ability* -0.291	 -0.512; +0.015 -0.847; +0.068 -0.808; +0.107 -0.506; +0.405 -0.364; +0.162 -0.738; +0.176 -0.622; +0.290 -0.684; +0.228 -0.447; +0.079 -0.669; +0.243 -0.554; +0.357 -0.747; +0.166

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

Appendix B: Studies on within-class homogeneous ability grouping

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% CI*
Condron, 2008 (ECLS-K study)	Within-class ability grouping	USA	K–1: 13,625 students (ungrouped: 4,718, low group: 2,219, average group: 3,380, high group: 3,308) Grade 1–3: 13,010 students (ungrouped: 6,873, low group: 1,436, average group: 2,067, high group: 2,634)	Growth from kindergar-ten to the end of grade 1 and from grade 1 to the end of grade 3	Achievement	Propensity score matching was used to estimate the effects of placement in a high-, average-, or low-ability group in comparison with non-grouped instruction. Effect of grouping on reading was investigated, measured using ECLS-K tests. General effects cumulated over the various strata are reported.	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
Crijnen et al. 1998	Classrooms using mastery learning vs business as usual	USA, Baltimore	363 students (207 intervention and 156 control). Each participating school had at least one intervention and one control classroom.	1 year interventi-on	Achievement	Pretest-posttest, follow-up design. Schools were randomly selected. Within schools, existing classrooms were randomly assigned to the intervention or control condition. Effects on reading were measured using the California Achievement Test (CAT).	+0.138	-0.093; +0.370
Hunt, 1996	Classrooms using homogeneous grouping vs classrooms using heterogeneous grouping	USA, Southwest	Grade 6. 100 students in classrooms using homo-geneous grouping (10 gifted (state), 9 gifted (pretest), 51 average, 30 low). 108 in classrooms using hetero-geneous grouping (5	12-week interventi-on	Achievement	Randomized pre-posttest design. Comparison of students' mathematics achievement (TOMA test) in homogeneously grouped classrooms vs heterogeneously grouped classrooms	*Low ability* +0.013 *Average ability* +0.137 *Gifted 1 (identified by state)* +1.061 *Gifted 2 (high pretest scores)* +0.183	 -0.460; +0.487 -0.244; +0.519 -0.078; +2.200 -0.771; +1.138

			gifted (state), 8 gifted (pre-test), 55 average, 40 low)					
Leonard, 2001	Within-class heterogeneous small groups vs within-class homogeneous small groups	USA, Maryland	177 students from 3 classrooms: 88 students heterogeneous cohort (15 low, 31 average, 42 high); 89 students homogeneous cohort (35 low, 28 average, 26 high)	Fall – spring	Achievement	Comparison of students' mathematics achievement (measured using Maryland Functional Mathematics Test) 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
Nomi, 2010 (ECLS-K study)	Within-class ability grouping	USA	13,512 students from 900 schools. Ungrouped: 3,922 students. Ability grouped: 9,590 students	Achieve-ment from kindergar-ten to end of grade 1	Achievement	Propensity score matching was used to estimate the effects on reading scores (ECLS-K measures) of placement in a high-, average-, or low-ability group in comparison with 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 (ECLS-K study)	Within-class ability grouping	USA	grade 1 sample: Total 3,113 classrooms (10,747 students). Ability grouped: 2,241 classrooms	Achieve-ment from kindergar-ten to the end of grade 1	Achievement	Multilevel analyses were used to determine the effects of ability grouping on students' reading performance (measured using ECLS-K test)	-0.191*	-0.261; -0.120

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

Appendix C: Studies on computerized systems

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% CI*
Connor et al., 2007	Within-class differentiated instruction	USA	10 schools, 47 classrooms (22 treatment, 25 control), 616 students	Fall-spring	Achievement	A cluster-randomized field trial was used in which students in experimental schools were compared with students in matched control schools on a language and literacy test (Woodcock Johnson Tests of Achievement-III).	+0.183*	+0.025; +0.342
Connor et al., 2011a	Within-class differentiated instruction	USA	7 schools. Experimental group: 16 schools, 219 students. Control group: 17 schools, 229 students.	Fall-spring	Achievement	Multilevel modeling was used to analyze the effects of differentiated instruction using a computer program compared with a vocabulary instruction intervention on reading comprehension and vocabulary (subtests from Woodcock Johnson Tests of Achievement-III).	*Overall (both measures)* +0.187* *Reading comp.* +0.191* *Vocab.* +0.033	 +0.001; +0.373 +0.005; +0.377 -0.153; +0.219
Connor et al., 2011b	Within-class differentiated instruction	USA	7 schools. Experimental group: 14 classrooms, 222 students. Control group: 11 classrooms, 174 students.	Fall-spring	Achievement	Multilevel modeling was used to analyze the effects of differentiated instruction using a computer program compared with a control group on language and literacy (Woodcock Johnson Tests of Achievement-III).	+0.249*	+0.050; +0.448
Ysseldyke et al., 2003	Within-class differentiated instruction	USA	Experimental group: 397 students. Within-school control group: 484 students	September - June	Achievement	An analysis of variance of the mean scores on two mathematics tests (Northwest Achievement Levels Test (NALT), standardized; and STAR Math, standardized and computer-adaptive) of the experimental and the control group.	*Math (NALT)* +0.189* *Math (STAR)* +0.268*	 +0.030; +0.349 +0.109; +0.428
Ysseldyke et al., 2004	Within-class differentiated instruction	USA	Comparison of performance of gifted and non-gifted students	4 months	Classified as gifted or talented as defined by the	Four-group pretest-posttest control group design. Performance in Mathematics	*Gifted* +0.456* *Not gifted*	 +0.059; +0.853

			in experimental classrooms (48 gifted students, 743 non-gifted) and control classrooms (52 gifted, 736 non-gifted).		state in which student was enrolled	measured using STAR Math test.	+0.369*	+0.266; +0.472
Ysseldyke et al., 2007	Within-class differentiated instruction	USA	Experimental condition: 8 schools, 41 classrooms; Control condition: 8 schools, 39 classrooms	October - May	Achievement	An analysis of variance of the mean scores on two mathematics tests (Terra Nova and STAR Math) of the experimental and the control group in primary education.	*Overall (both measures)* +0.290* *Math (Terra Nova)* +0.324* *Math (STAR)* +0.491*	 +0.206;+0.373 +0.213; +0.435 +0.375; +0.607

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

Appendix D: Studies on differentiation as part of a broader program

Article	*Type of differentiation*	*Location*	*Sample size*	*Duration*	*Grouping criteria*	*Design*	*Effect sizes (d)*	*95% CI*
Borman et al., 2007	Ability grouping across grades for reading, as part of a whole-school comprehensive reform (SfA)	USA	Experimental: 18 schools, 7,920 students,. Control: 17 schools, 7,395 students.	3 years, from kindergar-ten to grade 2	Achievement measured every 9 weeks	Cluster randomized design. Language and literacy outcomes measured using 3 subtests of the Woodcock Reading Mastery Tests-Revised (WMTR).	*Word reading* +0.220* *Non-word decoding* +0.330* *Reading compr*. +0.210*	 +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 Nether-lands	Experimental: 21 schools, 567 students. Control: 16 schools, 454 students	December -May	Achievement	Quasi-experimental design. Reading performance measured using the Dutch tests DMT (accurate word decoding) and AVI (reading fluency)	*Accuracy* +0.280* *Fluency* +0.620*	 +0.156; +0.404 +0.494; +0.746
Reis et al., 2007	SEM-R (School-wide Enrichment Model in Reading Framework): differentiated, individual instruction ans tasks for reading.	USA	2 schools, grades 3–6. Experimental: 1 school, 7 teachers, 110 students. Control: 1 school, 7 teachers, 116 students.	12 weeks	Teacher's judgment	Randomized design. Performance on two domains of reading were measured: reading comprehension (subtest of the Iowa Tests of Basic Skills) and oral reading fluency (Curriculum-based measure)	*Reading compr*. +0.220 *Fluency* +0.299*	 -0.529; +0.970 +0.005; +0.594
Reis et al., 2011	SEM-R (School-wide Enrichment Model in Reading Framework): differentiated individual instruction and tasks for reading.	USA	5 schools, grades 2-5. Experimental: 37 classrooms, 649 students. Control: 33 classrooms, 543 students.	24 weeks	Teacher's judgment	Cluster-randomized design. Performance on two domains of reading were measured: reading comprehension (subtest of the Iowa Tests of Basic Skills) and oral reading fluency (curriculum-based measure).	*Reading compr*. +0.145 *Fluency* +0.254	 -0.096; +0.386- 0.063; +0.571

Sterbinsky et al., 2006	Four types of school reforms were compared, all using differentiated instruction.	USA (regions: Kentucky, Tennessee Virginia and West Virginia)	19 schools (10 intervention, 9 control), in total approx. 350-400 teachers (variation per year), in total approx. 550-707 students (complete 3-year data available for 170 students)	3 years	Achievement	3-year quasi-experimental study using a matched treatment-control group. Reading performance was measured using subtests of the Woodcock–Johnson Reading Mastery Test (word reading and passage comprehension) and the Durrell Oral Reading Test.	*Word reading* +0.308* *Reading compr.* +0.286* *Oral reading* +0.429*	 +0.023; +0.592 +0.001; +0.570 +0.142; +0.715
Stevens & Slavin, 1995	Students worked in heterogeneous learning teams but received instruction in relatively homogeneous teaching groups, as part of a whole-school reform program.	USA	5 schools, grades 2–6. Experimental group: 2 schools, 21 classrooms, 411 students. Control group: 3 schools, 24 classrooms, 462 students.	After 1 and 2 years (only data after 2 years used)	Achievement	Quasi-experimental. Reading, language, and math performance was measured using subtests from the California Achievement Test (CAT)	Reading *Vocabulary* +0.210* *Comprehension* +0.280* *Language* *Mechanics* +0.100 *Expression* +0.210* Mathematics *Computation* +0.290* *Concept & applic.* +0.100	 +0.075; +0.345 +0.128; +0.432 -0.069; +0.269 +0.069; +0.351 +0.139; +0.441 -0.058; +0.258

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