Explanatory latent variable modeling of mathematical ability in primary school : crossing the border between psychometrics and psychology

Hickendorff, M.

CHAPTER 1

Performance outcomes of primary school mathematics programs in the Netherlands: A research synthesis

This chapter is based on research I have done for the KNAW Committee on Primary School Mathematics Teaching, reported in KNAW (2009). Note that this report is written in Dutch, and the reproduction of ideas in English is on my account.

ABSTRACT

The results of a systematic quantitative research synthesis of empirical studies addressing the relation between mathematics education and students' mathematics performance outcomes is presented. Only studies with primary school students carried out in the Netherlands were included. In total, 25 different studies were included: 18 intervention studies in which the effects of different mathematics interventions (instructional programs) were compared, and 7 curriculum studies in which differential performance outcomes with different mathematics curricula (usually textbooks) were assessed. In general, the review did not allow drawing a firm univocal conclusion on the relation between mathematics education and performance outcomes. Some more specific patterns emerged, however. First, performance differences were larger within a type of instructional approach than between different instructional approaches. Second, more time spent on mathematics education resulted in better performance. Third, experimental programs implemented in small groups of students outside the classroom had positive effects compared to the regular educational practice. Fourth, low mathematics performers seemed to have a larger need for a more directing role of their teacher in their learning process.

1.1 INTRODUCTION

1.1.1 Background

Recently, there has been a lot of criticism on mathematics education in primary school in the Netherlands, originating in growing concern on children's mathematical proficiency. This public debate – both in professional publications as well as in more mainstream media – is characterized by its heated tone and its polarizing effect. That caused the *Royal Netherlands Academy of Arts and Sciences* (KNAW) to set up a Committee on Primary School Mathematics Teaching in 2009. When the State Secretary, Ms. Sharon Dijksma, announced a study on mathematics education, these two initiatives were combined. The Committee's mission was *"To survey what is known about the relationship between mathematics education and mathematical proficiency based on existing insights and empirical facts. Indicate how to give teachers and parents leeway to make informed choices, based on our knowledge of the relationship between approaches to mathematics teaching and mathematical achievement."* (KNAW, 2009, p. 10).

The current chapter is based on the systematic quantitative review of empirical studies addressing the relation between mathematics education or instruction and children's mathematical proficiency in the Netherlands, one of the core parts of the committee's report (KNAW, 2009, ch. 4[1]). In the remainder of the Introduction, first a short overview of the state of primary school students' mathematical proficiency level is presented, based on findings of national and international large-scale educational assessments. Then a brief discussion of existing international reviews and meta-analyses of research on the effects of mathematics instruction follows. In the main part of this chapter, the methodology and results of the current systematic quantitative review are presented. This review is largely along the lines of what Slavin (2008) proposed as a *best-evidence synthesis*: a procedure for performing syntheses of research on educational programs that resembles meta-analysis, but requires more extensive discussion of key studies instead of primarily aiming to pool results across many studies (Slavin & Lake, 2008). In the current review into the effect of primary school mathematics programs in the Netherlands, a distinction is made between *intervention studies* in which the researchers intervened in the educational practice, and *curriculum studies* in which no intervention took place, the mathematics programs compared were self-selected by schools. This chapter ends with a summary of the research synthesis, conclusions, and implications.

1.1.2 The state of affairs of Dutch students' mathematical performance

To describe the state of Dutch primary school students' mathematical performance level, empirical quantitative results of national and international assessments were used. Such large-scale educational assessments aim to report on the outcomes of the educational system in various content domains such as reading, writing, science, and mathematics. At least two aspects are important (Hickendorff, Heiser, Van Putten, & Verhelst, 2009a). The first aspect is a description of students' learning outcomes: what do students know, what problems can they solve, to what extent are educational standards reached, and to what extent are there differences between subgroups (such as different countries in international assessments, or boys and girls within a country)? The second aspect concerns trends: to what extent are there changes in achievement level over time?

1 I carried out this research review at request of the KNAW Committee, for which I worked as an associate researcher.

At the national level, CITO carried out educational assessments – PPON [*Periodieke Peiling van het Onderwijsniveau*] – of mathematics education in grade 3 (9-year-olds) and in grade 6 (12-year-olds) in cycles of five to seven years since 1987. In the current overview only the results for grade 6 are discussed, because these concern students' proficiency at the end of primary school. At the international level, there is TIMSS (*Trends in International Mathematics and Science Study*): an international comparative study in the domains of science and mathematics, carried out in grade 4 (10-year-olds) and in grade 8 (14-year-olds, second grade of secondary education in the Netherlands), with assessments in 1995, 2003, and 2007. Only the grade 4 results concern primary school, so we focus on those.

Dutch national assessments: PPON in grade 6

Van der Schoot (2008) presented an overview of the grade 6 mathematics assessment results. Thus far, there have been four cycles: 1987, 1992, 1997, and 2004 (the next assessment is planned in 2011). The domain of mathematics is structured in three general domains: (a) numbers and operations, (b) ratios/fractions/percentages, and (c) measures and geometry. In each general domain, several subdomains are distinguished. In total, there were 22 different subdomains in the most recent assessment of 2004 (J. Janssen et al., 2005).

Students' results were evaluated in two ways: the trend over time since 1987, and the extent to which the educational standards were reached. For the latter evaluation, the standards set by Dutch Ministry of Education, Culture, and Sciences (1998) were operationalized by a panel of approximately 25 experts, ideally consisting of 15 primary school teachers, 5 teacher instructors, and 5 educational advisors. In a standardized procedure, these panels agreed upon two performance levels: a *minimum* level that 90-95% of the students at the end of primary school should reach, and a *sufficient* level, that should be reached by 70-75% of all students. Table 1.1 presents the relevant results. First, it shows the effect size (ES, standardized mean difference) of the performance difference between the baseline measurement (usually 1987), interpreted as $.00 \leq |\text{ES}| < .20$ negligible to small effect, $.20 \leq |\text{ES}| < .50$ small to medium effect, $.50 \leq |\text{ES}| < .80$ medium to large effect, and $|\text{ES}| \geq 0.80$ large effect. Second, it shows the percentage of students reaching the educational standards of minimum and sufficient level.

The trends over time show varying patterns, with the most striking developments

TABLE 1.1 *Dutch mathematics assessments results, from Van der Schoot (2008, p. 20-22).*

	trend in ES (baseline 1987 = 0)			reaching standard in 2004	
	1992	1997	2004	min.	suff.
numbers and operations					
numbers and number relations	+.28	+.46	+.94	96%	42%
simple addition/subtraction	*	−.11	+.24	92%	76%
simple multiplication/division	*	−.30	−.20	90%	66%
mental addition/subtraction	n.a.	+.49	+.53	92%	50%
mental multiplication/division	n.a.	−.12	−.11	92%	66%
numerical estimation	n.a.	+.94	+1.04	84%	42%
complex addition/subtraction	−.12	−.17	−.53	62%	27%
complex multiplication/division	−.17	−.43	−1.16	50%	12%
combined complex operations	−.40	−.44	−.78	50%	16%
calculator	*	+.29	+.26	73%	34%
ratios/fractions/percentages					
ratios	+.11	+.26	+.14	92%	66%
fractions	+.09	+.23	+.15	95%	60%
percentages	+.12	+.28	+.51	88%	58%
tables and graphs	n.a.	*	+.10	84%	50%
measures and geometry					
measures: length	+.00	−.03	−.13	79%	38%
measures: area	−.32	−.04	+.05	67%	21%
measures: volume	+.10	.00	−.03	67%	21%
measures: weight	+.02	+.20	+.33	88%	58%
measures: applications	−.05	−.21	−.25	92%	50%
geometry	.00	+.12	−.08	95%	62%
time	+.17	+.23	.00	92%	50%
money	−.21	−.31	n.a.	84%	42%

* Earlier results not available, alternative baseline.

in the domain of numbers and operations. Differences were negligible to medium-sized ($|ES| < .50$) on 14 of the 21 subdomains for which trends could be assessed. Positive developments of at least medium size ($ES \geq .50$) were found in percentages, mental addition/subtraction, numbers and number relations, and numerical estimation. Negative trends of at least medium size ($ES \leq -.50$), however, were found for complex addition and subtraction, combined complex operations, and complex multiplication and division.

Regarding attainment of the educational standards, Table 1.1 shows that on only one subdomain (simple addition/subtraction), the desired percentage of 70% or more students attaining the sufficient level was reached. On eleven domains, this percentage was between 50% and 70%, and on five domains it was between 30% and 50%. Finally, on five domains the percentage of students attaining sufficient level did not exceed 30%. So, in particular performance in the *complex operations* (addition/subtraction, multiplication/division, and combined operations; all concern multidigit problems on which the use of pen and paper to solve them is allowed) and in the *measures* subdomains *weight* and *applications* is worrisome according to the expert panels.

International assessments: TIMSS in grade 4

The Netherlands participated in the grade 4 international mathematics assessments in 1995, 2003, and 2007 (Meelissen & Drent, 2008; Mullis, Martin, & Foy, 2008). Worldwide, 43 countries participated in TIMSS-2007. In this TIMSS cycle there were mathematics items from three mathematical content domains – number, geometric shapes and measures, and data display – crossed with three cognitive domains – knowing, applying, and reasoning. Curriculum experts judged 81% of the mathematics items suited for the intended grade 4 curriculum in the Netherlands. Conversely, only 65% of the Dutch intended curriculum was covered in the TIMSS-tests.

Dutch fourth graders' mathematics performance level was in the top ten of the participating countries; only in Asian countries performance was significantly higher. Interestingly, the spread of students' ability level was relatively low, meaning that students' scores were close together. Another way to look at this is to compare performance to the TIMSS International Benchmarks: the advanced level was attained by 7% of the Dutch students, high level by 42%, intermediate level by 84%, and low level by 98% of the students. Although these percentages were all above the international median, compared to other countries that had such a high overall performance as the Netherlands, there were relatively many students attaining the low performance level, but relatively few students reaching the advanced level. Furthermore, developments over time showed a small but significant negative trend in total mathematics performance since 1995 (average score 549), via 2003 (average score 540), towards 2007 (average score 535). Internationally, more countries showed improvements in fourth grade performance than declines, so the Netherlands stand out in this respect.

Students' attitudes toward mathematics were investigated with a student questionnaire with questions on positive affect toward mathematics and self-confidence in own mathematical abilities (Mullis et al., 2008). Students reported a slightly positive affect toward mathematics, although it showed a minor decrease compared to 2003. Moreover, in the Netherlands there were proportionally many students (27%; international average 14%) at the low level of positive affect, and proportionally few students (50%; international average 72%) at the high level. Dutch students had quite high levels of self-confidence, and the distribution was comparable to the international average distribution.

Finally, we discuss some relevant results on the teacher and the classroom characteristics and instruction. Dutch fourth grader teachers were at the bottom of the international list in participating in professional development in mathematics. Still, they reported to feel well prepared to teach mathematics for 73% of all mathematics topics (international average 72%). Furthermore, Dutch fourth grade teachers reported experiencing much fewer limitations due to student factors than the international averages. A last relevant pattern was that Dutch students reported relatively frequently to work on mathematics problems on their own, while they reported explaining their answer relatively infrequently.

Summary national and international assessments

The national assessments (PPONs) were tailor-made to report on the outcomes of Dutch primary school mathematics education. Results showed that in many subdomains there were only minor changes in sixth graders' performance level between 1987 and 2004, and opposed to subdomains where performance declined there were subdomains in which performance improved. International assessments (TIMSS) showed that Dutch fourth graders still performed at a top level from an international perspective.

However, these results do not justify complacency (KNAW, 2009). In TIMSS, too few students reached the high and advanced levels, there was a small performance decrease over time causing other countries to come alongside or even overtake the position of the Netherlands, and students too often reported low positive affect toward mathematics. Moreover, it seems unwise to cancel out the positive and negative developments that were found in PPON. In addition, students' performance level lagged (far) behind the educational standards for primary school mathematics in most subdomains, also in the

subdomains showing improvement over time.

1.1.3 International reviews, research syntheses, and meta-analyses

We briefly review some patterns that emerge from international reviews and meta-analyses into the effects of mathematics instruction on achievement outcomes[2]. Note that this discussion is by no means exhaustive. Moreover, the findings are to a large extent based on studies carried out in the US. A first important observation is that the authors of most of the reviews stated that there are few studies that meet methodological standards that permit sound, well-justified conclusions about the comparison of the outcomes of different mathematics programs. The number of well-conducted (quasi-)experimental studies is low, and in particular studies meeting the 'golden standard' of randomized controlled trials are rarely encountered. For example, the US *National Mathematics Advisory Panel*, that had a similar assignment as the Dutch KNAW Committee, reviewed 16,000 research reports and concluded that only a very small portion of those studies met the rigorous methodological standards that allowed conclusions on the effect of instructional variables on mathematics learning outcomes (National Mathematics Advisory Panel, 2008). This review, however, has been heavily criticized for its stringent inclusion criteria that resulted in exclusion of relevant research findings, as well from its narrow cognitive perspective on mathematics education (see Verschaffel, 2009, for an overview of reactions in the US).

We primarily focus on two recent research syntheses: one by Slavin and Lake (2008) of research on achievement outcomes of different approaches to improving mathematics in regular primary education, and the other by Kroesbergen and Van Luit (2003) of research on the effects of mathematics instruction for primary school students with special educational needs.

Slavin and Lake (2008) conducted a 'best-evidence synthesis' of research on the achievement outcomes of three types of approaches to improving elementary mathematics: mathematics curricula, computer-assisted instruction (CAI), and instructional process programs. In total, 87 studies were reviewed, meeting rather stringent methodological criteria based on the extent to which they contribute to an unbiased, well-justified quantitative estimate of the strength of the evidence supporting each program.

2 This section is partly based on contributions of prof. dr. Lieven Verschaffel to chapter 3 of the KNAW (2009) report.

Regarding mathematics curricula, the results of the synthesis showed that there was little empirical evidence for differential effects. A noteworthy shortcoming of these studies was that they mainly used standardized tests that focused more on traditional skills than on concepts and problem solving that are addressed in reform-based mathematics curricula. However, in the cases when outcomes on these 'higher-order' mathematics objectives were considered, they do not suggest a differential positive effect of reform-based curricula. This observation contrasts with that of Stein, Remillard, and Smith (2007), who reviewed US-studies comparing 35 different mathematics textbooks (written curricula), of which approximately half could be characterized as reform-based or constructivistic, and the other half as traditional or mechanistic. They concluded that students trained with reform-based textbooks performed at about an equal level on traditional skills, but did better on higher-order goals such as mathematical reasoning and conceptual understanding, compared to students trained with traditional textbooks. An important remark, however, is that Stein et al. found that variation in teacher implementation of traditional curricula was smaller than in teacher implementation of reform-based curricula, hampering sound conclusions on differential effects of mathematics curricula.

CAI-supplementary approaches had moderate positive effects on students learning outcomes, especially on measures of computational skills (Slavin & Lake, 2008). Although the effects reported were very variable, the fact that in no study effects favoring the control group were found, and that the CAI-programs usually supplement the classroom instruction by only about 30 minutes a week, Slavin and Lake claimed that the effects were meaningful for educational practice. CAI primarily adds the possibility to tailor the instruction to individual students' specific strengths and weaknesses. In a meta-analysis of intervention research of word-problem solving in students with learning problems, Xin and Jitendra (1999) also found that CAI was a very effective intervention, but Kroesbergen and Van Luit (2003) found negative effects of CAI compared to other interventions in their meta-analysis of mathematics intervention studies in students with special educational needs.

Finally, Slavin and Lake (2008) found the largest effects for instructional process programs, that primarily focus on what teachers do with the curriculum they have, not changing the curriculum. The programs reviewed were highly diverse. Programs with positive effects either used various forms of cooperative learning, focused on classroom management strategies, used direct instruction models, or supplemented

traditional classroom instruction (including small group tutoring). These are quite general characteristics of how teachers use instructional process strategies. In line with these findings are results from a recent investigation of the Dutch Inspectorate of Education (2008) into school factors that are related to students' mathematics performance in primary school. They found that the educational process (quality control, subject matter, didactical practice, students' special care) was of lower quality in mathematically weak schools than in mathematically strong schools. In particular, there were nine school factors in which mathematically weak schools lagged behind: (a) yearly systematic evaluation of students' results; (b) quality control of learning and instruction; (c) the number of students for whom the subject matter is offered up to grade 6 level; (d) realization of a task-focused atmosphere; (e) clear explaining; (f) instructing strategies for learning and thinking; (g) active participation of students; (h) systematic implementation of special care; and (i) evaluation of the effects of special care.

Slavin and Lake (2008, p. 475) concluded their research synthesis with stating that *"the key to improving math achievement outcomes is changing the way teachers and students interact in the classroom."* The central and crucial role of the teacher in improving mathematics education is also subscribed to by others, such as Kroesbergen and Van Luit (2003) and Verschaffel, Greer, and De Corte (2007). An important concept is teachers' *Pedagogical Content Knowledge* (PCK), a blend of content knowledge and pedagogical knowledge of students' thinking, learning, and teaching. Fennema and Franke (1992) and Hill, Sleep, Lewis, and Ball (2007) pointed at the potential of pre-service and in-service training programs to improve teachers' mathematical PCK, but at the same time they acknowledge that there is little empirical evidence about the causal relation between teachers' PCK and students' achievement outcomes.

A lot of research attention is devoted to interventions for students with special educational needs, sometimes distinguished in students with learning disabilities (LD) and students with (mild) mental retardation (MR). Kroesbergen and Van Luit (2003) carried out a meta-analysis into the effects of mathematics interventions for these students, reviewing 58 studies addressing three mathematical domains: preparatory arithmetic, basic skills, and problem solving. The meta-analysis showed that intervention effects were largest in the domain of basic skills, implying that it may be easier to teach students with mathematical difficulties basic skills than problem-solving skills. Further relevant conclusions were that regarding treatment components of the interventions, self-instruction and direct instruction (more traditional instructional approaches) were more

effective than mediated/assisted instruction (more reform-based approach). The results favoring direct instruction were in in line with other meta-analyses of intervention studies with students with learning disabilities (e.g., Gersten et al., 2009; Swanson & Carson, 1996; Swanson & Hoskyn, 1998), stressing the importance of the role of the teacher to help students with special educational needs and to evaluate their progress. Similarly, the National Mathematics Advisory Panel (2008) also concluded that explicit instruction is effective for students struggling with mathematics. Apart from this instructional component, Kroesbergen and Van Luit's meta-analysis did not find effects of other characteristics of Realistic Mathematics Education. Kroesbergen and Van Luit therefore concluded that the mathematics education reform does not lead to better performance for students with special educational needs.

Another review worth mentioning is that of Hiebert and Grouws (2007) into the effects of classroom mathematics teaching on students' learning. Their first conclusion was that *opportunity to learn*, which is more nuanced and complex than mere exposure to subject matter, is the dominant factor influencing students learning. Secondly, they distinguish between teaching for skill efficiency and teaching for conceptual understanding. In teaching that facilitates skill efficiency, the teacher plays a central role in organizing, pacing, and presenting information or modeling to meet well-defined learning goals; in short: teacher-directing instruction. Teaching that facilitates conceptual understanding, however, is characterized by an active role of students and explicit attention of students and teachers to concepts in a public way.

1.2 Method of the current review

The basic approach of the current review was along the lines of Slavin's (2008) best evidence synthesis procedure. This technique *"seeks to apply consistent, well-justified standards to identify unbiased, meaningful, quantitative information from experimental studies"* (Slavin & Lake, 2008, p. 430). Slavin contended that the key focus in synthesizing (educational) program evaluations is minimizing the bias in reviews of each study, because there are usually only a small number of studies per program. The scarceness of studies also precludes pooling of results over studies and statistically testing for effects of study characteristics or procedures like in meta-analysis (Lipsey & Wilson, 2001). Instead, a more extensive discussion of the nature and quality of each study is incorporated. For each qualifying study not only effect sizes are computed, but also the context, design,

and findings of each are discussed (Slavin & Lake).

The objective of the current review was to *"investigate what is known scholarly about the relation between instructional approaches and mathematical proficiency"* (KNAW, 2009, p. 12). To that end, a quantitative synthesis of achievement outcomes of alternative mathematics programs was carried out. In this synthesis, quantitative results of other outcomes such as motivation or attitudes were not included, although relevant findings are discussed in the text. Two types of empirical studies addressing this objective are distinguished, similar to Slavin and Lake (2008): *intervention studies* and *curriculum studies*.

Intervention studies aim to assess the effect of one or more mathematics programs that are implemented with an intervention in the regular educational practice. These programs either replace or supplement (part of) the regular curriculum, and usually address a specific delimited content area such as addition and subtraction below 100. The programs are highly diverse. Furthermore, the implementation of the (experimental) programs is under researcher control, but the extent of control varies. It may be that external trainers implement the programs – yielding much control – or that the regular teacher was trained to implement the program. Combinations are also possible. Assignment to conditions (i.e., programs) may be either on individual student level or at the level of whole classrooms or schools. Furthermore, assignment may be random (experimental design) or non-random (quasi-experimental design). Finally, in most studies a pretest is administered before start of the program under study, in others not.

Curriculum studies aim to investigate differential achievement outcomes of different mathematics curricula, usually operationalized as mathematics textbook (series). The researchers have no control on assignment to curricula or on the implementation of the curriculum, and therefore these are observational studies. A disadvantage is that selection effects cannot be ruled out: factors that determine which mathematics textbook a school uses are likely to be related to achievement, biasing the results. Moreover, there is usually only one measurement occasion, so that correcting for differences between groups is also not possible.

1.2.1 Search and selection procedures

A number of inclusion criteria for a study to qualify for the review were set up, based on their potential to address the review's objective. The criteria were:

1. the study specifically addresses mathematics, or at least it should be possible to parcel out the mathematics results;
2. it should be possible to examine the results for children in the age range 4-12 years;
3. the study is executed less than 20 years ago[3];
4. the study is carried out in the Netherlands, with Dutch classes and students, or in case of an international study it should be possible to parcel out the effects for the Netherlands;
5. the study is empirical, meaning that conclusions are based on empirical data;
6. the study's results are published, preferably in (inter)national journals, books, and doctoral theses;
7. at least two different mathematics programs are compared,
8. there is enough statistical information in the publication to compute or approximate the effect size (see section 1.2.2)[4].

Compared to Slavin and Lake (2008) and Slavin's (2008) recommendations, we were less strict in excluding studies. Specifically, we were less stringent in excluding studies based on the research design (i.e., studies with non-random assignment and without matching were not excluded), based on pretest differences (i.e., studies with more than half a standard deviation difference at pretest were not excluded *per se*, but rather were marked as yielding unreliable effect sizes), based on study duration, and based on outcome measures. Our approach to including studies was this liberal because we argue that compromises on study quality are necessary, because there are so few studies in number. Moreover, by including studies liberally but clearly describing each study's limitations, readers have a comprehensive overview of the existing literature and can judge the studies' quality themselves.

To search for relevant studies, the KNAW Committee asked 50 experts in mathematics education research in the Netherlands to give input on studies to include. This resulted in 76 proposed publications, 17 of which met the inclusion criteria as set in the current chapter. Additional literature searches resulted in a total of 25 different studies (18 intervention studies and 7 curriculum studies) that met the inclusion criteria, reported in 29 different publications.

3 We were more strict on this criterion than in KNAW (2009), thereby excluding one study that was included in that report.

4 This was not one of the original inclusion criteria in KNAW (2009, p. 43-44), and thereby one more study was excluded.

1.2.2 Computation of effect sizes

To compare and synthesize quantitative results from many different studies they need to be brought to one common scale. To that end, results are reported in effect sizes (ES): the standardized mean difference between conditions (e.g., Lipsey & Wilson, 2001). The difference in mean posttest achievement scores in condition or program 1 ($\overline{X}_1$) and condition 2 ($\overline{X}_2$) is divided by the pooled standard deviation s_p, i.e.,

$$\text{ES} = \frac{\overline{X}_1 - \overline{X}_2}{s_p}, \tag{1.1}$$

with

$$s_p = \sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-1}}, \tag{1.2}$$

with n_1 and n_2 the number of students in program 1 and 2, respectively, and s_1 and s_2 the standard deviation in program 1 and 2. Guidelines for interpreting these effect sizes are commonly: $.00 \leq |\text{ES}| < .20$ negligible to small effect, $.20 \leq |\text{ES}| < .50$ small to medium effect, $.50 \leq |\text{ES}| < .80$ medium to large effect, and $|\text{ES}| \geq .80$ large effect, see for example Cohen (1988). Furthermore, Slavin (2008) qualified an ES of at least .20 as practically relevant in educational research. If there were multiple achievement outcomes, effect sizes were computed and reported for each measure separately. For studies that did not report means and standard deviations, other statistical information was used to compute and approximate the mean difference and the pooled standard deviation (e.g., Kroesbergen & Van Luit, 2003).

An important possible threat to the validity of comparisons of program outcomes is the influence of pre-existing group differences. These differences were accounted for in the following ways. If the study reported posttest means that were corrected for pretest measures or background variables (for example from an analysis of covariance or a multiple regression analysis), these adjusted means were used in computing the effect size. If such adjusted means were not reported, correction was approximated by subtracting the standardized mean difference in pretest scores from the standardized mean difference in posttest scores, as recommended by Slavin (2008). If no data from before the start of the program were reported, statistically correcting for pre-existing differences was not possible, and this should be held in mind in evaluating the reported effect sizes.

1.2.3 Study characteristics coded

For each study, several characteristics were coded, and they are described in the Summary Tables in Appendices 1.A and 1.B. The characteristics were:

1. *reference*: the publication reference(s) in which the study is reported;
2. *domain*: the mathematical content domain the study addressed;
3. *participants*: several characteristics of the students participating in the study: the sample size N, the number of classes or schools they originated from, the type of primary school they attended (regular or special education), and whether all students or only low math performers participated;
4. *intervention* or *curriculum*: the programs evaluated [intervention studies] or the mathematics curricula used [curriculum studies];
5. *duration and implementation*: the duration of the mathematics programs or curricula and who implemented it [intervention studies only];
6. *design and procedure* [intervention studies only]: the study design (measurement occasions and intervention) and the procedure of assigning students to conditions;
7. *corrected*: per outcome measure, for which pre-existing differences the comparison was statistically corrected for;
8. *(posttest) results*: per outcome measure, the results of the comparison of posttest scores between programs [intervention studies] or of performance measures with different curricula [curriculum studies], in which it is indicated whether the difference was significant (indicated with $<$ and $>$) or not significant (n.s.);
9. ES: per outcome measure, the effect size computed (standardized mean difference on posttest), statistically corrected as indicated in column *corrected*.

If applicable, in the columns *(posttest) results* and ES the mean score in the least innovating program was subtracted from the mean score in the more innovating program. Furthermore, if the results were separated by subgroups of students in the original publication, this was also done in the *results* and ES.

1.3 Intervention studies

The didactical approach used can differ greatly between studies.Furthermore, in the programs studied it is very common that more than one didactical element is varied, such as the models used (e.g., the number line), the type of instruction and the role of

the teacher (varying from very directive to very open), the type of problems used (very open problem situations, contextual math problems, or bare number problems), and type of solution strategies instructed (standard algorithms or informal strategies). This mixing of program elements makes it impossible to investigate which of the elements caused the effect reported. The study characteristics of the intervention studies reviewed are displayed in the Summary Table in Appendix 1.A.

In discussing the relevant findings of the intervention studies, we distinguish the results according to the type of comparison that was made. The first type involved comparisons of outcomes of *two or more different experimental programs*, second, the second type comparisons of outcomes of an *experimental program with a control program* (the latter usually the self-selected curriculum), and the third type, comparisons of outcomes of a *supplementary experimental program with a control group* that did not receive any supplementary instruction or practice. In some studies, comparisons of more than one of these categories were made (for instance when there were two experimental programs and one control condition). The findings of these studies were split up accordingly.

1.3.1 Comparing the outcomes of different experimental programs

In this section, study findings regarding comparisons of achievement outcomes of at least two experimental mathematics instruction programs are discussed. For a comparison to qualify in this category, the programs had to be implemented similarly, i.e., by the same kind of instructor in the same kind of instructional setting with the same duration. Six studies compared two specific instructional interventions (guided versus direct instruction) in low mathematically achieving students, in regular education as well as in special education. In another study, two different remedial programs for low mathematics achievers in regular education were compared. Finally, two more studies addressed instructional programs for all students (not only the low achieving ones) in regular education.

Guided versus direct instruction in low mathematics achievers

Six studies focusing on low mathematics achievers, both in special education and in regular education, were quite comparable in their instructional interventions, and are therefore discussed together. Each of these studies compared *guided instruction* (GI)

versus *direct instruction* (DI)[5] in a particular content domain. Guided or constructivistic instruction involved either students bringing up possible solution strategies, or teachers explaining several alternative ways to solve a problem. Students choose a strategy to solve a problem themselves. By contrast, in direct (also called explicit or structured) instruction, students were trained in one standard solution strategy. In one study (Milo, Ruijssenaars, & Seegers, 2005), there were two direct instruction conditions: one (DI-j) instructing the 'jump' strategy (e.g., $63-27$ via $63-20=43$; $43-7=36$), and the other (DI-s) instructing the 'split' strategy (e.g., $63-27$ via $60-20=40$; $3-7=-4$; $40-4=36$, see also Beishuizen, 1993).

The intervention programs consisted of between 26 and 34 lessons. One study (Van de Rijt & Van Luit, 1998) addressed 'early mathematics' in preschoolers, the other studies addressed the domain of multiplication (Kroesbergen & Van Luit, 2002; Kroesbergen, Van Luit, & Maas, 2004) or addition and subtraction below 100 (Milo et al., 2005; Timmermans & Van Lieshout, 2003; Timmermans, Van Lieshout, & Verhoeven, 2007) with students between 9 and 10 years old. With respect to the outcomes, often a distinction was made in automaticity/speed tests, performance measures (achievement on the content domain addressed in the program), and transfer tests (performance on problems that students were not exposed to in the intervention programs). All six studies had a pretest - intervention - posttest design, thereby making statistical correction for pre-existing group differences possible. Either whole classes were randomly assigned to programs, or students within classes were matched and then assigned to programs (however, in Milo et al. (2005) the assignment procedure was unclear). Table 1.2 synthesizes the main findings of these six comparable studies.

In four studies, *automaticity* was an outcome measure. In two studies, a small to medium disadvantage of guided instruction was found, while in the other two studies, differences were negligible. Thus, guided instruction resulted in comparable or lower automaticity outcomes than direct instruction.

All six studies reported on *performance* in the domain of study. Two studies reported a small to medium advantage for guided instruction, two studies found negligible to small advantage of guided instruction, and two studies reported a small to medium advantage for direct instruction. Two additional patterns are worth mentioning. First, in Milo et al. (2005) there were two direct instruction conditions: one (DI-j) instructing the

5 If reported, the comparisons between outcomes of the GI and DI conditions on the one hand and a control condition on the other hand, are discussed in section 1.3.2.

TABLE 1.2 *Synthesis of results from six studies comparing guided instruction (GI) and direct instruction (DI) in low mathematics performers.*

		effect size GI - DI		
study	school type	automaticity	performance	transfer
Kroesbergen & Van Luit (2002)	reg. + spec.	[–.51]	+.43	+.52
	special	*[–2.42]*	*+.32*	*+.36*
	regular	*[+.61]*	*+.86*	*+.95*
Kroesbergen et al. (2004)	reg. + spec.	+.03	–.30	n.a.
Milo et al. (2005)	special	n.a.	–.73 (DI-j)	+.07* (DI-j)
		n.a.	–.21 (DI-s)	+.59* (DI-s)
Timmermans & Van Lieshout (2003)	special	–.23#	.00#	–.57*
Timmermans et al. (2007)	regular	+.05	+.13	n.a.
	girls	*+.07*	*+.84*	*n.a.*
	boys	*+.03*	*–.53*	*n.a.*
Van de Rijt & Van Luit (1998)	regular	n.a.	+.20	n.a.

Note. ES between [ ]: pretest difference > .5 SD, adequate statistical correction not possible.
* no statistical correction for pre-existing differences possible.
# mean difference approximated with available data, in which ES was set to 0 if the only information reported was that the difference was not significant.

'jump' strategy and the other (DI-s) instructing the 'split' strategy. Although in both DI-conditions outcomes were better than in the GI-condition, direct instruction in the jump strategy led to better performance than direct instruction in the split strategy (ES = .52). Second, in Timmermans et al. (2007) differential instruction effect for boys and girls were observed. For girls, guided instruction resulted in better performance, while for boys, direct instruction had better performance outcomes.

Finally, three studies reported results on *transfer*. Again, results were mixed: small to medium differences were found favoring guided instruction as well as favoring direct instruction.

Next to achievement outcomes, other outcomes investigated (not reported in the Summary Table) were strategy use and motivational/affective variables. With respect to strategy use (Kroesbergen & Van Luit, 2002, 2005; Milo & Ruijssenaars, 2005; Timmermans & Van Lieshout, 2003; Timmermans et al., 2007), findings showed that students who received direct instruction in a standard strategy more frequently used that strategy than students who received guided instruction. However, the latter students were not more flexible in their strategy use, meaning that they did not use their larger strategy repertoire adaptively to solve different problems. Finally, there were only minor instruction effects found on variables regarding motivation and affect (Kroesbergen et al., 2004; Milo, Seegers, Ruijssenaars, & Vermeer, 2004; Timmermans et al., 2007).

Remedial programs for low mathematics achievers in regular education

Willemsen (1994, study 2) compared two experimental remedial programs[6] for low mathematics achievers in regular education (grade 4) in the domain of written subtraction. These programs were the 'mapping' program aiming to remediate misconceptions that are at the basis of systematic computational errors, and the 'columnwise' program introducing an alternative strategy replacing the traditional subtraction algorithm. Students trained with the mapping program performed better than students trained with the columnwise program at posttest (ES = +.92) and at retention test (ES = +.64), medium to large differences. Furthermore, students in the mapping program made fewer systematic computational errors than students in the columnwise program (not in the Summary Table). In conclusion, the mapping program for remediating misconceptions that are at the basis of systematic computational errors had small to medium positive effects on written subtraction performance, compared to the columnwise program in which an alternative for the traditional algorithm was instructed.

Other instructional programs in regular education

Two studies compared the outcomes of two experimental programs in regular education students: Klein (1998) compared two instructional programs for addition and subtraction in grade 2, while Terwel, Van Oers, Van Dijk, and Van Eeden (2009; see also Van Dijk, Van Oers, Terwel, & Van Eeden, 2003) compared two instructional programs on 'mathematical modeling' in grade 5.

6 The comparisons with the control program are discussed in section 1.3.2.

First, Klein (1998; see also Blöte, Van der Burg, & Klein, 2001; Klein, Beishuizen, & Treffers, 1998) compared the *Realistic Program design* (RPD) with the *Gradual Program Design* (GPD) in instruction of 2-digit addition and subtraction. In the RPD, the focus was on letting students create and discuss their solution strategies. Realistic contexts for mathematics problems were used, and flexible strategy use was emphasized. Note that the authors contended that this program differed from the principles of realistic mathematics education, with instruction in the RPD being more directive and with students having more opportunity to practice. In the GPD, instruction was more traditional with knowledge being built up stepwise, starting from one basic addition and subtraction procedure: the jump strategy (see before).

No pretest was administered before the program started, so it was not possible to correct for pre-existing group differences. On the posttest, the performance differences (RPD - GPD) in speed tests (ES = +.19), strategy test (ES = +.15), paper-and-pencil addition and subtraction test (ES = +.10), standardized mathematics test LVS (CITO's Student Monitoring System - Mathematics; ES not estimable, difference was not significant), transfer test (ES = –.03), and retention test (ES = +.20) were all negligible to small favoring the RPD. On the speed tests, strategy test, and paper-and-pencil test, the program effects were assessed separately for low and high mathematics achievers. In the low achieving group, students in the RPD program performed better than those in the GPD, with a small to medium effect size (ES +.57, +.31, and +.36, respectively). In the high achieving group, students in the RPD performed better on the speed test (ES = +.47), almost the same on the strategy test (ES = +.02), and lower on the paper-and-pencil test (ES = –.15) than their counterparts in the GPD. However, before the start of the program, the high achievers in the RPD program performed better at the standardized mathematics test LVS (ES = +.50) than the high-achievers in the GPD, a pre-existing difference that could not be statistically accounted for. Furthermore, students in the RPD (low and high achievers) showed more flexible strategy use (not in the Summary Table) than students in the GPD. Finally, there were negligible to small differences in diverse affective and motivational outcomes, usually in the advantage of the RPD.

In summary, achievement outcomes differences were minor to small in favor of the Realistic Program Design over the Gradual Program Design. In addition, the RPD resulted in more flexible strategy use than the GPD, as well as in slightly better outcomes on affective and motivational measures.

Second, Terwel et al. (2009; see also Van Dijk et al., 2003) compared the outcomes of

two instructional programs on mathematical modeling in the domain of percentages and graphs. In the 'co-constructing/designing' program, students were instructed how to make models or representations of the open, complex problem situations that were offered, in co-operation with their classmates and under guidance of their teacher. In the 'providing' program, students were instructed to work with ready-made models that the teacher provided. Furthermore, students worked individually on the problems, followed by a classroom discussion. Note that the authors contended that this latter condition resembles common practice in Dutch education. Results showed that students in the co-constructing/designing program performed better than students in the providing program on problems on percentages and graphs (ES $= +.32$) and on transfer problems (ES $= +.55$). The co-constructing/designing program thus appeared to have a small to medium positive effect on achievement, compared to the providing program.

Summary

First, results of six studies on achievement outcomes of guided versus direct instruction in low mathematics performers (special and regular education) were mixed. Differences were found in both directions, and that even within a particular study on different outcome measures as well as between studies within one outcome measure. It seems that factors that were not measured or controlled for, such as the teacher, the composition of the class, and the program implementation, were more important than the instructional approach. The differential gender effect merits further research: in only one study, program effects were reported separately for boys and girls, and large differences in instruction effects were found. Finally, students receiving guided instruction showed a larger strategy repertoire than students receiving direct instruction, but did not use these strategies more adaptively or flexibly.

Second, for low mathematics achievers in regular education, a remedial program based on remediating misconceptions that are at the basis of systematic computational errors had medium to large positive effects on written subtraction performance, compared to a program in which an alternative (RME-based) solution strategy was instructed as replacement of the traditional algorithm. Finally, two studies in regular education showed that the more RME-based instructional programs (RPD in Klein, 1998, and co-constructing/designing program in Terwel et al., 2009) had negligibly small to medium positive effects on achievement, compared to the more traditional instructional

programs.

1.3.2 Experimental programs versus a control program

In this category of intervention studies, we discuss studies in which performance of students who followed an experimental program was compared to performance of students who followed a control program, commonly the regular mathematics curriculum. The majority of the programs addressed low mathematics achievers, both in special and in regular education. There were results of four studies in preschoolers (three with low math achievers), in three studies experimental remedial programs for low mathematics achievers were evaluated, and in the remainder four studies (three with low math achievers) experimental programs for 9 to 10 year-olds were compared to a control program. It is worth noting that besides the instructional program, usually also the instructor (external person in experimental program versus regular teacher in control group) and the instructional setting (small groups of students outside the classroom in experimental program versus whole class in the control group) differed between conditions. Therefore, it is not possible to assign found differences to any of these elements separately.

Preschoolers

In four studies, outcomes of students trained in an experimental program addressing early mathematical skills for preschoolers were compared with outcomes of peers in the regular preschool mathematics curriculum, that in practice was or was not characterized by the use of a specific mathematics textbook. Two studies were carried out in regular education (Poland & Van Oers, 2007; Van de Rijt & Van Luit, 1998), and the other two in special education (Schopman & Van Luit, 1996; Van Luit & Schopman, 2000).

Poland and Van Oers (2007; see also Poland, 2007) developed an experimental program for preschoolers in which schematizing activities were taught in meaningful situations. Preschoolers (not selected on their mathematics achievement level) who followed the program performed at about equal level as their control group peers on a mathematics test halfway the intervention (ES = –.05) and at the end of the intervention (ES = +.02). Eight months after the intervention, they performed better than the controls (ES = +.57), a medium to large difference. At the end of first grade (twelve months after the intervention), this difference reduced (ES = +.18) to a small advantage of the

experimental group. Furthermore, preschoolers in the experimental program showed more schematizing activities during and after the intervention than the controls (not in the Summary Table). In conclusion, the experimental program for preschoolers in which schematizing activities were taught in meaningful situations had a negligibly small to medium sized positive effect on first grade mathematics performance, compared to the control group.

In Van de Rijt and Van Luit (1998; see also section 1.3.1), low achieving preschoolers trained with the Additional Early Mathematics (AEM) program (either in the guided instruction or in the direct instruction variant) outperformed their control group peers in early mathematics skills, with large differences (ES = +1.06 and ES = +1.26, respectively). Thus, the AEM-program had a large positive effect on low achieving preschoolers' early mathematics skills.

There were two intervention studies with programs for preschoolers with low mathematics achievement level in special education. Schopman and Van Luit (1996) investigated the effect of an intervention program addressing counting to 10 as preparation for formal mathematics education that starts in first grade in special education. Preschoolers with a low mathematics level who were trained with this experimental program[7] performed better on a test of preparatory arithmetic skills (ES = +1.07) than preschoolers in the control group, a large effect. In the second study, Van Luit and Schopman (2000) extended the intervention program to more sessions and to numbers up to 15. Again, preschoolers in the experimental program performed better than their peers in the control group on a test of early numeracy (ES = +.73), and also on a transfer test (ES = +.22). In conclusion, in both studies, preschoolers who followed a preparatory program on counting skills to 10 or 15 performed better on a test of early numeracy than preschoolers in the control group, with medium to large differences.

Remedial programs

In three studies (one in special education, and two in regular education) the effects of an experimental remedial program compared to the regular mathematics curriculum were addressed.

7 In Schopman and Van Luit (1996) there were actually two experimental conditions: one with guiding instruction, and one with directing instruction. However, these instructional variants appeared not to differ from each other in practical implementation. Therefore, the results of these two experimental conditions were combine in the current review.

Harskamp and Suhre (1995) developed a remedial program for instruction in addition and subtraction below 100 for low mathematics achievers (10-11 years old) in special education. The program aimed to build on students' individual solution strategies, and it replaced two regular mathematics lessons a week. The program turned out to have a large positive effect compared to the control group that followed just the regular lessons on posttest and retention test achievement in addition and subtraction (ES = +3.22, but adequate statistical correction not possible), also separately for students with learning disabilities (LD) (ES = +3.13, but adequate statistical correction not possible) and for students with learning difficulties (MR) (ES = +3.69). Furthermore, the program also had a large positive effect on application problems in LD students (ES = +3.58, but adequate statistical correction not possible) and MR students (ES = +3.58). In conclusion, the experimental remedial program had large positive effects on addition and subtraction performance in LD and MR students in special education, compared to the control group.

Willemsen (1994) compared one (study 1) or two (study 2) experimental remedial programs[8] for low mathematics achievers in regular education (grade 4) in the domain of written subtraction with a control program, in which the subject matter was systematically rehearsed and practiced. In study 1, students in the 'mapping' program performed better at posttest than students in the control program (ES = +.32), a small to medium difference. In study 2, students in the mapping program again performed better than students in the control program at posttest (ES = +.74) and at retention test (ES = +.84), medium to large differences. Students in the columnwise program, however, performed somewhat less well than students in the control program at posttest (ES = –.17), but somewhat better at retention test (ES = +.20). Furthermore, students in the mapping program made fewer systematic computational errors than students in the control program (study 1 and 2, not presented in the Summary Table). In conclusion, the mapping program for remediating misconceptions that are at the basis of systematic computational errors had small to medium positive effects on written subtraction performance compared to the control program (systematic rehearsal and training). By contrast, the outcomes differences of the other experimental remedial program 'columnwise' versus the control program were only small and in both directions.

8 See section 1.3.1 for the comparison of the outcomes of the two experimental remedial programs.

Other studies

The results of four studies in which the outcomes of an experimental program were compared with the outcomes of a control group who followed the regular curriculum remain.

Keijzer and Terwel (2003; see also Keijzer, 2003) developed a program for instruction in fractions in fourth grade. This program was innovating compared to the RME-based textbook *Wereld in Getallen* (WIG) used in the control group on two aspects: the fractions model (number line versus circles or bars in WIG) and the instructional approach ('negotiation of meaning' in whole class discussions versus students working individually in WIG). On standardized LVS mathematics tests, differences between the groups were negligible in the domain of numbers and operations (ES = –.01), but students in the experimental group performed better than the controls in the domain of measures and geometry (ES = +.35), a small to medium difference. On fraction problems that were administered in interviews with standardized support, students in the experimental program performed better than the controls (uncorrected ES = +.52). In conclusion, the fractions program had no effects to medium sized positive effects on fourth graders' mathematics performance, compared to the control group.

Van Luit and Naglieri (1999) developed the MASTER program for students (age 10-12 years) in special education, focused on the development of solution strategies for multiplication and division up to 100. The program used principles of self-instruction, discussion, and reflection. Students who followed this program performed much better than students from the control group (ES = +2.16), which also held separately for LD students (ES = +2.50) and for MR students (ES = +3.08). Furthermore, there were also positive effects on a follow-up test (LD and MR students) and far transfer (only LD students; not in the Summary Table). In conclusion, the MASTER-training, aimed at development of strategies for multiplication and division below 100 making use of self-instruction, discussion, and reflection, had very large positive performance effects compared to the control group.

Finally, in both studies of Kroesbergen (Kroesbergen & Van Luit, 2002; Kroesbergen et al., 2004) from section 1.3.1 a modified version of the MASTER program was used. The comparisons between the experimental conditions (GI and DI) on the one hand and the control conditions on the other hand fit in the current section. In Kroesbergen and Van Luit (2002), posttest differences between students in the GI-condition and control

students were zero to large, with ES .00, +.89, and +.96 in automaticity, multiplication ability, and transfer, respectively. Comparisons between students in the DI-condition and control students should be evaluated with caution because pretest differences were too large to adequately statistically account for, but nevertheless all results favored the experimental program with ES +.51, +.46, and +.44 in automaticity, multiplication ability, and transfer, respectively. Similarly, in Kroesbergen et al. (2004) students in the experimental programs variant performed better than control students in automaticity (ES +.35 for GI and +.32 for DI) and in multiplication ability (ES = +.23 for GI and +.53 for DI). In conclusion, there were small, medium, and large positive effects of the program found compared to the regular curriculum, both in special education students and in regular education students.

Summary

The experimental programs investigated had negligibly small to large positive effects on mathematics performance, compared to the control group in which students usually followed the regular curriculum implemented by the regular teacher. These experimental programs each incorporated aspects of RME: development of solution strategies by self-instruction, discussion, and reflection; schematizing in meaningful situations; the number line as model; and whole-class discussion aiming at 'negotiation of meaning'. However, it is impossible to disentangle the effects of these elements from the general implementation differences between experimental and control conditions, such as instructor and instructional setting.

1.3.3 Supplemental programs for low mathematics achievers

There were two studies in which the effects of supplemental remedial or training programs for low mathematics achievers in regular education were investigated.

Harskamp, Suhre, and Willemsen (1993) compared performance of regular education students (grade 2 and 3) in six different combinations of a mathematics textbook based on RME principles on the one hand (*Wereld in Getallen*, *Operatoir Rekenen*, or *Rekenen & Wiskunde*), and a remedial program that was either structuralistic (more traditional: *Rekenspoor* or *Gouds Rekenpakket*) or RME-based (*Remelka*) on the other hand, with performance of students in the control group who did not receive this supplemental remedial training. Because the practical implementation of the six different combination

appeared not to differ from each other, we will not differentiate between them here. Supporting this equivalence was the result that performance of students in the six combinations of RME-textbook and RME-based or structuralistic remedial program did not differ from each other on either the number problems or the application problems. Compared to the control group, however, posttest performance in bare number problems was higher in the six remedial conditions in grade 2 (ES = +1.18, but pretest differences to large for adequate statistical correction) and in grade 3 (ES = +.39). On application problems, small positive effects of the remedial conditions compared to the control condition were found in grade 2 (ES = +.17) and in grade 3 (ES = +.24). In conclusion, the remedial programs seemed mainly to improve low mathematics achievers' abilities in number problems, irrespective of the didactical characteristics of the remedial program and the combination with didactical characteristics of the regular mathematics textbook.

Finally, Menne (2001) developed a supplemental 'productive practice' (in contrast to 'reproductive practice') program. This program addressed basic counting with units and tens, aiming to make students jump fluently and flexibly on the (empty) number line with varying step lengths. She implemented this program in grade 2 of regular education, and compared it to a control group of students who only followed their regular lessons. Students following the supplemental training program performed better than their control group peers: on LVS tests the ES was approximately +.44, and the performance difference between students who did and who did not follow the training program was larger for ethnic minority students (approximated ES = +.59) than for native Dutch students (approximated ES = +.41). In conclusion, the supplemental productive practice program had a small to medium positive effect on mathematics performance compared to the control group, in particular for ethnic minority students.

Summary

In these two studies, a positive effect of supplemental programs on students' achievement was found, compared to the control students who followed their regular mathematics lessons and did not receive extra training.

1.4 Curriculum studies

As said, curriculum studies are observational studies aiming to investigate differential achievement outcomes of different mathematics curricula, usually different mathematics

textbooks. They are discussed in three sections: domain-specific studies that address one specific delimited content domain of mathematic, large-scale curriculum studies carried out in 1980s that addressed general mathematics achievement, covering a range of mathematical domains, and differential outcomes by mathematics textbook in the Dutch national assessments. All study characteristics are in the Summary Table in Appendix 1.B.

1.4.1 Domain-specific curriculum studies

Two studies analyzed performance difference between students with different mathematics curricula on a specific content domain: one on addition and subtraction in special education (Van Luit, 1994) and the other on division in regular education (Van Putten, Van den Brom-Snijders, & Beishuizen, 2005).

Van Luit (1994) compared special education students' (age 9-11 years) addition and subtraction performance who followed a structuralistic or an RME-based curriculum. On the posttest[9] involving addition and subtraction without crossing tens, MR-students in the RME-based curriculum performed somewhat worse (ES = –.22; a small difference) than MR-students in the structuralistic curriculum, while in addition and subtraction with crossing tens there was only a negligible difference (ES = +.04). In LD-students, performance differences were in disadvantage of the RME-based curricula, with respectively ES = –.62 and ES = –1.00. On problems involving a realistic context, performance differences between LD-students in structuralistic or RME-based curricula were minor (ES = –.08). In conclusion, addition and subtraction performance of special education students (MR and LD) in RME-based curricula was equal to or lower than in structuralistic curricula.

Van Putten et al. (2005) compared fourth graders' division performance with two different textbooks, *Rekenen & Wiskunde* (R & W) and *Wereld in Getallen* (WIG) in regular education. Both textbooks are based on RME-principles, but WIG has a more (pre-)structured learning trajectory for division than R & W. Halfway fourth grade, R & W students had lower performance than WIG students (ES = –.43), while at the end of grade four the performance difference was reversed (ES = +.35). Furthermore, strategy use (not in the Summary Table) developed positively over time on the aspects schematizing (R & W more increase than WIG) and number relations (R & W and WIG same increase,

9 Although a pretest was administered, differences were not corrected for, because at the time the pretest was administered the students already had six months instruction in addition and subtraction according to a structuralistic or RME-based curriculum.

but WIG higher overall score). These results from Dutch students were also compared with UK students from the same age (Anghileri, Beishuizen, & Van Putten, 2002; not in the Summary Table). In the UK, the learning trajectory for division is characterized by a rather abrupt transition of informal solution strategies to the traditional long division algorithm. By contrast, in the Dutch RME-based textbooks R & W and WIG, informal strategies are progressively schematized toward more structured and efficient strategies (not the traditional algorithm). At the end of fourth grade, Dutch students outperformed the UK students, an indication that the progressive schematization of informal solution strategies was effective. In summary, Dutch students with mathematics textbook *Rekenen & Wiskunde* had a lower division performance than students with the textbook *Wereld in Getallen* halfway fourth grade, but reversed this to an advantage at the end of grade four. Both groups of Dutch students outperformed UK counterparts.

1.4.2 Large-scale curriculum studies from the 1980s

In the 1980s, two large-scale curriculum studies were carried out, in which modern (at that time) mathematics textbooks were compared to traditional textbooks.

First, the MORE-project was carried out at the Freudenthal Institute, a study in which students were longitudinally followed from first to third grade (Gravemeijer et al., 1993). The two mathematics curricula compared were the traditional textbook series *Naar Zelfstandig Rekenen* (NZR) and the modern RME-based textbook series *Wereld in Getallen - edition 1* (WIG-1). Results were corrected for students' mathematics level in first grade, socio-economical background, and intelligence scores. On general mathematics, WIG-1 students performed at approximately equal level as NZR students in grade 1 (ES = –.02), but were outperformed in grade 2 (ES = –.10; negligible to small difference) and in grade 3 (ES = –.32; small to medium difference). On automatization WIG-1 students were outperformed by NZR students with medium to large differences, in grade 2 (ES = –.60) and in grade 3 (ES = –.58). Furthermore, investigation of the implementation of the two textbooks (not in the Summary Table) showed that the instruction in NZR-teachers was reasonably mechanistic (traditional), while the instruction by WIG-teachers was RME-based only to a limited extent. Thus, the implemented curriculum in NZR-teachers was more in accordance with the didactical theory in the textbook series than in WIG-teachers. In conclusion, in grade 1 to 3, students in the RME-based *Wereld in Getallen - edition 1* curriculum performed negligibly to substantially lower than students in the

traditional *Naar Zelfstandig Rekenen* curriculum, in particular on automatization.

Second, Harskamp (1988) compared sixth graders' achievement outcomes with 8 different mathematics curricula (textbook series), 3 of which he classified as 'modern' (NB. not *Wereld in Getallen*) and 5 as 'traditional' (among which *Naar Zelfstandig Rekenen*). Corrected for intelligence scores, performance differences on the CITO End of Primary School Test (ES = +.09) and at mathematics tests developed at RION (ES = +.06) were negligible to small in the advantage of modern textbooks. Furthermore, there were implementation factors associated with mathematics performance (not in the Summary Table): two general factors with a positive relation with performance were the number of mathematics lessons per week and the percentage of students for whom the basic subject matter from the textbook was covered, and two content-specific factors were variation in subject matter (positive relation) and differentiation in subject matter (negative relation). In summary, students with 'modern' mathematics textbook series performed somewhat better than students with 'traditional' mathematics textbook series.

1.4.3 Differential mathematics outcomes by mathematics textbook in national assessments

In CITO's national assessments of mathematics education, usually performance differences between students who were taught with different textbook series are reported.

The most recent assessment halfway primary school (grade 3) carried out in 2003 analyzed performance differences between students with seven different mathematics textbooks, corrected for several background variables (Kraemer, Janssen, Van der Schoot, & Hemker, 2005). Students who were instructed with *Talrijk* and *Rekenrijk* performed best, students with *Wereld in Getallen - edition 1* and *Rekenen & Wiskunde* had the lowest performance level. The difference between highest and lowest average performance by textbook was medium to large (ES = +.64). Because all textbooks used were based on RME-principles, it was not possible to compare performance between RME-based and traditional curricula. What was reported, though, is that the shift in mathematics textbooks shares between the cycles of 1997 and 2004 had in general a small positive effect on students' mathematics performance (ES = +.18). It is hard to characterize this shift in terms of RME-based versus traditional curricula, because in the 1997 cycle less that 5% of the textbooks used was still traditional.

The most recent assessment at the end of primary school (grade 6) carried out in 2004

did not report performance differences between different mathematics textbooks used, because 80% of all schools started using a new textbook in reaction to the introduction of the new currency (the euro) in 2002, and therefore sixth graders had experienced a change in textbook in their primary school trajectory (J. Janssen et al., 2005). Only the summative effects of shifts in mathematics textbooks shares were reported. In the period 1997 to 2004 this summative effect was positive but very small (ES = +.12), in the period 1992 to 2004 this effect was also positive but small (ES = +.18). Thus, in total, the shift in market share of mathematics textbooks between 1992 and 2004, from 37% to 100% RME-based textbooks, had a negligible to small positive effect on sixth graders' mathematics performance.

In the third assessment cycle (1997) at the end of primary school, performance differences by mathematics textbook used were still reported (J. Janssen, Van der Schoot, Hemker, & Verhelst, 1999). Students instructed with *Wereld in Getallen - edition 2* performed best, students with *Niveau Cursus Rekenen* and *Naar Zelfstandig Rekenen* performed at the lowest level. The difference between highest and lowest average performance by textbook was medium (ES = +.53). Importantly, differences *within* a curriculum type (RME-based or traditional) were larger than *between* the two curriculum types.

In summary, in third grade, mathematics performance of students instructed with one of seven different RME-based mathematics textbooks differed from each other. Similarly, in sixth grade, the 1997 cycle showed performance differences by mathematics textbook, in which differences within a curriculum type (RME-based or traditional) were larger than between curriculum types. Both in third grade and in sixth grade, the shift in market shares of mathematics textbooks over time had a very small to small positive effect on mathematics performance. In third grade, this shift in textbooks used was almost entirely within the spectrum of RME-based curricula. In sixth grade, this entailed both shifts from traditional and hybrid textbooks toward RME-based textbooks, as well as shifts within the different RME-based textbooks.

Summary

The domain-specific curriculum studies showed that in special education, students in a structuralistic curriculum outperformed students in an RME-based curriculum on addition and subtraction. In regular education, fourth graders' division performance

with two RME-based textbooks showed a varying pattern over the school year, but both curricula seemed more effective than the UK approach that was more traditionally oriented.

Within the large-scale curriculum studies covering many domains of mathematics, one of the studies from the 1980s showed that students with the RME-based textbook *Wereld in Getallen - edition 1* (WIG-1) were outperformed by students with the traditional textbook *Naar Zelfstandig Rekenen* (NZR). By contrast, the other curriculum study from the 1980s showed that students with 'modern' textbooks slightly outperformed students with 'traditional' textbooks.

The Dutch national assessments of mathematics education showed first and foremost that there are no univocal results in comparing students' performance with RME-based and traditional mathematics curricula. There were substantial performance differences within both curricula types. Furthermore, the shift from older to newer RME-based textbooks resulted in somewhat better performance in grade 3. In grade 6, the shift from traditional, hybrid, and (older) RME-based textbooks toward (newer) RME-based textbooks also had a small positive effect on mathematics performance.

1.5 SUMMARY, CONCLUSIONS, AND IMPLICATIONS

The results of all empirical studies reviewed taken together do not give an unequivocal picture on the relation between mathematics instruction and mathematics performance in the Netherlands. There is remarkably little research that allows for well-grounded and univocal conclusions, a similar conclusion as was reached in other research syntheses (e.g., Kroesbergen & Van Luit, 2003; National Mathematics Advisory Panel, 2008; Slavin & Lake, 2008). *Intervention studies* compare the effects of different mathematics interventions, i.e., instructional programs. The intervention studies reviewed do not yield firm conclusions, because they are limited in content domain, sample size, duration or magnitude of the intervention, and range of outcome variables. In addition, usually several didactical and instructional aspects were varied simultaneously, making it impossible to disentangle their effects. *Curriculum studies* make a large-scale comparison between performance of students who were instructed with different mathematics curricula or textbook series. However, these studies are limited in the amount of control on the practical implementation of the curricula and in correction for confounding variables: they are carried out in everyday educational practice.

Furthermore, the results of the curriculum studies did not point univocally in one direction either. More generally speaking, the idea of extending the concept of 'evidence-based medicine' towards the field of education (e.g., National Mathematics Advisory Panel, 2008) is hampered by many practical limitations.

Regarding the public debate on RME-based mathematics education versus traditional education, it is important to note that – even if it would be possible to speak of 'the' RME approach and 'the' traditional approach – the studies reviewed rarely compared these instructional approaches directly. Rather, specific elements that can be characterized as reform-based/constructivistic or more traditional/mechanistic were varied. The effects found were, generally speaking, small. Synthesizing all results, one could say that a very small advantage of RME-based programs was found, in particular regarding the higher-order goals such as flexibility. This advantage, however, was too small to draw firm conclusions. In addition, there are studies that favored more traditional programs instead.

The lack of a firm general conclusion, however, does not preclude conclusions on some more specific patterns. First, it is striking that *within* a type of instructional approach performance differences were larger than *between* instructional approaches. Apparently, didactical principles play a less important role than the practical implementation by the teacher and the teacher-student interaction, implications that agree with the findings of for example Slavin and Lake (2008). Second, more time spent on mathematics education leads to better performance, as comes forward from the positive results of remedial and training programs that are supplemental to the regular curriculum, which is congruent with the notion of 'opportunity to learn' being the most important predictor of mathematics achievement outcomes (Hiebert & Grouws, 2007). Third, with educational time held equal, experimental programs implemented in small groups of students outside the classroom had positive effects compared to the regular educational practice. Similarly, Slavin and Lake (2008) also reported a positive effect of small-group tutoring. Fourth, in the studies reviewed there was a lot of attention for low mathematics performers. These students seemed to benefit less from a free form of instruction and to have a larger need for a more directing role of their teacher in their learning process, similar to findings of international reviews (Gersten et al., 2009; Kroesbergen & Van Luit, 2003; Swanson & Carson, 1996; Swanson & Hoskyn, 1998). Much less research has been done into the relation between mathematics instruction and mathematics performance in medium and high performers. Likewise, little is known

about differential instruction effects for boys and girls.

The main implication of the current research synthesis may be that the key to improving mathematics education seems to be in the teacher quality (KNAW, 2009). The crucial role of the teacher in mathematics education also comes forward from international reviews (Hiebert & Grouws, 2007; Hill et al., 2007; Kroesbergen & Van Luit, 2003; Slavin & Lake, 2008; Verschaffel et al., 2007). It is widely accepted that teachers differ in their effectiveness, although empirical evidence is weak (Nye, Konstantopoulos, & Hedges, 2004). Furthermore, the teacher is the pivot in the learning/teaching process. We put high demands on our teachers, in particular in instructional approaches that focus on the input of and interaction with the students, such as in mathematics education reform (see also Stein et al., 2007). These demands necessitate an investment in teacher training to consolidate and improve teachers' (pedagogical) content knowledge. Noteworthy in this light is the finding from TIMSS-2007 that the Dutch teachers participated the least in professional development in mathematics from all participating countries (Mullis et al., 2008). Furthermore, Dutch pre-service teacher training programs has received a lot of criticism, but promising developments (e.g., future teachers need to show a sufficient level of mathematical content knowledge and skills in a mandatory test, and a national framework has been developed of knowledge that future mathematics teachers need to have) have taken place recently. A final conclusion is that more research into effective instructional elements and ways to improve mathematics instruction and teachers' quality is indispensable for further educational recommendations.

APPENDIX 1.A STUDY CHARACTERISTICS OF INTERVENTION STUDIES

references	domain	participants	intervention	duration and implementation	design / procedure	corrected	posttest results	ES
Harskamp & Suhre (1995)	addition and subtraction < 100	*N* = 24 4 schools M = 10.2 years (LD) M = 11.4 years (MR) special education low math performers	1. remedial program 2. control (regular math curriculum; all structuralistic or traditional textbooks)	remedial program: 13 weeks, two 45-minutes lessons per week replacing two regular math lessons implementation remedial program by remedial assistant with 2 students outside the classroom implementation control curriculum by regular teacher in whole class	pretest - intervention - posttest selection and assignment procedure unclear, but assignment to conditions within schools	pretest add. and subtr. < 100	addition and subtraction < 100 (posttest and retention test) remedial > control [+3.22] *LD students [+3.13]* *MR students +3.69*	
						pretest add. and subtr. < 100	application problems remedial > control ?? *LD students [+3.58]* *MR students +3.58*	
Harskamp, Suhre & Willemsen (1993)	grade 2: addition & subtraction < 20 grade 3: addition & subtraction < 100, and multiplication tables	grade 2: *N* = 357 89 schools grade 3: *N* = 242 81 schools regular education low math performers	1. remedial: 6 combinations of 1 out of 3 remedial programs and 1 out of 3 math textbooks (self-selected) 2. control (no remedial program)	duration remedial program: 12 blocks of lessons (in total 16 hours) in 4-5 months implementation remedial program by trained remedial assistants	pretest - intervention - posttest assignment to remedial programs at school level, but precise procedure unclear	pretest number problems	numerical problems add. and subtr. (and mult. in grade 3) gr. 2: remedial > control [+1.18] gr. 3: remedial > control +.39	
						pretest application problems	performance application problems gr. 2: remedial – con. n.s. +.17 gr. 3: remedial – con. n.s. +.24	

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Keijzer (2003) Keijzer & Terwel (2003)	fractions	*N* = 20 2 classes (1 school) grade 4 (9-10 years old) regular education	1. experimental fractions program (exp) 2. control (regular math textbook *Wereld in Getallen*)	one school year experimental program implemented by researcher control curriculum by regular teacher	pretest - intervention - posttest assignment to conditions on class level, not random post-hoc matching on student level	pretest LVS – numbers and operations	LVS – numbers and operations exp − control n.s.	 −.02
						pretest LVS – measuring and geometry	LVS - measuring and geometry exp − control n.s.	 +.35
						no correction	skills in fractions (interviews) exp > control	 +.52
Klein (1998) Klein, Beishuizen & Treffers (1998) Blöte, Van der Burg & Klein (2001)	two-digit addition and subtraction < 100	*N* = 275 10 classes (9 schools) grade 2 regular education	1. Realistic Program Design (RPD) 2. Gradual Program Design (GPD)	one school year, program replacing 75% of regular math curriculum implementation by regular teacher in whole class	intervention - intermediate test - posttest random assignment to programs on class level matched pairs of classes (based on LVS general math scores)	no correction	speed-tests add. and subtr. RPD − GPD n.s. *low math performers* *high math performers*	 +.19 *+.57* *+.47*
						no correction	strategy-test add. and subtr. RPD − GPD n.s. *low math performers* *high math performers*	 +.15 *+.31* *+.02*
						no correction	p & p test add. and subtr. RPD − GPD n.s *low math performers* *high math performers*	 +.10 *+.36* *−.15*
						no correction	LVS-test End Grade 2 RPD − GPD n.s.	 ??
						no correction	transfer (add. / subtr. > 100): RPD − GPD n.s.	 −.03
						no correction	retention (add. / subtr < 100) RPD > GPD	 +.20

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Kroesbergen & Van Luit (2002)	multiplication up to 10 x 10	*N* = 75 7 schools M = 9 years regular education + special education low math performers	1. guided instruction (GI) 2. structured instruction (SI) 3. control (regular math curriculum, mixture of different instructional approaches)	4 months, GI and SI involved 30 sessions of 30 minutes, replacing 2 regular math lessons a week GI and SI implemented by researcher in small groups of students control condition by regular teacher in whole class	pretest - intervention - posttest random assignment to experimental condition (GI or SI) vs. control condition on student level random assignment within experimental conditions to either GI or SI on class level	pretest multiplication automaticity	multiplication automaticity GI – control: n.s. SI > control GI < SI *regular ed.: GI > SI* *special ed.: GI < SI*	.00 [+.51] [–.51] *[+.64]* *[–2.42]*
						pretest multiplication ability	multiplication ability GI > control SI > control GI > SI *regular ed.: GI > SI* *special ed.: GI > SI*	+.89 [+.46] +.43 *[+.85]* *+.32*
						pretest transfer	transfer (up to 10 x 20) GI > control SI > control GI > SI *regular ed: GI > SI* *special ed: GI > SI*	+.96 [+.44] +.52 *[+1.03]* *+.36*
Kroesbergen, Van Luit & Maas (2004)	multiplication	*N* = 265 23 schools M = 9.7 years regular education + special education, only low math performers	1. constructivist instruction (CI) 2. explicit instruction (EI) 3. control (regular math curriculum)	4 months, GI and SI involved 30 sessions of 30 minutes, replacing 2 regular math lessons a week GI and SI implemented by externals in small groups of students control implemented by regular teacher in whole class of students	pretest - intervention - posttest random assignment to experimental condition vs. control condition on student level random assignment within experimental conditions to either GI or SI on class level	pretests, months of prior multiplication instruction, gender, general math level, IQ, school type (regular / special)	multiplication automaticity CI > control EI > control CI – EI: n.s.	+.35 +.32 +.03
							multiplication ability CI > control EI > control CI < EI	+.23 +.53 –.30

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Menne (2001)	mental calculation < 100	*N* = 225 12 schools (with sufficient number of low math achievers) grade 2 regular education	1. productive training program 2. control (regular math curriculum, no extra training sessions)	whole school year, at least 3 times a week 15 minutes training program implemented by regular teacher	pretest - intervention - posttest assignment to program at school level, but precise procedure unclear	pretest LVS (general math)	LVS-tests (general math) [NB. means and SDs approximated] training > control *ethnic minorities* *native Dutch*	 +.44 *+.59* *+.41*
Milo, Ruijssenaars & Seegers (2005)	addition and subtraction < 100	*N* = 70 3 schools M = 9.8 years special education	1. directing instruction - jump (DI-j) 2. directing instruction - split (DI-s) 3. guiding instruction (GI)	6 months, 30 lessons, replacing 2 lessons a week implemented by trained university students in groups of 3-5 students outside the class	pretest -intervention - posttest assignment of groups of students (uniform MR or LD) to programs procedure selection and assignment unclear	pretest add. and subtr. < 100	performance add. and subtr. GI < DI-j GI − DI-s: n.s. DI-j − DI-s: n.s.	 −.73 −.21 +.52
						no correction	transfer (add. / subtr. > 100) GI − DI-j: n.s. GI − DI-s: n.s. DI-j − DI-s: n.s.	 +.07 +.59 +.52
Poland (2007) Poland & Van Oers (2007)	general mathematics	*N* = 133 6 schools (with 'develop-mental education' approach) preschool through grade 1 regular education	1. experimental program 'schematizing' 2. control (standard preschool curriculum)	one school year (2nd preschool year) experimental program implemented by regular teacher supported by researcher control by regular teacher	pretest - intervention - posttest matching schools in pairs, assignment procedure within pairs unclear	pretest number sense	general math: halfway exp − control n.s.	 −.05
						pretest number sense	general math: end exp − control n.s.	 +.02
						pretest number sense	general math: 8 months after exp > control n.s.	 +.57
						pretest number sense	general math: 12 months after exp − control n.s.	 +.18

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Schopman & Van Luit (1996)	preparatory arithmetic skills	*N* = 60 unknown number of preschools preschoolers (5-7 years old) special education low math performers	1. experimental program (guiding or directing instruction*) 2. control (standard curriculum)	experimental program: 3 months, 13 lessons, ½ hour sessions twice a week implementation experimental program: unclear by whom, in groups of 4 students control curriculum by regular teacher	pretest - intervention - posttest assignment to program at student level with matching of students, but precise assignment procedure unclear	pretest preparatory arithmetic skills	preparatory arithmetic skills exp > control	+1.07
***** In the practical implementation, the instructional approaches did not differ from each other. Therefore, these groups were combined here.								
Timmermans & Van Lieshout (2003)	subtraction < 100	*N* = 16 2 schools M = 10.5 years special education students low performing in subtraction	1. guided instruction (GI) 2. direct instruction (DI)	34 lessons (of which 24 specific GI vs. DI) all lessons implemented by same trainer in groups of 4 students	pretest - intervention - posttest students within classes assigned to GI vs. DI students in DI and GI matched on pretest and age	pretest speed-test	speed-test addition (without regrouping) GI < DI	−.99
						pretest speed-tests	remaining speed-tests add. and subtr. GI − DI n.s.	??
						pretest performance	performance test GI − DI n.s.	??
						no correction	transfer (add. / subtr. > 100 without regrouping) GI < DI	−.96
						no correction	transfer (add. / subtr. > 100 with regrouping) GI − DI n.s.	−.18

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Timmermans, Van Lieshout & Verhoeven (2007)	subtraction < 100	*N* = 40 5 schools M = 9.3 years regular education students low performing in subtraction	1. guiding instruction (GI) 2. directing instruction (DI)	34 lessons (of which 24 specific GI vs. DI), 2 lessons a week within a school, all GI / DI lessons implemented by same trainer in groups of 4 students	pretest -intervention - posttest assignment: matched pairs of students within classes, within each pair random assignment to GI vs. DI	pretest speed-tests	speed-tests add. and subtr. GI − DI n.s. *girls: GI − DI n.s.* *boys: GI − DI n.s.*	 +.05 *+.07* *+.03*
						pretest power test	power test add. and subtr. GI − DI n.s. *girls: GI > DI* *boys: GI < DI*	 +.13 *+.84* *−.51*
						no correction	strategy- test subtraction GI − DI n.s. *girls: GI > DI* *boys: GI − DI n.s.*	 +.17 *+.60* *−.25*
Van de Rijt & Van Luit (1998)	early mathematics	*N* = 136 20 schools M = 5.9 years regular education low math performers	1. AEM program - guiding instruction (AEM-GI) 2. AEM program - structured instruction (AEM-SI) 3. control (regular math curriculum)	13 weeks, AEM-GI and AEM-SI twice a week 30 minutes lesson replacing the regular math curriculum AEM-GI and AEM-SI: instruction in groups of 4-5 students, probably (?) by regular teacher	pretest -intervention - posttest matching of students on pretest, age, and gender assignment to conditions on student level	pretest early mathematics	early mathematics AEM-GI > control AEM-SI > control AEM-GI − AEM-SI: n.s.	 +1.06 +1.26 +.20
Van Dijk, Van Oers, Terwel & Van Eeden (2003); Terwel, Van Oers, Van Dijk & Van Eeden (2009)	mathe-matical modeling: percentages and graphs	*N* = 238 10 classes (8 schools) grade 5 10-11 years regular education	1. program 'co-constructing / designing' (CCD) 2. program 'providing' (P)	3 weeks, 13 daily lessons of 1 hour implementation by regular teacher	pretest - intervention - posttest random assignment to programs on class level	standardized pretest general math	percentages and graphs CCD > P	 +.33
						standardized pretest general math	transfer: CCD > P	 +.55

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Van Luit & Naglieri (1996)	multiplication and division < 100	*N* = 84 unknown number of classes special education M = 10.8 years (LD) M = 12.7 years (MR) low math performers	1. experimental MASTER program 2. control (standard curriculum)	MASTER program: 17 weeks, 3 times a week 45-minutes lesson, replacing all regular math lessons implementation MASTER program by remedial teacher in groups of 5-6 students outside the class control by regular teacher	pretest - intervention - posttest random assignment to program at student level, within type of student (LD or MR)	pretest multiplication and division	multiplication and division < 100 exp > control *LD students* *MR students*	+2.16 *+2.50* *+3.08*
Van Luit & Schopman (2000)	early numeracy	*N* = 124 9 schools preschoolers (5-7 years old) special education low math performers	1. experimental intervention program 2. control (standard curriculum)	exp program: 6 months, 20 lessons, ½ hour sessions twice a week, replacing all regular math lessons implementation experimental program by trained assistant, in groups of 3 students control curriculum by regular teacher	pretest - intervention - posttest assignment to program at student level with matching of students, but precise assignment procedure unclear	pretest early numeracy no correction	early numeracy exp > control transfer exp – control n.s.	+.75 +.22

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Willemsen (1994) - study 1	written subtraction	*N* = 40 5 classes (5 schools) grade 4 regular education low math performers	1. remedial program 'mapping' 2. control remedial program 'systematic practice'	10 weeks, 10 50-minutes lessons mapping program implemented by remedial teacher in groups of 3-5 students control program implemented by regular teacher	pretest - intervention - posttest random assignment to conditions on class level	pretest written subtraction	written subtraction mapping > control	 +.32
Willemsen (1994) - study 2	written subtraction	*N* = 34 3 schools grade 4 regular education low math performers	1. remedial program 'mapping' 2. remedial program 'columnwise' 3. control remedial program 'systematic practice'	10 weeks, 10 50-minutes lessons remedial programs (1+2) implemented by remedial teacher in groups of 3-5 students control program implemented by regular teacher	pretest -intervention - posttest random assignment to conditions on student level	pretests written subtraction & mental computation	written subtraction mapping > control column − control: n.s. mapping > column	 +.75 −.17 +.92
						pretests written subtraction & mental computation	retention written subtraction mapping > control column − control: n.s. mapping > column	 +.84 +.20 +.64

APPENDIX 1.B STUDY CHARACTERISTICS OF CURRICULUM STUDIES

references	domain	participants	curriculum	duration	corrected	results	ES
Gravemeijer et al. (1993) [MORE-study]	general mathematics	*N* = 430 18 schools longitudinal setup grade 1 – grade 3 regular education	1. RME-based curriculum *Wereld in Getallen – edition 1* (WiG-1) 2. traditional curriculum *Naar Zelfstandig Rekenen* (NZR)	one to three school years	grade 1 math level, SES, intelligence	general mathematics grade 1 grade 2 grade 3	-.02 -.10 -.32
					grade 1 math level, SES, intelligence	automatizing grade 2 grade 3	-.60 -.58
Harskamp (1988)	general mathematics	*N* = 2579 120 schools grade 6 regular education	math textbook used 1. modern (3 textbook series) 2. traditional (5 textbook series)	school years up to measurement	intelligence	CITO test (general math) modern vs. trad. n.s.	+.09
					intelligence	RION-tests (general math) modern vs. trad. n.s.	+.06
Janssen, Van der Schoot, Hemker & Verhelst (1999) = PPON grade 6 - 1997	general mathematics	*N* = 7890 (321 schools) in 1987 *N* = 4335 (241 schools) in 1992 *N* = 5314 (253 schools) in 1997 grade 6 regular education	math textbook used (8 different)	school years up to measurement	students' SES, gender, nr. of school years; school's SES; assessment cycle	general mathematics performance compared to *NZR* *Wereld in Getallen - ed. 2* *Nieuw Rekenen* *Pluspunt* *Rekenen & Wiskunde* *Operatoir Rekenen - ed. 1* *Wereld in Getallen - ed. 1* *Niveau Cursus Rekenen* *Naar Zelfstandig Rekenen*	+.53 +.31 +.29 +.25 +.23 +.22 +.02 ref = 0
Janssen, Van der Schoot, & Hemker (2005) = PPON grade 6 - 2004	general math	*N* = 4335 (241 schools) in 1992 *N* = 5314 (253 schools) in 1997 *N* = 3078 (122 schools) in 2004 grade 6, regular education	math textbook used (NB. 80% new textbooks in 2004)	school years up to measurement	students' SES, gender, nr. of school years; school's SES;	general math performance: summative effect of shift in textbooks 2004 vs. 1997 2004 vs. 1992 1997 vs. 1992	+.12 +.18 +.06

references	domain	participants	curriculum	duration	corrected	results	ES
Kraemer, Janssen, Van der Schoot, & Hemker (2005) = PPON grade 3 - 2003	general math	*N* = 3350 (164 schools) in 1992 *N* = 5972 (130 schools) in 1997 *N* = 2032 (77 schools) in 2004 grade 3 regular education	math textbook used (7 different)	school years up to measurement	students' origin, gender, nr. of school years; assessment cycle	general math performance compared to *R&W* *Talrijk* *Rekenrijk* *Wereld in Getallen - eds. 2+3* *Pluspunt - eds. 1+2* *Wereld in Getallen - ed. 1* *Rekenen & Wiskunde*	 +.64 +.62 +.46 +.44 +.18 ref = 0
						summative effect of shift in textbooks 2003 vs. 1997	 +.18
Van Putten, Van den Brom-Snijders & Beishuizen (2005)	multidigit division	*N* = 256 10 schools grade 4 regular education	1. math textbook *Rekenen & Wiskunde* (R&W) 2. math textbook *Wereld in Getallen* (WiG)	school years up to measurement	speed-test performance	halfway grade 4 R&W < WiG	 -.43
					speed-test performance	end grade 4 R&W > WiG	 +.35