



Implementation quality of principles of reciprocal teaching in whole-classroom settings: a two-year study with low-achieving adolescents

M. Okkinga, A. J. S. van Gelderen, E. van Schooten, R. van Steensel & P. J. C. Sleegers

To cite this article: M. Okkinga, A. J. S. van Gelderen, E. van Schooten, R. van Steensel & P. J. C. Sleegers (2021) Implementation quality of principles of reciprocal teaching in whole-classroom settings: a two-year study with low-achieving adolescents, *Reading Psychology*, 42:4, 323-363, DOI: [10.1080/02702711.2021.1887019](https://doi.org/10.1080/02702711.2021.1887019)

To link to this article: <https://doi.org/10.1080/02702711.2021.1887019>



Published online: 15 Mar 2021.



Submit your article to this journal [↗](#)



Article views: 353




View related articles [↗](#)



View Crossmark data [↗](#)



Implementation quality of principles of reciprocal teaching in whole-classroom settings: a two-year study with low-achieving adolescents

M. Okkinga^a , A. J. S. van Gelderen^{a,b}, E. van Schooten^{a,b},
R. van Steensel^{c,d} and P. J. C. Slegers^e

^aResearch Centre Urban Talent, Rotterdam University of Applied Sciences, Rotterdam, The Netherlands; ^bKohnstamm Institute, Amsterdam, The Netherlands; ^cDepartment of Psychology, Education & Child Studies, Erasmus University, Rotterdam, The Netherlands; ^dFaculty of Humanities, Free University Amsterdam, Amsterdam, The Netherlands; ^eBMC Advice, Amersfoort, The Netherlands

ABSTRACT

Low-achieving adolescents are known to have difficulties with reading comprehension. This article discusses whether principles of reciprocal teaching can improve low-achieving adolescents' reading comprehension in whole-classroom settings and to what extent treatment effects are dependent on implementation quality. Over the course of two years, experimental teachers ($n = 10$) were given training and coaching aimed at using principles of reciprocal teaching, while control teachers ($n = 10$) used their regular teaching method. Observations of teacher implementation were focused on instruction of reading strategies, modeling, and support of group work, and were performed in both experimental and control classes, comprising a total of 238 students (grade 7). The study shows that overall, there is no effect of the treatment on adolescent low-achievers' reading comprehension. Interestingly however, the principle of modeling positively moderated the effect of reciprocal teaching. In addition, results suggest that the quality of implementation of reciprocal teaching in whole-classroom settings should receive more attention.

ARTICLE HISTORY

Received 12 September 2019
Accepted 1 February 2021

Reading comprehension is an essential skill for all students in their school careers. However, many students in secondary education, especially low-achieving students, struggle with reading comprehension (e.g., Dutch Education Inspectorate, 2008; Kordes, Bolsinova, Limpens, & Stolwijk, 2013; Organization for Economic Co-operation and Development [OECD], 2003; OECD, 2014), resulting in difficulties with several school subjects. Not being able to comprehend texts can have serious implications for students' educational success and for their later societal careers. Long-term, evidence-based reading comprehension programs that target low-achieving adolescents are thus of vital importance (Edmonds et al., 2009; National Reading Panel (US), 2000; Slavin et al., 2008). A well-known evidence based method for teaching reading comprehension for low-achieving adolescents is called reciprocal teaching (Palincsar & Brown, 1984). This method was successfully tested in numerous experiments in which researchers or other experts were instructing small groups of students (e.g., Palincsar & Brown, 1984; Palincsar, Brown, & Martin, 1987; Rosenshine & Meister, 1994; Spörer, Brunstein, & Kieschke, 2009).

Therefore, there are many attempts to include principles of reciprocal teaching in reading comprehension curricula. However, evidence concerning the effectiveness of reciprocal teaching in whole-classroom settings in which students' regular teachers are responsible for delivering the intervention is mixed. In this study, in which we used an existing program, we investigated reading comprehension instruction including principles of reciprocal teaching in whole-classroom settings for low-achieving adolescents over the course of two years (from grade 7 to grade 8). From previous studies it appears that the quality of implementation of reading comprehension programs in such whole-classroom settings is an important determinant of success (Chiu, 1998; De Boer, Donker, & Van der Werf, 2014; Swanson, Wanzek, Haring, Ciullo, & McCulley, 2013; Vaughn et al., 2013). Therefore, we analyzed the moderating effects of implementation quality of principles of reciprocal teaching on students' reading comprehension.

Principles of Reciprocal Teaching

Reciprocal teaching (Palincsar & Brown, 1984) is a widely used method of instructing and guiding learners in reading comprehension. It consists of a set of three related instructional principles: a) teaching comprehension-fostering reading strategies, including predicting, question-generating, summarizing, and clarifying; b) expert modeling, scaffolding and fading; and c) students practicing and discussing reading strategies with other students, coached by the teacher. Reciprocal teaching is based upon a gradual shift of responsibility for the learning process from

teacher to student, which includes the teacher explicitly modeling the use of reading strategies (Rosenshine & Meister, 1994) as well as scaffolding the application of reading strategies within the groups of students working together. It is assumed that by gradually fading teacher's support, students become increasingly more capable of regulating their own reading process.

Many studies have confirmed the positive effects of reciprocal teaching (Kelly, Moore, & Tuck, 1994; Rosenshine & Meister, 1994; Spörer et al., 2009). In a review by Rosenshine and Meister (1994), sixteen experimental studies were analyzed. The authors found an overall positive effect on reading comprehension, with a small Cohen's effect size value ($d = .32$) for standardized tests and a large Cohen's effect size value ($d = .88$) for researcher-developed tests.

Reciprocal teaching was originally designed by Palincsar and Brown (1984) for small-group tutoring under the guidance of experts, in which small groups of students were taken out of the classroom. From the literature, there are some indications that replacing experts by the students' regular teachers is not without problems. Most of the studies under investigation in the review by Rosenshine and Meister (1994) were small experimental studies in which students were taken out of the classroom and reciprocal teaching was delivered by researchers or research assistants. However, seven studies in the review were teacher-led and the effects on reading comprehension for those studies were ambiguous, with two studies with positive significant results, three studies with mixed results and two studies with non-significant results. Thus, there is sufficient ground for investigating implementation quality of principles of reciprocal teaching by teachers because they may influence treatment effects. In particular, it is of interest to study implementation quality when teachers are delivering the treatment in their own classroom, overseeing multiple groups of students' practicing reading strategies through reciprocal teaching. Such settings are from here on indicated as whole-classroom settings.

Implementation Quality

Even though many researchers in the field of reading comprehension underscore the need to take into account implementation quality as moderator in the analysis of treatment effects on students outcomes, especially in whole-classroom settings (Andreassen & Bråten, 2011; Hulleman & Cordray, 2009; Larsen & Samdal, 2007; Swanson et al., 2013; Vaughn et al., 2013), such studies have not been carried out yet in the context of reciprocal teaching. Qualitative studies, however, show that teachers in whole-classroom settings face problems in the implementation of

reciprocal teaching or similar interventions (Duffy, 1993; Hacker & Tenent, 2002; Seymour & Osana, 2003). Results of those studies show that teachers find it hard to induce strategic thinking in students by modeling the use of strategies and explicitly relating strategy-use to text comprehension (Duffy, 1993). In addition, teachers found the didactic principles of reciprocal teaching and the specific reading strategies that had to be taught hard to understand (Seymour & Osana, 2003). Finally, Hacker and Tenent (2002) found that when teachers implemented reciprocal teaching in their classrooms, students showed poor application of reading strategies. The teachers felt obliged to extend whole-classroom instruction of reading strategies and to provide more scaffolding of strategy use, and therefore they were hindered in changing from a teacher-centered to a student-centered approach (Hacker & Tenent, 2002). In addition, teachers found that students exhibited poor discourse skills while collaborating, hampering the implementation of collaborative group work in discussing and practicing reading strategies.

Quantitative studies in whole-classroom settings that focus on teaching reading strategies show that positive effects on students' reading comprehension are often not found (Okkinga et al., 2018; De Corte, Verschaffel, & Van de Ven, 2001; Fogarty et al., 2014; McKeown, Beck, & Blake, 2009; Muijselaar et al., 2018; Simmons et al., 2014). Many of these studies used principles of reciprocal teaching in whole-classroom settings (such as modeling or group work). Non-significant results on reading comprehension were obtained, which may be explained by problems with treatment adherence in whole-classroom settings (De Corte et al., 2001; Simmons et al., 2014). It is thus of importance to understand more thoroughly why such treatments do not seem to work.

The Present Study

In this study, the final results are presented of a two-year experiment, following low-achieving students from grade 7 to 8. Effects of principles of reciprocal teaching in whole-classroom settings and moderation effects of implementation quality were investigated. Teachers were trained and coached in all three principles of reciprocal teaching. In the first year, more attention was spent on the training and coaching of instruction in reading strategies than on modeling and group work (Okkinga et al., 2018). The focus on reading strategies was more important in the first year, since both teachers and students had to be familiarized with each of the strategies. In the second year, more attention was spent on coaching of teachers in their modeling behavior and their guidance of group work, than on strategy-instruction, assuming that teachers already mastered strategy-instruction.

The results of the first year (Okkinga et al., 2018), which followed students through 7th grade, revealed that there were no overall treatment effects on reading comprehension. However, a significant interaction between implementation quality of strategy-instruction and the treatment was found. This effect implied that in the experimental condition more elaborate explanations of the nature, function, importance, and application of reading strategies positively contributed to students' reading comprehension. The effect was substantial: it explained an additional 37 per cent of the differences between classes after individual and class-level variables had been taken into account. A few conclusions can be drawn from these results. First, the results underscore the importance of including implementation quality in the analyses. Neglecting such variation can result in overlooking meaningful effects. Second, no moderation effects were found for two principles of reciprocal teaching: modeling and group work. An explanation of this lack of moderator effects might be that these two principles are hard for teachers to master, as suggested by the study of Hacker and Tenent (2002). Additionally, several authors point to the problem that mastering multi-component treatments fostering reading comprehension is quite difficult for teachers (Roberts, Fletcher, Stuebing, Barth, & Vaughn, 2013; Scammacca et al., 2007; Scammacca, Roberts, Vaughn, & Stuebing, 2015). In the case of principles of reciprocal teaching, modeling requires that teachers model strategies in a fashion that is adaptive to the students' capacities (both reading skills and word knowledge) and is able to empathize with students' thinking processes. Group work requires active participation from the students with the teachers transferring control to their students. It is plausible that it requires more time to master the necessary skills underlying modeling and the guidance and supervision of group work effectively in a whole-classroom setting. Thus, it may be necessary for teachers to spend more time to become familiar with implementing both modeling and group work effectively.

The present study adds to the research base by analyzing moderation effects of implementation quality of principles of reciprocal teaching in whole-classroom settings over a two-year period, in addition to treatment effects on students' reading comprehension. This allows insight into the conditions and necessary duration under which the treatment will be effective in whole-classroom settings with low-achieving adolescents. An explorative study of Chambers Cantrell et al. (2016) suggests that after a first year a second year of intervention directed at reading strategies of 6th and 9th grade low achieving students did show promise for improvement in reading comprehension for those students that did not profit from a first year of intervention. In addition, they showed that teachers' quality of implementation improved from year to year (in a course of four years). Therefore, combining implementation quality and students'

reading comprehension in the analysis of effects of our two-year treatment is of interest.

In this study, we will answer the following research questions:

1. Is a treatment based on principles of reciprocal teaching in the context of whole-classroom settings, over a period of two school years, effective in fostering reading comprehension of adolescent low achievers?
2. Does the quality of implementation of the three main principles of reciprocal teaching (strategy instruction, modeling and group work) moderate effects on reading comprehension?

Method

Design

A two-year longitudinal design with a randomized controlled trial was used in this study (Shadish, Cook, & Campbell, 2002). The study was situated in the Netherlands. Randomization took place at the class level. At every participating school two classes, each with their own Dutch language teacher, took part in the study. Classes within each school were randomly assigned to either the control or treatment condition. The dependent variable, reading comprehension, was measured at four time points and was used as repeated measure.

We included five control variables on the student level. First, we included gender, because girls generally show greater reading skills than boys (Logan & Johnston, 2009; Schaffner, Philipp, & Schiefele, 2016). Additionally, we included non-verbal IQ, vocabulary knowledge and metacognitive knowledge, since theoretical models and empirical evidence suggest that reading comprehension draws heavily on these variables (Van Gelderen et al., 2004; Van Gelderen et al., 2007; Trapman et al., 2014; Just & Carpenter, 1976, 2004; LaBerge & Samuels, 1974; Ouellette & Beers, 2010; Rumelhart, 2004; Samuels, 2004; Verhoeven & Van Leeuwe, 2008). Finally, age was included as a control variable.

Additionally, we included two control variables on the class level: teacher replacement and canceled classes. Six teachers (three treatment and three control teacher) were replaced during the study (see teacher replacements and attrition). For some schools, it was difficult to find replacements immediately. Therefore, we also included a class-level control variable “cancelled class” to account for the missed classes. This concerned two treatment classes in total. Those classes missed at least 6 weeks of Dutch language teaching before a replacement was found.

Finally, we included three moderator variables, covering the three didactic principles behind our treatment: whole-classroom instruction of reading strategies, teacher and student modeling, and group work. For these variables, two class observations were performed in each year, resulting in a total of four observations (over two years) in each class.

Sample Selection and Description

Our study focused on low achievers. Our operationalization of low achievement was based on educational track. The Netherlands have a tracked system of secondary education. After primary school, students are placed in one of three main tracks: prevocational secondary education, senior general secondary education, or pre-university education. This decision is based on their scores on a general attainment test (directed at language, reading and mathematics) and their educational performance as assessed by their primary school teachers (Ministry of Education, Culture, & Science, 2006). Since students in prevocational education are generally characterized by poor reading skills (Dutch Education Inspectorate, 2008; Gille, Loijens, Noijons, & Zwitser, 2010), we selected our sample from schools offering this type of education.¹

We recruited schools in two ways. First, we contacted schools that had participated in a previous study on low-achieving readers. Second, we contacted schools via a digital community of Dutch language teachers. Schools had to meet the following five criteria:

- Willingness to participate in a two-year treatment study.
- They had (at least) two seventh grade classes.
- Each class had its own Dutch language teacher.
- The teachers were prepared to take part in the randomization procedure, implying that a) if their class was assigned to the treatment condition, they were prepared to take part in our training and coaching program and to weekly give our experimental lessons; and b) if their class was assigned to the control condition, they were prepared to not use our program nor discuss its contents with the colleague in the treatment condition.
- Control teachers were to use their regular language program during the language classes.

Ten different schools in different parts of the Netherlands were willing to participate. Within each school, two teachers volunteered. Thus, in total twenty classes participated in the study. Randomization was done at the class level within each school, resulting in a total of ten experimental and ten control classes, each with their teacher, divided over the ten

schools. At the start of the study, these classes comprised 369 students, of which 189 were in the treatment condition (51%) and 180 in the control condition (49%). The students' mean age was 13.01 years ($SD = 0.52$) at the start of the project. There was no statistically significant difference between the two conditions on this variable ($t(366) = -1.27, p = .20$). There were relatively more girls in the sample ($n = 200$; 54%) than boys ($n = 169$; 46%), with relatively more girls than boys (59 vs 41%) in the treatment condition. The distribution in the control condition, however, was more equal (49 vs 51%). The difference in distribution between the two conditions was statistically significant ($\chi^2(1) = 3.99, p = .046$).

More female than male teachers participated in the study ($n = 15$ vs $n = 5$), with two male teachers in the treatment group and three males in the control group. The mean age of the teachers was 46.40 years ($SD = 11.12$). On average they had 13.50 ($SD = 13.73, \min = 1, \max = 38$) years of teaching experience in secondary education. No differences were found between the conditions on either variable, ($t(14) = -.45, p = .66$) and ($t(14) = .053, p = .96$), respectively.

Teacher Replacements and Attrition

There was considerable attrition among the students. From a total of 369 students at the start of the project, 44 students changed schools, of which 19 students in the treatment condition and 25 in the control condition. Six students (5 from the treatment) switched classes within their school and three students were ill for a long period of time, of which two were in the treatment condition. In all these cases, there were no posttest data therefore, in total, 53 students dropped out of the study. The distribution of these categories (students staying, changing schools, switching classes, and illness) over the treatment and control condition was not statistically significant at .05 level, $\chi^2(3) = 4.78, p = .19$.

During the two school years, six (3 control and 3 experimental) of a total of twenty teachers were replaced during the study due to pregnancy, illness or a new job. Two of the three teachers in the control condition were immediately replaced. The third control teacher was replaced after the start of the second school year because of scheduling issues. It took the schools a few weeks to find a replacement for two treatment teachers. Finally, it was not possible to replace another treatment teacher in the second school year with a teacher who was willing to participate in the study. Therefore, this class dropped out ($n = 24$ students).

After the collection of all data, the dataset contained data of 292 students ($369 - 53 - 24$).

Lastly, we imputed missing data within individual tests at the item level using the EM procedure from SPSS missing value analysis. The

missing data in this procedure never exceeded 7% of the data matrix. If a student was not present during a test session, all tests from that session were regarded as missing and these missing data were not imputed. This resulted in an additional loss of 54 students for the final analysis and included one experimental class as a whole ($n = 11$). In this class it was not possible to schedule class observations in the second year of treatment. Therefore, the final dataset contained a total of 238 students (110 experimental students and 128 control students), with a total of 18 teachers (8 experimental, 10 control).

Treatment

Our treatment was based on the following three principles of reciprocal teaching (Palincsar & Brown, 1984), that is:

1. Whole-classroom instruction of reading strategies, focusing on procedural knowledge. This implies that for each strategy, it was emphasized what the strategy entailed, how to use the strategy, when to use the strategy and why to use the strategy (Veenman, Hout-Wolters, & Afflerbach, 2006).
2. Teacher and student modeling. Teachers were trained to model the use of reading strategies during plenary instruction by thinking aloud when reading text. They encouraged students to take over this role, both plenary and in small group sessions. The teacher read the first paragraph of the text while thinking aloud and modeled how to use the central reading strategy. Then, the teacher invited one of the students to read the next sentences while thinking aloud and use the central strategy. The teacher supported the student giving feedback and asking questions.
3. Group work. The primary objective of encouraging students to work in groups was to have them collaboratively apply reading strategies while thinking aloud during text reading. Teachers were given instructions on how to give feedback to the groups of students working together. For example, if a teacher noticed that the students were struggling with the application of a reading strategy, the teacher was required to model this strategy again and encourage and aid the students in doing this themselves.

With respect to strategy instruction, the intervention focused on five strategies that were shown to be related to reading comprehension in previous research (Dole, Duffy, Roehler, & Pearson, 1991; Palincsar & Brown, 1984; Pressley & Afflerbach, 1995; Van Silfhout, Evers-Vermeul, Mak, & Sanders, 2014):

1. *Predicting*. On the basis of text features such as title, subheadings, and pictures, students are instructed to make predictions about text content before reading, and to check their predictions while reading.
2. *Summarizing*. Students are instructed to summarize sections of text, encouraging them to focus on main ideas and ignore irrelevant details as well as to check their understanding of the text so far.
3. *Self-questioning*. Students are instructed to generate questions about the text being read, helping them to focus on main ideas as well as to monitor understanding.
4. *Clarifying*. When confronted with a word or passage they do not understand, students are instructed to reread, read ahead, or, in the case of an unknown word, analyze it, and see whether its meaning can be inferred by looking at parts of the word.
5. *Interpreting cohesive ties*. Students are instructed to look for relationships between sentences or paragraphs that are connected, e.g. by using ‘signal words’ (connectives signaling conceptual relations).

Students received weekly lessons over a period of two school years, from October until June in the first year and from September until June in the second year. The treatment was offered in the context of an existing program called “Nieuwsbegrip”[®], developed by an educational consultant organization, the CED-Group. Lessons were developed weekly by a team of developers at the CED Group. They were based on recent news texts (i.e., texts that had been issued the week before) about subjects close to students’ everyday life (e.g., sugar in energy drinks, abdication of the Dutch queen, or 20 years of text messaging). The use of interesting texts aimed to increase students’ task motivation (Guthrie & Wigfield, 2000; Schiefele, 1999). The lessons could be downloaded by teachers from the program website (www.nieuwsbegrip.nl) every week, starting Monday evening.

Lessons were provided in sequences of six weeks. Each sequence consisted of six weekly lessons (approximately 45 minutes per lesson). In each of the first five lessons, the focus was on one reading strategy that was practiced in a central strategy assignment that was provided on a work sheet. In addition, students worked on assignments such as answering questions about the text on the work sheet.

Students practiced each of the five strategies several times during the year. This cyclical approach was assumed to result in the consolidation of strategy use. In the final lesson of each sequence all strategies were practiced simultaneously. The idea behind this was that students have to be able to apply all strategies during the reading process, selecting an appropriate strategy depending on their own needs. [Appendix A](#) provides

examples of translated assignments from the program for each reading strategy.

Training and Coaching of Treatment Teachers

Treatment teachers took part in a training and coaching program that was provided by teacher trainers from the Rotterdam University of Applied Sciences, who had, in turn, been trained by three of the authors.

In the training phase (October 2011–January 2012), teachers participated in three one-hour training sessions. In *Session 1*, they received general, practical information about the program (e.g., how to use the program website), theoretical information about the reading process and its components, and basic information about the program's didactic principles (direct instruction of reading strategies, teacher and student modeling, and group work). In *Session 2*, in-depth information was provided about the nature, function, importance, and application of the five central strategies and on the way teachers could model the use of these strategies. Examples of modeling were provided by means of video clips and lesson protocols. In *Session 3*, the focus was on reciprocal teaching and how, by means of scaffolded instruction, the use of reading strategies is transferred to students. Attention was given to how the teacher can give feedback to groups of students and how his or her expert role is gradually faded.

Teachers were given a template for the lessons that would help them keeping focused on the reading strategies (see [Figure 1](#)).

In the coaching phase (February 2012–May 2013), teachers participated in six coaching sessions; three coaching sessions during February–June 2012 and three coaching sessions during September 2012–May 2013. A coaching session involved a classroom observation conducted by the trainer during a treatment lesson, followed by a feedback session of approximately twenty minutes on the same day. During the classroom observations, trainers used an observation scheme comparable to the one used by the researchers (see below). This scheme directed the trainers' attention to the three principles of the treatment (whole-classroom instruction of reading strategies, modeling, and group work). During the first year, coaching was mainly directed at instruction of the reading strategies and to a lesser extent to the teacher modeling those reading strategies. During the second year, the focus of coaching was on modeling and group work.

Control Classes

Control classes were “business as usual”. Teachers in the control classes used their regular textbook for Dutch language arts. Among our schools, three different language textbooks were used. The teacher manuals were

Introduction	<ul style="list-style-type: none"> ▪ Write the subject of the text and the central strategy of the lesson on the blackboard. ▪ Introduce the subject and the central strategy with a whole-class approach and activate prior knowledge. ▪ Write down questions students have about the text during orientation. ▪ Read the first paragraph together and model the central strategy. ▪ Invite a student to read the next paragraph while thinking aloud and applying the central strategy. Give support when necessary, that is, ask questions that stimulate the use of the reading strategy.
Processing	<ul style="list-style-type: none"> ▪ Instruct the students to work together in groups of two or three. Let them work on the remainder of the work sheet (example in Appendix). ▪ Walk around to give the groups of students feedback. Focus on the central strategy and motivate the students to apply the strategy while thinking aloud. If necessary, model the strategy again.
Reflection	<ul style="list-style-type: none"> ▪ Reflect with the students on the reading process as well as the content. ▪ Together with the students, answer the questions they had before reading the text. Did reading the text answer those questions?

Figure 1. Template for the lessons.

analyzed to determine whether the three central principles of reciprocal teaching were present. No attention was given to modeling by teachers or students or group work in the teacher manuals of the control classes. Some reading strategies were mentioned in two teacher manuals (for example, “some assignment require activating prior knowledge”), but no guidelines were given for how to instruct reading strategies. The textbooks for students were analyzed for presence of the five reading strategies of the treatment program (predicting, summarizing, self-questioning, clarifying and interpreting cohesive ties). Attention was given to reading strategies in all three textbooks. However, not all strategies that were covered in the treatment condition were also covered in the

control textbooks. Reading strategies that were often referred to were: predicting, clarifying, and attention to cohesive ties. This occurred in all three textbooks with similar frequency. Self-questioning did not occur in any of the textbooks. Summarizing only occurred as a specific assignment after reading texts, but was not used as a reading strategy during reading. Almost all of the assignments were individual and there were only a few instances where students were instructed to work together on an assignment in all three textbooks.

Measures

Reading comprehension

Reading comprehension was measured by means of the SALT-reading, a test that was validated for use among low-achieving adolescents (Van Steensel, Oostdam, & Van Gelderen, 2013). The SALT-reading comprises eight tasks, each consisting of one or two texts and comprehension questions about those texts. The texts cover different genres (narrative, expository, argumentative, and instructive). They were selected from media students supposedly come across regularly in their daily lives: (school) books, newspapers, magazines, and official documents (such as regulations in a youth hostel). The eight tasks comprised a total of 59 test items, that were divided into three categories: items requiring students to retrieve relevant details from the text, items requiring students to make inferences on a local level (e.g., draw cause-effect relationships between sentences), and items requiring students to show their understanding of the macrostructure of the text (e.g., by inferring the main idea of the text or the intention of the author). The test consisted mainly of multiple-choice questions but contained also five open-ended questions. The SALT-reading was administered at four time points. The Cronbach's alpha coefficients were .82, .83, .82, and .85 respectively, indicating sufficient reliability (Field, 2009).

Vocabulary knowledge

Vocabulary knowledge was assessed with a 73-item multiple-choice test, measuring the knowledge of nouns, verbs, adjectives, and adverbs belonging to the 23,000 words in a dictionary for junior high school students (see Hazenberg & Hulstijn, 1996, for details). Each item consists of a neutral carrier sentence with a bold-faced target word and four answer options, one of which represents a correct synonym. Vocabulary knowledge was administered two times and the average of both was used as a measure for vocabulary knowledge. The Cronbach's alpha coefficients were .86 and .85, respectively indicating sufficient reliability (Field, 2009).

IQ

Intellectual ability was measured by administering the Raven Progressive Matrices, a nonverbal IQ test. The total test consists of 60 items, divided into 5 sets of 12 items. Each item represents a logical reasoning puzzle. The items become more difficult within a set and the sets become increasingly difficult as well (Raven, Raven, & Court, 1998). For students from the lowest tracks of prevocational education the last set was assumed to be too difficult and for this reason this set was omitted. The Cronbach's alpha coefficient was sufficient: .82 (Field, 2009).

Metacognitive knowledge

Metacognitive knowledge was measured by a questionnaire consisting of 45 statements about text characteristics, reading and writing strategies (Trapman et al., 2014). It consisted of a selection of items from a metacognitive knowledge test destined for students in grades 8 to 10 (Van Gelderen et al., 2003, 2007). Items consisted of correct or incorrect statements and students had to agree or disagree with each statement. An example of an incorrect statement is "When you read, it is sensible to put the most effort into memorising details". A correct statement was for example "It is sensible to think beforehand why you are going to read a text". The Cronbach's alpha coefficient was .51. Although this indicates a rather low level of reliability (Field, 2009) we maintained the measure because in previous research it still predicted significant variance in reading comprehension (Trapman et al., 2017).

Classroom variables and treatment fidelity

To measure the moderator variables, we conducted classroom observations in both the experimental and control conditions twice each year, resulting in a total of four observations for each class over two years. Our aim was to examine a) whether the treatment teachers provided the lessons in the way we instructed during the training and coaching program and b) whether the control teachers applied the three treatment principles, even though they were not trained to do so. Therefore, we devised an observation scheme focusing on the three main principles: whole-class teaching of reading strategies, teacher and student modeling, and group work. This was done in the following manner, resulting in three four-point scales (0–3) to be used for further analysis:

1. Whole-class teaching of reading strategies. We distinguished four categories:
 - a. Teachers provided no information on reading strategies (0 points).

- b. Teachers introduced the central strategy of the lesson (in the treatment condition) or any strategy (in the control condition), but provided no further explanation (1 point).
 - c. Teachers introduced a strategy and explained its nature, function, importance, and/or application (2 points).
 - d. Teachers introduced a strategy, explained its nature, function, importance, and/or application and discussed this strategy with the class (3 points).
2. Teacher and student modeling. We distinguished four categories of behavior:
- a. Teachers did not use any modeling of strategy use (0 points).
 - b. Teachers modeled strategy use (1 point).
 - c. Teachers modeled strategy use and asked students to think aloud while using reading strategies, either individually (i.e., in front of the class) or in groups (2 points).
 - d. Teachers modeled strategy use, asked students to think aloud, and provided them with feedback (3 points).
3. Group work. The following four categories were distinguished:
- a. Teachers did not order students to work in groups (0 points).
 - b. Teachers ordered students to work in groups, but did not provide feedback (1 point).
 - c. Teachers ordered students to work in groups, provided feedback, but focusing on students' understanding of the assignment, their answers to questions, or on unknown words (2 points).
 - d. Teachers ordered students to work in groups and provided feedback on collaboration itself or collaboration directed to any of the previous issues (3 points).

The scales were constructed in such a way that a score of 3 indicated optimal realization of the treatment principle.

Before the start of the classroom observations, the observation scheme was piloted during two lessons, one in an experimental class and one in a control class. Two researchers filled out the observation scheme during the lessons, after which they compared their codes and discussed causes for any differences. The coding scheme was adjusted when needed.

The adjusted scheme was used for all observations. In order to be able to check codes after the observation, the lessons were recorded using an audio-recorder carried by the teacher. Means were calculated over the four classroom observations per class. Inter-rater reliability was calculated by means of observed agreement between two observers. In total, 30 from a total of 76 classroom observations were performed independently by two coders. Across these 30 observations, 93.89% agreement was obtained.

Procedure

Tests were first administered in the fall of 2011, just before the start of the treatment. It concerned the SALT reading, vocabulary knowledge and non-verbal IQ. At the end of the first school year (May–June 2012), the SALT reading, vocabulary knowledge and metacognitive knowledge were administered. At the start of the second school year (September–October 2012), and at the end of the second school year (May–June 2013) the SALT-reading was administered. All test administrations took place in classroom settings. The test sessions were introduced by a trained test leader. A familiar teacher was present to maintain order. Questions were answered by the test leaders following a standardized protocol.

Classroom observations took place during January–February 2012 and during April–May 2012 in the first school year. In the second school year, classroom observations were performed during October–November 2012 and April–May 2013. During the classroom observations the researcher(s) sat at the back of the classroom to observe the teacher. See [Appendix B](#) for an overview of all research activities.

Exit Interviews

After the treatment was completed, the first author held exit interviews with 7 of the 8 treatment teachers. The interviews were semi-structured and covered the following topics: how did teachers look back on the implementation of the treatment (e.g., did they encounter difficulties, if so, how did they solve these; what were advantages and disadvantages of the treatment; how did they perceive the training and coaching), how did their view of the principles of reciprocal teaching (whole-class instruction of reading strategies; modeling and group work) change during the course of the treatment, and did they see any changes for their students (e.g., in their views about reading or their learning outcomes). Interviews lasted 45–60 minutes. All interviews were recorded and transcribed.

Analyses

Repeated measures multilevel analyses were performed to account for the hierarchical structure of the data (using MLwiN 2.16; Rasbash, Steele, Browne, & Goldstein, 2009). The time variable ‘Occasion’ (variance within students across times of measurement) was defined in months; with the first measurement of reading comprehension at month zero, and subsequent measurements at months 9, 12, and 22, respectively. These months correspond to the time points of the SALT-reading: September

2011, June 2012, September 2012, and June 2013. Thus, growth is measured as a repeated measure.

We tested a) whether the treatment had a significant, positive effect on growth in reading comprehension by testing the interaction between treatment (yes or no) and occasion, and b) whether the classroom variables (strategy-instruction, modeling, and group work) moderated the treatment effect.

Adding predictors was done in the order Hox (2010) suggests. First, we tested whether adding a class or school level to the model significantly improved model fit. Levels significantly improving model fit were added to the model. Second, we tested whether a model with random slopes both at the student or class level for the occasion variable improved model fit. The treatment variable is a class level variable, random slopes at class level indicate differences in growth between classes. If a treatment effect exists, we would expect significant model fit improvement by adding random slopes at class level to the occasion variable.

Third, we added the class level variables 'teacher replacement' and 'canceled classes' to check whether we should include these variables as covariates. Fourth, we tested whether the student-level predictors gender, IQ, age, vocabulary knowledge, and metacognitive knowledge significantly improved model fit.

To answer the first research question, the treatment variable and the interaction between treatment and occasion were added to the model. A treatment effect implies a greater learning gain in the treatment group and thus a significant interaction effect between occasion and treatment.² To answer the second research question, we started with a model containing the significant predictors of the model resulting from the first research question. For each of the three moderator variables (strategy-instruction, group work and modeling), we checked separately whether adding the moderator variable and its interactions with the occasion and treatment variables significantly improved model fit. The interaction between occasion and moderator variable is indicative of an effect of the moderator variable on growth. The three way interaction (occasion \times moderator \times treatment) indicates a differential effect on growth of the moderator variable on students in the experimental and the control group.

Dichotomous independent class variables (teacher replacement and canceled classes) and student variables (gender) are always scored 0 and 1. All continuous independent variables (IQ, age, vocabulary knowledge and metacognitive knowledge) are centered around their grand mean before adding them to the model (Hox, 2010). The number of levels needed in the analyses was tested by comparing nested models with one-sided Chi-square significance tests (Hox, 2010). Significance of predictors

was tested both with Wald-tests (coefficient divided by the standard error) and by means of comparing nested models (with and without the predictors) with a Chi-square test.³ Regression coefficients for class-level variables were tested with number of classes as sample size ($df = \text{number of classes} - \text{number of predictors} - 1$) (Hox, 2010).

Results

Descriptive Statistics

Table 1 shows the mean student scores for all student level variables (the pretest and the three reading comprehension, IQ, vocabulary, and meta-cognitive knowledge) and t-tests for the differences between experimental and control students.

No significant differences were found (according to the t-tests) between the treatment and the control condition between any of the variables. This means that there were no significant differences between the control and experimental classes before the start of the treatment (time 1) on all student level variables including vocabulary knowledge, IQ, meta-cognitive knowledge, and reading comprehension. In addition, no significant differences were found between the control and experimental classes at all subsequent measurements for reading comprehension (time 2–4).

The development of instructional principles (strategy-instruction, modeling, and group work) was tested with three repeated measures ANOVA's.⁴ For each of the instructional principles no main effects over time were found (strategy-instruction: $F(3,48) = 1.84, p = .15$; modeling: $F(3,48) = 2.77, p = .05$; and group work: $F(3,48) = .73, p = .54$), nor were interaction effects between instructional principles and treatment found (strategy-instruction: $F(3,48) = .78, p = .51$; modeling: $F(3,48) = .06, p = .82$; and group work: $F(3,48) = .95, p = .42$), suggesting that there was no systematic difference between the treatment and control teachers in growth of use of the principles in their lessons. To give a more precise impression of the development of the instructional

Table 1. Descriptives student-level variables.

Variable	Treatment ($n = 110$)	Control ($n = 128$)	t-value
	Mean (SD)	Mean (SD)	
Reading comprehension (time 1)	35.47(7.21)	34.67(8.38)	.79
Reading comprehension (time 2)	37.72(6.81)	36.72(8.69)	.97
Reading comprehension (time 3)	36.85(7.10)	36.93(8.60)	.08
Reading comprehension (time 4)	37.77(8.46)	39.28(8.53)	1.36
Vocabulary	49.66(6.80)	49.56(7.85)	.97
IQ	36.01(5.08)	35.36(5.24)	1.28
Metacognitive knowledge	26.26(4.19)	25.59(4.53)	1.19

* $p < .05$; no statistical significant differences at pretest for all variables between treatment and control.

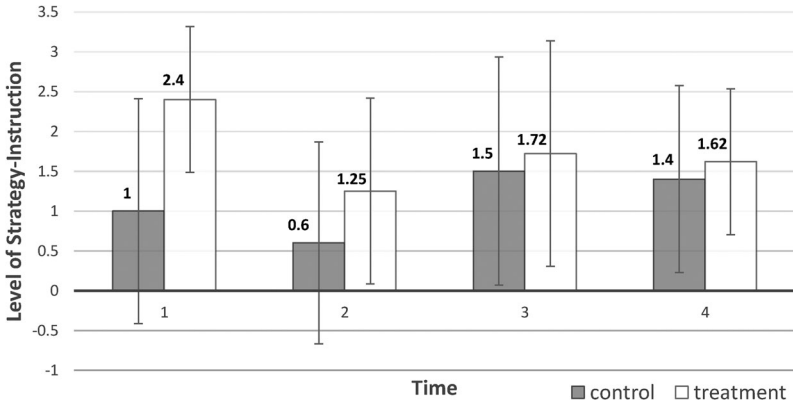


Figure 2. Mean observations of level of strategy-instruction over time, for both the control ($n=10$) and treatment ($n=8$) teachers. Error bars represent standard deviations.

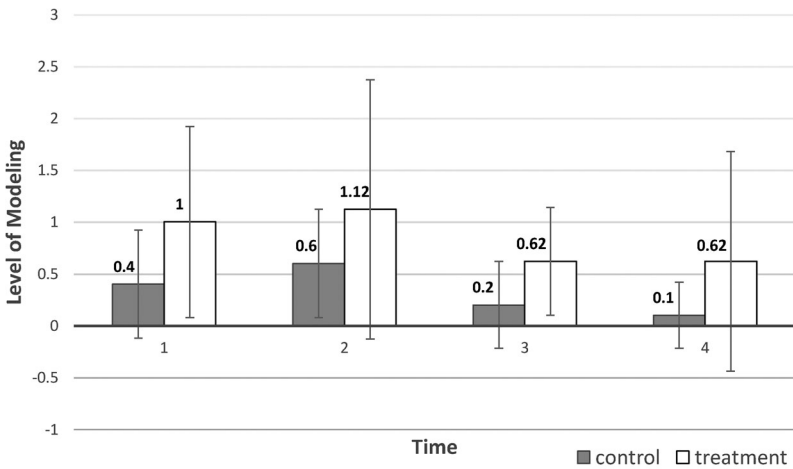


Figure 3. Mean observations of level of modeling over time, for both the control ($n=10$) and treatment ($n=8$) teachers. Error bars represent standard deviations.

principles over time in both the control and the treatment classes, [Figures 2–4](#) are presented. It appears that strategy instruction was practiced more in the start of each academic year than at the end for both the control and the treatment teachers. Modeling was practiced more at the end of the first academic year and seems to decrease somewhat thereafter. Group work, however shows a slight increase in the second year of the treatment. Overall, the figures show that in the treatment group differences in application of each of the principles between the 8 teachers are quite large. Although the means are considerably higher than for the controls,

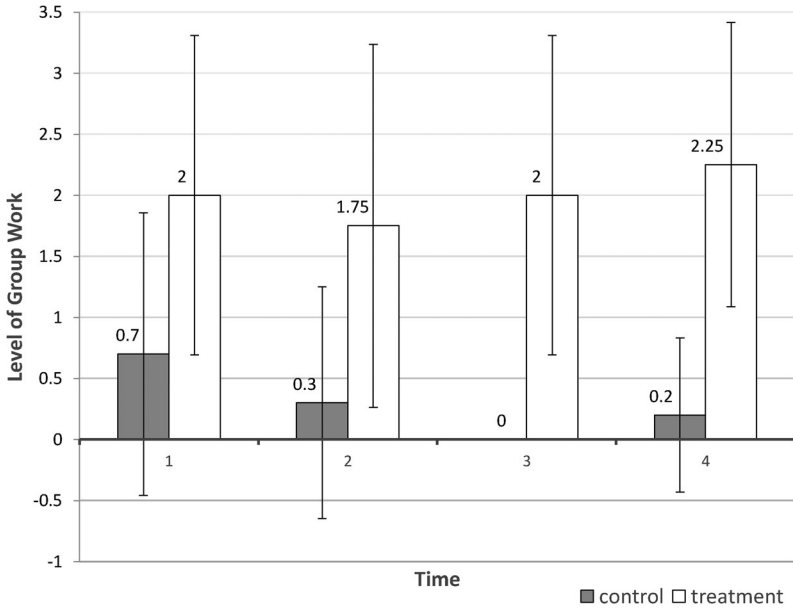


Figure 4. Mean observations of level of group work over time, for both the control ($n = 10$) and treatment ($n = 8$) teachers. Error bars represent standard deviations.

Table 2. Descriptives teacher-level variables.

Variable	Treatment ($n = 8$)		Control ($n = 10$)	
	Mean (SD)	Mean (SD)	t-value	p-value
Strategy-instruction	1.81(.80)	1.13(.65)	2.02	.061
Modeling	.84(.65)	.33(.35)	2.15	.047
Group work	2.00(1.14)	.30(.33)	4.51	<.001

Note. Scoring between the three variables cannot be compared one-on-one. The meaning of the scoring (0–3) is different for each variable. See Classroom variables and treatment fidelity for an explanation of each variable.

that does not mean that each treatment teacher implements reciprocal teaching optimally. Ideally each of them should reach the maximum score, which is obviously not the case.

For each of the instructional principles, mean scores were calculated over the four time points. In Table 2, means and standard deviations are presented for the variables resulting from the classroom observations. As expected, the mean scores of the treatment group are higher than those of the control group, indicating that in the experimental classrooms modeling, strategy-instruction and group work were more often observed than in the control classrooms. The difference between both groups is statistically significant on the .05 level for all variables, except for strategy-instruction.

Multilevel Analyses

As no significant random variance on reading comprehension was found at the school level (see Appendix C), models with three levels were tested (occasion-, student-, and class level). Appendix C also shows that random slopes were tested and found significant for both the class and student level. Next, the control variables ‘teacher replacement’ and ‘canceled classes’ were entered in the model (see Appendix D). Both variables did not significantly contribute to the model and were omitted from all further analyses. Subsequently, the student-level control variables were entered to control for differences between students at pretest. Inclusion of age and gender did not improve model fit (see Appendix E), whereas vocabulary knowledge, metacognitive knowledge, and IQ did. Model E-5 (see Appendix E) is therefore the baseline model in subsequent analyses.

In the next step, the interaction between occasion and treatment was entered (research question 1). This effect was not significant implying there was no effect of the treatment on growth in reading comprehension (see Table 3, Model 3-2; $\Delta IGLS = 2.131$, $df = 1$, $p > .05$).

Table 3. Multilevel analyses with reading comprehension (repeatedly measured) as dependent variable to establish influence of treatment (main effect of treatment over time), after correcting for control variables (N = 952 cases/238 student).

Model	Model D-5 ^a	Model 3-1	Model 3-2
Fixed part			
Intercept	35.058 (.739)	35.377 (.889)	34.671 (.986)
Occasion (in months)	.155 (.040)	.155 (.040)	.207 (.051)
IQ ^b	.219 (.062)	.219 (.062)	.219 (.062)
Vocabulary ^b	.479 (.048)	.480 (.048)	.479 (.048)
Metacognitive knowledge ^b	.274 (.079)	.276 (.079)	.277 (.079)
Treatment (1 = treatment, 0 = control)		-.710 (1.033)	.871 (1.471)
Treatment × occasion			-.114 (.076)
Random part (variances)			
Class	7.628 (3.251)	8.157 (3.425)	7.477 (3.182)
Class slope variance occasion	.022 (.010)	.022 (.010)	.019 (.009)
Class covariance slope × intercept	-.325 (.154)	-.346 (.160)	-.298 (.143)
Student	13.091 (2.665)	13.081 (2.663)	13.095 (2.666)
Student slope variance occasion	.018 (.010)	.018 (.010)	.018 (.010)
Student covariance slope × intercept	.117 (.123)	.117 (.123)	.116 (.124)
Occasion (rep. measures)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)
Deviance testing			
-2*loglikelihood (deviance)		5916.529	5914.398
Difference between -2*loglikelihood		.016	2.131
Difference df		1	1
Compared to model		D-5	3-1

^aModel D-5 from Appendix D.

^bVariable is grand mean centered.

Bold and italicized = $p < .001$.



Table 4. Multilevel analyses with reading comprehension (repeatedly measured) as dependent variable to establish influence of interaction between modeling, occasion and treatment, after correcting for control variables (N = 952 cases/238 students).

Model	Model 3-2 ^a	Model 4-1	Model 4-2	Model 4-3	Model 4-4
Fixed part					
Intercept	34.671 (.986)	34.798 (1.033)	35.578 (.962)	35.381 (.968)	36.099 (.965)
Occasion (in months)	<i>.207</i> (.051)	<i>.206</i> (.051)	<i>.206</i> (.051)	<i>.178</i> (.055)	<i>.276</i> (.057)
IQ ^b	<i>.219</i> (.062)	<i>.219</i> (.062)	<i>.218</i> (.062)	<i>.218</i> (.062)	<i>.271</i> (.062)
Vocabulary ^b	<i>.479</i> (.048)	<i>.480</i> (.048)	<i>.470</i> (.048)	<i>.470</i> (.048)	<i>.471</i> (.048)
Metacognitive knowledge ^b	<i>.277</i> (.079)	<i>.273</i> (.079)	<i>.260</i> (.079)	<i>.260</i> (.079)	<i>.261</i> (.079)
Treatment (1 = treatment, 0 = control)	.871 (1.471)	.601 (1.587)	.269 (1.369)	.691 (1.404)	.257 (1.347)
Treatment × occasion	-.114 (.076)	-.114 (.076)	-.115 (.076)	-.156 (.082)	-.345 (.094)
Modeling ^b		.543 (1.077)	3.808 (2.030)	2.983 (2.168)	5.950 (2.392)
Treatment × modeling			-4.909 (2.364)	-4.939 (2.365)	-9.009 (2.789) ^c
Occasion × modeling				.083 (.075)	-.211 (.121)
Occasion × modeling × treatment					.403 (.141)
Random part (variances)					
Class	7.477 (3.182)	7.772 (3.298)	4.800 (2.297)	4.572 (2.225)	3.946 (2.015)
Class slope variance occasion	.019 (.009)	.019 (.009)	.018 (.008)	.016 (.008)	.009 (.005)
Class covariance slope × intercept	-.298 (.143)	-.309 (.146)	-.204 (.116)	-.182 (.108)	-.115 (.083)
Student	13.095 (2.666)	13.080 (2.663)	13.029 (2.658)	13.043 (2.659)	13.001 (2.655)
Student slope variance occasion	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)
Student covariance slope × intercept	.116 (.124)	.116 (.123)	.114 (.123)	.113 (.124)	.115 (.123)
Occasion (rep. measures)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)
Deviance testing					
-2* ^c loglikelihood (deviance)	5914.398	5914.165	5911.381	5910.219	5903.398
Difference between		.233	2.784	1.162	6.827
-2* ^c loglikelihood					
Difference df		1	1	1	1
Compared to model		3-2	4-1	4-2	4-3

^aModel 3-2 from Table 3.

^bVariable is grand mean centered.

^cThis significant interaction effect indicates a 9 points higher score for the control group at the first measurement of reading comprehension for each point scored higher on modeling. Modeling was measured after the first measurement of reading comprehension (between the first and second measurement moment). For interpretation of the effect of modeling, the results of all main and interaction effects concerning modeling should be taken into account. This can therefore best be done by looking at Figures 5-7.

Italicized = $p < .01$; **Bold and italicized** = $p < .001$.

Moderator effects of the teacher variables were tested subsequently (research question 2). We did not find a statistically significant relationship between strategy-instruction and growth in reading comprehension: the interaction between occasion and strategy-instruction was not significant (see Appendix F, Table F1, Model F1-3; $\Delta IGLS = 2.165$, $df = 1$, $p > .05$). In addition, over time, no moderator effect of the level of strategy-instruction on the treatment effect was found (i.e., there was no significant interaction between strategy-instruction, occasion and treatment; see Table F1, Model F4-4; $\Delta IGLS = .025$, $df = 1$, $p > .05$).

Next, no statistically significant effect of group work at the 0.5 level on growth in reading comprehension was found (the interaction of group work and occasion), nor was a moderator effect of group work found (the interaction of group work, occasion and treatment; see Appendix F, Table F2, Model F2-3; $\Delta IGLS = .315$, $df = 1$, $p > .05$ and Model F2-4; $\Delta IGLS = .007$, $df = 1$, $p > .05$). Table 4 shows the results for modeling. There was no significant effect of modeling on growth in reading comprehension (the interaction of modeling and occasion; see Table 4, model 4-3; $\Delta IGLS = 1.162$, $df = 1$, $p > .05$; $b = .083$, $SE = .075$, $p > .05$), but modeling did significantly moderate the effect of the treatment over time (the interaction of modeling, occasion and treatment). It appeared that in the treatment condition more elaborate modeling positively contributed to students' growth in reading comprehension (see Table 4, model 4-4; $\Delta IGLS = 6.821$, $df = 1$, $p < .01$; $b = .403$, $SE = .141$, $p < .05$). The moderator effect of modeling explains 13.69% of the variance at the class level and 43.75% of the variance in slopes at class level.

The interpretation of the moderating effect of modeling on growth in reading comprehension becomes clear when looking at regressions for

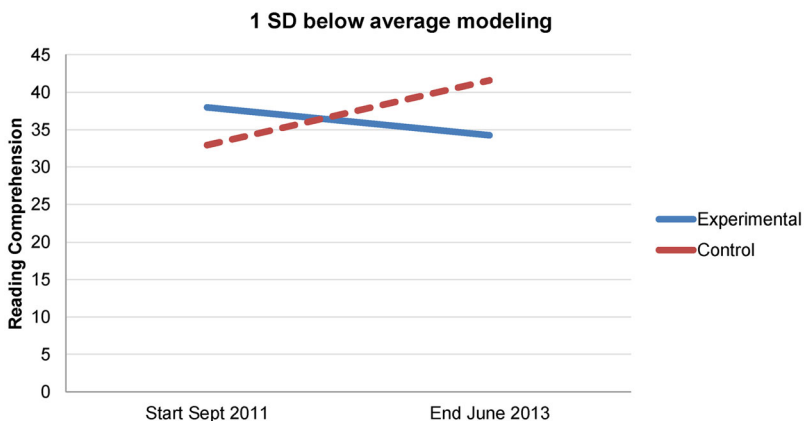


Figure 5. Calculated regression lines (based on regression weights of Model 4-4, Table 4) for the experimental and control condition for 1 SD below average modeling.

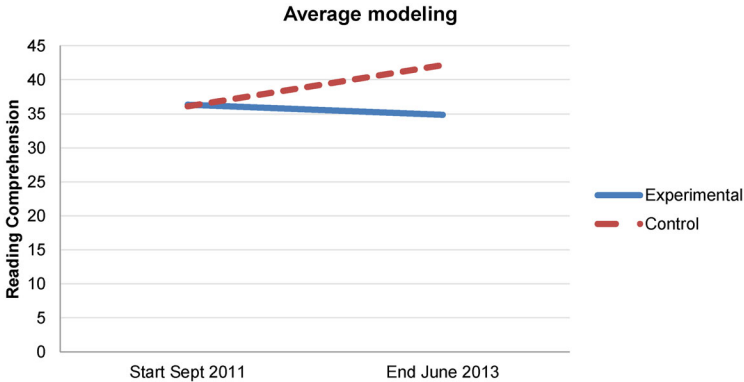


Figure 6. Calculated regression lines (based on regression weights of Model 4-4, Table 4) for the experimental and control condition for average modeling.

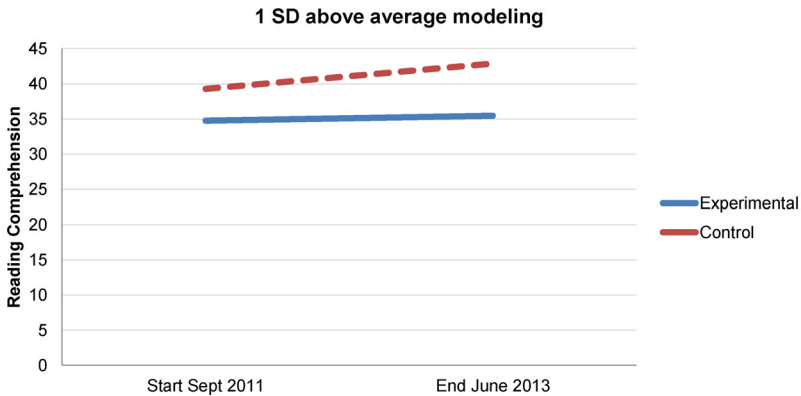


Figure 7. Calculated regression lines (based on regression weights of Model 4-4, Table 4) for the experimental and control condition for 1 SD above average modeling.

different combinations of scores on the independent variables (Hox, 2010). For treatment we used two scores (0 and 1), for occasion we used the scores 0 and 3 for the first and the last time of measurement and for modeling we used three scores: one standard deviation below the mean, on the mean and one standard deviation above the mean. The resulting six regression lines are presented in Figures 5–7. These figures show that in cases where modeling is less elaborate or moderately elaborate, growth in treatment students' reading comprehension is less than in the control group. Only in the case of more elaborate modeling (Figure 7) the development in reading comprehension of the two groups appears to be similar.

Conclusions and Discussion

Our study set out to analyze how principles of reciprocal teaching can improve low-achieving adolescents' reading comprehension in whole-classroom settings and to what extent treatment effects are dependent on implementation quality. Apart from analyzing the overall effects of the treatment in a whole-classroom setting, our aim was to examine whether effects were larger when teachers provided more elaborate instruction of reading strategies, engaged more in teacher modeling, promoted more student modeling, and supported more collaboration during group work.

With regard to our first research question, our study revealed no overall treatment effects. No significant differences were found between students in the treatment classes compared to the control classes on growth in reading comprehension. In this respect, our result is similar to what was found in several other studies analyzing effects of reading strategy instruction on reading comprehension in whole-classroom settings (De Corte et al., 2001; Fogarty et al., 2014; McKeown et al., 2009; Simmons et al., 2014; Vaughn et al., 2013). Answering our second research question, we did not find significant moderation effects of strategy-instruction and group work on students' reading comprehension growth. Modeling (i.e., thinking aloud during reading, with the purpose of showing students how reading strategies can be used), however, moderated the effect of the treatment over time.

Observations of Implementation Quality

Implementation quality may have played an important role in preventing an overall positive effect of the treatment. First, our observational data of implementation quality showed that quality of strategy instruction, modeling and group work differed quite a lot between treatment teachers. These data also showed that the average scores for the use of these principles of reciprocal teaching, although higher than for the controls, were still not as high as could be wished for. Even on the last observations at the end of the second year, averages were only slightly higher – no significant differences were found – than at the end of the first year for strategy-instruction and group work, and even a little lower for modeling. These results are quite disappointing. It is plausible that a certain level of treatment-adherence should be met for the treatment to sort effect, as Simmons, Fogarty et al. (2014) suggested. In their study, as in ours, the treatment involved multiple components. The authors argue that in such cases it is difficult to ensure that teachers implement each of the different components as intended (see also Roberts et al., 2013). It seems that even two years of training and coaching is not sufficient to improve implementation quality of principles of reciprocal teaching in a decisive way.

Main Treatment Effects

There are at least two explanations for the lack of success of our treatment (research question 1). The first has to do with its length, the second with advantages of the business-as-usual control. The fact that we deal with a two-year treatment may not have worked out beneficially for our students. It is possible that a long-term intervention results in disappointing effects, because it is difficult to maintain experimental control ensuring high implementation quality over a longer period of time. This is similar to an effect noted by Roberts et al. (2013), who point out the negative effects of scale in several reading comprehension interventions. According to the authors, large scale interventions have a disadvantage in terms of internal validity and experimental control. Long-term interventions are also sensitive to such disadvantages. An example of how this might work was found in the exit interviews with teachers carried out after the treatment was completed. One of the main comments made was that teachers found it hard to maintain the motivation of the students in the second year of the treatment, and as a result found it hard to keep motivated themselves, as one teacher put it: “Two years of the same thing is a long time. The novelty effect is gone and they [students] want something different, something new.” This attitude of students and teachers might have resulted in a less than inspired way of bringing forward the classroom practices necessary for reciprocal teaching. Short-term interventions do not meet this problem of decrease in motivation and quality of implementation, as they appear novel and interesting.

The second explanation for our disappointing result is the possibility that “business as usual” as practiced in the control group may have had some advantages over the treatment. For example, in the control group, teachers were free to select their own materials, adapt these materials or their teaching practices and had more opportunities to motivate their students by adapting their approach. The experimental treatment was quite rigid in the sense that it was prescribed what the teachers should teach, what materials and texts they should use and how the pedagogical and didactic procedures should be realized (i.e., there were specific guidelines for strategy use, modeling and group work). This may have had negative side effects such as that our treatment teachers were less able to adapt their teaching to the needs of students in their classroom.

Moderation Effects

Below, we will discuss the moderation effects for each of the three principles of reciprocal teaching (research question 2) and how they can be interpreted.

Strategy-instruction

We did not find a moderation effect for strategy-instruction on the basis of the two-year intervention, whereas there was a significant positive interaction for strategy-instruction on reading comprehension in the first year. This difference can be explained by looking at the different roles that strategy instruction might have played in the first and second year of the intervention. The strategy component of the treatment was obviously more important in the first year, because students had to be familiarized at the start with the nature and function, of each strategy. In the second year of the treatment, however, it was not necessary to spend much attention to explaining the characteristics of the different reading strategies to the students, as they were already quite familiar with them. Students knew the characteristics of the different reading strategies, but now particularly needed to learn how to apply those strategies. Elaborate attention to teaching of the nature and function of the strategies in the second year can therefore be expected not to be more successful in fostering reading comprehension. Thus, it is not surprising that we found no overall moderation effect of strategy instruction after the two-year intervention.

Modeling

The moderation effect of modeling shows that growth in reading comprehension of the treatment and control students depends on teachers' attention to modeling (both by themselves and by students). It appeared that growth in reading comprehension in the case of less elaborate modeling (one standard deviation below average or on average) in the treatment group was less than in the control group. Only in the case of more elaborate modeling (one standard deviation above average) the growth in reading comprehension between the two groups appeared to be the same. Thus, elaborate modeling only served to prevent that the experimental students performed worse in reading comprehension.

In order to model reading strategies, a certain theatricality is needed. If the teacher is not comfortable with acting, this theatricality is difficult to master, as one teacher put it in our exit interviews: "It [modeling] didn't work, the students thought it was strange when I tried to model". Another teacher pointed out that she found it difficult to teach the students how to model. In both cases, modeling was jeopardized. The impressions of these two teachers were representative for the majority of the treatment teachers. This can be seen in the mean scores for modeling in our observations which were rather low. However, the significant moderation effect shows that a few of the treatment teachers nevertheless could make a difference by more elaborate modeling of reading strategies. Although the moderation did not result in more growth in reading

comprehension for the treatment group, it shows promise for reciprocal teaching in the future. Given that even the best teachers in our study did not perform optimally in modeling, there is reason to expect that when modeling is practiced optimally, results on students' reading comprehension will be better than we have found.

Group work

Regarding group work, we can conclude that the specific focus on coaching teachers in implementing this component of reciprocal teaching did result in a slight improvement in the quality of observed group work. Whereas in the first year teachers did not reach a higher average than 2 (meaning that group work was organized, but students received no teacher feedback on cooperation), in the second year the average score was higher (2.25 with a maximum of 3), meaning that feedback on collaboration was provided more frequently. Nevertheless, no positive effects of this improved group work were obtained in terms of improved reading comprehension, given that no significant moderation effect for group work was found.

In the original format of reciprocal teaching, small groups of students were taken out of the classroom. Under the guidance of a tutor, who had optimal control over the students' behavior, they practiced reading strategies while reading a text (Palincsar & Brown, 1984). In our treatment, the teachers were to manage up to five groups of students and provide guidance to all simultaneously. This means that compared to the original format, there is much less teacher control on collaboration and the quality of strategy use by the students in each group. The strength of reciprocal teaching may lie in the fact that there is enough time for the needs of each group of students. However, in a whole-classroom setting this time needs to be shared among multiple groups. This disadvantage of whole-classroom settings is supported by observations of Hacker and Tenet (2002). According to their in-depth analysis of teacher practice in reciprocal teaching, they concluded that group work was the most vulnerable component. The collaboration process between students was hampered because students did not practice reciprocal teaching in a productive way. Their discussions about the texts were rather superficial, and therefore did not reach a higher level of comprehension monitoring. In order to compensate for this problem, teachers often returned to whole-classroom instruction, thereby jeopardizing one of the most important aspects of reciprocal teaching: the fading of responsibility for the reading process to the students.

In addition, we have to consider the fact that our students were low-achievers and therefore may have needed much more guidance in group work than higher achieving students. Our students have to be supported

in comprehension monitoring, because many of them are not used making inferences and practicing other types of deeper comprehension processes (Trapman et al., 2014; Oakhill & Cain, 2007; Rapp, Van den Broek, McMaster, Kendeou, & Espin, 2007).

Moreover, it seems that the students needed more support for their collaboration process. From our own classroom observations of the treatment teachers we may conclude that the majority of them experienced serious problems with group work guidance, resulting in insufficient collaborative practice in reading strategies in the whole-classroom settings. In the exit interviews, teachers acknowledged that working in small groups is quite difficult in a whole-classroom setting. They found it hard to keep order and to keep the students motivated in the groups that the teacher was not supporting at that particular moment. Thus, one teacher asked herself: “Do I keep trying to work in groups, even if that is at the expense of learning outcomes?” Nevertheless, some of our teachers recognized the added value group work may have. In our exit interviews, a teacher, who was already proficient in applying group work before the start of the treatment, said: “If you mainly keep focused on whole-class instruction, it is difficult to get a glimpse of the reading process of the students. You keep repeating the reading strategies and hope that the students pick up what they need. For me, that is not enough. I want to exert more control on the [reading] process of the students.” Another teacher added that a big advantage of group work is that all students take multiple turns, and that they can react immediately to each other. In a whole-classroom approach, students have to wait longer to take turns, and there is no time for every student to take a turn. But, “you need to be able to steer the group work in such a way that they [the groups of students] work effectively”. These remarks about group work point to the fact that more intensive coaching may be needed to make this principle of reciprocal teaching successful. Accordingly, many teachers mentioned that they would have appreciated more coaching of group work. Alternatively, it may be needed to include extra classroom assistance as group tutors.

Suggestions for Further Research

The results of our study emphasize the importance of taking into account different aspects of quality of implementation as moderators in the analysis of treatment effects. Not taking into account quality of implementation may lead to overlooking meaningful effects, in particular in whole-classroom settings (Hulleman & Cordray, 2009; Larsen & Samdal, 2007; Swanson et al., 2013; Vaughn et al., 2013). Incorporating quality of implementation may also give clues to which treatment components

contribute to the treatment effects. To our knowledge, our study is the first that systematically analyzes the moderating role of implementation quality of reading comprehension instruction using principles of reciprocal teaching. We strongly recommend that future studies incorporate such moderation analyses in order to enhance our knowledge of conditions of successful application. In particular, it is of interest to find out what differences exist in successful implementation for different student populations (e.g., low ability vs high ability, younger vs older etc.).

Implications for Practice

Our observations of classroom practices and exit interviews revealed large differences between treatment teachers in how they implemented reciprocal teaching. Therefore, it is recommended that prior insight into classroom practices of individual teachers are used to adapt the contents of training and coaching to their specific needs. For example, for teachers who have no prior experience with managing multiple groups in a whole-classroom setting may need support dividing their attention among the groups in an efficient and effective manner. Aspects of group work, such as group composition, group-size and ability grouping can then be discussed in detail and adapted to the needs of teachers and their classes. Such prior knowledge of individual teachers' classroom practices is useful in optimizing conditions for experimental research into principles of reciprocal teaching, but it also may be useful for educational practice. Programs that use principles of reciprocal teaching in educational practice will certainly profit from such tailored training and coaching to the individual needs of teachers.

Finally, we need to acknowledge the fact that it is difficult to implement principles of reciprocal teaching for low-achieving adolescents in whole-classroom settings. Even in our two-year treatment, treatment teachers did not succeed in an optimal implementation of these principles. The main reason for this seems to be that the whole-classroom setting makes it difficult to attend to multiple groups of students at the same time and give them the guidance they need. However, given that the central objective of reading comprehension instruction is that students take more responsibility for their comprehension processes, there is no doubt that the quality of group work should be a prime concern for educational practice. It is important that such quality can be guaranteed, so that students may experience that the use of reading strategies is not the goal but the means for using textual information for reaching their *own* goals. In order to help students achieving these goals, a program should provide a basis for students, on which they can build their reading comprehension competency in a flexible way, thus stimulating motivation and self-efficacy. Students'

reading goals may be strictly related to the school context (such as content area learning), but they are also relevant in a much wider context, such as their future professional and societal careers.

Notes

1. The prevocational track is subdivided in three types. We selected our sample from the two lowest of these, representing about the 30% lowest scoring on the general attainment test.
2. A significant main effect of the ‘treatment variable’ indicates a significant difference between treatment and control group on the dependent variable at the start of the study, whereas the interaction between occasion and treatment indicates a difference in growth between treatment and control group on the repeatedly measured dependent variable (reading comprehension), which can be seen as the effect of the treatment.
3. The difference in $-2 \times$ Loglikelihood of nested models has a Chi-square distribution with a number of degrees of freedom equal to the difference in number of estimated parameters between both models.
4. The multilevel structure was tested but there was no significant variance at school level, therefore unilevel analyses were carried out.

Funding

This work was funded by Ministry of Education, Culture, and Science (OCW), the Netherlands.

ORCID

M. Okkinga  <http://orcid.org/0000-0002-0612-3242>

References

- Chambers Cantrell, S., Almasi, J.F., Rintamaa, M., & Carter, J.C. (2016). Supplemental reading strategy instruction for adolescents: A randomized trial and follow-up study. *The Journal of Educational Research*, 109(1), 7–26. 10.1080/00220671.2014.917258.
- Chiu, C. W. T. (1998, April). *Synthesizing metacognitive interventions: What training characteristics can improve reading performance?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- De Boer, H., Donker, A. S., & Van der Werf, M. P. C. (2014). Effects of the attributes of educational interventions on students’ academic performance: A meta-analysis. *Review of Educational Research*, 84(4), 509–545. doi:10.3102/0034654314540006
- De Corte, E., Verschaffel, L., & Van de Ven, A. (2001). Improving text comprehension strategies in upper primary school children: A design experiment. *The British Journal of Educational Psychology*, 71(Pt 4), 531–559. doi:10.1348/000709901158668

- Dole, J. A., Duffy, G. G., Roehler, L. R., & Pearson, P. D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research*, 61(2), 239–264. doi:10.3102/00346543061002239
- Duffy, G. G. (1993). Teachers' progress toward becoming expert strategy teachers. *The Elementary School Journal*, 94(2), 109–120. doi:10.1086/461754
- Dutch Education Inspectorate. (2008). *Basisvaardigheden taal in het voortgezet onderwijs: Resultaten van een inspectieonderzoek naar taalvaardigheid in de onderbouw van het vmbo en praktijkonderwijs* [Basic Language Skills in Secondary Education: Results of an Inspectorate Study into Language Skills in the First Two Years of Prevocational Secondary Education and Practical Training]. Utrecht: Dutch Education Inspectorate.
- Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading Comprehension Outcomes for Older Struggling Readers. *Review of Educational Research*, 79(1), 262–300. doi:10.3102/0034654308325998
- Field, A. P. (2009). *Discovering statistics using SPSS: And sex and drugs and rock 'n' roll* (3rd ed.). London: Sage Publications.
- Fogarty, M., Oslund, E., Simmons, D., Davis, J., Simmons, L., Anderson, L., ... Roberts, G. (2014). Examining the effectiveness of a multicomponent reading comprehension intervention in middle schools: A focus on treatment fidelity. *Educational Psychology Review*, 26(3), 425–449. doi:10.1007/s10648-014-9270-6
- Gille, E., Loijens, C., Noijons, J., & Zwitser, R. (2010). Resultaten PISA-2009. *Praktische kennis en vaardigheden van 15-jarigen*. Arnhem: Cito.
- Guthrie, J. T., & Wigfield, A. (2000). Engagement and motivation in reading. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Volume III* (pp. 403–422). New York: Erlbaum.
- Hacker, D. J., & Tenen, A. (2002). Implementing reciprocal teaching in the classroom: Overcoming obstacles and making modifications. *Journal of Educational Psychology*, 94(4), 699–718. doi:10.1037/0022-0663.94.4.699
- Hazenbergh, S., & Hulstijn, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics*, 17(2), 145–163. doi:10.1093/applin/17.2.145
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.
- Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, 2(1), 88–110. doi:10.1080/19345740802539325
- Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4), 441–480. doi:10.1016/0010-0285(76)90015-3
- Just, M. A., & Carpenter, P. A. (2004). A theory of reading: From eye fixations to comprehension. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 1182–1218). Newark, DE: International Reading Association.
- Kelly, M., Moore, D. W., & Tuck, B. F. (1994). Reciprocal teaching in a regular primary school classroom. *The Journal of Educational Research*, 88(1), 53–61. doi:10.1080/00220671.1994.9944834
- Kordes, J., Bolsinova, M., Limpens, G., & Stolwijk, R. (2013). *PISA resultaten 2012: Praktische kennis en vaardigheden van 15-jarigen*. Nederlandse uitkomsten van het Programme for International Student Assessment (PISA) op het gebied van leesvaardigheid, wiskunde en natuurwetenschappen in het jaar 2012 [PISA

- results 2012: Skills and knowledge of 15-years olds. Results of Dutch students in PISA in regard to reading, mathematics and science in the year 2012*. Arnhem, The Netherlands: Cito.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2), 293–323. doi:10.1016/0010-0285(74)90015-2
- Larsen, T., & Samdal, O. (2007). Implementing second step: Balancing fidelity and program adaptation. *Journal of Educational and Psychological Consultation*, 17(1), 1–29. doi:10.1080/10474410709336588
- Logan, S., & Johnston, R. (2009). Gender differences in reading ability and attitudes: Examining where the differences lie. *Journal of Research in Reading*, 32(2), 199–214. doi:10.1111/j.1467-9817.2008.01389.x
- McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, 44(3), 218–253. doi:10.1598/RRQ.44.3.1
- Ministry of Education, Culture, & Science. (2006). *The education system in the Netherlands*. The Hague: Ministry of Education, Culture, & Science/Dutch Eurydice Unit.
- Muijselaar, M., Swart, N., Steenbeek-Planting, E., Droop, M., Verhoeven, L., & de Jong, P. (2018). The effect of a strategy training on reading comprehension in fourth-grade students. *The Journal of Educational Research*, 111(6), 690–703. doi:10.1080/00220671.2017.1396439
- National Reading Panel (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.
- Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions and technologies* (pp. 47–72). New York: NY: Erlbaum.
- Okkinga, M., Van Steensel, R., Van Gelderen, A., & Slegers, P. (2018). Effects of reciprocal teaching on reading comprehension of low-achieving adolescents. The importance of specific teacher skills. *Journal of Research in Reading*, 41(1), 20–41. <https://doi.org/10.1111/1467-9817.12082>.
- Okkinga, M., Van Steensel, R., Van Gelderen, A., Van Schooten, E., Slegers, P., & Arends, L.R. (2018). Effectiveness of reading-strategy interventions in whole classrooms: A meta-analysis. *Educational Psychology Review*, 30, 1215–1239. <https://doi.org/10.1007/s10648-018-9445-7>.
- Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, 23(2), 189–208. doi:10.1007/s11145-008-9159-1
- Organisation for Economic Co-operation and Development [OECD]. (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.
- Organisation for Economic Co-operation and Development [OECD]. (2014). Profile of student performance in reading. In *PISA 2012 Results: What students know and can do (Volume I, Revised edition, February 2014): Student performance in mathematics, reading and science*. OECD Publishing. doi:10.1787/9789264201118-en

- Palincsar, A. S., & Brown, A. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*, 117–175. doi:10.1207/s1532690xci0102_1
- Palincsar, A. S., Brown, A., & Martin, S. M. (1987). Peer interaction in reading comprehension instruction. *Educational Psychologist, 22*(3), 231–253. doi:10.1080/00461520.1987.9653051
- Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.
- Rapp, D. N., Van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*(4), 289–312. doi:10.1080/10888430701530417
- Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A user's guide to MlwiN. Version 2.10*. Bristol: University of Bristol, Centre for Multilevel Modelling.
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales. Section 1: General overview*. San Antonio, TX: Harcourt Assessment.
- Roberts, G., Fletcher, J. M., Stuebing, K. K., Barth, A. E., & Vaughn, S. (2013). Treatment effects for adolescent struggling readers: An application of moderated mediation. *Learning and Individual Differences, 23* (1), 10–21. doi:10.1016/j.lindif.2012.09.008
- Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research, 64*(4), 479–530. doi:10.3102/0034654306400447
- Rumelhart, D. E. (2004). Toward an interactive model of reading. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 719–747). Newark, DE: International Reading Association.
- Samuels, S. J. (2004). Toward a theory of automatic information processing in reading, revisited. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 1127–1148). Newark, DE: International Reading Association.
- Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., & Torgesen, J. K. (2007). *Interventions for adolescent struggling readers: A meta-analysis with implications for practice*. Portsmouth, NH: Center on Instruction.
- Scammacca, N., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4-12: 1980-2011. *Journal of Learning Disabilities, 48*(4), 369–390. doi:10.1177/0022219413504995
- Schaffner, E., Philipp, M., & Schiefele, U. (2016). Reciprocal effects between intrinsic reading motivation and reading competence? A cross-lagged panel model for academic track and nonacademic track students. *Journal of Research in Reading, 39*(1), 19–18. doi:10.1111/1467-9817.12027
- Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading, 3*(3), 257–279. doi:10.1207/s1532799xssr0303_4
- Seymour, J. R., & Osana, H. P. (2003). Reciprocal teaching procedures and principles: Two teachers' developing understanding. *Teaching and Teacher Education, 19*(3), 325–344. doi:10.1016/S0742-051X(03)00018-0
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

- Simmons, D., Fogarty, M., Oslund, E. L., Simmons, L., Hairrell, A., Davis, J., ... Fall, A.-M. (2014). Integrating content knowledge-building and student-regulated comprehension practices in secondary English arts classes. *Journal of Research on Educational Effectiveness*, 7(4), 309–330. doi:10.1080/19345747.2013.836766
- Slavin, R.E., Cheung, A., Groff, C., & Lake, C. Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, (2008). 43(3), 290–322.
- Spörer, N., Brunstein, J. C., & Kieschke, U. (2009). Improving students' reading comprehension skills: Effects of strategy instruction and reciprocal teaching. *Learning and Instruction*, 19(3), 272–286. doi:10.1016/j.learninstruc.2008.05.003
- Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, J. (2013). Intervention fidelity in special and general education research journals. *The Journal of Special Education*, 47(1), 3–13. doi:10.1177/0022466911141951
- Trapman, M., Van Gelderen, A., Van Steensel, R., Van Schooten, E., & Hulstijn, J. (2014). Linguistic knowledge, fluency and meta-cognitive knowledge as components of reading comprehension in adolescent low achievers: differences between monolinguals and bilinguals. *Journal of Research in Reading*, 37(S1), S3–S21. doi:10.1111/j.1467-9817.2012.01539.x.
- Trapman, M., van Gelderen, A., van Schooten, E., & Hulstijn, J. (2017). Reading comprehension level and development in native and language minority adolescent low achievers: Roles of linguistic and metacognitive knowledge and fluency. *Reading & Writing Quarterly*, 33(3), 239–257. doi:10.1080/10573569.2016.1183541
- Van Gelderen, A., Schoonen, R., De Gloppe, K., Hulstijn, J., Snellings, P., Simis, A., & Stevenson, M. (2003). Roles of linguistic knowledge, metacognitive knowledge and processing speed in L3, L2 and L1 reading comprehension. *International Journal of Bilingualism*, 7(1), 7–25. doi:10.1177/13670069030070010201.
- Van Gelderen, A., Schoonen, R., Stoel, R.D., De Gloppe, K., & Hulstijn, J. (2007). Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology*, 99(3), 477–491. <https://doi.org/10.1037/0022-0663.99.3.477>.
- Van Gelderen, A., Schoonen, R., De Gloppe, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic Knowledge, Processing Speed, and Metacognitive Knowledge in First- and Second-Language Reading Comprehension: A Componential Analysis. *Journal of Educational Psychology*, 96(1), 19–30. <https://doi.org/10.1037/0022-0663.96.1.19>.
- Van Silfhout, G., Evers-Vermeul, J., Mak, W. M., & Sanders, T. J. (2014). Connectives and layout as processing signals: How textual features affect students' processing and text representation. *Journal of Educational Psychology*, 106(4), 1036–1048. doi:10.1037/a0036293
- Vaughn, S., Roberts, G., Klingner, J. K., Swanson, E. A., Boardman, A., Stillman-Spisak, S. J., ... Leroux, A. J. (2013). Collaborative strategic reading: Findings from experienced implementers. *Journal of Research on Educational Effectiveness*, 6(2), 137–163. doi:10.1080/19345747.2012.741661
- Veenman, M. V. J., Hout-Wolters, B. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, 1(1), 3–14. doi:10.1007/s11409-006-6893-0
- Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology*, 22(3), 407–423. doi:10.1002/acp.1414

Appendix A

Table A. Examples of strategy assignments, translated from several assignment sheets from the program “Nieuwsbegrip”

Strategy	Example
Predicting	This text has five subheadings. Write down for each subheading a) which thoughts it evokes and b) what you already know about the subject addressed in the subheading.
Summarizing	Read the text. Read paragraph by paragraph and underline in each paragraph the most important information. For each paragraph, write one or two sentences summarizing it. Use the words you underlined.
Self-questioning	Read the text. Note at least five questions that spring to mind while reading.
Clarifying	Search the text for difficult words. Try to uncover their meaning using these hints: a) reread the previous piece of text or read on, b) look at the illustrations in the text, c) look at the word: you might know part of the word, d) sometimes you have to use your own knowledge to figure out word meanings, or e) use a dictionary.
Interpreting cohesive ties	Read the text. Underline the signal words. Answer the questions, while noting the signal words: <ul style="list-style-type: none"> • Which contrast is explained in lines 16-17? [signal word = however] • Why are energy boosters unfit as sports drinks? [signal word = hence]

Appendix B

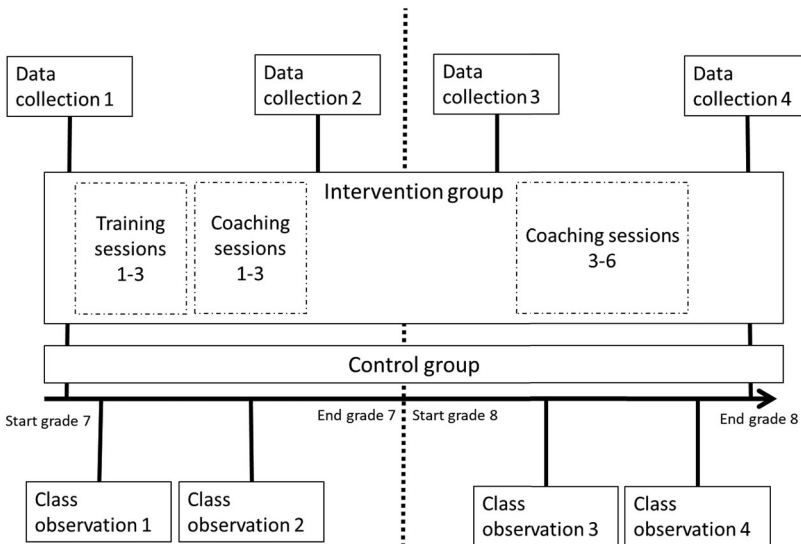


Figure B. Overview of the timeline of all research activities during the experiment.

Appendix C

Table C. Multilevel Analyses with Reading Comprehension (repeatedly measured) as Dependent Variable to Establish Multilevel Structure of Data (N = 952 cases/238 students).

Model	model C-0	model C-1	model C-2	model C-3	model C-4	model C-5	model C-6a
Fixed part							
Intercept	35.232 (.499)	34.667 (.965)	34.385 (1.159)	34.633 (.985)	34.635 (.983)	34.608 (1.050)	34.593 (1.094)
Occasion (in months)	.150 (.019)	.150 (.019)	.150 (.019)	.150 (.021)	.150 (.021)	.152 (.038)	.156 (.040)
Random part (variances)							
School			8.497 (6.123)				
Class		12.912 (5.315)	4.599 (3.730)	14.069 (5.651)	13.968 (5.636)	16.506 (6.575)	18.218 (7.174)
Class slope variance occasion						^b 0	.022 (.010)
Class covariance slope × intercept							-.355 (.209)
Student	43.576 (4.486)	32.237 (3.586)	32.302 (3.592)	28.643 (3.485)	29.228 (4.139)	27.849 (4.020)	27.553 (3.994)
Student slope variance occasion				.036 (.009)	-.038 (.011)	.019 (.010)	.018 (.010)
Student covariance slope × intercept				^b 0	-.041 (.161)	.121 (.146)	.142 (.144)
Occasion (rep. measures)	21.050 (1.114)	21.050 (1.114)	21.050 (1.114)	17.949 (1.101)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)
Deviance testing							
-2*loglikelihood (deviance)	6132.566	6100.015	6097.977	6077.942	6077.883	6058.501	6054.507
Difference between		32.551	2.038	22.037	.059	19.382	3.994
-2*loglikelihood							
Difference df		1	1	1	1	1	1
Compared to model	C-0	C-1	C-1	C-1	C-3	C-4	C-5

^aModel C-6 shows that the multilevel structure of the data consists of three levels: Class, Student and Occasion. Random slopes are applied for both the class and student level, as slopes differ significantly for both classes and students

^b= fixed at zero

Bold = p < .05; italicized = p < .01; bold and italicized = p < .001.

Appendix D

Table D. Multilevel Analyses with Reading Comprehension (repeatedly measured) as Dependent Variable to Verify Influence of Teacher Replacement (0 = no, 1 = yes) and Cancelled Classes (0 = less than 6 weeks, 1 = 6 weeks or more) (N = 952/238)

Model	model C-6a	model D-1	model D-2b	model D-3	model D-4c
Fixed part					
Intercept	34.593 (1.094)	35.292 (1.914)	34.940 (1.227)	34.961 (1.149)	34.666 (1.121)
Occasion (in months)	.156 (.040)	.157 (.040)	.178 (.044)	.157 (.040)	.173 (.038)
Teacher replacement (yes = 1)		-3.160 (2.153)	-1.558 (2.619)		
Teacher replacement × occasion			-1.101 (.094)	-6.995 (3.997)	-1.392 (4.889)
Cancelled classes					-333(.169)
Cancelled classes × occasion					
Random part (variances)					
Class	18.218 (7.174)	18.306 (7.196)	17.829 (7.044)	19.654 (7.650)	18.106 (7.136)
Class slope variance occasion	.022 (.010)	.022 (.010)	.020 (.009)	.022 (.010)	.017 (.008)
Class covariance slope × intercept	-.355 (.209)	-.410 (.216)	-.380 (.206)	-.466 (.229)	-.378 (.198)
Student	27.553 (3.994)	27.542 (3.993)	27.548 (3.993)	27.559 (3.994)	27.555 (3.994)
Student slope variance occasion	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)
Student covariance slope × intercept	.142 (.144)	.142 (.144)	.142 (.144)	.142 (.144)	.142 (.144)
Occasion (rep. measures)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)
Deviance					
-2*loglikelihood (deviance)	6054.507	6052.474	6051.360	6051.972	6048.400
Difference between		2.033	1.114	2.535	3.572
-2*loglikelihood					
Difference df		1	1	1	1
Compared to model		C-6	D-1	C-6	C-3

^a = Model C-6 from Appendix C.

^b Model D-2 shows that there is no significant differential growth for classes in which teacher replacement took place compared to classes in which this was not the case. ^c Model D-4 shows that cancelled classes did not account for significant differential growth in classes with more than 6 weeks of cancelled classes compared to less than 6 weeks of cancelled classes. Therefore, both teacher replacement and cancelled classes were omitted from further analyses.

Bold = $p < .05$, italicized = $p < .001$.

Appendix E

Table E. Multilevel Analyses with Reading Comprehension (repeatedly measured) as Dependent Variable to Establish Influence of Student-Level Variables (IQ, Gender, Vocabulary Knowledge, Metacognitive Knowledge, and Age (N = 952 cases/238 students))

Model	Model C-6b	Model E-1	Model E-2	Model E-3	Model E-4	Model E-5	Model E-6
Fixed part							
Intercept	34.593 (1.094)	34.651 (1.023)	33.628 (1.108)	34.550 (.842)	35.019 (.761)	35.058 (.739)	35.057 (.740)
Occasion (in months)	.156 (.040)	.155 (.040)	.155 (.040)	.155 (.040)	.155 (.040)	.155 (.040)	.155 (.040)
IQ ^a		.391 (.074)	.374 (.074)	.235 (.063)	.240 (.063)	.219 (.062)	.219 (.062)
Gender (0 = male, 1 = female)			1.904 (.770)	.858 (.646)			
Vocabulary knowledge ^a				.521 (.047)	.531 (.047)	.479 (.048)	.479 (.048)
Metacognitive knowledge ^a						.274 (.079)	.274 (.079)
Age (in days/365) ^a							.084 (.570)
Random part (variances)							
Class	18.218 (7.174)	15.791 (6.263)	16.015 (6.332)	8.311 (3.485)	8.185 (3.466)	7.628 (3.251)	7.694 (3.254)
Class slope variance occasion	.022 (.010)	-.022 (.010)	.022 (.010)	-.022 (.010)	.022 (.010)	-.022 (.010)	.022 (.010)
Class covariance slope × intercept	-.355 (.209)	-.291 (.192)	-.286 (.192)	-.318 (.157)	-.321 (.157)	-.325 (.154)	-.325 (.154)
Student	27.553 (3.994)	23.962 (3.660)	23.574 (3.624)	14.164 (2.763)	14.159 (2.761)	13.091 (2.665)	13.081 (2.665)
Student slope variance occasion	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)
Student covariance slope × intercept	.142 (.144)	.126 (.140)	.093 (.140)	-.092 (.126)	.107 (.125)	.117 (.123)	.117 (.123)
Occasion (rep. measures)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)
Deviance testing							
-2* ^b loglikelihood (deviance)	6054.507	6028.443	6022.513	5926.841	5928.564	5916.945	5916.923
Difference between		26.064	5.930	95.692	1.723 ^c	11.619 ^d	.022
-2* ^b loglikelihood							
Difference df		1	1	1	1	1	1
Compared to model	C-6	E-1	E-1	E-2	E-3 ^c	E-4 ^d	E-5

^aVariable is grand mean centered.

^b—Model 6 from Appendix C.

^cModel E-4 is more parsimonious than Model E-3, therefore gender is omitted.

^dModel E-5 is more parsimonious than Model E-6, therefore Model E-5 is used as the baseline model in subsequent analyses.

Bold = p < .05; italicized = p < .01; bold and italicized = p < .001.

Appendix F

Table F1. Multilevel Analyses with Reading Comprehension (repeatedly measured) as Dependent Variable to Establish Influence of Interaction between Strategy-Instruction, Occasion and Treatment, after Correcting for Control Variables (N = 952 cases/238 student)

Model	Model F1-1	Model F1-2	Model F1-3	Model F1-4
Fixed part				
Intercept	34.671 (.986)	34.494 (1.005)	34.174 (1.003)	34.137 (1.030)
Occasion (in months)	.207 (.051)	.207 (.051)	.232 (.050)	.234 (.054)
IQ ^a	.219 (.062)	.221 (.062)	.221 (.062)	.221 (.062)
Vocabulary ^a	.479 (.048)	.477 (.048)	.477 (.048)	.477 (.048)
Metacognitive knowledge ^a	.277 (.079)	.278 (.079)	.278 (.079)	.278 (.079)
Treatment (1 = treatment, 0 = control)	.871 (1.471)	1.322 (1.491)	2.006 (1.524)	2.008 (1.523)
Treatment × occasion	-.114 (.076)	-.155 (.076)	-.168 (.079)	-.145 (.169)
Strategy-instruction ^a		-.659 (.742)	-.1567 (1.283)	-.1.681 (1.474)
Treatment × strategy-instruction			-.173 (1.486)	.036 (1.995)
Occasion × strategy-instruction			.079 (.052)	.088 (.077)
Occasion × strategy-instruction × treatment				-.016 (.104)
Random part (variances)				
Class	7.477 (3.182)	6.612 (2.909)	6.126 (2.732)	6.125 (2.732)
Class slope variance occasion	.019 (.009)	.019 (.009)	.015 (.007)	.015 (.007)
Class covariance slope × intercept	-.298 (.143)	-.270 (.135)	-.231 (.121)	-.231 (.121)
Student	13.095 (2.666)	13.098 (2.665)	13.101 (2.666)	13.101 (2.666)
Student slope variance occasion	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)
Student covariance slope × intercept	.116 (.124)	.115 (.123)	.115 (.124)	.115 (.124)
Occasion (rep. measures)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)
Deviance testing				
-2* ^a loglikelihood (deviance)	5913.722	5913.708	5911.543	5911.518
Difference between	.676	.014	2.165	.025
-2* ^a loglikelihood				
Difference df	1	1	1	1
Compared to model	3-2	F1-1	F1-2	F1-3

^aVariable is grand mean centered, ^b = Model 3-2 from Table 3. **Bold = p < .05; italicized = p < .01; bold and italicized = p < .001.**

Table F2. Multilevel Analyses with Reading Comprehension (repeatedly measured) as Dependent Variable to Establish Influence of Interaction between Group Work, Occasion and Treatment, after Correcting for Control Variables (N = 952 cases/238 students)

Model	Model 3-2b	Model F2-1	Model F2-2	Model F2-3	Model F2-4
Fixed part					
Intercept	34.671 (.986)	34.242 (1.090)	31.946 (1.859)	31.667 (1.922)	31.799 (2.480)
Occasion (in months)	.207 (.051)	.207 (.051)	.207 (.051)	.228 (.063)	.218 (.134)
IQ ^a	.219 (.062)	.215 (.062)	.219 (.062)	.219 (.062)	.219 (.062)
Vocabulary ^a	.479 (.048)	.478 (.048)	.482 (.048)	.482 (.048)	.482 (.048)
Metacognitive knowledge ^a	.277 (.079)	.279 (.079)	.289 (.079)	.289 (.079)	.289 (.079)
Treatment (1 = treatment, 0 = control)	.871 (1.471)	1.917 (1.819)	3.839 (2.211)	4.462 (2.470)	4.347 (2.826)
Treatment × occasion	-.114 (.076)	-.115 (.076)	-.115 (.076)	-.161 (.112)	-.152 (.153)
Group work ^a		-.565 (.665)	-.3599 (2.114)	-.3964 (2.211)	-.3792 (3.017)
Treatment × group work			3.346 (2.219)	3.336 (2.218)	3.145 (3.170)
Treatment × group work × occasion				.028 (.049)	.015 (.163)
Occasion × group work × treatment					.014 (.171)
Random part (variances)					
Class	7.477 (3.182)	7.064 (3.046)	6.602 (2.896)	6.533 (2.874)	6.527 (2.872)
Class slope variance occasion	.019 (.009)	.019 (.009)	.019 (.009)	.018 (.008)	.018 (.008)
Class covariance slope × intercept	-.298 (.143)	-.288 (.140)	-.288 (.137)	-.282 (.136)	-.282 (.136)
Student	13.095 (2.666)	13.089 (2.665)	13.085 (2.664)	13.082 (2.664)	13.083 (2.664)
Student slope variance occasion	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)	.018 (.010)
Student covariance slope × intercept	.116 (.124)	.116 (.123)	.115 (.123)	.115 (.123)	.115 (.123)
Occasion (rep. measures)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)	17.861 (1.158)
Deviance testing					
-2*loglikelihood (deviance)	5914.398	5913.699	5911.510	5911.195	5911.188
Difference between		.699	2.189	.315	.007
-2*loglikelihood		1	1	1	1
Difference df		3-2	F2-1	F2-2	F2-3
Compared to model					

^aVariable is grand mean centered, ^b = Model 3-2 from Table 3.
Bold = p < .05; italicized = p < .01; bold and italicized = p < .001.