ZDM (2019) 51:691–701
https://doi.org/10.1007/s11858-018-1006-5

ORIGINAL ARTICLE

Measuring metacognitive skills for mathematics: students' self-reports versus on-line assessment methods

Marcel V. J. Veenman[1] · **Dorit van Cleef**[1]

Accepted: 12 October 2018 / Published online: 22 October 2018

Abstract
Various instruments for assessing metacognitive skills and strategy use exist. Off-line self-reports are questionnaires and interviews administered either before or after task performance, while on-line measures are gathered during task performance through thinking aloud or observation. Multi-method studies in reading have shown that off-line methods suffer from serious validity problems, whereas the validity of on-line methods is adequate. Little is known, however, about the validity of methods for assessing metacognition in mathematics. Five instruments were administered to 30 secondary-school students: two prospective questionnaires (MSLQ and ILS) before a mathematics task, two on-line methods (observation and thinking aloud) concurrent to the mathematics task, and a task-specific retrospective questionnaire after the mathematics task. Mathematics performance was assessed by a posttest and GPA. Results confirm that prospective questionnaires have poor convergent and predictive validity in mathematics. Although the retrospective questionnaire does slightly better than prospective questionnaires, the validity of both on-line methods stands out. It is concluded that on-line instruments should be preferred over off-line instruments for the assessment of metacognitive skillfulness in mathematics.

1 Introduction

This paper focuses on the validity of methods for assessment of metacognitive skill and strategy use in mathematics. Metacognition has been recognized as the most important predictor of learning outcomes, surpassing other cognitive and motivational characteristics of students (Wang et al. 1990). A variety of instruments have been used to assess metacognition in mathematics, such as questionnaires, interviews, observations, thinking-aloud protocols, eye-movement registration, computer-logfile registration, note taking, and stimulated recall (Desoete and Veenman 2006; Gascoine et al. 2017). Too often, it is taken for granted that certain measurement methods are suitable for assessing metacognition (Veenman 2005). Therefore, the most prevalent methods will be scrutinized and discussed here.

In conceptions of metacognition, knowledge of cognition often is distinguished from regulation of cognition (Brown 1987; Schraw and Dennison 1994; Veenman et al. 2006). Metacognitive knowledge pertains to the declarative or descriptive knowledge one has about the interplay between person characteristics, task characteristics and the available strategies in a learning situation (Flavell 1979). This self-knowledge, however, is not necessarily correct. Students may under- or overestimate their competences due to a subjective appraisal of task complexity (Veenman et al. 2006). Moreover, metacognitive knowledge does not automatically lead to appropriate strategic behavior (Veenman 2017). For instance, students may *know* that checking the outcome of a mathematics problem reduces the number of errors being made and yet refrain from performing this activity for various reasons. Student may find mathematics uninteresting or too difficult, they may overestimate their calculation accuracy, or they may lack the necessary knowledge and skills for recalculating the problem. According to Winne (1996), knowledge has no effect on behavior until it is actually being used and put to the test. Although prerequisite to the acquisition of metacognitive skills, metacognitive knowledge often is a poor predictor of learning outcomes (Veenman 2017).

Metacognitive skills refer to the executive function of metacognition (Brown 1987), that is, the procedural knowledge that is required for the actual regulation of and control over one's learning activities (Flavell 1976; Veenman 2017). Task orientation, planning, monitoring, evaluation, recapitulation, and reflection typically are manifestations of

✉ Marcel V. J. Veenman
mvjveenman@snelnet.net

[1] Institute for Metacognition Research, Mesdaglaan 50, 2182SX Hillegom, The Netherlands

metacognitive skills. These skills can be represented as a series of internalized self-instructions, prescribing the student *what* to do, *when, why*, and *how* in the course of task performance (Veenman 2013, 2017). Thus, when solving a mathematics problem, metacognitively proficient students address themselves first with reading the entire task assignment and mathematics problem to extract information given and detect what is asked for. Subsequently, they activate prior knowledge needed for understanding and solving the problem. Before acting, they set goals and design a plan of action. While executing their plan, they monitor their problem-solving activities for the detection and fixing of errors and for noticing progress made towards their goals. Before giving the answer, they recapitulate their findings and evaluate their outcomes. Whenever they get stuck, they return to the task assignment to re-orient on the mathematics problem and make a fresh start. Finally, they reflect on their problem-solving behavior in order to learn for future occasions (Veenman 2006, 2013). This overview of metacognitive skillful behavior is not exhaustive. Meijer et al. (2006) derived a taxonomy of 65 detailed activities for problem solving from students' thinking-aloud protocols. Metacognitively poor students tend to skip most of these activities. They often read a small part of the mathematics problem and immediately start calculating whenever they encounter numbers. Consequently, their problem-solving behavior is characterized by trial-and-error and muddling through without monitoring control for repairing errors (Veenman 2017).

Metacognitive skills directly affect leaning behavior and, as a consequence, learning outcomes. Veenman (2008) estimated that metacognitive skillfulness accounts for about 40% of variance in learning performance for a broad range of tasks, including mathematics performance (Veenman 2006). Moreover, the causal relation of metacognitive skillfulness with learning performance has been corroborated by training studies, showing that metacognitive training results in both improved metacognitive behavior as well as enhanced learning outcomes (Azevedo et al. 2007; Dignath and Büttner 2008; Pressley and Gaskins 2006; Veenman et al. 1994). Similar results were obtained with metacognitive training for mathematical problem solving (Kramarsky and Mevarech 2003; Mevarech and Fridkin 2006; Veenman et al. 2005). In conclusion, instruction should focus on facilitating metacognitive skills in order to improve mathematics performance.

1.1 Assessment of metacognitive skills

Although a steep incremental development in both frequency and quality of metacognitive skills occurs from late childhood to early adulthood (Li et al. 2015; Van der Stel et al. 2010; Van der Stel and Veenman 2014; Veenman et al. 2004), huge individual differences can be observed within each age group. Some students hardly employ metacognition, while others are ahead of their peers (Veenman et al. 2004) and they maintain their relative positions in the course of development (Van der Stel and Veenman 2014). Students who lag behind in metacognitive development are at risk and, eventually, they may suffer from study delay or drop out of school (Veenman 2015). Assessment of metacognitive skills is required to discern metacognitively poor students at an early stage and provide them with proper instruction and training. Metacognitively proficient students also need to be identified, however, in order to exclude them from training that interferes with their spontaneous use of adequate metacognitive skills (Veenman 2013).

1.2 Off-line versus on-line assessment methods

A distinction is made between off-line and on-line methods in the assessment of metacognitive skills (Dent and Koenka 2016; Veenman et al. 2006). Off-line methods mainly refer to questionnaires (e.g., MSLQ, Pintrich and De Groot 1990; MAI; Schraw and Dennison 1994) and interviews (SRLIS, Zimmerman and Martinez-Pons 1990) that are administered to students either prior or retrospective to task performance. Students are addressed with questions about (the frequency of) their strategy use and skill application. On-line methods, on the other hand, pertain to assessments during actual task performance through observations, thinking-aloud protocols, and computer-logfile registrations (Azevedo and Cromley 2004; Veenman 2013; Veenman et al. 2000; Winne 2014). Student behavior is then coded or rated according to a standardized coding system. The crucial difference between off-line and on-line methods is that off-line measures merely rely on self-reports from individual students, whereas on-line measures concern the coding of all student behavior on externally defined criteria.

Off-line methods have their pros and cons. Questionnaires can be easily administered to large groups and, therefore, they are widely used in metacognition research (Dent and Koenka 2016; Dinsmore et al. 2008; Gascoine et al. 2017; Veenman 2005). Interviews, on the other hand, need to be individually administered, which is time-consuming. Off-line self-reports of metacognitive skills may suffer from three validity problems (Veenman 2011, 2017). A first validity problem emanates from the off-line nature of self-reports. While answering questions, students have to consult memory in order to reconstruct their earlier behavior, which reconstruction process might suffer from memory failure and distortions (Ericsson and Simon 1993). When off-line assessments are administered prospectively (i.e., prior to actual performance), memory problems increase because students have to base their answers on earlier experiences in the past. A second validity problem with off-line methods pertains to the prompting effect of questions. Questions may interfere with spontaneous self-reports of metacognitive

activity by students (Veenman 2011). Obviously, questions may elicit socially desirable answers. Questions, however, may also evoke an illusion of familiarity with strategies or skills that are queried and students may be tempted to label their behavior accordingly. Thus, questions may prompt the recall of strategy use or skill application that in fact never occurred. Especially students with poor metacognitive knowledge are likely to be susceptible to prompting effects (Veenman 2017). The last validity problem relates to questions about the relative frequency of certain activities ("How often do/did you…?"). In order to answer these questions, students have to compare themselves with others, such as classmates, teachers, and parents. The individual reference point chosen, however, may vary from one student to the other, or even within a particular student from one question to the other (Veenman et al. 2003). When each individual student consistently chooses the same reference point, measurement reliability may be high. Even so, disparate data may arise from variation in reference points among students (Veenman 2017).

On-line assessments of metacognitive behavior have their own merits and limitations. Contrary to off-line self-reports or introspection, thinking aloud requires the mere verbalization of ongoing thoughts during task performance. Students do not reconstruct or interpret their thought processes. Consequently, thinking aloud does not interfere with thought processes in general (Ericsson and Simon 1993) or with regulatory processes in particular (Bannert and Mengelkamp 2008; Veenman et al. 1993), although task performance slightly slows down due to verbalization. In case students frequently fall silent, however, protocols may be incomplete. This is referred to as the tip-of-the-iceberg phenomenon (Ericsson and Simon 1993). The thinking-aloud method is time-consuming, because assessments are individually based and protocols need to be transcribed and analyzed. Yet, thinking aloud is the only method giving access to students' thoughts and metacognitive deliberations concurrent to task performance.

In on-line observation, observers judge the student's metacognitive behavior, either directly while the student performs the task, or indirectly from video-recordings. Observational methods are often used when the task does not lend itself to verbalization or when young students are not sufficiently verbal proficient (Gascoine et al. 2017). Similarly to thinking aloud, observation is time-consuming. On-line logfile registration demands that students perform a learning task on a computer. All student activities are recorded in a logfile, which data may be automatically analyzed on frequencies of metacognitive activities and meaningful patterns in activity sequences (Veenman 2013; Winne 2014). Logfile registration is hardly intrusive to students and can be easily administered to groups. Both observation and logfile registration, however, only assess the concrete, overt behavior of students without giving access to mental processes underlying that behavior. Therefore, the metacognitive nature of activities in both coding systems needs to be verified and validated against other on-line measures (Veenman 2013; Veenman et al. 2014).

1.3 Validity of metacognitive assessments

Three validity issues are relevant to the assessment of a construct (De Groot 1969; Nunnally and Bernstein 1994; Veenman 2007). The first validity issue concerns the *internal consistency* of a measure. Internal consistency not only refers to standard reliability measures, such as Cronbach's Alpha, but also to inter-rater reliabilities when scores are rated or judged from assessment materials. Research usually reports reliability indices and factorial structures of questionnaires, but often fails to do so for other methods (Gascoine et al. 2017; Veenman 2007). In the same vein, inter-rater reliabilities are usually disclosed in research with thinking aloud or observations, but rarely for ratings of interviews (Gascoine et al. 2017). Internal consistency is relevant to statistical interpretations, especially when significant effects fail to occur, but it does not provide information about *what* is being measured (Veenman 2011).

The second validity issue consists of *construct validity*. A first aspect of construct validity is content validity, that is, the extent to which key-concepts and key-processes from metacognitive theory are represented and operationalized in the assessment instrument (Veenman 2011). Meaningful assessment instruments are designed through the selection of relevant metacognitive activities on rational grounds and knowledge from the literature. Once a construct is operationalized, construct validity can be substantiated by convergent validity (Veenman 2007). The latter means that an assessment method should point in the same direction as other assessment methods for the same construct, leading to high correlations between scores obtained with different methods in a multi-method design (Veenman et al. 2006). Not many studies on metacognitive self-regulation with a multi-method design were conducted in the past (Dinsmore et al. 2008; Veenman et al. 2006). Veenman (2005) distinguished across-method comparisons from within-method comparisons in multi-method designs. Across-method comparisons pertain to contrasts between off-line and on-line methods. In a review study, Veenman (2005) concluded that off-line self-reports of metacognitive self-regulation hardly correspond to on-line metacognitive behavior on a reading task. This divergence between off-line and on-line methods has been corroborated by later multi-method studies for reading (cf. Bannert and Mengelkamp 2008; Cromley and Azevedo 2006; Veenman et al. 2003; Winne and Jamieson-Noel 2002), although Schellings (2011) reported a higher, but non-significant correlation of 0.51 between thinking-aloud

data and scores on a task-specific retrospective questionnaire. So far, results pertained to reading tasks. Desoete (2008) compared prospective and retrospective self-reports with thinking aloud during a mathematics task. On the average, self-report data correlated 0.18 with thinking-aloud scores. Jacobse and Harskamp (2012) reported a correlation of 0.16 between self-reports on the MSLQ and thinking aloud during mathematical problem solving. Similarly, Li et al. (2015) found an overall correlation of 0.18 between self-reported planning and logfile-registration of planning activities during a puzzle task. Summarizing these results for across-method comparisons, correlations ranged from −0.07 to 0.51 (mean $r = 0.22$). On the average, off-line and on-line measures have less than 5% of variance in common. Apparently, students do not actually do what they earlier said to do, nor do they accurately report in retrospect what they have done (Veenman 2013, 2017). Despite the behavioral basis of on-line methods, the evidence of divergence between off-line and on-line methods does not prove which method is preferred over the other.

In within-method comparisons, either off-line or on-line methods are contrasted. Research shows that correlations among different off-line measures range from 0.02 to 0.49 ($r = 0.31$ on the average; cf. Muis et al. 2007; Sperling et al. 2012; Veenman et al. 2003), while correlations among on-line measures vary from 0.41 to 0.92 ($r = 0.76$ on the average; cf. Cromley and Azevedo 2006; Veenman et al. 1993, 2000, 2005, 2014). These within-method correlations confirm that off-line measurements show less mutual convergence, relative to on-line measurements. Only the Veenman et al. (2000, 2005) studies, however, pertain to mathematics.

The third validity issue relates to *external or predictive validity*. An assessment instrument should behave as expected by its theoretical foundation in relation to other variables. Most theories on metacognition postulate that better metacognitive self-regulation leads to better learning outcomes (Brown 1987; Veenman 2017; Wang et al. 1990). Consequently, any assessment instrument of metacognitive skills or strategy use should substantially predict learning outcomes. Correlations with learning performance range from slightly negative to 0.36 for off-line measures ($r = 0.17$ on the average; cf. Cromley and Azevedo 2006; Dent and Koenka 2016; Pintrich and De Groot 1990; Schraw and Dennison 1994; Sperling et al. 2012; Veenman 2005; Winne and Jamieson Noel 2002; and specific to mathematics:; Aydin and Ubuz 2010; Jacobse and Harskamp 2012; Pape and Wang 2003), while stretching from 0.40 to 0.88 for on-line measures ($r = 0.61$ on the average; cf. Bannert and Mengelkamp 2008; Cromley and Azevedo 2006; Dent and Koenka 2016; Veenman 2008, 2006; Veenman et al. 2014, 2005; Winne and Jamieson Noel 2002; and specific to mathematics:; Jacobse and Harskamp 2012; Van der Stel and Veenman 2014; Van der Stel et al. 2010). Apparently, off-line methods fall short of external validity, contrary to on-line methods.

1.4 Aims of the present study

The majority of multi-method studies merely addressed metacognitive self-regulation in reading or text studying. Only two multi-method studies reported across-method data for mathematics (Desoete 2008; Jacobse and Harskamp 2012), while another two studies reported within-method data for mathematics (Veenman et al. 2000, 2005). Despite fitting in with the overall picture emerging from reading studies, results of these mathematics studies do not cover all comparisons between assessment methods. In order to allow for within- and across-method comparisons simultaneously, multi-method studies in mathematics need to contrast multiple off-line and on-line methods.

The present study intends to triangulate data from two frequently used questionnaires (MSLQ and ILS) administered prospectively, two on-line assessments (thinking aloud and observation) concurrent to solving mathematics problems, and a task-specific questionnaire administered retrospectively. Scores on the self-report questionnaires are expected to correlate poorly with on-line assessments. Moreover, off-line measures are anticipated to reveal low within-method correlations, whereas on-line methods should converge. Finally, it is hypothesized that on-line methods have a higher predictive value for mathematics performance than off-line methods.

2 Method

2.1 Participants

Thirty third-grade secondary-school students from an urban town in The Netherlands participated in the study. They were equally distributed over three different tracks in Dutch secondary education (pre-academic, higher general, and vocational). Two-thirds of the participants were female, while one-third was male. Their age ranged from 14 to 15 years. Parental consent was requested and granted.

2.2 Tasks

In individual sessions, participants had to solve two series of five mathematics word problems while thinking aloud. These mathematics problems were adapted from a book frequently used in mathematics education (Vuijk et al. 2003) and they were tested for suitability and time duration in a pilot study with another group of third-grade students beforehand (Van der Stel et al. 2010). Problems were deliberately chosen from the curriculum 1 year ahead in order to elicit a learning

process of solving new mathematics problems, rather than solving problems the participants were already familiar with. For instance, a word problem from the first series was:

> The air pollution in the center of town on a given day is represented by the formula $V = -0.2t^2 + 3.1t + 1.7$, where V is the air pollution in grams per m^3 and t is the time in hours. How much was the air pollution at a quarter past eight in the morning? With what percentage did the air pollution change that day between 7 and 11 a.m.?

The first series of five word problems represented a learning phase. Apart from a sheet with the problems, some blank paper, and a calculator, participants were provided with a help-sheet that could be consulted for inspecting the step-by-step solution of each problem. Participants had to hand in all materials after a time limit of 20 min.

The second series consisted of five parallel problems, that is, with the same deep structure as the first series of problems, but with different surface characteristics. For instance, a parallel problem from the second series was:

> The length of a burning candle is represented by the formula $L = 11.5 - 5\sqrt{t}$, where L is the length of the candle in cm and t is the number of burning-hours. How tall is the candle after burning for 2 h? After how many hours the candle is burned down?

For this second series, no help-sheet was available. Therefore, the second series was a posttest for mathematical problem-solving adequacy. Again, participants had a time limit of 20 min. to complete this series of problems.

2.3 Metacognitive skillfulness

Five different methods were used for assessing metacognitive skillfulness, three off-line methods (two prospective questionnaires and one retrospective questionnaire) and two on-line methods (systematical observation and the analyses of think-aloud protocols).

2.3.1 Prospective off-line assessments

Participants completed two questionnaires, one consisting of items selected from the Motivated Strategies for Learning Questionnaire (MSLQ), and the other one with items from the Inventory Learning Styles (ILS). For practical reasons of time constraints, only scales for strategy use and regulation were selected from both questionnaires.

The MSLQ is a widely used self-report instrument for assessing self-efficacy, intrinsic value, test anxiety, strategy use, and self-regulation (Pintrich and De Groot 1990). For the purpose of this study, only the 13 items of the Cognitive Strategy Use scale (CSU) and the 9 items of the Self-Regulation scale (SR) were administered. An item from the CSU scale is, for instance: "I use what I have learned from old homework assignments and the textbook to do new assignments". An item from the SR scale is: "I work on practice exercises and answer end of chapter questions even when I don't have to". Answers are given on a Likert-scale, ranging from 1 ("not at all true for me") to 7 ("very true of me"). These items were translated from English into Dutch and back into English again, by two proficient translators with a Masters degree in English. The translators then scrutinized both translations to detect any discrepancies. On content level, none were found. On word level, only a few differences occurred because some words in English (e.g., "class") have a broader meaning, while requiring a more specific translation in Dutch. Cronbach's alphas were 0.76 and 0.67 for CSU and SR, respectively.

The ILS is a Dutch self-report instrument for assessing cognitive processing strategies, metacognitive regulation strategies, and learning conceptions (Vermunt 1998). For the purpose of this study, only items regarding metacognitive regulatory activities were used. The scale of self-regulation (SRi) consists of 11 items about controlling one's own the learning process. For instance, an item is: "I thoroughly practice with assignments for applying the methods that are taught in the course". The scale for external regulation (ERi) comprises another 11 items for the student's dependency on instructions by the teacher, textbooks, and assignments. An example is: "While studying, I follow the instruction given in the study materials or given by the teacher". Finally, the scale of Lack of regulation (LRi) is composed of 6 items on difficulties with regulation of the learning process. For instance, an item is: "I always study the subject matter in the same way." Thus, LRi is a *negative* indicator of self-regulation. Answers are given on a Likert-type scale, ranging from 1 ("I never or rarely do this") to 5 ("I-almost-always do this"). Unfortunately, one male participant did not manage to complete the ILS items within the time given. Therefore, analyses of ILS data include 29 participants. Cronbach's alphas were 0.77, 0.59, and 0.53 for SRi, ERi, and LRi scales, respectively.

2.3.2 On-line assessments

While solving the first series of mathematics problems and thinking aloud concurrently, the participants' behavior was observed and scored by the experimenter on the occurrence of metacognitive activities. This Systematical Observation scale (SO) entailed 15 activities that were scored for each problem separately, according to criteria established by Veenman et al. (2000, 2005): (1) entirely reading the problem statement, (2) selection of relevant data, (3) paraphrasing what was asked for (goal setting), (4) making a drawing related to the problem, (5) estimating a possible

outcome, (6) designing an action plan before actually calculating, (7) systematically executing such plan, (8) precision in calculation, (9) avoiding negligent mistakes, (10) orderly note-taking of problem-solving steps, (11) monitoring the ongoing process, (12) checking the answer, (13) drawing a conclusion (recapitulating), (14) evaluating the answer against the problem statement, and (15) relating to earlier problems solved (reflection). These activities are characteristic of metacognitive skillfulness in general (Schraw and Moshman 1995; Veenman 2013), but in particular of metacognitive skillfulness for mathematics (Desoete and Veenman 2006; Kramarski and Mevarech 2003). Activities 1 through 6 represent the participant's orientation on the problem before acting, activities 7 through 10 depict the systematic execution of plans and actions, activities 11 and 12 delineate the evaluation activity during and after problem solving, while activities 13 through 15 refer to reflections after solving the problem. Two points were granted if the activity was clearly present, one point was granted if the activity was initiated but not completed, and zero points were granted if the activity was absent. The experimenter practiced this rating procedure beforehand in order to reach an adequate level of rating fluency. Moreover, scores were checked afterwards by replaying the thinking-aloud tapes. For each participant an average score for each activity was calculated over the five problems of the first series. Next, for each participant a sum score was calculated over the 15 activities (with Cronbach's alpha = 0.75). Previously, research has revealed sufficient inter-rater reliabilities for this SO method (Veenman et al. 2000, 2005).

The thinking-aloud protocols of the first series of mathematics problems were transcribed verbatim. Two independent judges analyzed the protocols on the quality of metacognitive activities with respect to four subscales (Van der Stel and Veenman 2014; Veenman et al. 2000, 2005). Orientation was scored on activating prior knowledge, analyzing the problem (task analysis), setting goals, and estimating outcomes. Planning was scored on generating a plan of actions, systematically acting according to that plan, and time management. Evaluation consisted of detecting and repairing errors, monitoring progress, and checking outcomes. Finally, reflection pertained to drawing conclusions while referring to the problem statement, recapitulating the problem-solving process, and learning from the task for future occasions. Judgments were not merely based on the presence of metacognitive activity, but also accounted for the quality of executed metacognitive activities. For instance, one may thoroughly read the problem statement while selecting relevant problem elements, or one may read it superficially while ignoring the relevance of information given. Similarly, evaluation activities may be constrained to passively noticing that 'something is wrong', or it may expand to actively repairing mistakes or misunderstandings. Moreover, it must be emphasized that protocols were judged on the quality of performing regulatory activities, *not* on the correctness of information these activities produced. For instance, evaluating one's answer would contribute to one's evaluation score, even though the outcome of this evaluation might eventually prove wrong. Scores on each subscale ranged from 0 to 4. Mean scores for each subscale were calculated over the five problems, and a total TA score was computed from the four subscales (with Cronbach's alpha = 0.91). As the two judges substantially converged in their scores ($r = 0.85$, $p < 0.01$), the average of judgments was taken as the final TA score.

2.3.3 Retrospective off-line assessment

Immediately after completing the mathematics tasks a retrospective questionnaire (RQ) was administered to all participants. The content of the 21 items matched the SO activities. Moreover, all items explicitly referred to the mathematics tasks. For instance, an item was: "I planned my activities before starting to calculate the solution of a problem". A reversed item was: "I forgot to check the solution to a problem." Answers were given on a Likert-type scale, ranging from 1 ("I have not done this") to 5 ("I have done this every time"). After converting reversed items, a sum score was computed (with Cronbach's alpha = 0.67).

2.4 Mathematics performance

A first measure of mathematics performance concerned the second series of mathematics problems (Posttest). All five posttest problems were scored on correctness of the answer and correctness of the procedure leading to that answer. Participants received two points for each problem if both the answer and procedure was correct, one point if they used the correct procedure but arrived at a wrong answer due to a small miscalculation, and zero points if both the answer and procedure were wrong. A total Posttest score was computed over the five problems (with Cronbach's alpha = 0.59).

As a second indicator of mathematics performance, the mathematics grades of all participants were collected from the school administration. GPA was based on the mathematics grades for four terms of the school year (with a range of 0–10; Cronbach's alpha = 0.83).

2.5 Procedure

The prospective questionnaires were administered during class prior to sessions with the mathematics tasks. During individual sessions, which took place in a quiet room at school, each participant solved the two series of mathematics word problems while thinking aloud. Beforehand, participants received a thinking-aloud instruction about verbalizing their ongoing thoughts. Whenever a participant

fell silent, the experimenter urged him/her to continue thinking aloud, using a standard instruction ("Please, keep on thinking aloud"). The experimenter refrained from offering help. While participants solved the problems, the experimenter concurrently scored SO activities. After 20 min. all materials were taken away and the Posttest series of mathematics problems was presented without help-sheet. Finally, after another 20 min. the retrospective questionnaire was administered. This questionnaire was presented last in order to prevent potential prompting effects of the questions on posttest performance.

3 Results

3.1 Descriptives

As no gender effects emerged in the data, gender was omitted from the analyses. In order to check for sufficient variance in all measures, descriptives are depicted in Table 1. Apparently, none of the measures suffers from a lack of variance, or from bottom or ceiling effects.

3.2 Convergent validity

Correlations among the different assessment methods were calculated (see Table 2). As was expected, across-method comparisons show that scores on the prospective questionnaires hardly correlate with on-line data. When corrected for the inverse relation with LRi, prospective questionnaires correlate 0.15 on the average with on-line data (2% of shared variance). Similarly, the correlation between RQ and SO is low (less than 3% of shared variance). Off-line RQ and on-line TA, however, are moderately correlated, sharing 16% of variance. This correlation between RQ and TA, however, is not significantly different from correlations of the prospective questionnaires with TA (Fisher-z ratios < 1.43, n.s.; Guilford 1965).

Within-method comparisons show that CSU from the MSLQ is significantly correlated to the SRi and ERi scales from the ILS, but not to the LRi scale. SR from the MSLQ is hardly related to any of the ILS scales. After correction for the inverse relations with LRi, the average correlation between MSLQ and ILS scales is 0.28 (less than 8% of shared variance). Moreover, correlations of the prospective questionnaires with RQ are low, except for the ERi scale of the ILS. On the average, MSLQ scales have 3% of variance in common with RQ, while the ILS scales share 6% of variance with RQ. On-line assessments of SO and TA, on the other hand, are strongly correlated (69% of shared variance). This correlation between SO and TA is significantly higher

Table 1 Descriptives

	Mean	SD	Max
CSU	57.80	9.89	91
SR	37.63	4.93	63
SRi	26.41	6.92	55
ERi	32.76	5.36	55
LRi	16.17	3.65	30
SO	14.44	2.47	30
TA	7.40	3.10	16
RQ	52.27	8.56	105
Posttest	6.83	1.86	10
GPA	6.23	0.94	10

CSU cognitive strategy use (MSLQ), *SR* self-regulation (MSLQ), *SRi* self-regulation (ILS), *ERi* external regulation (ILS), *LRi* lack of regulation (ILS), *SO* systematical observation, *TA* thinking aloud, *RQ* Retrospective Questionnaire, *Posttest* Posttest with mathematics problems, *GPA* grade point average for mathematics, All $N = 30$, except for SRi, ERi, and LRi where $N = 29$

Table 2 Correlations among assessment methods

	CSU	SR	SRi	ERi	LRi	SO	TA
SR	0.59**						
SRi	0.54**	0.06					
ERi	0.32*	0.10	0.24				
LRi	−0.23	0.12	−0.14	0.25			
SO	0.13	0.09	0.06	0.18	−0.26		
TA	0.04	−0.09	0.16	0.23	−0.19	0.83**	
RQ	0.22	0.13	0.25	0.36*	−0.03	0.16	0.40*

CSU cognitive strategy use (MSLQ), *SR* self-regulation (MSLQ), *SRi* self-regulation (ILS), *ERi* external regulation (ILS), *LRi* lack of regulation (ILS), *SO* systematical observation, *TA* thinking aloud, *RQ* Retrospective Questionnaire; All $N = 30$, except for correlations with SRi, ERi, and LRi where $N = 29$; *$p < 0.05$, **$p < 0.01$

than within-method correlations among the three questionnaires (Fischer-z ratios > 2.98, $p < 0.01$).

3.3 External validity

Posttest scores correlate 0.45 ($p < 0.01$) with GPA. Correlations between assessments methods and mathematics performance are depicted in Table 3. Overall, off-line prospective assessments correlate low with mathematics performance on either the Posttest or GPA (on the average accounting for less than 1% of variance). Both on-line SO and TA assessments appear to correlate highly with Posttest mathematics performance (accounting for resp. 27% and 50% of variance), and to a lesser extent with GPA (accounting for resp. 6% and 11% of variance). Finally, off-line retrospective RQ is moderately correlated to Posttest mathematics performance (accounting for 12% of variance), while correlating low with GPA (accounting for less than 2% of variance).

Correlation of TA with posttest mathematics performance and GPA are not different from correlations of SO (Fisher-z ratios < 1.15, n.s.). The correlation between TA and the posttest mathematics scores is significantly higher than correlations of the three off-line questionnaires with posttest scores (Fisher-z ratios > 1.96, $p < 0.05$). The correlation between SO and posttest scores is significantly higher than correlations between MSLQ scales and posttest scores (Fisher-z ratios = 2.12, $p < 0.05$), significantly higher than correlations of both SRi and ERi scales with posttest scores (Fisher-z ratios = 1.99, $p < 0.05$), but not significantly different from the correlation between LRi and posttest scores (Fisher-z ratio = 1.64, n.s.) or the correlation between RQ and posttest scores (Fisher-z ratio = 0.82, n.s.). Despite the significant correlation between TA and GPA, correlations with GPA are too low and compressed for detecting differences (all Fischer-z ratios < 1.05, n.s.).

4 Discussion

Self-reports on prospective questionnaires show poor across-method convergence with on-line thinking-aloud and observational data, obtained from students solving mathematics problems. These results are in line with earlier multi-method studies for reading and mathematics (see above). Likewise, self-reports on the retrospective questionnaire do not converge with observational data. Retrospective self-reports, however, appear to be moderately correlated to thinking-aloud data. The magnitude of this 0.40 correlation is in line with results from some earlier studies (Schellings 2011; Veenman 2005), but not with the slightly negative correlations between retrospective self-reports and thinking-aloud data obtained in other studies (Desoete 2008; Winne and Jamieson Noel 2002). Further discussion of the relation between retrospective self-reports and thinking-aloud measures is resumed below.

Within-method comparisons show that both on-line methods strongly converge in the assessment of metacognitive behavior, even though SO is a frequency measure of metacognitive activities and TA is a qualitative assessment of metacognitive skillfulness. These results are consistent with earlier multi-method studies, including studies using the same on-line methods for assessing metacognition in mathematics (Veenman et al. 2000, 2005). Within-method comparisons among the prospective off-line methods reveal a positive correlation of the CSU scale (MSLQ) with the SRi scale (ILS), but not for the SR scale (MSLQ) with the SRi scale (ILS). These correlations reflect different conceptualizations of self-regulated learning on which the scales of the MSLQ and ILS are based (cf. Gascoine et al. 2017), which different perspectives are confirmed by a close inspection of items in both questionnaires. The CSU scale (MSLQ) represents the use of cognitive strategies for deep processing, such as relating and critical processing information, while the SR scale (MSLQ) mainly refers to self-regulation of attention, effort, and task persistence (Pintrich and De Groot 1990). The SRi scale (ILS), however, pertains to regulation of learning through deep-processing strategies (Veenman et al. 2003; Vermunt 1998). Hence, it comes as no surprise that the SRi scale (ILS) is related to the CSU scale, but not to the SR scale of the MSLQ. Remarkably, both the CSU and SR scales (MSLQ) are not (inversely) related to the Lack-of-Regulation scale (LRi of ILS). Lack of regulation in the ILS denotes a disposition to inflexible strategy use and monitoring deficiency (Veenman et al. 2003; Vermunt 1998), attributes that apparently are not discerned by the MSLQ.

Moreover, within-method comparisons between prospective, general questionnaires and the retrospective, task-specific questionnaire indicate that retrospective self-reports

Table 3 Correlations of assessment methods with mathematics performance

	CSU	SR	SRi	ERi	LRi	SO	TA	RQ
Posttest	−0.09	−0.03	0.03	0.03	−0.13	0.52**	0.71**	0.34*
GPA	0.06	0.06	−0.08	0.17	−0.21	0.24	0.33*	0.13

CSU cognitive strategy use (MSLQ), *SR* self-regulation (MSLQ), *SRi* self-regulation (ILS), *ERi* external regulation (ILS), *LRi* lack of regulation (ILS), *SO* systematical observation, *TA* thinking aloud, *RQ* Retrospective Questionnaire, *Posttest* Posttest with mathematics problems, *GPA* grade point average for mathematics, All $N = 30$, except for correlations with SRi, ERi, and LRi where $N = 29$; *$p < 0.05$, **$p < 0.01$

neither converge with the MSLQ scales, nor with the SRi and LRi scales (ILS). Both the retrospective questionnaire and the CSU scale (MSLQ), however, are moderately correlated to the ERi scale (ILS). In the ILS, external regulation (ERi) is characterized by a dependency on teacher instruction and surface processing through memorizing and rehearsal (Veenman et al. 2003; Vermunt 1998). Thus, external regulation ought to be incompatible with deep processing in the CSU scale (MSLQ) and regulatory activities in the retrospective questionnaire, especially because the ILS scales are designed as orthogonal dimensions (Veenman et al. 2003; Vermunt 1998). In conclusion, the inconsistent pattern of correlations and the low within-method convergence of most self-report scales indicate that these questionnaires do not systematically tap one single construct.

The very low correlations of prospective questionnaires with posttest mathematics performance and GPA signify that self-reports, when administered prior to or entirely separated from actual task performance, have virtually no predictive value for mathematics achievements. Once more, these results corroborate findings from earlier studies (see above). In summary, self-reports on prospective questionnaires lack convergent validity, both across and within methods, and they fall short of external validity for mathematics performance. The inevitable conclusion is that prospective questionnaires do not assess metacognitive skills or strategy use in mathematics. Some researchers argue that questionnaires assess metacognitive knowledge that is prerequisite to applying metacognitive skills (Mevarech and Fridkin 2006; Schraw and Moshman 1995). In particular, they refer to conditional knowledge as a constituent of metacognitive knowledge. Conditional knowledge is declarative knowledge about what to do and when (Schraw and Moshman 1995), but it does not include the procedural knowledge for how to execute a metacognitive skill (Veenman 2017). Within-method divergence among questionnaires in the present study, however, shows that off-line self-reports do not unequivocally assess the construct of self-regulation. The bottom line is that we do not know what off-line self-reports are measuring. Nevertheless, the use of prospective questionnaires is omnipresent in metacognition research (Dinsmore et al. 2008; Gascoine et al. 2017; Veenman et al. 2006) or in metacognition research for mathematics (Dent and Koenka 2016; Dignath and Büttner 2008). Yet, one should have reservations about the utility of off-line methods for assessment of metacognitive skills (Veenman 2017).

At first glance, the retrospective questionnaire seems to do slightly better than prospective questionnaires. Scores on the retrospective questionnaire are moderately correlated to on-line thinking-aloud assessments and to posttest mathematics performance. Schellings (2011) concluded, after obtaining a similar correlation between thinking-aloud and retrospective-questionnaire data, that the task-specific nature of that retrospective questionnaire allowed students to more accurately report on their earlier use of concrete metacognitive activities. In the present study, items of the retrospective questionnaire were modeled after concrete activities in the codebook for on-line observations of mathematical problem solving. Scores on the retrospective questionnaire, however, appear to be hardly related to observational data. Consequently, task specificity of questionnaires is an unsatisfactory explanation for the relatively modest correlation between thinking-aloud and retrospective-questionnaire data. An alternative explanation is that such a modest correlation may be due to the implicit feedback students received when performing the mathematics task. Performing a mathematics task may have affected their mathematics self-esteem and, consequently, either moderated or augmented their retrospective estimation of metacognitive activities employed during the task. The bottom line is that posttest mathematics performances may have confounded the students' retrospective self-reports of metacognition. Indeed, partialing out posttest mathematics performance from the retrospective-questionnaire scores reduced the correlation between thinking-aloud and retrospective-questionnaires data to a semi-partial correlation of 0.17 (n.s.; Nunnally and Bernstein 1994). Thus, while thinking-aloud and observational methods were causally separated in time from posttest mathematics performance, self-reports on the retrospective questionnaire may have been blended with students' subjective experiences of earlier mathematics mastery in the learning phase and on the posttest. Obviously, this explanation needs further investigation before making final judgments about the utility of retrospective self-reports, for instance, by including immediate and delayed judgments of learning (JOLs; Dunlosky and Nelson 1992) in a multi-method design.

Both on-line assessments demonstrate substantial predictive validity for posttest mathematics performance. Correlations of on-line assessments with GPA were attenuated, which is commonly found (Dent and Koenka 2016). These results are in line with earlier studies using the same methodology (Veenman et al. 2000, 2005), although Veenman et al. (2005) obtained a higher correlation of observational data with mathematics GPA ($r = 0.40$, $p < 0.01$). Despite the absence of significant Fisher-z ratios, thinking aloud tends to be the better predictor over observational methods in terms of variance accounted for in mathematics performance. Perhaps, qualitative assessments of metacognitive skillfulness through thinking aloud render more information about the adequacy of metacognitive behavior, relative to frequency rates in the observational method (Veenman et al. 1993, 2000, 2004). Nevertheless, the within-method convergence and substantial external validity of on-line methods warrant the conclusion that on-line methods should be preferred over

off-line methods for the assessment of metacognitive skills in mathematics.

A limitation of the study concerns the representativeness of participants. Due to the labor-intensive method of thinking aloud, only a relatively small number of participants could be included. Earlier thinking-aloud studies in mathematics involved similar numbers of participants (cf. Desoete 2008; Jacobse and Harskamp 2012; Van der Stel and Veenman 2014; Van der Stel et al. 2010; Veenman et al. 2000, 2005). Increasing the number of observations would enhance the power of tests, but it hardly affects the magnitude of correlations. Secondly, the present study was limited to participants in the age of 14–15 years. The overall results of this study, however, were in line with various results obtained in other mathematics studies for participants of 8–9 years (Desoete 2008), 10–11 years (Jacobse and Harskamp 2012), 11–12 years. (Pape and Wang 2003), 12–13 years (Veenman et al. 2000, 2005), and 12–15 years (Van der Stel and Veenman 2014; Van der Stel et al. 2010; Veenman 2006). Although more multi-method studies in mathematics are appreciated, the present results endorse on-line methods for the assessment of metacognitive skills in mathematics.

References

Aydin, U., & Ubuz, B. (2010). Structural model of metacognition and knowledge of geometry. *Learning and Individual Differences, 20*, 436–445.

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*, 523–535.

Azevedo, R., Greene, J. A., & Moos, D. C. (2007). The effect of a human agent's external regulation upon college students' hypermedia learning. *Metacognition and Learning, 2*, 67–87.

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalization method affect learning? *Metacognition and Learning, 3*, 39–58.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation and understanding* (pp. 65–116). Hillsdale: Erlbaum.

Cromley, J. G., & Azevedo, R. (2006). Self-report of reading comprehension strategies: What are we measuring? *Metacognition and Learning, 1*, 229–247.

De Groot, A. D. (1969). *Methodology, foundations of inference and research in the behavioral sciences*. The Hague: Mouton.

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review, 28*, 425–474.

Desoete, A. (2008). Multi-method assessments of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning, 3*, 189–206.

Desoete, A., & Veenman, M. V. J. (2006). Introduction. In A. Desoete & M. V. J. Veenman (Eds.), *Metacognition in mathematics education* (pp. 1–10). Hauppauge: Nova Science Publishers.

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*, 231–264.

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*, 391–409.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*, 374–380.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. Cambridge: MIT Press.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231–235). Hillsdale: Erlbaum.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906–911.

Gascoine, L., Higgins, S., & Wall, K. (2017). The assessment of metacognition in children aged 4–16 years: a systematic review. *Review of Education, 5*, 3–57.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Jacobse, A. E., & Harskamp, E. G. (2012). Towards efficient measurement of metacognition in mathematical problem solving. *Metacognition and Learning, 7*, 133–149.

Kramarski, B., & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. *American Educational Research Journal, 40*, 281–310.

Li, J., Zhang, B., Du, H., Zhu, Z., & Li, Y. M. (2015). Metacognitive planning: Development and validation of an online measure. *Psychological Assessment, 27*, 260–271.

Meijer, J., Veenman, M. V. J., & van Hout-Wolters, B. H. A. M. (2006). Metacognitive activities in text-studying and problem-solving: Development of a taxonomy. *Educational Research and Evaluation, 12*, 209–237.

Mevarech, Z., & Fridkin, S. (2006). The effects of IMPROVE on mathematical knowledge, mathematical reasoning and meta-cognition. *Metacognition and Learning, 1*, 85–97.

Muis, K. R., Winne, P. H., & Jamieson-Noel, D. (2007). Using a multitrait-multimethod analysis to examine conceptual similarities of three self-regulated learning inventories. *British Journal of Educational Psychology, 77*, 177–195.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edn.). New York: McGraw-Hill.

Pape, S. J., & Wang, C. (2003). Middle school children's strategic behavior: Classification and relation to academic achievement and mathematical problem solving. *Instructional Science, 31*, 419–449.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated leaning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33–40.

Pressley, M., & Gaskins, I. (2006). Metacognitive competent reading is constructively responsive reading: How can such reading be developed in students? *Metacognition and Learning, 1*, 99–113.

Schellings, G. (2011). Applying learning strategy questionnaires: Problems and possibilities. *Metacognition and Learning, 6*, 91–109.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460–475.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*, 351–371.

Sperling, R. A., Ramsay, C. M., Richmond, A. S., & Klapp, M. (2012). The measurement and predictive ability of metacognition in middle school learners. *Journal of Educational Research, 105*, 1–7.

Van der Stel, M., & Veenman, M. V. J. (2014). Metacognitive skills and intellectual ability of young adolescents: A longitudinal study from a developmental perspective. *European Journal of Psychology of Education, 29*, 117–137.

Van der Stel, M., Veenman, M. V. J., Deelen, K., & Haenen, J. (2010). Development of metacognitive skills in mathematics. *ZDM International Journal on Mathematics Education, 42*, 219–229.

Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und Metakognition: Implikationen für Forschung und Praxis* (pp. 75–97). Berlin: Waxmann.

Veenman, M. V. J. (2006). The role of intellectual and metacognitive skills in math problem solving. In A. Desoete & M. V. J. Veenman (Eds.), *Metacognition in mathematics education* (pp. 35–50). Hauppauge: Nova Science Publishers.

Veenman, M. V. J. (2007). The assessment and instruction of self-regulation in computer-based environments: A discussion. *Metacognition and Learning, 2*, 177–183.

Veenman, M. V. J. (2008). Giftedness: Predicting the speed of expertise acquisition by intellectual ability and metacognitive skillfulness of novices. In M. F. Shaughnessy, M. V. J. Veenman & C. Kleyn-Kennedy (Eds.), *Meta-cognition: A recent review of research, theory, and perspectives* (pp. 207–220). Hauppage: Nova Science Publishers.

Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instruments: A discussion. *Metacognition and Learning, 6*, 205–211.

Veenman, M. V. J. (2013). Training metacognitive skills in students with availability and production deficiencies. In H. Bembenutty, T. Cleary & A. Kitsantas (Eds.), *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman* (pp. 299–324). Charlotte: Information Age Publishing.

Veenman, M. V. J. (2013). Assessing metacognitive skills in computerized learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 157–168). New York: Springer.

Veenman, M. V. J. (2015). Metacognition: 'Know thyself'. Use that knowledge especially to regulate your own behavior. *De Psycholoog, 50*, 8–18 **(special ed.)**.

Veenman, M. V. J. (2017). Learning to self-monitor and self-regulate. In R. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 233–257). New York: Routledge.

Veenman, M. V. J., Bavelaar, L., De Wolf, L., & Van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized environment. *Learning and Individual Differences, 29*, 123–130.

Veenman, M. V. J., Elshout, J. J., & Busato, V. V. (1994). Metacognitive mediation in learning with computer-based simulations. *Computers in Human Behavior, 10*, 93–106.

Veenman, M. V. J., Elshout, J. J., & Groen, M. G. M. (1993). Thinking aloud: Does it affect regulatory processes in learning. *Tijdschrift voor Onderwijsresearch, 18*, 322–330.

Veenman, M. V. J., Kerseboom, L., & Imthorn, C. (2000). Test anxiety and metacognitive skillfulness: Availability versus production deficiencies. *Anxiety, Stress, and Coping, 13*, 391–412.

Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills at the onset of metacognitive skill development. *Instructional Science, 33*, 193–211.

Veenman, M. V. J., Prins, F. J., & Verheij, J. (2003). Learning styles: Self-reports versus thinking-aloud measures. *British Journal of Educational Psychology, 73*, 357–372.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14.

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14*, 89–109.

Vermunt, J. D. H. M. (1998). The regulation of constructive learning processes. *British Journal of Educational Psychology, 68*, 149–172.

Vuijk, R. A. J., Reichard, L. A., Rozemond, S., Dijkhuis, J. H., Admiraal, C. J., Aalmoes, H., et al. (2003). *Getal en Ruimte [Number and Space]*. Houten: EPN.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1990). What influences learning? A content analysis of review literature. *Journal of Educational Research, 84*, 30–43.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*, 327–353.

Winne, P. H. (2014). Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning, 9*, 229–237.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibrations of self reports about study tactics and achievement. *Contemporary Educational Psychology, 27*, 551–572.

Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology, 82*, 51–59.