Assessing metacognitive activities: the in-depth comparison of a task-specific questionnaire with think-aloud protocols.

Eur J Psychol Educ (2013) 28:963–990
DOI 10.1007/s10212-012-0149-y

Assessing metacognitive activities: the in-depth comparison of a task-specific questionnaire with think-aloud protocols

Gonny L. M. Schellings · Bernadette H. A. M. van Hout-Wolters · Marcel V. J. Veenman · Joost Meijer

Received: 9 January 2012 / Revised: 11 July 2012 / Accepted: 16 July 2012 /
Published online: 10 August 2012

Abstract Teaching and assessing metacognitive activities are important educational objectives, and teachers are calling for efficient instruments. The advantages of questionnaires in measuring metacognitive activities are obvious, but serious validity issues appear. For example, correlations of questionnaire data with think-aloud measures are generally moderate to low. An explanation may be that these questionnaires are not constructed in line with the metacognitive activities measured by the think-aloud method. In the present study, a questionnaire is constructed based directly on a taxonomy for coding think-aloud protocols. Twenty ninth-graders studied a text while thinking aloud, after which they immediately received the questionnaire. The overall correlation between the questionnaire and the think-aloud protocols (r=0.63) was promising. However, scale and item analyses clearly demonstrate some new validity issues. Comparing the questionnaire and the think-aloud results, the students seem to report overt metacognitive activities corresponding more with their behavior reported in the protocols than covert ones. In-depth explorations are presented.

G. L. M. Schellings (✉) · B. H. A. M. van Hout-Wolters
Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, The Netherlands
e-mail: g.l.m.schellings@uva.nl

B. H. A. M. van Hout-Wolters
e-mail: b.h.a.m.vanhout-wolters@uva.nl

M. V. J. Veenman
Institute for Psychological Research, Department of Developmental and Educational Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands
e-mail: Veenman@fsw.leidenuniv.nl

J. Meijer
SCO-Kohnstamm Institution of the Faculty of Social and Behavioural Sciences, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands
e-mail: j.meijer@uva.nl

Keywords Metacognition · Strategy questionnaire · Think-aloud method · Convergence validity

Metacognition is a powerful determinant in learning results (Hattie 2009; Veenman and Alexander 2011). Recognition of the important role of metacognition is paralleled by the construction of various assessment methods and instruments (Veenman et al. 2006). However, different instruments or methods may be aimed at assessing different facets in metacognitive learning, and it is important to evaluate disparate measures in relation to metacognitive theory (Muis et al. 2007). Furthermore, many researchers are involved in an 'online versus offline' measuring debate (cf. Dinsmore et al. 2008; Schellings and Van Hout-Wolters 2011; Veenman and Alexander 2011; Van Hout-Wolters 2000; Winne and Perry 2000). Researchers are discussing the preference of measuring metacognitive activities during the learner's learning (online) or apart from it (offline, which is when the learner is not learning). While the use of online methods is favoured in this discussion, we want to argue that more research is needed to evaluate validity issues surrounding the comparison of online with offline methods. Overall, the convergent validity (e.g., correlations) between questionnaires and online assessment methods, such as the think-aloud method, is regularly reported to be low, ranging from −0.07 to 0.31 (Veenman 2005). These low correlations between questionnaires and think-aloud measures are often explained by the low self-reporting ability of respondents. Self-reports rely on the accuracy and validity of the respondent's knowledge of his or her behavior. In interpreting the commonly low correlations between the think-aloud and questionnaire results, we consider two other explanations. First, questionnaire and think-aloud methods may measure different metacognitive activities. Second, the methods aim at different learning tasks, or they (especially offline methods) are not directed at a specific learning task. For example, offline measures can be aimed at general learning (cf. Richardson 2004). Task-specific measuring connects to ideas and research, from which it appears that learners' learning strategies differ for types of learning tasks or subjects (cf. Bråten and Samuelstuen 2007; Broekkamp and Van Hout-Wolters 2007; Hadwin et al. 2001). Online measuring (measuring learning strategies during task performance) is, by definition, bound to the task performed within the assessment.

We start with an overview of defining metacognitive activities. As theoretical refinement of metacognitive terms is continuously taking place, a great many methods for measuring metacognition have been developed over the years, which has resulted in different categories or scaling of metacognitive activities. These expansions may form the core explanation of noncorrespondence between questionnaire scales and coding schemes.

Defining metacognitive activities

Metacognition was originally referred to as the knowledge about and the regulation of one's cognitive activities in learning processes (Flavell 1979). Through the years both the knowledge and the regulation components have been specified in sophisticated, more detailed metacognitive terms. All metacognitive regulation activities described in the literature stem from Flavell's original three-fold categorisation: planning, monitoring and evaluation (Schraw and Moshman 1995). According to these researchers, planning involves the selection of appropriate strategies and the allocation of resources that affect performance (e.g., making predictions before reading, strategy sequencing and allocating time or attention before beginning a task). Monitoring refers to a learner's online awareness of comprehension

and task performance (e.g., comprehension checking while learning). Evaluation refers to appraising the learning outcome and regulatory processes of one's learning (e.g., re-evaluating one's goals and conclusions).

Pressley and Afflerbach (1995) present an empirically driven list of many metacognitive activities based on an extensive review of think-aloud studies of (expert) readers. All activities actually fit into three main types of activities: constructing the text's meaning, monitoring and evaluating activities. Activities needed in the construction of meaning of a text involve, for example, skimming, identifying important text information and reflecting on text information. The second main category includes perceptions during reading and monitoring, which is immediately followed by fix-it processes. The third category concerns evaluation. Some evaluative reactions are provoked by the readers' predisposition, while others are specific and concern particular parts of the text or the text as a whole.

When we compare the theoretical distinction of Schraw and Moshman (1995) and the empirically driven distinction of Pressley and Afflerbach (1995), we see that monitoring and evaluating activities are mentioned in both, but the distinctions seem to differ as to the first main category. Schraw and Moshman consider the first category to be that of planning activities, whereas Pressley and Afflerbach broaden this main category by including activities of construction, and they label this category 'constructing the meaning of text'. In addition to the activities that really fit the category of planning, for example, skimming the text, Pressley and Afflerbach also mention activities that seem to be more text-oriented in nature, such as identifying important information and reflecting on text information. One explanation for the differences between the theory-driven categories of Schraw and Moshman and the data-driven categories of Pressley and Afflerbach may be that the distinction of metacognition from cognition is quite clear from theory (cf. Nelson 1996), but it is sometimes hard to distinguish the theoretically different levels of cognition in the data. Apparently, this might be the case for both think-aloud data (cf. Meijer et al. 2006) and questionnaire data (cf. Wernke et al. 2011).

In our research, we follow the theoretical framework of Veenman and Beishuizen (2004), who distinguished four main categories of metacognitive skills. The first category is that of *orientation and planning* activities. Orientation activities concern activating prior knowledge and establishing task demands. These activities may actually precede planning activities, which refer to formulating an action plan and subgoaling. The second category of metacognitive activity refers to *execution* activities. These activities correspond with many activities in Pressley and Afflerbach's main category of 'construction of text's meaning'. Execution activities are, for example, executing an action plan, reading only a part of a text and note-taking. Some of these activities appear to be of a more cognitive nature rather than a metacognitive nature (such as reading a part of a sentence or re-reading). Yet, according to Meijer et al. (2006), the metacognitive activities in execution are inferred from the more overt cognitive activities as they reflect a metacognitive decision for action. For instance, reading a particular section of a text is a cognitive activity in itself, but the decision to select only that particular section to read is of a metacognitive nature. The third main category to be distinguished in metacognitive activity includes *monitoring* activities, e.g. such as detecting comprehension failure, error detection and noticing unfamiliar words. The fourth and final category of activities concerns *elaboration* and *evaluation* activities. Activities such as concluding, connecting by reasoning and summarising are construed to be elaborative. Evaluative activities refer to checking whether the text is understood or judging whether reading goals are attained.

Assessing metacognitive activities: questionnaires or thinking-aloud method?

Assessing metacognitive activities by questionnaires

Questionnaires, either predominating measures or specially constructed instruments, are often administered for examining metacognition. While completing questionnaires, the respondents themselves deduce their activities performed during studying. Questionnaires include questions or statements which are focused on the learner's activities during actual learning. The learners are regularly instructed to rate the frequency of their activities in learning situations, or they have to indicate the importance of different learning activities. Learners' rating of the statements is often done on five or seven-point Likert-type scales.

Many questionnaires are aimed at *general* learning strategies. These questionnaires (e.g. Approaches to Studying Inventory, Study Process Questionnaire, Learning and Study Strategies Inventory (LASSI), the Motivated Strategies for Learning Questionnaire (MSLQ)) should reveal how learners usually learn, irrespective of the learning environment, learning task, content and objectives (cf. Winne and Perry 2000).

Many questionnaires are constructed following interview-based research, in which the learners are questioned about conducting learning tasks. In this research, it is assumed that the validity of the self-reports depends on the fact that the mental episodes in performing the task persist as objects of focal attention in short-term memory. Interviews obtained immediately after task completion may be considered to give an accurate reflection of online cognitive processing (cf. Richardson 2004). On the other hand, according to Baddeley (2003), the phonological loop allows the rehearsal of sentences that lasts 2 s. In this way, the validity of self-reports depends on the data in at least episodic memory or more probably in long-term memory.

In questionnaires on *general* learning strategies, the learners are asked to give cumulative and retrospective accounts of how they conduct academic tasks. Hereto, the learners should access long-term memory, and it is quite unlikely that they have retained an accurate record in long-term memory of the mental activities that were involved (Veenman and Alexander 2011). In this case, the learner's perceptions of his or her strategies are measured. More specifically, the learners are expected to abstract one general characterisation of executing learning strategies over multiple occurrences and events of strategic learning within 'general' types of situations (Richardson 2004; Samuelstuen and Bråten 2007).

Because of these validity issues for 'general' assessment measures, questionnaires may be tailored to particular contexts. However, the level of specificity may differ: A questionnaire can be constructed to assess the activities applied in one school domain, or more sophisticated, the instrument can be constructed to assess the specific activities applied in a defined learning task. For example, the respondents may be asked about their learning strategies in a particular course or class (e.g. *When reading for this class, I …;* Blom and Severiens 2008).

Mokhtari and Reichard (2002) have constructed the Metacognitive Awareness of Reading Strategies Inventory (MARSI), which is especially aimed at the context of academic reading. The questionnaire is designed to assess students' awareness and perceived use of reading strategies. Although this theory-based instrument is a general measure, meaning that the items are asking about "what do you *generally* do …" (cf. Veenman 2005), the context of the items is defined. All items ask about: "What do you generally do when *reading academic or school-related materials*". Yet, the items do not refer to a specific reading task or a detailed reading assignment.

A task-specific questionnaire has been constructed by Samuelstuen and Bråten (2007). They designed a questionnaire in conjunction with a concrete reading task (reading an

expository text about socialisation). The task-specific items are indirectly based on items from the general LASSI instrument. But these items are adapted in the sense that they point to the reading task just completed as a frame of reference.

Pros and cons of questionnaires In educational practice, the assessment of metacognitive activities calls for the use of efficient instruments. Using questionnaires is least labor-intensive in both administering and processing the raw data. Often, adequate reliabilities are reported (e.g. Mokhtari and Reichard 2002; Muis et al. 2007). However, the assessment discussion is focusing on the abovementioned validity issues.

Clearly, questionnaires differ in their level of specificity, as they do in goals, content, aimed population, reliability, validity, etc. (cf. Van Hout-Wolters 2000; 2009). Because of the differences in metacognitive questionnaires, it is not strange to find different or low correlations between questionnaires and other measures of metacognitive activities or, more specifically, with think-aloud measures. In the next paragraph, we describe assessing metacognitive activities by means of thinking aloud.

Assessing metacognitive activities by thinking aloud

The think-aloud method is often used to assess metacognitive activities in educational research (see e.g. Azevedo 2005; Bannert and Mengelkamp 2008; Hofer 2004). Participants in think-aloud studies perform a particular task while continuously reporting whatever thoughts pass through working memory (Ericsson 1988; Ericsson and Simon 1984, 1993, 1994; Veenman 2005). These verbalisations are perceived as related to the underlying thinking activities or processes. An interpreter deduces the kind of thinking activities that seem present in the verbalisations according to a well-defined coding system. It is important to prohibit the participants from theorising about their own thinking.

In comparison with retrospective reports, the protocols resulting from online think-aloud methods may be less distorted by interpretations, expectations or memory errors from the participants (Breuker et al. 1986; Van Someren et al. 1993). First, the protocols are considered to be fairly reliable because thinking aloud happens almost simultaneously with the thinking process. The thinking activities are closely followed in very small steps. Second, the protocols are 'pure' because the actual thinking activities are afterwards deduced by an interpreter and not by the participant himself or herself during task performance or during thinking. "Merely verbalizing one's ongoing thoughts differs from non-concurrent self-reports or introspections, as the latter two require a reconstruction of, or a reflection upon one's thought processes" (Veenman 2005).

According to Ericsson and Simon (1984), every person has direct access to the information present in working memory, and the verbalisation about it is accurate and without any interpretations. These authors further conclude that verbalising thoughts does not change the structure or course in which the cognitive activities take place. A study by Veenman et al. (1993) showed that metacognitive processes were not affected by thinking aloud, and Bannert and Mengelkamp (2008) also found that thinking aloud did not affect learning performances. In their meta-analysis based on 94 independent data sets, Fox et al. (2011) demonstrated that thinking aloud is a nonreactive measuring method, that is to say, instructing participants to merely verbalise their thoughts during a task does not alter their task performance. Thinking aloud does not interfere with task performance, but the time to complete the task is increased.

Pros and cons of think-aloud method Data interpreted from think-aloud protocols may give way to new theoretical insights into metacognitive activities and their interplay with learner

characteristics and situational variables (Afflerbach 2000). The think-aloud method may be experienced especially by practitioners as an "eye-opener", since unexpected learner's activities (or problems) may be revealed in the protocols.

However, think-aloud data only provide information about activities or behaviors that are not (yet) automatised and so occupy space in working memory. Besides, the think-aloud method is time-consuming. For example, due to the huge effort of scoring verbal protocols, Hill and Hannafin (1997) analysed the data of only four participants in their explorative study on assessing metacognitive strategies. Think-aloud protocols have to be interpreted by trained raters using well-developed (sophisticated) coding systems. Veenman and Alexander (2011) conclude that the quality of online assessment (i.e. interpreting think-aloud protocols) depends on the adequacy of the coding system. For example, in a study of examining task awareness (Schellings and Broekkamp 2011), two raters jointly categorised all verbalisations. Consulting a third rater without training, the interrater reliability of 0.69 was considered to be fair but modest, which means that this coding system leaves much room for discussion. In all, the actual protocols employed in different think-aloud studies can vary quite dramatically, for example, by using different think-aloud procedures or different coding systems (Afflerbach 2000; Fox 2009).

Although most researchers who analyse think-aloud protocols (cf. Ericsson and Simon 1994; Veenman 2005) assume that little information from the thinking activities is lost, observation during the think-aloud sessions indicates that some information might still remain covert. For example, students remain silent while pulling faces; they seem to skim the text, and sometimes they mumble, which makes recording and transcribing very difficult (Schellings et al. 2007). Verbal reports might not be complete. Furthermore, a reader may not be able to say everything that comes to mind and may edit or omit some thoughts that do come to mind. A researcher can only analyse the protocol's content, and unspoken processes that give rise to these verbalisations must be inferred (Magliano and Graesser 1991; Magliano et al. 2011). The think-aloud method seems pre-eminently suitable to tap *conscious* reflections. But any set of verbalisations still only provides an approximation of what a reader (or readers) actually does (or do).

In conclusion, the think-aloud method is a valuable metacognitive assessment method with strong and weak points. In order to produce the most optimally informative research findings, both the procedure and the coding system should be described in greater depth (cf. Afflerbach 2000; Fox 2009). The think-aloud method seems particularly suited for research purposes rather than for practical aims because of its labor-intensive nature.

Comparing questionnaires with the think-aloud method

Because the number of methods for assessing metacognitive strategies is increasing, researchers stress the need for validity research (Veenman and Alexander 2011; Winne and Perry 2000) and preferably that questionnaires and think-aloud measures should be compared in *one* research design (Veenman 2005).

In a study by Veenman et al. (2003), a general questionnaire, i.e. the Inventory Learning Styles (ILS), was administered to 30 university students prior to studying a text about earth sciences while thinking aloud. The Self-regulation scale from the ILS correlated 0.22 with think-aloud measures of activities corresponding to the ILS scale. Veenman and Van Cleef (2007) administered the MSLQ and ILS to 30 secondary school students prior to mathematical problem-solving while thinking aloud. The Cognitive strategy use and Self-regulation scales from the MSLQ and the Self-regulation scale from the ILS correlated 0.11 on average with measures for metacognitive skillfulness, rated from think-aloud

protocols. Additionally, a retrospective questionnaire was administered immediately after solving the math problems. Scores on this retrospective questionnaire correlated 0.28 with protocol measures, although both instruments addressed the same broad set of metacognitive skills for problem-solving in mathematics.

Cromley and Azevedo (2006) administered three parallel strategy-use measures in one study: a general questionnaire (MARSI; mentioned above), think-aloud protocols and a multiple-choice strategy-use measure. This multiple-choice strategy-use measure was designed to capture *actual* use of strategies. In taking this measure, the participants were asked to read a short expository passage and then to apply a strategy to the passage by choosing the best answer from four options. The multiple-choice strategy-use measure and the think-aloud protocols both significantly correlated with two comprehension measures of reading comprehension and with each other. The MARSI had non-significant correlations with all the other measures. To be more specific, a −0.02 correlation was found between the think-aloud measures and the MARSI, and a 0.41 correlation was found between the think-aloud measures and the multiple-choice strategy-use questionnaire.

Van Hout-Wolters (2009) notes that questionnaires display a variable picture when compared with think-aloud measures (correlations from −0.07 to 0.42). She differentiates between general questionnaires (reaching correlations up to 0.22 with think-aloud measures) and task-specific questionnaires (reaching correlations up to 0.42). Often, in comparing instruments, only correlations of total scores to the measuring instruments are reflected. Correlations between two measuring methods might turn out differently if analyses of specific learning strategies took place, as Verheul and Yang already reported in 1986 (the correlation for 'structuring', for instance, was 0.56 between the questionnaire and the thinking-aloud scores in their first study). Bannert and Mengelkamp (2008) find that observational data do not correspond with the separate questionnaire's scales (the values of the correlations are not reported), with the exception of the elaboration scale, i.e. one scale of the questionnaire (r=0.54).

Based on his review of multi-method research, Veenman (2005) concludes that several questionnaires, irrespective of the many adequate reliabilities, do not adequately represent respondents' real activities. This lack of correspondence could be explained by memory failure in recollecting activities that have been performed (cf. Schwarz 1999). The low correspondence may also be explained by varying reference points, i.e. while answering a questionnaire the respondents are comparing themselves to different persons, such as the teacher or the best/worst student in class. Failure of correspondence between the measures might also be caused by the social desirability of answers in completing questionnaires. Students may report more strategic processing than was actually performed because they believe this will meet the approval of the test administrator, their teachers or other students. Another possibility is that students report more activities because they believe some activities to be effective, not because they actually used those activities to any great extent. Yet another problem with questionnaires may be a tendency to report using activities described by the items even when other activities were actually used. Students' reports of reading-strategy use may be restricted to activities prelisted in a questionnaire. Differences in the instructions how to rate the activities may also lead to noncorrespondence between questionnaires and think-aloud protocols. A frequency scale aims at measuring the estimated occurrence of specific activities, whereas an importance scale is more appropriate for measuring the perceived usefulness of an activity. And usefulness is not measured by the think-aloud method.

As mentioned above, there may still be other reasons why online measures and offline measures do not yield the converging results of previous studies (cf. Schellings 2011; Van Hout-Wolters 2009). If one wants to compare two measuring methods, the multi-method study must be executed meticulously well, meaning that the individual methods should be aimed at the same metacognitive activities and at the same learning situation. However, by maximising the possible alignment between two instruments, there is always an upper limit in the extent that two measures are correlated. That is, a correlation cannot exceed the square root from the product of the reliabilities (see classical test theory). Assuming that the *average* reliability of the questionnaire method[1] is the same as that was found in this study (0.87, see "Results" section and Table 4), and assuming that the reliability of the thinking aloud method is quite low, i.e. 0.60, one can expect a maximum correlation of 0.72 between both measuring methods. On the other hand, if the reliability of the thinking-aloud method is estimated at 0.80, the maximum correlation would amount to 0.83. Furthermore, comparing *online* methods with each other generally renders highly correlating results (0.64–0.89), for example, when thinking aloud is compared with observation of behaviour and the logfile method (Veenman 2005). Therefore, we will set our expectations about the correlations to be reached in comparing think-aloud protocols with the questionnaire to at least 0.70.

In a previous study, we have constructed a metacognitive strategy questionnaire which contains the two methodological prerequisites just mentioned. The questionnaire taps the same metacognitive activities that are included in a hierarchical taxonomy for coding think-aloud protocols (Schellings 2011). Moreover, the questionnaire is administered immediately after completing a learning task, and all items refer to the preceding task performance. As expected, we found a higher correlation (r=0.51) between the questionnaire and think-aloud protocols than is regularly reported (Schellings 2011). We consider this correlation reasonably high, but not high enough to exclude the possibility that both instruments may still measure different constructs. Importantly, this value concerned the *overall* correlation. Analyses including the subscales from both instruments showed a variable picture. The Elaboration and Evaluation scale of the questionnaire reached a significant correlation (r=0.60) with the same scale in the think-aloud method, whereas the Orientation and Planning scale did not (r=0.24). Moreover, questions arose about the students' ability of reporting about the distinct, specific activities within the scales. These specific activities are in fact distinct activities to be instructed in acquiring the superordinate metacognitive skills. Using the original data set (16 students) and including the data set of four new students, the present study is aimed at examining the correspondences and differences more in depth.

The present study This study includes a deepening of our prior study. First, by engaging more students, we examine if a questionnaire will reach an acceptable correspondence with a thinking-aloud measure in case that both methods measure the same activities performed at the same learning task. If so, questionnaires will be more reasonably adequate alternatives for measuring metacognition in educational practice than regularly assumed. Additionally, previous results yielded new validity issues concerning the ability of reporting about specific activities. By performing analyses at item level, we examine whether students report some kinds of learning activities corresponding more with their thinking-aloud behavior.

[1] We can only estimate this theoretical maximum because questionnaire's reliabilities were not presented in the cited studies.

Method

Participants

Participants were four boys and 16 girls (15-year-olds) recruited from five different schools (third grade of their senior secondary general education). In The Netherlands, History is a compulsory subject in the first three grades of secondary education, and there is one national History curriculum. Participation was voluntary; the students were paid for their contribution.

Materials

Text While thinking aloud, the students studied a history text previously used in a think-aloud study (Meijer et al. 2006). The topic of the text concerned the arrival of the first Africans in the United States of America, slavery and the causes and course of the American civil war. The text was divided into six sections and seven subsections. Each (sub) section was indicated by a (sub) heading. In total, the text contained 1,650 words; the mean number of words per sentence was 17.64. No bold words or words in italics were added. Two open-ended questions were inserted in the text, which the participants also had to answer while thinking aloud. These questions were intended to explicitly trigger metacognitive activity, because simply reading the text aloud could have resulted in meager protocols.

The metacognitive questionnaire The questionnaire was straightforwardly constructed parallel to the taxonomy of metacognitive activities that Meijer et al. (2006) constructed for coding think-aloud protocols of students in secondary education in the domains of text studying (history text) and problem solving (physics). In this taxonomy, superordinate categories of metacognitive activity were theoretically driven and specified with the activities mentioned in the Pressley and Afflerbach overview (1995; see above). An elaborate taxonomy resulted and was evaluated in subsequent protocol analyses. The initial taxonomy appeared to be too highly specified (resulting in disappointing interrater reliabilities), so fewer categories were formulated and tested on new protocols in a cyclic fashion. First, the protocols were re-analysed by using a bottom-up method, i.e. the activities of the students were described very concretely, resulting in subordinate categories of a new model. Second, these descriptions were allocated to the original superordinate categories of the taxonomy. Third, several protocols were analysed by three independent raters. The results were compared and discussed until a satisfactory interrater agreement was reached on a new set of protocols, and the final taxonomy appeared (see Meijer et al. 2006). This final taxonomy, which we also used for coding the protocols in our present study, discerned 56 specific metacognitive activities that adhered to four superordinate scales: *orientation and planning* (16 items), *executing* (11 items), *monitoring* (14 items), *elaboration and evaluation* (15 items).

Furthermore, the revised taxonomy formed the basis for the questionnaire (Schellings 2011; and present study). Besides the descriptions and examples in the taxonomy, protocol fragments from the Meijer et al. study (2006) were used in order to formulate questions on the questionnaire that fitted the vocabulary of secondary school students. In order to construct a task-specific questionnaire, all items were formulated in a way that referred to the text that was read before answering the questionnaire (cf. Samuelstuen and Bråten 2007). So the questionnaire was both task-specific and aimed at the same metacognitive activities that could be coded in the think-aloud protocols. Moreover, in our study, there was a direct

relationship at the *item level* between the two different methods: Each individual category (i.e. activity) in the taxonomy was tapped by one item in the questionnaire. Two categories "skipping word(s)" and "summarising by rereading (sub) headings or important words" were represented by two items each. In all, the questionnaire consisted of 58 items.

The items were scored on a three-point frequency scale. The participants rated whether they performed the activity '*almost never*' (=1); '*sometimes*' (=2), or '*often*' (=3). Because of the task-specificity of the questionnaire's items, it might be very hard for the respondents to pinpoint the gradual differences within this particular task on a more elaborate scale, for example, between "often" and "very often". Another point favouring the three-point scale was raised by Veenman et al. (2003), who suggested that the assessment of learning activities through the method of self-reporting may bring about a serious problem inherent to that method. Self-reports consistently reflect the students' conceptions of the activities they have performed. While reporting on those conceptions, students may choose various reference points for comparing their conceptions of their performance (e.g., their own individual standard, the (alleged) viewpoint of their teacher, a standard referring to the (alleged) ideal student or, conversely, a standard referring to poor students). Therefore, a stable response pattern may be found *within* students consistently choosing one reference point while reporting. High reliability coefficients and stable component structures may be the result. However, variation in reference points *between* students may account for the often-found low correspondence to think-aloud process measures, rated by independent judges against an invariable standard *over* students (Veenman 2005). A three-point scale may then be used to reduce the variation in the choice of a reference point amongst students, though it cannot fully abolish it.

Procedure

Participants were required to read the history text and study it carefully, as they would in preparation for a test. There was no time limit for text studying. In the think-aloud instruction, students were told to read the text aloud and to verbalise any thoughts that arose during studying; as soon as a thought popped up into the student's head, he or she should talk about it. The experimenter concluded the explanation by asking students whether they had understood what thinking aloud was all about. When a student simply read the text aloud, or in cases when the student remained mute, the experimenter stimulated him or her to think aloud, although such prompting was done as little as possible (standardised prompts were, for example: "Keep on thinking aloud", "What are you thinking now?").

After it had been studied, the text was removed, and the students were presented with the metacognitive questionnaire. Answering the questionnaire took about 10 min.

Protocol analysis

The 20 think-aloud protocols were audio-taped and transcribed. Every utterance (as well as an omission or a repetition) made by the participants that deviated from the literal text was interpreted by means of the categories in the taxonomy of Meijer et al. (2006). First, two raters (i.e. a research assistant and the first researcher) defined each 'unit of analyses' in the utterances. The size of this unit was determined by the presence of thought units or ideas. Each unit should contain a more or less complete idea; consequently, some utterances were split by the raters, and the single parts would receive a separate code (cf. Coté et al. 1998; Meijer et al. 2006).

In Table 1, some fragments from the think-aloud protocols are given. Three protocols were used in training a research assistant to segment the units and to apply the categories.

Table 1 Examples of the categories in the metacognitive taxonomy and the corresponding typical verbalisations in the think-aloud protocols

Categories	Fragments from think-aloud protocols
Orientation and planning	
Activating prior knowledge	Student 13:
	I already knew that he was murdered, eh ….
Establishing task demands	Student 03:
	Do I have to read the questions?
Formulating action plan	Student 05:
	First I am going to read the text, then I will be noting what I don't understand and … (laughing) … and yeah, what I do understand.
Executing	
Commenting on (explanation in) text	Student 10:
	Okay, I think this sentence is a bit strange.
Re-reading	Student 13:
	Five days later, Lincoln was murdered by a fanatic southerner (just reading text; not coded) … Five days later, Lincoln was murdered by a fanatic southerner (re-reading text)
Monitoring	
Comprehension failure	Student 05:
	I don't really understand what they mean with this.
Noticing unfamiliar words or terms	Student 10:
	Interpretation, I think that is a difficult word.
Elaboration and Evaluation	
Inferring, i.e. drawing conclusions beyond literal text.	Student 07:
	They were important, they were really very important, because, let's say to… in order to contribute to that eh so to speak.
Interpreting, i.e. deducing the meaning of a word from other text parts	Student 05:
	Abolition (reading text).
	Okay that means, eh, they are the men who believe that slavery has to end.
Uncertainty about conclusion	Student 13:
	Eh.. because, I don't know that for sure.

First, one protocol coded by the first author was discussed and taken as an example protocol to illustrate the different categories. Then the assistant scored a second protocol in discussion with the first author. Finally, a third protocol was independently scored and discrepancies in categorisations were discussed afterwards. The inter-rater reliability was established for 13 other protocols; it concerns the coding of the units and also indirectly concerns the segmentation in units. Pair-wise categorisations were summarised for 13 protocols in a cross-table, and Pearson's contingency coefficient was calculated. This coefficient is found by calculating χ^2 by summing up the categorisations of both judges across all participants and leaving out the empty categories[2]. The Pearson's contingency coefficient is recommended to establish association between two nominal variables if the number of rows and

[2] The formula of the Pearson's contingency coefficient is $C = \sqrt{\frac{\chi^2}{N+\chi^2}}$

columns is large. Pearson regarded the coefficient as a nominal approximation of the product–moment correlation for interval variables (Garson 2004; Meijer et al. 2006). The contingency coefficient between the researcher and the assistant was 0.97, so the interrater reliability was highly acceptable. The remaining four protocols were scored by the first author.

Results

Before presenting the results concerning the relation between the questionnaire and the think-aloud data, we present some descriptive results of the separate measuring methods.

Data concerning the think-aloud protocols The 20 (individual) research sessions resulted in a total of 447 min of thinking aloud and 3,436 activities were counted, interpreted and coded. These results clearly demonstrate the time- and labor-intensive methodology of the think-aloud method. The mean time of studying the text aloud was 22.35 min (SD=7.56). The number of activities that were discerned in the 20 protocols varied considerably. In a meager protocol (student 06), only 50 activities were observed whereas, in the richest protocol (student 15), 384 activities were coded. The mean number of activities per protocol was 171.8 (SD=90.95). Four of the 56 categories in the taxonomy were not found in the 20 protocols: "Change of strategy by reversing arguments", "reading ahead", "fine tuning previously given answer" and "identifying cause and effect". All other categories were observed at least once, with a maximum number of 697 occurrences for the subcategory "error detection (plus correction), keeping track" that belonged to the superordinate category of "monitoring". Four categories, "error detection (plus correction), keeping track", "deviating from literal text", "skipping words" and "noticing unfamiliar terms" occurred at least once in *all* protocols. As Table 2 indicates, the distribution of activity frequencies counted in the protocols was highly skewed. Half of the coding categories (n=23) were not present in half of the protocols (range, 0–9), and some activities (n=15) were rarely performed (occurrences $n \leq 5$). Additionally, some students performed an activity more often than their peers. Most executing activities (n=8) were present in more than half of the protocols (range, 12–10). However, the relatively low frequencies of some activities and the activities found in but a few protocols trigger the question of whether the coding system may be too fine-grained by including rather idiosyncratic student behaviors.

In Table 3, the cumulative numbers of activities found per superordinate category are given. The majority of activities (70.5 %) concern executing (39.9 %) and monitoring (30.6 %) activities. Although the "Orientation and Planning" scale included the most subcategories (16), these activities were counted less (9 %). In learning from text, the activities of "Orientating and Planning", "Executing", "Monitoring" and "Elaboration and Evaluation" that we separated on theoretical grounds (see above) were far from equal in number.

Data concerning the questionnaire Because of the think-aloud part of this study, the sample size was rather small, yet the reliabilities (Cronbach's alpha) for the questionnaire were established (see Table 4). The full instrument resulted in a fair reliability (alpha=0. 87), and also, the Elaboration and Evaluation scale reached a high reliability (alpha=0.83). However, the reliabilities for the other three subscales were low (Orientation and Planning alpha=0.45; Executing alpha=0.15; Monitoring alpha=0.36).

Table 2 Descriptives of the coding category: the cumulative number of occurrences of the activity, the number of protocols in which the activity is counted, the maximum and minimum frequency of the activity found in one protocol, mean and standard deviation

Orientation	Cumulative count	*N*	Max	Min[a]	Mean	SD
Establishing task demands	58	16	8		2.90	2.49
Studying question carefully.	27	12	4		1.35	1.50
Activating prior knowledge	21	9	7		1.05	1.70
Identifying important text	16	5	7		0.80	1.88
Hypothesising	5	5	1		0.25	0.44
Planning						
Formulating action plan	60	17	22		3.00	4.89
Resuming	37	10	16		1.85	2.21
Deciding to read again	28	7	11		1.40	2.76
Organising thoughts	26	9	10		1.30	2.25
Reading notes	23	2	22		1.15	4.91
Looking for information	18	9	5		0.90	1.37
Selecting particular text part	8	5	3		0.40	0.82
Keep on reading for clarity	6	6	1		0.30	0.47
Using external source	5	4	2		0.25	0.55
Subgoaling	3	2	2		0.15	0.49
Change of strategy	0	0	–		–	–
Executing						
Deviating from literal text	423	20	64	7	21.15	15.07
Skipping word(s)	259	20	32	1	12.95	9.80
Re-reading	204	17	56		10.20	13.08
Repeating parts of sentences	122	13	28		6.10	9.20
Reacting to experimenter	119	19	11		5.95	3.14
Error in technical reading	94	18	26		4.70	5.70
Note-taking, underlining	49	5	18		2.45	5.46
Conclusion without checking	46	18	7		2.30	1.81
Commenting on explanation	37	12	10		1.85	2.52
Empathising	17	7	8		0.85	1.84
Reading ahead	0	0			–	–
Monitoring						
Error detection+correction	697	20	65	13	34.85	14.64
Noticing unfamiliar terms	191	20	17	2	9.55	4.82
Comprehension failure	48	13	15		2.40	3.56
Claiming understanding	42	14	11		2.10	3.24
Noting lack of knowledge	26	13	7		1.30	1.63
Noticing inconsistency	13	4	5		0.65	1.46
Checking memory capacity	9	6	3		0.45	0.83
Comment on task demands	8	8	1		0.40	0.51
Noticing retrieval failure	5	5	1		0.25	0.44
Information not found	4	3	2		0.20	0.52
Found required information	3	3	1		0.15	0.37
Deliberately pausing	3	3	1		0.15	0.37

Table 2 (continued)

Orientation	Cumulative count	*N*	Max	Min[a]	Mean	SD
Referring to text to comprehend	1	1	1		0.05	0.22
Fine-tuning previous answer	0	0			–	–
Elaboration						
Paraphrasing	269	18	64		13.45	18.76
Inferring	138	19	37		6.90	8.73
Concluding	50	15	11		2.50	3.17
Summarising by re-reading	36	4	17		1.80	4.38
Connecting text parts	33	7	25		1.65	5.54
Summarising	20	1	20		1.00	4.47
Comprehensive summarising	2	2	1		0.10	0.31
Distinguishing cause–occasion	1	1	1		0.05	0.22
Identifying cause and effect	0	0			–	–
Evaluation						
Interpreting	61	13	14		3.05	4.27
Reading goals accomplished	26	13	3		1.30	1.13
Explaining strategy	19	11	4		0.95	1.15
Uncertain about conclusion	10	6	3		0.50	0.89
Finding similarities	7	4	2		0.35	0.75
Checking	3	2	2		0.15	0.49

[a] With the exception of four activities, the minimum number of occurrence in *one* protocol was zero

Relating the activities measured by the questionnaire to the think-aloud protocols First, the overall correlation between the two measuring methods was computed. Because of the skewed distribution of activities within the think-aloud protocols, a nonparametric correlation (Spearman's rho) was used. Spearman test works by first ranking the data and then applying Pearson's equation to those ranks. It should not be used if many scores have the same rank (Field 2005). In our data set, we did not find a large number of tied ranks (see Fig. 1). The two methods showed a correlation of 0.63 that reached significance ($p<0.01$). This overall correlation is higher than the correlations reported in previous research (cf. Veenman 2005).

In this study, we were also able to compare the measuring methods on both the scale and item level because every item on the questionnaire corresponded directly to one particular coding category. For this in depth analyses, we used the more conservative nonparametric correlation (Kendall's tau), because of the many tied ranks (Field 2005). A cautionary note is

Table 3 Number and percentages of counted activities per superordinate category

Superordinate category	Number of subcategories	Number of activities	Percentage of activities
Executing	11	1,370	39.9 %
Monitoring	14	1,050	30.6 %
Elaboration & Evaluation	15	675	19.6 %
Orientation & Planning	16	341	9.9 %
All	56	3,436	100

Table 4 Reliabilities found on the metacognitive questionnaire

Scale	Number of items	Scale mean	SD	Cronbach's α
Questionnaire	58			0.87
Orientation and planning	16	31.74	3.84	0.45
Executing	12	22.35	2.60	0.15
Monitoring	14	26.53	3.01	0.36
Elaboration and evaluation	16	39.20	6.28	0.83

that values of the Spearman's rho and Kendall's tau cannot be compared directly with one another. The absolute value of the rho correlation is about 1.5 times that of the tau correlation (Gilpin 1993). In the analyses on scale level, we only considered the two scales with 'modest' and 'high' reliability, namely the Orientation and Planning and the Elaboration and Evaluation scales, because low reliabilities limit the theoretical maximum of the correlation coefficient. It appeared that the correlation between the questionnaire and thinking aloud on the Elaboration and Evaluation scale reached significance (r=0.50; $p<$ 0.01), whereas the Orientation and Planning scale did not (r=0.10). This low correlation between the Orientation and Planning scales from the two measuring methods could have indicated major method variation (cf. Campbell and Fiske 1959; Muis et al. 2007). To support this claim, a multi-trait-multi-method analysis is needed, but to perform such an analysis, larger samples are needed. Still, the question arose of whether the ninth-graders within this study were more aware of elaboration and evaluation activities than of orientation and planning activities. Maybe this awareness can by explained by the number of occurrences. As was shown in Table 3, the elaboration and evaluation activities were executed more often (19.6 % of all activities) than the orientation and planning activities (9.9 %).

Acknowledging that correlational analyses on 58 items and 20 participants would be a bit of a stretch, we used correlational analyses at item level as an exploration. By examining the descriptives per thinking-aloud category and questionnaire's item together with the conservative Kendall's tau, we felt that these fine-grained analyses shed some light on the relationship between the instruments at a more precise level.

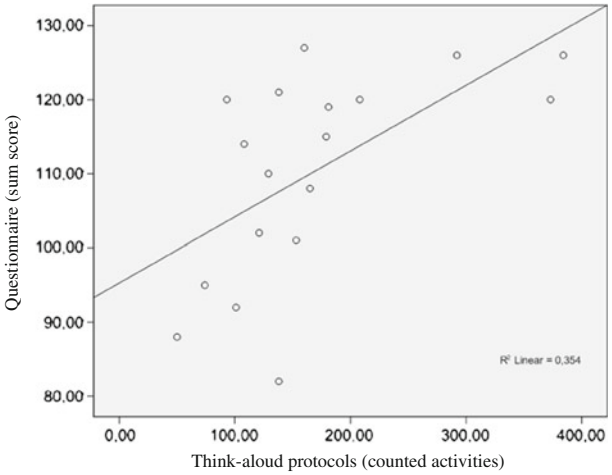

Fig. 1 Correlating think-aloud protocols with the questionnaire: scatter plot

Orientation and planning activities The strongest correlation (0.59) on the Orientation and Planning scale concerned the activity of "reading notes" on the planning section (see Table 5). The activity of "deciding to read again" also reached a high correlation (0.51). Four activities were hardly counted in the protocols ($n \leq 5$). Two negative correlations appeared. At the planning part, "selecting particular text parts" showed both a low occurrence in the think-aloud protocols (8) and a low mean on the questionnaire (M=1.85). On the orientation part, the activity of "activating prior knowledge" with a questionnaire's mean of 2.35 and 21 occurrences ended up with a negative correlation (−0.02).

Executing activities Three rather strong tau-correlations (>0.50) were found (see Table 6): "re-reading", "repeating parts of sentences" and "skipping words". The activities of "re-reading" and "repeating parts of sentences" both occurred relatively often in the think-aloud-protocols (resp. 204 and 122), and both had a rather high questionnaire's mean (M=2.30 each). On the other hand, "skipping words" occurred quite often in the protocols (259), but the questionnaire's means ("skipping "*headings*" M=1.55; skipping "*text parts*" M=1.20) remained moderate. No negative correlations appeared.

Monitoring activities As Table 7 shows, one strong correlation (>0.60) was found for "noticing inconsistency". Both the occurrence in the think-aloud protocols (13) and the questionnaire's mean (M=1.45) remained moderate. Six activities hardly appeared in the protocols. There were two negative correlations: "error detection and correction" (−0.07) and "claiming understanding" (−0.01). The first, "error detection and correction" was the activity most counted in the think-aloud protocols (697 occurrences), with the highest questionnaire's mean on this scale (M=2.85), but its correlation reached a negative value.

Elaboration and evaluation activities At the elaboration part of this scale, one strong correlation (0.63) was found for "paraphrasing" (see Table 8). This activity appeared most often in the protocols (269 occurrences) and showed a relatively high questionnaire's mean (M=2.15). No high correlations (>0.50) were found for the evaluation part of this scale. Two negative correlations were found for the whole scale and concerned "connecting text parts" and "explaining strategy" (resp. −0.04 and −0.06). The correlation was not calculated for four activities because of their low appearances in the protocols.

Summarising the results In all, seven significant and six negative correlations were found (see Table 9). In sorting the categories/items on the basis of the tau-correlations, the items from the four different superordinate scales became scattered across this list of descending values. Keeping in mind that 58 analyses will capitulate on chance, seven categories/items showed a relatively strong correlation (above 0.50) reaching significance at the 0.05 level. These activity items seemed especially to refer to 'elementary' reading activities, such as 'paraphrasing', 're-reading', 'skipping text's headings', 'reading notes', 'repeating parts of sentences' and 'deciding to read difficult parts of text again'. Students may be more aware of executing these particular activities.

Furthermore, six negative correlations were found, although the strengths of these correlations were moderate (−0.01 up to −0.15). Unexpectedly, the activity that was counted most in the think-aloud protocols, "error detection and correction" (697 occurrences), resulted in a negative correlation with the questionnaire's item (−0.07), whereas the questionnaire's mean for this item was still high (M=2.85). Maybe, for this activity, a ceiling effect took place. This explanation is supported by the low SD of .37 of this activity at the questionnaire.

Table 5 Item-correlations on the Orientation and Planning scale

Orientation	TA Count	TA Mean	TA SD	Q[a] Mean	Q SD	Kendall's Tau
TA: identifying important text	16	0.80	1.88	2.55	0.69	0.22
Q: During reading, I thought about whether some text parts were important to remember						
TA: establishing task demands	58	2.90	2.49	2.40	0.68	0.04
Q: I thought about what I was supposed to do.						
TA: studying question carefully	27	1.35	1.50	2.65	0.75	0.02
Q: I read the questions in the text carefully.						
TA: activating prior knowledge	21	1.05	1.70	2.35	0.81	−0.02
Q: I tried to relate the text to my previous knowledge						
TA: hypothesising	5	0.25	0.44	2.50	0.76	[b]
Q: Before reading the text, I thought about the text's topic.						
Planning						
TA: reading notes	23	1.15	4.91	1.47	0.70	0.59
Q: I read my notes about the text.						
TA: deciding to read again	28	1.40	2.76	2.20	0.83	0.51
Q: If I experienced a text part to be difficult, I read that part again.						
TA: organising thoughts	26	1.30	2.25	1.20	0.41	0.34
Q: To organise my thoughts, I questioned myself about the text.						
TA: keep on reading for clarity	6	0.30	0.47	2.50	0.76	0.27
Q: If I did not understand a particular text part, I read the text ahead, because I expected to find the explanation there.						
TA: formulating action plan	60	3.00	4.89	2.25	0.97	0.13
Q: I thought about the best way to read the text.						
TA: resuming	37	1.85	2.21	1.50	0.61	0.12
Q: At some moments, I thought about whether I should continue reading.						
TA: looking for information	18	0.90	1.37	1.85	0.81	0.10
Q: In order to answer the questions, I searched for the correct information in the text.						
TA: selecting particular text part	8	0.40	0.82	1.85	0.81	−0.15
Q: I searched for specific text parts because I knew I could find the answers to the questions there.						
TA: using external source	5	0.25	0.55	1.15	0.37	[b]
Q: I asked the supervisor to explain some text parts.						
TA: subgoaling	3	0.15	0.49	1.65	0.75	[b]
Q: During reading, I kept in mind why I had to read the text.						
TA: change of strategy	0	–	–	1.75	0.72	–
Q: Based on the text's information, I changed my pattern of thinking while reading.						

[a] The questionnaire items were scored on a three-point frequency scale

[b] The correlation is left out, because of the low count in the protocols (n≤5)

Table 6 Item-correlations on the Executing scale

Executing	TA Count	TA Mean	TA SD	Q[a] Mean	Q SD	Kendall's Tau
TA: Re-reading	204	10.20	13.08	2.30	0.73	0.63
Q: I read some parts in the text more than once.						
TA: Repeating parts of sentences	122	6.10	9.20	2.30	0.66	0.55
Q: During reading, I repeated some parts of sentences						
TA: Skipping word(s)	259	12.95	9.80	1.55	0.89	0.51
Q: I skipped the title and headings during reading.						
TA: Error in technical reading	94	4.70	5.70	2.56	0.59	0.26
Q: I noticed that I read some words incorrectly.						
TA: Commenting on explanation	37	1.85	2.52	2.15	0.75	0.23
Q: During reading, I made up my mind about explanations in the text.						
TA: Reacting to experimenter	119	5.95	3.14	2.10	0.79	0.18
Q: During reading, I responded to the supervisor's questions.						
TA: Skipping word(s)[1]	259	12.95	9.80	1.20	0.52	0.14
Q: I skipped some text parts during reading.						
TA: Empathising	17	0.85	1.84	1.95	0.83	0.14
Q: I empathised with the events described in the text.						
TA: Note-taking, underlining	49	2.45	5.46	1.25	0.64	0.12
Q: I underlined the text, marked some words, or I took notes.						
TA: Deviating from literal text	423	21.15	15.07	1.45	0.69	0.09
Q: I noticed that I had read some words that were not present in the text.						
TA: Conclusion without checking	46	2.30	1.81	2.10	0.72	0.02
Q: I drew my conclusions without searching in the text.						
TA: Reading ahead	0	-	-	1.35	0.19	-
Q: I read the text ahead.						

[a] The questionnaire items were scored on a three-point frequency scale

Discussion

Since it is widely accepted that metacognition is important in learning, the need for valid and efficient measuring methods to assess metacognition is strong. Using questionnaires is least labor-intensive, but correlations between questionnaires and think-aloud measures are reported to be moderate to low in multi-method research. However, if a questionnaire is meticulously constructed in line with the instrument to be compared, it reaches a higher correspondence (cf. Schellings 2011). In the present study, a questionnaire was compared with 20 think-aloud protocols, while the questionnaire met two criteria in the construction: The questionnaire was directed at the same metacognitive activities as assessed in the think-aloud method, and the questionnaire was directed at the same task as the think-aloud method. To be more specific, the metacognitive questionnaire was directly constructed parallel to a hierarchical taxonomy for coding think-aloud protocols (Meijer et al. 2006). An overall correlation (r=0.63) was found between the questionnaire and the think-aloud

Table 7 Item-correlations on the Monitoring scale

Monitoring	TA Count	TA Mean	TA SD	Q[a] Mean	Q SD	Kendall's Tau
TA: Noticing inconsistency	13	0.65	1.46	1.45	0.61	0.63
Q: I noticed that some things in the text did not match.						
TA: Comprehension failure	48	2.40	3.56	1.95	0.61	0.45
Q: During reading, I didnot comprehend some parts of the text.						
TA: Comment on task demands	8	0.40	0.51	1.40	0.75	0.24
Q: During reading, I thought about the remaining time.						
TA: Checking memory capacity	9	0.45	0.83	2.10	0.64	0.23
Q: During reading, I thought about whether I could remember the information.						
TA: Noticing unfamiliar terms	191	9.55	4.82	2.35	0.49	0.17
Q: I noticed words in the text that I did not know.						
TA: Noting lack of knowledge	26	1.30	1.63	1.55	0.76	0.16
Q: During reading, I noticed that my knowledge about the topic was insufficient.						
TA: Claiming understanding	42	2.10	3.24	2.45	0.69	-0.01
Q: During reading, I observed that I got a better understanding of text parts						
TA: Error detection+correction	697	34.85	14.64	2.85	0.37	-0.07
Q: If I read something wrong, I corrected myself.						
TA: Noticing retrieval failure	5	0.25	0.44	1.58	0.77	[b]
Q: During reading, I could not remember things.						
TA: Information not found	4	0.20	0.52	1.32	0.58	[b]
Q: In the text, I couldnot find the information I needed.						
TA: Found required information	3	0.15	0.37	2.10	0.72	[b]
Q: After searching the text, I found information I needed						
TA: Deliberately pausing	3	0.15	0.37	1.50	0.69	[b]
Q: I deliberately stopped reading to read previous text parts again.						
TA: Referring to text to comprehend	1	0.05	0.22	2.50	0.69	[b]
Q: By thinking about text parts I had already read, I understood future text parts.						
TA: Fine-tuning previous answer	0	-	-	1.55	0.69	-
Q: I elaborated on my previously given answer with information that I read later in the text.						

[a] The questionnaire items were scored on a three-point frequency scale

[b] The correlation is left out, because of the low count in the protocols ($n \leq 5$)

protocols. This correlation might be seen as moderate, but in comparison with the low regularly reported correlations between questionnaires and think-aloud measures, a 0.63 correlation is promising indeed.

One of the strengths of this study, strictly parallel construction of instruments at the level of items, can be considered as a weak point at the same time. In order to analyse the correspondence between the instruments, the study has implemented a design where every

Table 8 Item-correlations on the Elaboration and Evaluation scale

Elaboration	TA Count	TA Mean	TA SD	Q[a] Mean	Q SD	Kendall's Tau
TA: paraphrasing	269	13.45	18.76	2.15	0.88	0.63
Q: After reading some sentences, I tried to say the contents in my own words.						
TA: summarising	20	1.00	4.47	1.45	0.83	0.41
Q: In order to prepare for the test, I summarised the text first.						
TA: summarising by re-reading	36	1.80	4.38	1.15	0.49	0.28
Q: After reading the text, I read the headings again.						
TA: summarising by re-reading[1]	36	1.80	4.38	1.65	0.75	0.26
Q: After reading the text, I read the important words again.						
TA: concluding	50	2.50	3.17	2.20	0.70	0.14
Q: I drew conclusions during reading the text.						
TA: inferring	138	6.90	8.73	1.90	0.64	0.07
Q: I tried to draw conclusions beyond the literal text.						
TA: connecting text parts	33	1.65	5.54	2.05	0.69	−0.04
Q: I connected text parts.						
TA: comprehensive summarising	2	0.10	0.31	2.25	0.91	[b]
Q: During reading I summarised large parts of the text.						
TA: distinguishing cause–occasion	1	0.05	0.22	1.95	0.69	[b]
Q: During reading, I noticed the difference between cause and occasion.						
TA: identifying cause and effect	0	–	–	1.85	0.67	–
Q: During reading, I tried to extract the cause from effects.						
Evaluation						
TA: finding similarities	7	0.35	0.75	1.85	0.75	0.39
Q: During reading, I thought about things that resemble the things presented in the text.						
TA: interpreting	61	3.05	4.27	2.00	0.65	0.32
Q: I inferred from the text the real meaning of some text parts.						
TA: reading goals accomplished	26	1.30	1.13	2.10	0.91	0.06
Q: After reading, I thought about whether my goals were accomplished.						
TA: checking	3	0.15	0.49	2.35	0.81	0.05
Q: After reading, I checked whether I had understood the text.						
TA: uncertain about conclusion	10	0.50	0.89	1.90	0.64	0.03
Q: I was not sure about my conclusions drawn during reading.						
TA: explaining strategy	19	0.95	1.15	1.40	0.75	−0.06
Q: After reading, I thought about why I read the text in this way.						
TA: checking	3	0.15	0.49	2.35	0.81	[b]
Q: After reading, I checked whether I had understood the text.						

[a] The questionnaire items were scored on a three-point frequency scale

[b] The correlation is left out, because of the low count in the protocols ($n \leq 5$)

Table 9 The significant correlations and the negative correlations (Kendall's tau-value) found in comparing the activities in the protocols with the questionnaire

Category	Scale	TA Count	Q[a] Mean	Kendall's Tau	*p* value
Paraphrasing	Elaboration–evaluation	269	2.15	0.63	0.001
Re-reading	Executing	204	2.30	0.63	0.001
Noticing inconsistency	Monitoring	13	1.45	0.63	0.004
Reading notes	Orientation–planning	23	1.47	0.59	0.008
Repeating parts of sentences	Executing	122	2.30	0.55	0.005
Skipping word(s)	Executing	259	1.55	0.51	0.008
Deciding to read again	Orientation–planning	28	2.20	0.51	0.012
Claiming understanding	Monitoring	42	2.45	−0.01	
Activating prior knowledge	Orientation–planning	21	2.35	−0.02	
Connecting text parts	Elaboration–evaluation	33	2.05	−0.04	
Explaining strategy	Elaboration–evaluation	19	1.40	−0.06	
Error detection+correction	Monitoring	697	2.85	−0.07	
Selecting particular text part	Orientation–planning	8	1.85	−0.15	

[a] The questionnaire items were scored on a three-point frequency scale

participant first learns and thinks aloud and afterwards completes the questionnaire. This means that the questionnaire is always second. However, a balanced design, i.e. a change in the order of the instruments, is not feasible here. In that case, half of the students should complete a questionnaire before studying the text to report their expectations of what they *would* do in this specific task. A change in the order of instruments not only concerns changes in the wording of the questionnaire, but, more importantly, it changes the research question, as well. In comparing the think-aloud method with a task-specific questionnaire, we originally set out to examine the explanation, which is often mentioned for finding low correlations that questionnaires offer less grip on the learning activities the learner performed during the task execution. Our research question really concerned "are students able to verbally report what they have done while studying?" By changing the order of instruments, the research question becomes: "Do students foresee what they are going to do while learning a specific text?" Although both research questions are interesting in comparing the two measuring methods, this study was confined to the first question.

Another problem of this study concerns the small sample size (small number of students $n=20$; selected from one age group) and the use of one text (one study text from the History curriculum) which restricts the generalisation of the results. The method of think-aloud restricts the number of participants to small numbers. For the construction of valid instruments, there are usually needed large samples to estimate reliabilities, discrimination indices, conducting EFAs and CFAs and MTMM-analyses (cf. Muis et al. 2007). Yet our main goal was to compare the task-specific questionnaire with think-aloud protocols instead of constructing an instrument.

In multi-method research, correlational analyses between the different methods mostly concern complete instruments (see Comparing questionnaires with the think-aloud method). As in our previous study (Schellings 2011), we were able to perform analyses at the subscale level because the questionnaire was directly based on the think-aloud taxonomy. Finding different correlations at subscale level may imply that students report some kinds of learning activities that correspond more with their think-aloud behavior. Yet analyses including all

scales were still not carried out because of the reliability analyses on the questionnaire. The full instrument and one scale (Elaboration and Evaluation scale) obtained high Cronbach's alphas; the "Orientation and Planning" scale reached a rather moderate alpha, and the two remaining scales (Executing scale and Monitoring scale) showed low internal consistency, and these scales were not included in the correlational analyses. The Elaboration and Evaluation scale reached a significant correlation (r=0.50), whereas the Orientation and Planning scale did not. Students (ninth-graders) seem more aware of elaboration and evaluation activities than of orientation and planning activities.

A possible explanation might be that students are more aware of elaboration and evaluation activities because they performed these activities (scale frequency=675) more often than the orientation–planning activities (scale frequency=341). This different number of counted activities may reflect the differences in 'grain size' in the questionnaire's items. For example, an orientation item such as "*Before reading the text, I thought about the text's topic*" might be said to be of a larger grain size, relative to an item such as "*After reading some sentences, I tried to say the contents in my own words*" (cf. Samuelstuen and Bråten 2007). However, the frequency of an activity does not seem to play a decisive role in reaching students' awareness level, otherwise, the activity that was most counted (error detection plus correction, which was counted 697 times) should show a stronger correlation with the questionnaire's item *(If I read something wrong, I corrected myself),* and that did not happen.

Yet another explanation might be that it is not the frequency of the activity that is important in reaching the awareness level of ninth-graders, but rather the time spent to execute that particular activity. For example, hypothesising about the text topic (i.e. this text will be about civil war) is relatively quickly executed in comparison to paraphrasing one paragraph (i.e. in this text-part, it says that the black people lived in the South, the Irish lived in the North, the Chinese, etc. …).

One underlying point of concern is the distinction of metacognition into the different subscales, which may be queried by the poor scale reliabilities found on the questionnaire in this study. Questionnaires regularly reach adequate (scale) reliabilities (cf. Muis et al. 2007), however, we found poor reliabilities on two scales of the questionnaire (Executing scale and Monitoring scale). Our questionnaire was *empirically* constructed: Each activity had previously appeared in a think-aloud protocol. The distinction in four scales was *theoretically* based (Meijer et al. 2006). Students may agree with experts' labeling of individual activities while text studying, which may ultimately result in a fair overall correlation between the questionnaire and think-aloud method, but students may disagree with experts about the scales discerned; for instance, students may perceive different scales or relationships between the items (resulting in low reliabilities for two subscales). Noticeably, Wernke et al. (2011) point out that metacognitive processes are highly interwoven, making it hard to separate different 'metacognitive dimensions'. Exploring the dimensional structure of questionnaire items, they came up with a two-factorial solution instead of the four factors they had anticipated (planning, monitoring, regulation and evaluation). Indeed, by examining the factor structure of our questionnaire, we found it differed a lot from the anticipated structure (as was also the case in examining the factor structure found in the think-aloud protocols)[3]. Being aware of the low reliabilities, we found a

[3] Because of the small group of participants (n=20) combined with the relatively large number of items (n=56), we do not report in-depth the results of the performed factor analysis (principal component analysis; Varimax rotation; extraction=four factors). Because most of the distributions of frequencies of metacognitive activity were positively skewed, a square-root transformation was carried out on the think-aloud data (cf. Meijer et al. 2011). Both the factor structure of the think-aloud data (Explained Variance=49.75 %) and the structure of the questionnaire's data (Explained Variance=48.84 %) differed from the anticipated four-factor solution (Orientation and Planning; Executing; Monitoring; and Elaboration and Evaluation).

promising overall correlation, whereas an opposite phenomenon is more often reported, that is to say, many questionnaires reach adequate (overall) reliabilities but remain low in convergence validity (cf. Veenman 2005; Veenman and Alexander 2011).

In comparing the questionnaire to the think-aloud protocols, we further considered all items and did not confine the analyses to scale levels. For even within the scales, students report with more ease, or report some (meta) cognitive activities more validly than others. Moreover, separate items could be included in the analyses because each individual category in the taxonomy was re-phrased into a question. Although analyses on single items are not often carried out, Ainley and Patrick (2006) used single items to measure on-task states in self-regulating learning. At item level, there were some strong correlations among the items of the questionnaire and the categories of the taxonomy. A considerable part of these items seemed to concern rather 'elementary' reading activities, such as 'read again', 'repeating parts of sentences', 're-reading', 'skipping titles during reading' and 'paraphrasing'. These activities seemed to concern more 'overt' activities. Meijer et al. (2006) decided to include these activities in the taxonomy because one may infer 'covert' metacognitive activities from 'overt' activities. Additionally, many of these 'overt' activities could probably be detected by observational eye movement or log-file analyses research. For example, Bråten and Samuelstuen (2007) found a quite close correspondence between learners' task-specific self-reports and traces of the reported activities in the physical counterparts. In sum, metacognitive activities that are easily made observable might be easier for the students (ninth-graders) themselves to report on.

Still, the question remains why some activities seem more valid in self-reports. For example, it seems easier to recall that one has re-read text parts than that one has made some inferences from text, or easier to report that one has skipped the titles in reading (strong correlation) rather than that one has skipped some words in a sentence during reading (not a strong correlation).

One explanation may be found in the metacognitive development of the students of this age group. It is most likely that metacognitive knowledge and skills already develop during preschool or early school years (age 8 to 10 years) at a very basic level but become more sophisticated and academically oriented whenever formal educational requires the explicit utilisation of a metacognitive repertoire (Veenman et al. 2006). In research, it is found that certain metacognitive skills, such as monitoring and evaluation, appear to mature later than others (e.g. planning). However, in the present study, an opposite picture emerges: Students report more validly about elaboration and evaluation activities than about orientation and planning activities. Learning from texts may be mainly aimed at elaboration and evaluation activities because of the type of questions asked on an exam. In addition, education might also be focused more on the "overt" or "elementary" reading skills instead of the "covert" or "higher-order" activities (cf. Schellings and Van Hout-Wolters 2006).

In future research, it would seem worthwhile to register the activities in multiple ways (e.g. thinking aloud, eye-tracking) and to interview the students afterwards about what happened during task execution and how they filled in the questionnaire. In a previous study (Schellings 2011), four students also thought aloud while answering the questionnaire. The descriptive results pointed to a correspondence between the questionnaire and the four think-aloud protocols. However, the students showed some different answering patterns. Some task-specific questions were answered while referring to general learning; few questions were not understood, and some were answered from a social desirability perspective. With some questions the students seemed to feel uncertain in recollecting their behavior. The results found in this case study imply new possibilities in examining validity and reliability issues of questionnaires by having respondents answer a questionnaire while thinking aloud.

Another method is to extend the think-aloud sessions by interviewing the respondents afterwards about how they filled in the questionnaire.

Practical implications In this study, we originally set out to examine the validity issues surrounding questionnaires by using a research design that met two methodological prerequisites: The same learning activities were measured in the same learning situation. Our aim was not to construct a valid questionnaire; our aim was to compare the questionnaire with think-aloud protocols in order to learn whether students are able to report about their activities on questionnaires. In this way, we examine the explanation which is often mentioned for finding low correlations that questionnaires offer less grip on the learning activities the learner performed during the task execution. As a consequence, many researchers argue that questionnaires are doomed to fail assessing metacognitive activities. However, the observed 0.63 overall correlation was fairly high, and the questionnaire may provide a more valid assessment of metacognition than thought. Notably, students reported certain metacognitive activities corresponding more with their behavior reported in the protocols. Future research should be aimed at examining differences in reporting specific activities and constructing instruments directed on the most corresponding activities which can be used validly in practice. We also may eventually support the students in reporting more accurately about their actual behavior and to become more aware and adaptive of their metacognitive activities in their learning.

By focusing on task specificity, one might think that we suggest metacognition can only be measured in a specific context. Different studies do find support for the generality of metacognitive skills, whereas other studies provide evidence against such a general skill (cf. Broekkamp and Van Hout-Wolters 2007; Veenman and Alexander 2011). Both positions may be equally tenable, depending on the grain size of analyses (cf. Veenman and Alexander 2011). Moreover, one might argue that task-specific questionnaires are not authentic in real life: For each individual learning task, a specific questionnaire must be constructed. However, by carefully choosing one central learning task and carefully constructing the questionnaire, we may determine a starting point for assessing or instructing metacognitive activities.

Instead of focusing on the task-specific component of metacogition, McNamara and Magliano (2009) examine the dynamic part of metacognition. Specifically, they argue that metacognitive processes manifest themselves in a dynamic interaction between the reader and the text. Students who demonstrate high metacognitive skills are dynamically altering their processing given changing features of the texts. Questionnaires should be sensitive to this aspect of metacognition and reading. But the decontextualised nature of most self-report questionnaires makes them ill suited to the tasks of measuring metacognitive processes, i.e. questionnaires are not sensitive to the dynamic aspect of metacognitive processes. However, by constructing a task-specific questionnaire, we may have been capturing the dynamic processes and strategies more than other questionnaires do. Obviously, questionnaires that are assessing the dynamic interaction between the reader and the text are, by definition, task-specific (how was this text read by this reader). In future research, we should search ways for unraveling the idea of task-specific measuring into more aspects, such as measuring the dynamics in text processing.

The recent emphasis on the twenty-first-century skills calls for an improvement of the learner's thinking skills. These skills concern information literacy, ICT-skills, creative and critical thinking, problem solving and self-regulated learning. These skills should be taught in line with the twenty-first-century learning technologies. By constructing a questionnaire that measures metacognitive activities in a valid way, the questionnaire may be incorporated

into a digital learning environment for studying text. For example, the learner asks a pedagogical agent questions, such as to show text again, to infer unknown words or to connect different text parts. The answers of this agent should stimulate the learner to execute the proper task-specific activities. In this way, an off-line instrument (the questionnaire) is used in an online environment to support the metacognitive activities of the learner.

To conclude this discussion, we believe that task-specific measuring is not only a valuable trigger in supporting students in carefully reporting their performed metacognitive activities but also a means of obtaining valid information on how we can best teach metacognitive learning activities. After all, metacognitive learning activities can and should be successfully taught (cf. Hattie 2009; Dignath and Büttner 2008).

References

Afflerbach, P. (2000). Verbal reports and protocol analysis. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research (volume 3)* (pp. 163–181). New York: Longman.

Ainley, M., & Patrick, L. (2006). Measuring self-regulated learning processes through tracking patterns of student interaction with achievement activities. *Educational Psychology Review, 18*, 267–286.

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*, 199–209.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*, 829–839.

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.

Blom, S., & Severiens, S. E. (2008). Engagement in self-regulated deep learning of successful immigrant and non-immigrant students in inner city schools. *European Journal of Psychology of Education, 23*, 41–58.

Bråten, I., & Samuelstuen, M. S. (2007). Measuring strategic processing: Comparing task-specific self-reports to traces. *Metacognition and Learning, 2*, 1–20.

Breuker, J. A., Elshout, J. J., Van Someren, M. W., & Wielinga, B. J. (1986). Hardopdenken en protokolanalyse. [Thinking-aloud and protocol-analysis]. *Tijdschrift voor Onderwijsresearch, 11*, 241–254.

Broekkamp, H., & Van Hout-Wolters, B. H. A. M. (2007). Students' adaptation of study strategies when preparing for classroom tests. *Educational Psychology Review, 19*, 401–428.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitraitmultimethod matrix. *Psychological Bulletin, 56*, 81–105.

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes, 25*, 1–53.

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

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.

Ericsson, K. A. (1988). Concurrent verbal reports on text comprehension: A review. *Text, 8*, 295–235.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis. Verbal reports as data*. Cambridge: Institute of Technology Press.

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

Ericsson, K. A., & Simon, H. A. (1994). Verbal reports as data. *Psychological Review, 87*, 215–251.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: SAGE Publications.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*, 906–911.

Fox, E. (2009). The role of reader characteristics in processing and learning from informational text. *Review of Educational Research, 79*, 197–261.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*, 316–344.

Garson, D. (2004). *Nominal Association: Phi, Contingency Coefficient, Tschuprow's T, Cramer's V, Lambda, Uncertainty Coefficient*, from http://www2.chass.ncsu.edu/garson/pa765/assocnominal.htm.

Gilpin, A. R. (1993). Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement, 53*, 87–92.
Hadwin, A. L., Winne, P. H., Stockley, D. B., Nesbit, J. C., & Woszczyna, C. (2001). Context moderates students' self-reports about how they study. *Journal of Educational Psychology, 93*, 477–487.
Hattie, J. A. C. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
Hill, J. R., & Hannafin, M. J. (1997). Cognitive strategies and learning from the World Wide Web. *Educational Technology Research and Development, 45*(4), 37–64. http://www.education.auckland.ac.nz/staff/j.hattie/presentations.cfm.
Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online searching. *Educational Psychologist, 39*, 43–55.
Magliano, J. P., & Graesser, A. C. (1991). A three-pronged method for studying inference generation in literary text. *Poetics, 20*, 193–232.
Magliano, J. P., Millis, K. K., The, R. S. A. T., Team, D., Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). *Metacogntion and Learning, 6*, 131–154.
McNamara, D. S., & Magliano, J. P. (2009). Self-explanation and metacognition: The dynamics of reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 60–81). Mahwah: Erlbaum.
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.
Meijer, J., Veenman, M. V. J., & van Hout-Wolters, B. H. A. M. (2011). Multi-domain, multi-method measures of metacognitive activity: What is all the fuss about metacognition…indeed? *Research Papers in Education*. doi:10.1080/02671522.2010.550011.
Mokhtari, K., & Reichard, C. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology, 94*(2), 249–259.
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.
Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*, 102–116.
Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale: Erlbaum.
Richardson, J. T. E. (2004). Methodological issues in questionnaire-based research on student learning in higher education. *Educational Psychology Review, 16*, 347–358.
Samuelstuen, M. S., & Bråten, I. (2007). Examining the validity of self-reports on scales measuring students' strategic processing. *British Journal of Educational Psychology, 77*, 351–378.
Schellings, G. L. M. (2011). Applying learning strategies questionnaires: Problems and possibilities. *Metacognition and Learning, 6*, 91–109.
Schellings, G. L. M., & Broekkamp, H. (2011). Signaling task awareness in think-aloud protocols from students selecting relevant information from text. *Metacognition and Learning, 6*, 65–82.
Schellings, G. L. M., & Van Hout-Wolters, B. H. A. M. (2006). Leertaken in de mens-en maatschappijvakken in de tweede fase van het voortgezet onderwijs: Nieuwe onderwijsmethodes in de praktijk. [Learning tasks in the human and society courses in the secondary phase in secondary education: Instructional methods in practice. VELON. *Tijdschrift voor Lerarenopleiders, 27*, 23–34.
Schellings, G. L. M., & Van Hout-Wolters, B. H. A. M. (2011). Measuring strategy use with self-report instruments: Theoretical and empirical considerations. *Metacognition and Learning, 6*, 83–90.
Schellings, G.L.M., Van Hout-Wolters, B.H.A.M., Veenman, M.V.J., & Meijer, J. (2007). *Assessing metacognitive activities: Matching a questionnaire with thinking aloud.* Paper presented on the 12th Biennial EARLI-Conference, Budapest, Hungary, August 28th –September 1st, 2007.
Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*, 351–371.
Schwarz, N. (1999). Self-reports. How questions shape the answers. *American Psychologist, 54*, 93–105.
Van Hout-Wolters, B. H. A. M. (2000). Assessing self-directed learning. In P. R. J. Simons, J. van der Linden, & T. Duffy (Eds.), *New learning* (pp. 83–101). Dordrecht: Kluwer.
Van Hout-Wolters, B. H. A. M. (2009). Leerstrategieën meten. Soorten meetmethoden en hun bruikbaarheid in onderwijs en onderzoek. [Measuring Learning strategies. Different kinds of assessment methods and their usefulness in education and research]. *Pedagogische Studiën, 86*, 110–103.
Van Someren, M. W., Bernard, Y., & Sandberg, J. (1993). *The think-aloud method*. Amsterdam: University of Amsterdam.

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. 77–99). Münster: Waxmann.

Veenman, M. V. J., & Alexander, P. (2011). Learning to self-monitor and self-regulate. In R. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 197–218). New York: Routledge.

Veenman, M. V. J., & Beishuizen, J. J. (2004). Intellectual and metacognitive skills of novices while studying texts under conditions of text difficulty and time constraint. *Learning and Instruction, 14*, 619–638.

Veenman, M. V. J., & Van Cleef, D. (2007). Validity of assessing metacognitive skills for mathematic problem solving. In A. Efklides & M. H. Kosmidis (Eds.), *9th European Conference on Psychological Assessment. Program and abstracts* (pp. 87–88). Thessaloniki: Aristotle University of Thessaloniki.

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., 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.

Verheul, I., & Yang, J. (1986). Het traceren van leerprocessen via de hardop-leermethode en een procesvragenlijst. [Tracing learning processes through the learn-aloud method and a process questionnaire.]. In W. J. Van der Linden & J. M. Wijnstra (Eds.), *Ontwikkelingen in de methodologie van onderwijsonderzoek*. Lisse: Swets & Zeitlinger.

Wernke, S., Wagener, U., Anschuetz, A., & Moschner, B. (2011). Assessing cognitive and metacognitive learning strategies in school children: Construct validity and arising questions. *The International Journal of Research and Review, 6*, 19–37.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In P. R. Pintrich, M. Boekaerts, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–567). Orlando: Academic Press.

Gonny L.M. Schellings. Research Institute of Child Development and Education, University of Amsterdam Spinozastraat 55, 1018 HJ Amsterdam, The Netherlands. Phone: +20-525 1599, E-mail: g.l.m.schellings@uva.nl

Current themes of research:

Metacognition. Text studying. Learning activities.

Most relevant publications in the field of Psychology of Education:

Schellings, G.L.M. (2011). Applying learning strategies questionnaires: Problems and possibilities. *Metacognition and Learning, 6,* 91-109.

Schellings, G.L.M. & Broekkamp, H. (2011). Signaling task awareness in think-aloud protocols from students selecting relevant information from text. *Metacognition and Learning, 6,* 65-82.

Schellings, G.L.M. & Van Hout-Wolters, B.H.A.M. (2011). Measuring strategy use with self-report instruments: theoretical and empirical considerations. *Metacognition and Learning, 6*, 83-90.

Bernadette H.A.M. van Hout-Wolters. Research Institute of Child Development and Education, University of Amsterdam, Spinozastraat 55, 1018 HJ Amsterdam, The Netherlands. E-mail: b.h.a.m.vanhout-wolters@uva.nl

Current themes of research:

Metacognition. Text studying. Learning activities.

Most relevant publications in the field of Psychology of Education:

Van Hout-Wolters, B.H.A.M. (2000). Assessing self-directed learning. In P.R.J. Simons, J. van der Linden & T. Duffy (Eds.), *New Learning.* (pp. 83-101). Dordrecht: Kluwer.

Van Hout-Wolters, B.H.A.M. (2009). Leerstrategieën meten. Soorten meetmethoden en hun bruikbaarheid in onderwijs en onderzoek. [Measuring Learning strategies. Different kinds of assessment methods and their usefulness in education and research]. *Pedagogische Studiën, 86*, 110-103.

Marcel V.J. Veenman. Institute for Psychological Research, Department of Developmental and Educational Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands. E-mail: Veenman@fsw.leidenuniv.nl

Current themes of research:

Metacognition.

Most relevant publications in the field of Psychology of Education:

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. 77-99). Münster: Waxmann.

Veenman, M.V.J. (2011). Learning to self-monitor and self-regulate. In R. Mayer & P.Alexander (Eds), *Handbook of research on learning and instruction* (pp.197-218). New York: Routledge.

Joost Meijer. SCO-Kohnstamm Institution of the Faculty of Social and Behavioural Sciences, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam. E-mail: j.meijer@uva.nl

Current themes of research:

Metacognition.

Most relevant publications in the field of Psychology of Education:

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.

Meijer, J., Veenman, M.V.J., & van Hout-Wolters, B.H.A.M. (2011). Multi-domain, multi-method measures of metacognitive activity: What is all the fuss about metacognition…indeed? *Research Papers in Education,* DOI: 10.1080/02671522.2010.550011 1WR.