Working memory training improves reasoning skills in secondary social studies education: evidence from an experimental study

Roel J.F.J. Ariës, Joris Ghysels, Wim Groot & Henriette Maassen van den Brink
Working memory training improves reasoning skills in secondary social studies education: evidence from an experimental study

Roel J.F.J. Ariës; Joris Ghysels; Wim Groot; Henriëtte Maassen van den Brink;

Abstract

Evidence suggests that Working memory abilities and transfer of (content-based) WM training correlate with (school-based) deductive reasoning achievements. In this study, a combined WM-capacity and WM-reasoning strategies training is incorporated in secondary school social studies curricula to investigate its effects on reasoning achievements compared to control group conditions. Four secondary classes in three schools in the Netherlands participated in the experiment with a total of 81 general education medium track 16 year old students. WM-capacity and reasoning achievements improved significantly after 4 and 8 weeks of training and remained significant 8 weeks after training ended. The (significant) gain in reasoning abilities is demonstrated in both experimental subgroups, while both control group results did not improve. The study supports the notion that transfer of WM-training to deductive (school-based) reasoning can be successful when WM-capacity training is supported by strategy-training and both components contain content-based contexts. Furthermore, it supports preliminary evidence of accelerated learning following WM-training.

Key words: reasoning abilities, secondary education, social studies, brain based learning, working memory

Word count: 5,950

1 Corresponding author
2 Top Institute for Evidence Based Education Research (TIER), TIER-Maastricht University, Kapoenstraat 2, 6211 KW Maastricht, the Netherlands.
Introduction

This paper builds upon neuro-scientific training methods to create a brain based learning model that effectively improves reasoning abilities in social sciences courses. Secondary education students find it difficult to reason in social studies school tests, while these tests frequently contain a large number of reasoning questions (Sluijsmans, 2013). Neuropsychological research provides evidence that training of working memory (WM) improves reasoning skills in both math and reading comprehension (Conway, Kane & Engle, 2003; Cheshire, Ball & Lewis, 2005; Karbach, Strobach & Schubert, 2014; Nevo & Breznitz 2014). However, little research is conducted that tests whether working memory training improves reasoning abilities in secondary school classroom settings in social sciences education (e.g. history, geography, social studies, economics) (Olesen, Westerberg & Klingberg. 2004; Fuster 2005; Holmes, Gathercole & Dunning, 2009). Therefore more research has to be conducted in order to understand how these training methods can be successfully implemented into secondary social sciences curricula. A systematic review of the literature on school-based WM training (Ariës, Maassen van den Brink & Groot, 2014b) concluded that working memory training should consist of both a short term memory and a long term memory component. In the present study, we therefore implemented a two-component WM training method in a six week training period in secondary social studies education.

Reasoning in history education is defined by Leinhardt (1994) as ‘the process by which central facts and concepts are arranged to build a historical case’. According to Van Drie and Van Boxtel (2008), this requires analysis, synthesis, hypothesis, generalization and interpretation of questions, sources and retained knowledge. Van Drie
and Van Boxtel (2008) concluded that the main reasons for students to underachieve are attributed to students frequently being unable to (1) take into account alternative views, use sources extensively, (2) acquire detailed factual knowledge and a broader frame of reference, (3) judge the past by its own standards, (4) take into account the process of continuity and (5) understand many substantive concepts. Though research on the relevance of reasoning for social studies courses has not been found, we assume that the analyses of both Leinhardt (1994) and Van Drie and Van Boxtel (2008) on reasoning in history education can be transferred to social studies courses because of its similarities with history education. Both history and social studies are courses of the humanities. Students have to (1) develop both general and domain-specific thinking strategies that are used in all of the social sciences education, (2) address reasoning questions that meet the criteria of reasoning as defined by Van Drie and Van Boxtel (2008) and (3) address similar subjects. For instance, the main subjects of the social studies course in Dutch secondary education are 1. parliamentary democracy, 2. the Dutch constitutional state, 3. multicultural societies, 4. the welfare state (College voor Examens [Commission for exam standards], 2013). The same subjects are addressed in secondary history courses, but in a historical context (College voor Examens [Commission for exam standards], 2012). The social studies course addresses the four main subjects in a contemporary context. Because of the generality of its subjects, the social studies course addresses historic, geographic, economic, philosophic and art historic subjects with which secondary education courses it is closely related.

WM, the cognitive system that monitors higher cognitive processes in the brain, plays a crucial role in developing reasoning skills. It underlies several cognitive abilities,
including logical reasoning and problem solving (Klingberg, 2009). The relation between WM capacity and (deductive) reasoning abilities is causal and underlies the frequent inability of children with poor WM to make progress in higher cognitive skills, such as reasoning (Alloway & Gathercole, 2009; Holmes & Gathercole 2014; Süß, Oberauer, Wittmann, Wilhelm & Schulze, 2002). This process is affected by the two functions of WM. First, the cognitive function, which stores and manipulates information during reasoning tasks, depends strongly on the capacity of information that can be stored in WM (Fuster, 2005; Goldberg, 2010). When more information can be stored and, subsequently, manipulated, reasoning abilities are improved (Olesen et al., 2004). Short term storage and manipulation of information can be improved by training the cognitive WM function (Conway, Kane & Engle, 2003; Jaeggi, Buschkuehl, Jonides & Perrig, 2008). Second, pattern recognition of reasoning structures is regulated by the meta-cognitive function of WM (Gazzaniga, Ivry, Mangun & Steven, 2009). When frequently exposed to similar reasoning contexts, a student can develop and internalize matched response strategies (Gold, Berman, Randolph, Goldberg & Weinberger, 1996). In this view WM will form blueprints and will plan for rational analysis and analytical methods more effectively when trained via meta-cognitive training. This also causes WM to develop memories of previous patterns that successfully lead to solutions of problems (Goldberg, 2010). Consequently, this can be used to solve new reasoning problems, making reasoning processes more efficient and more effective (Kramarski & Mevarach, 2003; Cheshire et al., 2005).

A school based training method which contains domain specific content trained by both cognitive and meta-cognitive WM training integrates several general and
domain-specific reasoning processes. Combined training better addresses course-specific reasoning problems and therefore improves the effect of the training. The study on secondary history courses (Ariës et al., 2014a), which integrated general and domain-specific reasoning processes, showed that significant improvements in reasoning were caused by combined WM-training. This was a confirmation of conclusions from comparable research in which WM-capacity and meta-cognitive strategies were trained to benefit reasoning (Mevarach & Kramarski, 2003; Olesen et al. 2004; Holmes et al., 2009). Furthermore, it showed that reasoning structures were internalized which consequently improved achievements in reasoning tasks and strategies.

In the present study we investigate whether cognitive and meta-cognitive working memory training, based on subject-matter knowledge in the social studies course of secondary education school curriculum, results in an effect on students’ achievements in tests for which reasoning abilities are required. In three secondary schools in the Netherlands we conduct an experiment that is based on working memory training tasks. The brain-based learning strategy includes domain specific cognitive and meta-cognitive working memory tasks that are related to reasoning. The aim of the training intervention is to (1) investigate the effect of the complete training on achievement in reasoning tests and (2) its durability after completion of the training. The intervention is implemented in the existing social studies course curriculum. The subject-matter approach of the intervention contributes to a specific training that, according to earlier research, is expected to result in an improvement of reasoning achievements (Van der Sluis, De Jong & Van der Leij, 2007; Lee et al., 2009).
The experimental groups performed better on a battery of cognitive (Automated Working Memory Assessment) and meta-cognitive (IMPROVE) tasks after pretesting on reasoning abilities (Davidson, Amso, Anderson & Diamond 2006; Jolles, De Groot, Van Benthem, Dekkers, De Glopper et al., 2006). These tests were modifications of the original tests by Alloway and Gathercole (2009) and Mevarach and Kramarski (2003). After the training the students were post tested to evaluate improvements in cognitive and meta-cognitive working memory and reasoning abilities. Experimental groups were then compared with data of control groups that were trained by using ‘low-dose’ working memory tasks, such as evaluating reasoning questions using whole-class teaching methods.

Methods

Participants 81 students from four heterogeneous social studies classes in higher secondary general education medium track (HAVO), divided between three schools in the Netherlands, participated in the study. The classes were compiled by the school based on the students’ schedules. Because social studies is an extracurricular course, all students followed the course by choice. The control group (42 students) and the experimental group (39 students) both contained 1 or 2 pupils who are dyslectic, which can influence test outcomes caused by limitations in verbal im- and/or expression. The mean age (15.80; SD = .75) and the number of participants that repeated the class or came from a lower or higher track was approximately the same for both groups, as was the male/female ratio. All students lived in or near the urban areas of Maastricht, Heerlen
and Geleen, which are three neighboring urban areas in the most southern part of the Netherlands that are approximately 18 kilometers apart from each other.

The control group consisted of two classes (one Maastricht and one Geleen class). Two other classes (one Heerlen and one Geleen class) formed the experimental group. Three experienced teachers (all male) had been appointed to the classes prior to the intervention. All teachers used a conventional instruction method in their courses, which contains an introduction to new subjects and the use of whole class dialog and discussion to answer questions. The classes were randomly assigned to the control or experimental group. Only the Geleen teacher taught in both control and experimental groups. Therefore part of the effect in both the Heerlen and Maastricht groups could be attributed to ‘teacher effect’, ‘school effect’ or other context effects.

**Intervention** In this study, we duplicated Ariës et al.’s (2014a) combined training method and studied whether the training method, that has been tested in history courses, could be implemented in secondary education social studies courses. The following description of the tasks is derived from Ariës et al. (2014a) study on improving reasoning skills in secondary history education by working memory training.

In cognitive tasks, participants saw a series of twenty domain specific nouns, successively presented in a PowerPoint presentation at the rate of three seconds per stimulus in the n-back task. The task was to remember the sequence in which the nouns were presented. Most of the nouns that appeared were presented in black, but when a green colored noun appeared on the screen, participants had to remember the previous noun and write it on an answer form within five seconds. The tasks were presented in 1-
back, 2-back and dual 2-back batteries (Appendix A), which are derived from the standardized Automated Working Memory Assessment test battery (Alloway & Gathercole, 2008).

The Odd One Out task presented four domain specific nouns or images on a video screen in a PowerPoint presentation during 12 seconds. In the sequence, one noun or image was the ‘odd one out’, which participants had to write down on an answer form when all nouns or images were shown (Appendix B).

The experimental groups were trained on both tasks, n-back and Odd One Out. The content of these training tasks, which were content based social studies subjects, was the same as in the intermediate and posttests, but not the pretests which contained historical content. Furthermore, the dimensions of the cognitive training did not differ from the tests which contained 1-back, 2-back, dual 2-back, three verbal and three visual Odd One Out tasks during a section (30 minutes) of one lesson (50 minutes) per week. The content of the tasks was equivalent to the subjects at hand (Mass Media and Criminal Law) and to the conventional instruction method that was used in control group lessons.

We used the standardized IMPROVE-method (Appendix C) for meta-cognitive WM training (Mevarech & Kramarski 2003), which is used as a format to develop reasoning strategies. The training was performed in heterogeneous groups that contained three or four participants: one low, one or two middle, one high achieving student. Four reasoning questions were successively presented, 10 minutes per question. The question had to be addressed by one of the groups’ participants by using the IMPROVE-questionnaire to analyze the reasoning process. Directions by the other members of the group were given when the participant could not answer one of the questions of the
questionnaire. After 10 minutes, a new participant addressed a new problem. The teacher participated in each group’s reasoning process during two minutes per problem in the same way that the other members of the group participated. Tasks were trained in one 50 minute social studies lesson per week over a six week period.

**Data Collection** Combined effects of cognitive and meta-cognitive WM training were measured by a pretest, an intermediate test and two posttests in the intervention and control groups. Pretests measured the homogeneity of the experimental and control groups. One intermediate and two posttests measured short and long-term effects at the end of the training period and eight weeks after the first posttest to test for effect durability.

Tests to measure cognitive WM function levels were conducted on two tasks, the n-back and Odd One Out tasks (Jaeggi et al., 2008, Holmes et al., 2009). The tasks engage processes required for the management of different tasks, engage executive processes required for each task and discourages the development of task-specific strategies and the engagement of automatic processes. Furthermore, both tasks test the storage capacity and manipulation of information in WM. During the intervention period experimental groups were trained on altered, i.e. domain-specific, versions of the task standards and used nouns and verbs to train both verbal and visual components of cognitive WM (Ariës et al., 2014b). The tests were conducted on the same tasks, but the content differed from the content that was used in the training tasks: historical nouns and images were used to avoid that students of both groups were accustomed to the nouns and images that were shown in the intervention period, which could positively affect test
results. Participants could score a maximum of 58 points (one point per good answer) on the cognitive test. In this test, the participants were not tested for substantive knowledge, but for working memory capacity.

To test for acceleration of reasoning skills, all groups were tested by analyzing the answers on reasoning questions in official school tests. The tests were compiled by the three social studies teachers and contained a number of knowledge and reasoning questions. The participant’s reasoning achievements were analyzed by analyses of the correct responses on reasoning questions compared to the examination model with a maximum of 30 points per test.

**Design of the experiment** This experimental study compares the effects of the brain-based intervention with conventional instruction methods. The experiment employed pretests, intermediate tests and posttests to determine causal relationships. Research was conducted in four 4th-grade classes of Higher General Secondary Education in the Netherlands. Participants were nested in multiple classes in 3 different schools. Students of two classes were assigned to the experimental condition and two other classes to the control condition.

In a between subjects design, the two classes of the experimental group were trained by using the cognitive and meta-cognitive tasks to improve reasoning abilities. The two classes of the control group were trained by method of conventional instruction. In the latter, the teacher introduced new subjects and participants answered (reasoning) questions from the textbook that were verified by interaction between the teacher and all of the students.
To measure the effects of the intervention on cognitive working memory function levels and reasoning achievements, intermediate and post test scores from the experimental group and from the control group were analyzed independently by using paired t-tests. Reasoning achievement test scores of experimental and control groups were also compared by using independent t-tests, as the Kolmogorov-Smirnov test showed that the distribution in both experimental, $D(39) = 0.14, p > .05$, and control groups, $D(42) = 0.12, p > .05$, suggested no deviation from normality in the regular school pretest scores. Also, Levene’s test for equality of error variances showed no significant values $F(1, 79) = 0.006, p = 0.94$.

**Organization** Experimental and control groups studied different subjects from the same social studies textbook. The Geleen school, which included an experimental and a control group, studied Criminal Law, while the two other schools studied Mass Media. All classes spent three 50 minute lessons a week on social studies prior to and during the training. The structure of the lessons prior to the intervention was the same for all groups. All teachers used the aforementioned conventional instruction method as the dominant method of teaching. All groups were pretested on working memory capacity eight weeks after the academic year started. Prior to taking the test the teacher explained the reason, contents and structure of the test. The topics included historical nouns and general images that were not related to the topic at hand to get a baseline measurement. The pretest to measure meta-cognitive strategies and reasoning achievements took place in the next lesson by means of a regular school test. Like all other official school tests, the latter test was compiled by the teachers and based on the content of the textbook. The school tests
were evaluated by the teachers. After evaluation, the heterogeneous cooperative groups in the experimental groups were compiled to train the meta-cognitive working memory function. All groups consisted of one low, one or two middle and one high achiever.

A six week intervention period followed after pretesting in which both groups studied the subjects in the same time span in three lessons per week (total: 150 minutes). Control groups studied the subjects by using the conventional instruction method. The experimental groups studied these subjects in one week lesson cycles by using the brain-based method: an introduction on the subject was given by the teacher. After the cognitive working memory training was introduced during the first five minutes, training took place in the second lesson for twenty-five minutes (Jaeggi et al., 2008). In the remaining twenty minutes the textbook reasoning questions that were discussed in the first lesson were evaluated. In the third lesson, training of meta-cognitive reasoning skills took place by using the IMPROVE-method (Mevarech & Kramarski 2003). After a short introduction by the teacher, reasoning skills were trained in the next forty minutes (ten minutes per reasoning question). Process evaluation took place during the last five minutes in which questions of the students regarding the reasoning process are discussed by the teacher and students by means of Socratic debate.

The first posttests were conducted one week after the training ended. The second posttest was conducted eight weeks after the first posttest.

**Results**

The variances in pretest scores do not differ significantly between both groups as pre-test analysis showed homogeneity for both experimental and control groups as well
as for all subgroups (Table 1). The homogeneity of groups is supported by both tests in which the ‘reasoning achievements pretest’ was an official school test contrary to the ‘cognition pretest’ in which the motivation to achieve on the test is a far less dominant factor.

(Table 1 about here)

Note that we also have compared each subgroup’s results with the opposite general group scores to investigate individual group results. Test scores for the cognition pretest showed that the Geleen groups (one control and one experimental group) scored lowest, followed by the Maastricht control group and the Heerlen experimental group. However, there were no significant differences when subgroups were compared to opposite general group test scores. All groups used the same cognition pretest with the same historical content, which differed from the social studies subjects and was compiled specifically to exclude foreknowledge as a distinctive predictor for achievements.

The average control group scores in the reasoning achievements pretest were higher compared to the experimental group. While different tests between independently operating schools could explain the variance in test scores between groups, it does not explain the variance of the pretest score between the Geleen school’s control and experimental group which have been trained and tested by the same teacher. However, the subgroup differences were not significant compared to general group test scores. While both group test scores showed no significant variances, we analyzed the intermediate and (both of) the posttest scores based on the equality between both groups.
Independent t-tests have been performed on all intermediate and posttests (table 2). The analyses demonstrate that group differences are statistically significant for all intermediate and posttests between the control and experimental groups.

(Table 2 about here)

For more detailed information, we than compared each experimental subgroup (Geleen 1, Heerlen) to the control group general scores to investigate each experimental group’s achievements³. With regard to the intermediate and posttests that measure cognitive levels (table 2A), the Geleen experimental group test scores show significant effects compared to the control group. The effect of the Geleen group remains significant when compared to specifically the Geleen control group in both cognitive intermediate $t(37) = -2.37, p < .05, r = .36$ and posttest $t(37) = -4.51, p < .001, r = .60$ scores. While the Heerlen experimental group’s test scores also demonstrate statistically significant effects in the intermediate test, posttest scores do not significantly differ compared to the control group.

With regard to short term improvements in reasoning achievements (table 2B, Graph 1), achievements in the intermediate and first posttests demonstrate statistically significant effects for the Geleen experimental subgroup, but not for the Heerlen experimental group compared to the general control groups’ results.

Graph 1 about here

³ Specific analyses of control subgroups compared to experimental group test scores are not pertinent to the present discussion since the control group’s general test scores showed no significant variations in intermediate and posttests.
Comparisons between specific experimental and control groups are not shown in the table. The effect of the Geleen experimental group does not remain significant when compared to the Geleen control group in achievement intermediate test $t(37) = -1.77, p = .09, r = .28$. However, significant effects are found in the first posttest $t(37) = -3.02, p < .01, r = .44$ when Geleen experimental and control groups are compared. Non-significant improvements of the Heerlen experimental group could be attributed to the relatively low pre- and intermediate test scores as well as first posttest scores of the Heerlen group compared to general control group scores. Both developments could result in non-significant enhancement of reasoning achievements in the Heerlen group compared to the control group, but could result in a significant improvement when compared to its own pretest results. However, note that Heerlen intermediate $t(17) = -1.54, p = .14, r = 0.35$ and first posttest $t(17) = -2.00, p = 0.06, r = 0.44$ results also have not statistically significantly increased when compared to its own pretest results. During the experiment, the Heerlen group was not taught by the experienced teacher but by an intern teacher who trained the students during the experiment. Pretest results could therefore be attributed to the experienced teacher while intermediate and posttest results could be attributed to the intern. In this regard, experienced teacher vs. intern effects could affect the outcomes of Heerlen group’s test scores because of the teacher’s role of managing and directing reasoning processes (e.g. Kane, Rockoff & Staiger, 2008).

The second reasoning achievements posttest measured for durability of effects after completion of training. The second posttest was conducted 8 weeks after the training ended which resulted in significant effect sizes of the experimental group compared to the control group even when control subgroup results improved slightly compared to the
prior tests. While we can attribute the latter to natural learning processes, durable
significant results of the experimental group justify the analyses of the experimental
subgroups’ test results (Geleen 1 and Heerlen) to retrieve more information. We therefore
compared each experimental subgroup to their own pretest scores as well as the control
group’s general second posttest scores and conclude that both experimental subgroups
showed significant improvements in second posttest scores. The Geleen experimental
group posttest results reveal a significant improvement when compared to general control
group scores, as well as compared to Geleen control group results $t(37) = -4.44$, $p < .001$,
$r = .59$, and compared to its own pretest results $t(20) = -5.92$, $p < .001$, $r = 0.80$. Notably,
in line with results of intermediate and posttests, the Heerlen experimental group did not
achieve significantly better than control group results. The group did, however, improve
significantly when second posttest results are compared to its pretest results $t(17) = -2.48$,
$p < .05$, $r = .52$. We therefore conclude that test results remain durable due to training.

We can also conclude that both Geleen and Heerlen experimental group second posttest
scores improved (but not significantly) when compared to first posttest results. While
improvements in this regard could be attributed to maturation and natural learning
processes, several studies suggest that these improvements can also be caused by the
training and regarded as a long-term effect (Ariës et al., 2014a; Dahlin, 2010; Holmes et
al., 2009). In fact, Holmes and Gathercole (2014) state that there’s ‘preliminary evidence
of accelerated learning following training, with significant improvements in math […]
and improvements in reading comprehension reported post-training […].’ (p. 441), which
suggests that WM continues to take advantage from the training, even longtime after the
training has ended.
Discussion and conclusion

The experiment indicates that the combined domain-relevant training of WM-capacity and reasoning skills improves school curriculum-based test scores due to enhanced WM-skills and internalized reasoning structures.

While the experimental group’s training leads to (significantly) improved test results, the control group shows no significant improvement in both intermediate and posttest scores in WM-capacity and reasoning achievements. These improvements are comparable with conclusions of Ariës et al.’s (2014a) study in which similar groups underwent the combined training in history curricula, and also in earlier research in which either WM-capacity or metacognitive reasoning is trained (e.g. Kramarski & Mevarach 2003; Olesen et al., 2009). Control group reasoning achievement scores on their turn enhanced but not significantly during the intervention period, which can likely be explained by maturation effects since the control group had not trained reasoning skills differently compared to pre-test training.

Durability of training effects is indicated by experimental and control group test score differences which remain significant in the second posttest, eight weeks after the end of the intervention period. These results suggest improvements in WM-capacity and internalization of reasoning structures which, consequently, have durable effects on reasoning test achievements. However, durability of, specifically, WM-capacity improvements could not be explained by the WM tests in the current study for it did not contain a second cognitive posttest to measure for durable effects. Note that training and testing of WM-capacity has not been performed in an adaptive form, which is considered
to be crucial to continually improve WM-capacity and, as far as specifically n-back training is concerned, has proven to lead to (small) positive effects on fluid intelligence (Olesen et. al., 2009; Au, Sheehan, Duncan, Buschkuehl & Jaeggi, 2014). Also, WM-capacity training and testing did not contain standardized training batteries, e.g. Automated Working Memory Battery (Alloway & Gathercole, 2008), but were adopted to contain content-relevant subjects. Consequently, we argue that adoptive, non-standardized, types of the training and test batteries used for the current study are less valid for measuring WM-capacity and transfer effects. This lies in the fact that when course based content is implemented in training and test batteries, then the outcomes of the tests do not rely on WM exclusively, but also on e.g. knowledge on the subject and reading skills. Nonetheless, social studies content based WM-capacity batteries rather than standardized batteries were implemented to prevent both transfer effects of far transfer between training and tests (when standardized tests would have been used) and transfer effects of foreknowledge to the test (when tests have been conducted which would contain content of other school courses). However, retained knowledge of training type could explain higher WM-capacity test scores compared to control group results as similar tests have been used during training. This could provide some discussion of the extend of possible far transfer effects on content-based testing (whether transfer of knowledge or skills or a combination of both is concerned) of WM-capacity as well as the type of tests that could be used in classroom based interventions. In this regard, metacognitive training of reasoning skills could in itself provide significant effects on reasoning achievements since in earlier research gains in reasoning achievements are caused by stand-alone metacognitive training, which also proved to be durable and
therefore internalized. (e.g. Goldberg, 2010; Ariës et al., 2014a). Note that while the Heerlen experimental group’s WM-capacity test scores demonstrate statistically significant effects in the intermediate test, posttest scores do not significantly differ compared to the control group. This may be attributed to the fact that the Heerlen group’s cognitive posttest did not take place immediately after, but 4 weeks after cognitive training ended due to planning difficulties. The students’ cognitive WM-levels likely decreased during the 4 week time span in which they did not train WM capacity and spent much time on extracurricular activities and less on direct learning activities.

The school tests that measured reasoning achievements, were composed by the teachers and could provide for differential effects when groups are compared to each other. These differential effects can be caused by (1) the subjects that differed between groups, (2) the composition of the test in which both knowledge and reasoning questions are included and (3) the level of difficulty of sources on which a large amount of reasoning questions are based. However, pretest results and non-significant test improvements in the control groups could provide evidence for the homogeneity of test compositions between different teachers as well as in longitudinal comparisons.

**Acknowledgement**

This work was supported by the Netherlands Organisation for Scientific Research (NWO) under the Doctoral Grant for Teachers.
References


Appendices

**Appendix A: Example of a dual 2-back task**

3 sec [multinationals] 3 sec [terms of trade] 3 sec [microcredit] 3 sec [IMF]

1 stimulus = 4 seconds

1 sequence = 20 stimuli

Note: When participants only have to remember the previous word, then the n-back is called a 1-back task. The sequence represents a dual 2 back task. The noun ‘microcredit’ is green-colored which, in case of a dual 2-back task, means that the noun ‘multinationals’ and image of Africa should be written down on the answer form.

**Appendix B: Example of a visual Odd One Out task**

3000ms


1 stimulus = 3500ms

1 sequence = 4 stimuli

Note: Image ‘D’ is the Odd One Out. The image depicts a slum. Images ‘A’, ‘B’ and ‘C’ all depict examples of development aid.

**Appendix C: The IMPROVE-questionnaire**

IMPROVE-questionnaire (Mevarach and Kramarski 2003):

1. What is the problem/task about?? (comprehension question)
2. Which strategy/principle is appropriate for solving/addressing the problem/task? (Strategic question)
3. How is this problem/task different from what you have already solved? (Connection question)
4. What were the difficulties you felt during the solution process? (Reflection question)