Applying learning strategy questionnaires: problems and possibilities

Schellings, G.

[Link to publication](#)

Applying learning strategy questionnaires: problems and possibilities

Gonny Schellings

Abstract This article discusses measuring learning strategies by means of questionnaires. In 'multi-method' research, in which think-aloud measures are compared with questionnaires, low or moderate correlations are found. A conclusion often drawn is that learners are not able to verbally report on their learning activities. Alternative explanations concern two other possibilities: first, that different *learning strategies* may be measured by the two methods; second, that the measuring methods may be aimed at different *learning tasks*. Keeping these prerequisites in mind, we constructed a task-specific questionnaire directly based on a taxonomy for coding think-aloud protocols in text studying. We found a higher correlation (r=.51) between the questionnaire and think-aloud protocols than is regularly reported. A case-study, in which four students answered the questionnaire while thinking aloud, led to new insights into why a questionnaire may lead to somewhat different ratings of activities than the think-aloud method. Based on these results, task-specific questionnaires may be improved. Our studies involved a fair comparison between a questionnaire and think-aloud protocols. We cautiously conclude that if task-specific questionnaires are meticulously constructed and examined in new ways, they might become reasonably adequate alternatives for the labor-intensive think-aloud method in measuring learners' learning strategies.

Keywords Learning strategy questionnaires · Think-aloud method

Introduction

The ongoing focus on metacognition and learning strategies in general is resulting in various measuring methods and instruments (Veenman et al. 2006). Both in educational practice and in research different kinds of measuring methods are used to record learning strategies. Obviously, different instruments may lead to differences in data and conclusions. This article will principally address the use of questionnaires in measuring learning

G. Schellings (✉)
Research Institute of Child Development and Education, University of Amsterdam,
Spinozastraat 55, 1018 HJ Amsterdam, The Netherlands
e-mail: g.l.m.schellings@uva.nl

strategies. First, we will present some distinctions in learning strategies and in measuring methods. We consider the possibilities in task-specific measuring, then we concentrate on to the discussion about validity issues concerning questionnaires by presenting multi-method studies in which questionnaires are compared with other instruments. Finally, we discuss our empirical studies in constructing a task-specific learning strategy questionnaire.

Different learning strategies

Definitions of learning strategies vary. Different terms (e.g., methods, techniques, or skills) are used to describe the activities that are considered to constitute a learning strategy. Learning strategies may be defined as certain combinations of goal-oriented learning activities used by learners to improve learning (cf. Van Hout-Wolters 1992; Rachal et al. 2007). A learner can use all kinds of strategies to improve learning. These can be found in a multitude of divisions and overviews of learning strategies (see e.g., Alexander 2006; Pintrich 2004; Weinstein and Hume 1998; Winne and Hadwin 1998; Zimmerman 1994). A distinction is often made between cognitive, metacognitive and affective learning strategies. These strategies respectively include the executive part in acquiring knowledge and skills, the regulative part that directs the executive learning activities, and the affective strategies that are aimed at self-management. Besides divisions into the nature of learning strategies, there are divisions based on the moment of using learning strategies during task execution. The learner uses strategies before, during and after a learning task. Following these different distinctions, numerous learning strategies can be distinguished. Van Hout-Wolters et al. (2000), for instance, described forty-five different cognitive, metacognitive and affective learning strategies which could be used before, during and after a learning task. All these learning strategies differ from one another (e.g., finding missing prior knowledge, directing attention to the task, selecting information, persisting, monitoring learning processes, reflecting, attributing of outcomes). Consequently, measuring methods may be aimed at measuring *different* learning strategies at *different* phases in executing a learning task.

Distinctions in measuring methods: off-line versus on-line and task-specificity

Learning strategy questionnaires, oral interviews, stimulated recall methods, and portfolios belong to the off-line methods. The think-aloud method, eye movement measurement, computer logfile method, observation of behavior, trace analysis, and performance assessment (Van Hout-Wolters 2000) belong to the on-line methods. These methods differ in many ways. One clear distinction pertains to off-line versus on-line methods. This is a practical distinction indicating whether the measurement takes place during the learner's learning (on-line) or away from it (off-line, that is, when the learner is not learning). In other words, on-line assessments are obtained during task performance, whereas off-line methods are presented either before or after task performance (cf. Veenman 2005; Veenman et al. 2006). A great many measuring methods have been developed over the years following the many different learning strategies.

Both educational practice and educational research make much use of off-line methods: methods that measure learning strategies apart from concrete learning tasks. The main reason for this is that on-line methods are more difficult to use in groups and the gathering and processing of these data is more labor-intensive.

Another distinction in measuring methods concerns the context of measuring that is aimed by the methods. A measuring method can be especially attuned to a certain learning

task (Samuelstuen and Bråten 2007). On the other hand, instruments can be aimed at assessing learning strategies in general (i.e., without explicit reference to a specific task). These general or in Veenman's terms 'prospective' instruments are meant to record what a learner usually does when studying (cf. Veenman 2005).

More specifically, these instruments assess learners' predispositions to conducting learning in particular ways (Richardson 2004) and learning is said to be measured as an aptitude or trait (cf. Muis et al. 2007; Winne and Perry 2000).

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). The distinction in measuring within or outside the context (i.e., task-specific or general measures) also indirectly relates to the distinction in on-line versus off-line measuring. On-line measuring (measuring learning strategies during task performance) is, by definition, bound to the task performed within the assessment. Off-line measures can be aimed at general learning or learning from one specific task.

Measuring learning strategies with questionnaires

Self-report questionnaires are the most frequently used instruments for assessing learning strategies (cf. Winne and Perry 2000). Richardson (2004) points out that a great number of questionnaires (e.g., Approaches to Studying Inventory, *ASI* or Study Process Questionnaire, *SPQ*), are mainly aimed at general learning strategies. In his view, the assumption that learners are able to provide valid reports of their predispositions can be questioned. He illustrates that many questionnaires were constructed following interview-based research, in which the learners were questioned about conducting particular tasks, for example, reading an academic text. In interview-based self reports, the validity of the learners' reports depends on the fact that the mental episodes in performing the task persist as objects of focal attention in short-term memory. According to Richardson (2004), interviews obtained immediately after task completion may be considered to give an accurate reflection of on-line cognitive processing. In questionnaires on general learning strategies, the learners are asked to give cumulative and retrospective accounts of how they perform academic tasks. Hereto, the learners should access long-term memory and it is unlikely that they have retained an accurate record in long-term memory of the mental activities that were involved (Richardson 2004). In this case, the learner's perceptions about his or her strategies are measured. More specifically, the learners should abstract one general characterization of executing learning strategies over multiple occurrences and events of strategic learning within 'general' types of situations (Samuelstuen and Bråten 2007).

The number of questionnaires is still rising. New theoretical insights or educational innovations, and psychometric results obtained have resulted in adjusting or revising questionnaires (e.g., Approaches to Learning and Studying Inventory, *ALSI*; Entwistle and McCune 2004). Questionnaires (e.g., Motivated Strategies for Learning Questionnaire, *MSLQ*; Pintrich and De Groot 1990) are translated and attuned to the educational situation in different countries (Blom and Severiens 2008). Further, specific metacognition questionnaires are regularly being used that are not aimed at the wide spectrum of learning strategies, but only at measuring metacognitive aspects thereof (e.g., Metacognitive Awareness Inventory, *MAI,* Schraw and Dennison 1994; Self-efficacy and Metacognition Learning Inventory-Science, *SEMLI-S*, Thomas et al. 2008).

Appendix 1 presents a list of learning strategy questionnaires regularly used in both educational research and practice.

Because of the above validity issues for 'general' questionnaires, these instruments may be tailored to particular contexts (cf. Richardson 2004; Samuelstuen and Bråten 2007). However, the level of specificity may differ: a questionnaire can be constructed to assess the activities which are applied in one school domain, or even 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). One problem is that it is often not known (nor examined) whether the respondents indeed followed the instructions to answer a questionnaire with a particular domain or course in mind.

To direct the respondents in answering a questionnaire more specifically to a certain context, Mokhtari and Reichard (2002) constructed the Metacognitive Awareness of Reading Strategies Inventory (*MARSI*), which is especially aimed at the context of academic reading. The questionnaire was 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 "what do you *generally* do ..." (cf. Veenman 2005), the context of the items is precisely defined. On this instrument, the items really ask: "what do you generally do when *reading academic or school-related materials*" although the items do not refer to a specific reading task or a detailed reading assignment.

A real task-specific questionnaire was constructed by Samuelstuen and Bråten (2007). They constructed a questionnaire in conjunction with a concrete reading task (reading an expository text about socialization). The task-specific items were indirectly based on items from the general LASSI-instrument. The items were adapted in the sense that they referred to the just finished reading task as a frame of reference.

Learning strategy questionnaires may differ in goal, content, level of specificity, target group, type of questions/statements, number of scales and items, type of answer scales, time in which to complete the questionnaire, reliability and validity. However, what they have in common is that, by means of written verbal reports, they try to gain insight into the learners' cognitive, metacognitive and/or affective learning strategies.

Comparing learning strategy questionnaires with on-line measures

During the years, learning strategy questionnaires have been compared with other instruments or methods. For example, Muis et al. (2007) and Sperling et al. (2004) relate (general) questionnaires to other (general) questionnaires. The intercorrelations between the different subscales (either learning strategy subscales or motivational scales) vary from -.07 to .60. General questionnaires have also been compared with task-specific questionnaires. Veenman (2005) reports a correlation of .49 between the full-scale general questionnaire and a task-specific, retrospective questionnaire. For three different reading tasks, Samuelstuen and Bråten (2007) correlate four subscales of a general questionnaire with the subscales of a task-specific questionnaire. They find the range of intercorrelations varying from .24 to .63. Questionnaires may also be related to *other* off-line methods such as interviews (Hopfenbeck 2009), and they may be compared with *on-line* measures, such as observations, log-file measures or think-aloud measures (Anderson et al. 2009; Cromley and Azevedo 2006; Sins et al. 2008; Veenman et al. 2003).

Veenman (2005) gives an overview of multi-method research into the measurement of metacognitive strategies. His overview concerns the research in which relation measures (or correlations) are calculated between the data gathered with two measuring methods for the same learner. Elaborating upon Veenman's overview, and including other studies, Van Hout-Wolters (2009) concludes that comparing on-line methods generally renders highly correlating results (correlations from 64. to .89), whereas questionnaires display a variable picture when compared to think-aloud measures (correlations from -.07 to .42). In her conclusion, Van Hout-Wolters differentiates between general questionnaires (reaching correlations up to .22 with think-aloud measures) and task-specific questionnaires (reaching correlations up to .42). Correlations are calculated on the full instruments' scales. For example, Cromley and Azevedo (2006) find a low correlation (-.02) between their think-aloud measure and the MARSI.

The question arises of what happens to the correlations between questionnaires and think-aloud measures if analyses at scale level are executed instead of analyses at instrument level. Verheul and Yang (1986) report a correlation of .56 between the think-aloud measures and the questionnaire for the *structuring* scale. Bannert and Mengelkamp (2008) find that observational data do not correspond with the separate scales of a questionnaire (the precise values of the correlations are not reported) with the exception of the *elaboration* scale, i.e., one scale of the questionnaire in *one* out of three research conditions (r=.54).

The high correlations between the on-line methods point to the strength of the think-aloud method and to the possibilities of using different on-line methods within the same situation and for the same goal. The high correlations between the on-line measures are often contrasted to the clearly lower correlations between the think-aloud measures and questionnaires (e.g.,Veenman 2005). These low correlations are often explained by the fact that questionnaires offer *less grip on the learning activities* the learner actually performed during the task execution.

For example, the learners may have forgotten some learning activities or consider them to be too unimportant to mention them. Learners can also mention learning activities that did not actually take place. Moreover, a learner might be unaware of the learning activities he executed or be unable to reflect in a way necessary for the correct completion of the questionnaire (cf. Perry and Winne 2006). Another point is that from his answers to the questionnaire it is not clear with whom the learner compares himself (for instance, when he states that he asks himself many or few questions during studying, cf. Veenman et al. 2003). Social desirability could also play a part during the completion of the questionnaire. In short, the validity of the data gathered by means of self-reports may be questioned, and these instruments are qualified as measuring less valid. In other words, learners are said not to be able to verbally report on their learning activities. However, there are other explanations for the low correlations between think-aloud measures and self-reports that deserve close attention.

Alternative explanations for low correlations between questionnaires and think-aloud measures

The above-mentioned explanation aims at the low self-reporting ability of learners, however, explaining the low correlations may involve two other aspects: the *learning strategies* measured and the *learning task* to which the measurement is related.

The first explanation is that the measuring methods to be compared may be aimed at measuring learning strategies that are *different in content*. As already mentioned,

many learning strategies can be distinguished and they differ greatly. Questionnaires, for instance, can be aimed at 'orienting', 'structuring' and 'evaluating', whereas the think-aloud protocols, which are used in the comparison, might be analyzed for 'monitoring', 'guiding' and 'reflecting'. If this is the case, it is not surprising that there are low correlations between these measures. Furthermore, in multi-method studies, correlations of total scores on the measuring instruments are often reported (see the previous section), whereas most instruments consist of several subscales (see e.g., Muis et al. 2007). Correlations between self-reports and think-aloud measures might turn out differently if analyses at scale level were performed measuring the same learning strategies (see e.g., Bannert and Mengelkamp 2008; Veenman et al. 2003; Verheul and Yang 1986).

The second explanation for the low correlations between questionnaires and think-aloud measures considers the *learning task*. The think-aloud method measures the concrete approach to a specific learning task. If this method is compared with a questionnaire, the questionnaire will also have to be aimed at this specific learning task. Research and practice use many 'task-exceeding' or 'general' questionnaires. It is to be expected that these general questionnaires show low correlations with the think-aloud method. To record the learning activities learners use during a *specific* learning task the items in the questionnaire will have to explicitly refer to this very specific learning task (cf. Samuelstuen and Bråten 2007).

According to Messick (1995), construct validity is essential for all educational and psychological measurements, establishing the extent to which assessments actually elicit the knowledge, skills, and other abilities they are intended to measure. A fundamental feature of construct validity is construct representation. Messick points out two major threats to the construct validity of assessment: the construct-irrelevant variance and construct-underrepresentation (cf. Admiraal et al. 2011). In cases of *construct-irrelevant variance*, the assessment is too broad and thus contains variance associated with other – irrelevant- information. This may be a major threat for questionnaires: the respondents must comprehend the meaning of the different items before being able to answer them. In this sense a questionnaire about learning may in fact become a reading comprehension test. And, as already mentioned, the items may include questions about activities about which the respondent has not attended to during learning. In cases of *construct under-representation*, the assessment is too narrow and thus fails to capture critical aspects of the target construct. This may be the case in the think-aloud method when learners do not verbalize all thoughts that arose. All a researcher can analyze is the protocol's content, and unspoken processes that do give rise to these verbalizations must be inferred. Furthermore, a reader may not be able to say everything that comes to mind, and may edit or omit thoughts that do come to mind. The think-aloud method seems pre-eminently suitable to tap *conscious* reflections (Ericsson and Simon 1993). But any set of verbalizations is just an approximation of what a learner actually does (cf. Magliano et al. 2011). By tailoring the questionnaire to the task at hand, the construct-irrelevant variance might be reduced, and one may find a higher correlation between the questionnaire and the think-aloud method.

Constructing a task-specific questionnaire

The advantages of measuring learning strategies by means of questionnaires are obvious: the learners are not being disturbed during their learning activities and the

gathering and scoring of data usually are straightforward and cost little time, even for large groups of learners. Because of these reasons, we wanted to examine whether we could reach a fair correlation between a learning strategy questionnaire and the think-aloud method, when we meet the two prerequisites we previously mentioned in comparing these instruments: the questionnaire and the think-aloud method should be aimed at the same learning task *ánd* at the same learning activities. More precisely, we expected to find a correlation that was higher than previously reported (see "Comparing learning strategy questionnaires with on-line measures").

The questionnaire was straightforwardly constructed on the basis of a taxonomy for coding think-aloud protocols. Meijer et al. (2006a) developed this taxonomy to analyze the metacognitive activities secondary school students perform during text studying (history text). An initial taxonomy was constructed both theoretically based and data driven; this taxonomy was tested on new protocols in a cyclic fashion (see Meijer et al. 2006a). The revised and final taxonomy discerned 56 specific metacognitive activities that adhered to four superordinate scales, *orientation & planning, execution, monitoring,* and *elaboration & evaluation*. This revised taxonomy directly formed the basis for the questionnaire. Besides the descriptions and examples in the taxonomy, real protocol fragments from the Meijer et al. study were also used in order to formulate the questionnaire's items that fitted the vocabulary of secondary school students. In order to construct a task-specific questionnaire, all items were formulated in the way that they referred to the text that was read before completing the questionnaire (cf. Samuelstuen and Bråten 2007). So our questionnaire was both task-specific and was aimed at the same metacognitive activities that could be coded in the think-aloud protocols (see examples in Table 1). More specifically, there was a direct relationship at the *item-leve*l in the two different methods: Each individual category (i.e., activity) in the taxonomy was tapped by one item in the questionnaire. In all, the questionnaire consisted of 58 items.

The items were scored on a three-points-frequency scale. The participants rated whether they exerted 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 favoring the three-points-scale was raised by Veenman et al. (2003). They suggested that the assessment of learning activities through the method of self-reports 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 about 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. A three-points-scale may then be used to reduce the variation in the choice of a reference point amongst students (Veenman 2005).

Three studies were performed: a comparative study to relate the questionnaire to the think-aloud method, a large-group study to consider the internal consistency of the questionnaire, and a case study to examine the self-reporting ability of the student in more depth.

Table 1 Examples of the categories in the taxonomy, typical verbalizations in the think-aloud protocols, and the corresponding items on the questionnaire

Categories	Fragments from think-aloud protocols	Questionnaire's items
Formulate action plan	First, I am going to read the text, then I will note what I don't understand and … (laughing) … and yeah, what I do understand.	I thought about the best way to read the text.
Commenting on (explanation in) text	Okay I think this sentence is a bit strange, this.	During reading, I made my mind up about explanations in the text.
Empathizing	Yes, it is indeed fair that slavery was stopped, but equal rights, yes .. not quite.	I empathized with the events described in the text.
	Yes, it doesn't make sense that the black people did not have equal rights, that's my opinion.	
Comprehension failure	I don't really understand what they mean with this.	During reading, I did not comprehend some parts of the text.
Noticing unfamiliar words or terms	Interpretation, I think that is a difficult word.	I noticed words in the text that I did not know.
Inferring, i.e., drawing conclusions beyond literal text.	Okay, so .. he was actually not that great, Lincoln, because … actually, it concerned only his own gain.. he ..yeah.. ok	I tried to draw conclusions beyond the literal text.
Interpreting, i.e. deducing the meaning of a word from other text parts	*abolition* (*reading text*). Okay that means, eh, they are the men who thought that slavery has to stop.	I inferred from the text the real meaning of some text parts.

Study 1: comparative study

Participants Participants were 16 students (ninth-graders) from two different classes. The two different classes were taught history by the same teacher. Participation was voluntary; the students were paid for their contribution.

Procedure The participants were required to read aloud a history text and to think aloud. 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 contained 1,650 words; the mean number of words per sentence was 17.64. The students were instructed to study the text carefully, in the way they would prepare for a test. There was no time-limit for text studying. In explaining how to think aloud, the students were told to read the text aloud and to verbalize any thoughts that arose during studying; as soon as a thought popped into the students' mind, he or she should talk about it. The experimenter stimulated the students to continue thinking aloud, although the prompting was done as little as possible. After studying, they had to (silently) answer the task-specific questionnaire.

Results The 16 (individual) research-sessions resulted in a total of 370 min of thinking aloud. The mean time of studying the text aloud was 23.13 min (sd=8.37). The 16 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. (2006a). First time text reading

was not coded. In all, 2,817 activities were coded. Pairwise categorizations from two raters were summarized in a cross table and Pearson's contingency coefficient was .97, so the interrater reliability was highly acceptable.

The number of activities that were discerned in the 16 protocols varied considerably. In a meagre protocol only 50 activities were observed, whereas 384 activities were coded in the richest protocol. The mean number of activities per protocol was 176.06 (sd=92.63). Four categories in the taxonomy were not found in the 16 protocols. All other categories were observed at least once with a maximum number (541 occurrences) for the subcategory "error detection (plus correction), keeping track". The total number of activities counted for each superordinate scale differed; we counted 1,120 activities for the Executing scale, 842 activities for the Monitoring scale, 559 activities for the Elaboration & Evaluation scale, and 296 activities for the Orientation & Planning scale.

Although the sample size within this study was not large, the reliabilities (Cronbach's alpha) for the questionnaire were established. The full instrument (58 items) resulted in a fair reliability ($\alpha=.87$) and the Elaboration & evaluation scale also reached a fair reliability ($\alpha=.85$). However, the reliabilities for the other three subscales were moderate (Orientation & planning $\alpha=.52$) to low (Executing $\alpha=.26$; Monitoring $\alpha=.36$). The descriptives per scale are presented in Table 2.

In calculating the correlation between the questionnaire and the think-aloud protocols, the conservative nonparametric correlation (Spearman's rho) for skewed distributions was used. The two methods showed a correlation of .51 (that reached a one-tailed, marginal significance; $p=.05$). This correlation might be seen as moderate, but as shown in our overall picture concerning the comparison of questionnaires with think-aloud measures, where correlations between complete instruments range from -.07 to .42, a .51 correlation was promising. Because every item on the questionnaire directly corresponded to one individual coding category in the taxonomy for coding think-aloud protocols, correlational analyses at scale-level could also be performed. Yet, analyses including *all* scales were not executed because of the reliability analyses on the questionnaire: we just considered the two scales with moderate and high reliability. It appeared that the Elaboration & evaluation scale of the questionnaire reached a significant correlation ($r=.60$) with the same scale in the think-aloud method, whereas the Orientation & planning scale did not ($r=.24$). This low correlation across 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). Still, the question arose whether the ninth-graders within this study were more aware of elaboration and evaluation activities than of orientation and planning activities.

Table 2 Descriptives and internal consistencies for the questionnaire (study 1 and study 2)

Questionnaire		Study 1 (*n*=16)		Study 2 (*n*=190)	
Scale	N items	M (sd)	Cronbach's alpha	M (sd)	Cronbach's alpha
Orientation and planning	16	1.98 (.25)	.52	2.04 (.24)	.54
Executing	12	1.88 (.28)	.26	1.62 (.24)	.35
Monitoring	14	1.89 (.22)	.36	1.78 (.23)	.40
Elaboration and evaluation	16	1.91 (.40)	.85	1.80 (.31)	.69
Full instrument	58		.87		.79

As observed, the internal consistencies for the questionnaires' scales were moderate. Since the sample was rather small, we decided to administer the questionnaire within a larger group of students.

Study 2: larger-group study

Participants 190 students (ninth-graders) in secondary education (seven different classes). Participation was obligatory because the session was scheduled during their lessons.

Procedure The students silently studied the history text (about slavery and the American civil war). After studying, they were presented with the questionnaire.

Results The overall, internal consistency for the questionnaire was reasonably high ($\alpha=.79$). However, the internal consistency for each subscale appeared to be moderate to low: Orientation and planning ($\alpha=.54$), Executing ($\alpha=.35$), Monitoring ($\alpha=.40$), Elaboration and evaluation ($=.69$) (see Table 2). The reliabilities found in the present study correspond to the reliabilities found in the previous study.

Although, the internal consistency of the full instrument was reasonably high, the data concerning the subscales were not convincing. In sum, the subscales of the questionnaire were not clear indications of the constructs these subscales were intended to measure.

In Veenman's review of assessment methods, questionnaires reached adequate reliabilities but remained low in correlations with the think-aloud method, whereas we found moderate to low reliabilities at (sub)scale level on a questionnaire that had reached a provoking correlation with the think-aloud method. A possible explanation is that in our study the questionnaire was *empirically* constructed: each activity had previously appeared (albeit with different frequencies) in a think-aloud protocol. However, the distinction in four superordinate strategies (i.e., the four scales) was *theoretically* based (following the study of Meijer et al. 2006a). 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 not follow the experts about the subscales differentiation. In other words, students may perceive the scale items not to be interrelated, which might to lead to an inconsistent answering pattern according to the experts' view (resulting in low reliabilities for the subscales).

Both in study 1 and study 2, the full questionnaire scale reached acceptable reliabilities. Moreover, the correlation of .51 between the questionnaire and the think-aloud protocols was higher than the correlations regularly reported. This result indicates that questionnaires' data might be more validly interpreted than is often assumed. However, the correlation of .51 still indicates that both instruments also measure different aspects in learning. The question of what is happening during answering the questionnaire arises. By executing a case-study, in which a few respondents have to think aloud while completing the questionnaire, we want to get an idea of how they interpret each of the items and whether the questionnaire items affect the way in which the respondents recollect their actual behavior.

Study 3: case study

Participants Four ninth-graders. The students, who voluntarily participated, were paid for their contribution.

Procedure The students read and thought aloud about the history text. The questionnaire items were read aloud and the students still had to think aloud while answering the questionnaire.

Results The think-aloud protocols were audio-taped and transcribed. The verbalizations during text studying were coded in the same way as in the first comparative study. Every utterance (as well as an omission or a repetition) during the student's text reading that deviated from the literal text was interpreted by means of the categories in the taxonomy of Meijer et al. (2006a). In all, 619 activities were coded for text studying.

In Table 3, the scale descriptives from both instruments are presented for each student: the percentage of activities in the think-aloud protocols and the scale means on the questionnaire. It was expected beforehand that a student with many activities in the protocol would also have a relatively high mean on the full questionnaire scale (the items were scored on a frequency scale). As the last row of Table 3 shows, this trend was visible: the student executing the highest number of activities (student 2: 292 activities) showed the highest overall score on the questionnaire (M=2.17), whereas the student who executed the least number of activities (student 4: 74 activities) ended up with the lowest overall score on the questionnaire (M=1.64). Remarkably, on the Orienting & Planning scale all four students executed a relatively *low* percentage of activities according to their think-aloud protocols, whereas they rated relatively *high* frequencies of performing "Orienting & planning" activities on the questionnaire. The protocol percentages are in line with other studies, in which it is found that orienting activities are hardly performed by students while studying text (Meijer et al. 2006a; Molenaar et al. 2011). In educational environments, however, students are regularly pointed to the importance of orienting or planning activities (e.g., Butler and Cartier 2004). So this "Orienting and planning' scale in particular might be affected by social desirability.

The verbalizations in the think-aloud protocols while answering the questionnaire were interpreted if they referred to 'learning the history text within the research setting' (=taking a task-specific learning perspective) or to general learning activities, i.e., descriptions of activities the students normally exert in learning. A larger number of questions were answered by the students in a task-specific way in comparison with the 'general learning perspective' (see Table 4), although student 1 answered more questions in a general way. Table 5 presents some examples of answering in task-specific or general ways. However, many verbalizations were very short (*"no", "that is right", "often", "almost never"*, etc.).

Table 3 Scale descriptives in Study 3: percentage of activities in the think-aloud protocols (P) and the scale means on the questionnaire (Q)

	Student 1		Student 2		Student 3		Student 4	
	P	Q	P	Q	P	Q	P	Q
	%	M	%	M	%	M	%	M
Orienting and planning	9.8%	2.06	3.4%	2.31	11.6%	1.88	10.8%	1.81
Executing	33.3%	2.00	47.6%	1.75	33.1%	1.83	36.5%	1.58
Monitoring	40.9%	2.17	23.3%	2.21	52.1%	1.86	31.1%	1.71
Evaluation and elaboration	15.9%	1.88	25.7%	2.31	3.3%	1.50	21.6%	1.44
All	100 (=132)[a]	2.02	100 (=292)[a]	2.17	100 (=121)[a]	1.76	100 (=74)[a]	1.64

[a] The total number of activities counted in the protocols

Table 4 Perspectives of the students in answering the questionnaire (*n*=58 items)

Perspective	Student 1	Student 2	Student 3	Student 4
Task-specific learning	10	36	13	11
General learning	22	1	0	6
Indefinite	26	21	45	41

From these verbalizations it was not immediately clear which perspective the respondents had in mind while answering the questions.

By answering the questionnaire aloud, it turned out that most questions were properly understood by the students. However, some questions were unclear for the students (see Table 6). For example, students seemed to perceive terms in other ways than the author of the questionnaire did. Two students appeared to be confused by the term 'concepts'. Some difficulties were also found with questions including abstract words, such as "drawing conclusions", "finding information", and "things to be remembered".

In the cases where students experienced some trouble in interpreting the question they frequently chose the answer alternative in the middle of the scale as if this category referred to a 'neutral' answer instead of the frequency category of 'sometimes'. This approach of given 'neutral' answers was also used when the students did not know what they had done during studying.

Another striking example in choosing answer alternatives concerned the cases in which the students seemed to verbalize other frequencies than they actually rated on the frequency scale. For example the student who realized he made no notes in this task, but then gave an answer that did not match his actual behavior (student 1: "Which notes? I *haven't made any*

Table 5 Different excerpts in answering the questionnaire in a task-specific or in a general way

	Question	Excerpt
Task-specific learning verbalization		
4	During reading, I thought about the remaining time.	<*student 1*> I don't agree, because … there was no restriction in time here.
7	After reading the text, I read the concepts again.	<*student 4*> No. Well, normally would read it again, but this time I didn't. So I have to choose "almost never"?
17	I tried to draw conclusions beyond the literal text.	<*student 2*> Eh, did I do that? Eh I did not because often when I had a thought, then it was explained later on in the text. So I thought … eh.. not often
41	I noticed words in the text that I did not know.	<*student 3*> Yes <*smiles*> . If I, yeah, actually it is sometimes, because I only didn't know what ab-bo-lit-ion-something was.
General learning in verbalization		
2	I thought about the best way to read the text.	<*student 4*> I don't agree, I never do that. So. I always just read it. We'll see.
12	In order to prepare for the test, I summarized the text first.	<*student 2*> Well, I would do that if I had a real exam about the text. Then I would write down the important points, and the dates and everything. So I actually agree.
19	After reading some sentences, I tried to say the contents in my own words.	<*student 1*> Actually, I never do that… it takes too much time.

Table 6 Questions leading to different interpretations by or to uncertainty among the students

	Question	Excerpts
1	I thought about what I was supposed to do.	*<student 2>* Well, I have read the text and I answered the questions, so I believe that I have thought about this.
7	After reading the text, I read the concepts again .	*<student 1>* So. … they weren't present
		<student 2> Well, there really were no concepts present, at least not clearly marked. But if I came across a difficult idea, then I tried to look back in the text and search for what it meant, so "sometimes".
14	I drew my conclusions without searching in the text.	*<student 3:>* euh … at which … I drew conclusions without searching in the text .. Yeah, I think I chose sometimes, as I didn't really understand what to draw conclusions about.
17	I tried to draw conclusions beyond the literal text.	*<student 1>* I tried to draw conclusions beyond the … literal text. I don't understand this. Then, we put this in the middle, neutral
33	During reading, I could not remember things.	*<student 1>* What kind of things? I don't get it .. yes, I am… I don't understand this..
49	In the text, I could not find the information that I needed	*<student 1>* It depends which kind of information you need, so … I do not know how to answer this question …
		<student 3> Um, information for what? I *<speaking in soft voice>* don't think I did that

notes, so … in the *middle*"); this same student first verbalized to *frequently* think about the information to be remembered, but then chose for "*sometimes*" on the rating scale (student 1: Uh.. you're thinking about the information to be remembered, *really often* … well, then we just put *"sometimes"* here).

In all, the case study presents some insights in why a questionnaire may lead to somewhat different results in comparison with think-aloud protocols: sometimes students were not understanding the question, sometimes the students were 'rating' their general learning, and sometimes social desirability or uncertainty were probably playing a role. Because of the above descriptions, one might suppose that the four students experienced many problems in answering the questionnaire, but we have presented almost *all* notable verbalizations. No difficulties were detected concerning the remaining questionnaire items. Many questions were answered in a clear and task-specific way and the descriptives of the case-study were in line with the first think-aloud study (see "Study 1: comparative study") indicating the correspondence between the questionnaire and the think-aloud protocols.

Conclusions concerning the task-specific questionnaire

In our empirical work we constructed a task-specific questionnaire that we administered to students in three different studies. In our comparative study we found that a well-specified questionnaire, a questionnaire that is aimed at the same learning strategies and the same learning task as a think-aloud taxonomy resulted in a correlation ($r=.51$) between the questionnaire and the think-aloud protocols. This correlation is provocative because the correlations regularly reported (see our overall picture in "Measuring learning strategies with questionnaires" are indeed lower. However, the correlation of .51 still indicates that both instruments may also measure different constructs.

In Veenman's review (2005), questionnaires reach adequate reliabilities but remain low in convergence validity, i.e., low correlation between the measures. Yet in the first study we found a promising overall correlation, and a fairly acceptable internal consistency for the full instrument, but in both the first and second studies ($n_1=16$ students and $n_2=190$ students), the reliabilities at scale level were moderate to low. As already explained, the distinction in the scales was *theoretically* based (Meijer et al. 2006a), whereas the students may have perceived different links among the items.

The case study, in which four students thought aloud while answering the questionnaire, showed some different answering patterns of the students. Some task-specific questions were answered while referring to general learning, few questions were not understood, and some may have been answered from a social desirability perspective. With some questions the students seemed to feel uncertain in recollecting their behavior. However, the descriptive results in Table 3 pointed to a correspondence between the questionnaire and the think-aloud protocols. The results found within this case-study imply new possibilities in examining validity and reliability issues of questionnaires by having respondents answer a questionnaire while thinking aloud. Since many verbalizations concerning the questionnaire items were short, one may advise extending the think-aloud session by interviewing the respondents afterwards about how they filled in the questionnaire. For example, Karabenick et al. (2007) use cognitive interviewing, that is to say, systematically interviewing the respondent after answering each individual item of a questionnaire.

One of the advantages of the studies - strictly parallel construction of instruments at the level of items, is a disadvantage at the same time. In order to analyze the correspondence between the questionnaire and the think-aloud method, the studies (1 and 3) have implemented a design where every participant first learns and thinks aloud and afterwards completes the questionnaire. This means that the questionnaire is always second. The students filled out the questionnaire after having verbalized their thoughts directly beforehand. 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 in a questionnaire before studying the text to report their *expectations* of what they would do in this specific task (i.e. "prospective" questionnaire answering). However, a change in the order of instruments actually concerns changes in both the questionnaire and in the research question. Our research question really concerned "are students able to verbally report what they have done while studying?" 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. By changing the order of instruments for half of the students the research question for these students becomes: "do students *foresee* what they are going to do while learning a specific text?" Although both research questions are interesting in meticulously comparing the two measuring methods, this study was confined to the first question.

Overall, the three presented studies form a relevant step in considering the possibilities in using questionnaires to measure metacognitive activities. Yet numerous new research issues arise for example, what happens to the correlation between the questionnaire and the think-aloud protocols in different age groups and with different study tasks (e.g., problem solving); and what happens to the scale-reliabilities when answer scales other than the 3-points-scale are used. Additionally, some task-specific questions were answered while referring to general learning. Maybe all questions should be adjusted in their phrasing in

such a way that *each and every* item explicitly refer to the text and task at hand. For example, the original question: "*In order to prepare for the test, I summarized the text first*" may be replaced by "*In this case, I made a summary of the specific text in order to be prepared for the test*". Another possible adjustment concerns the precise phrasing of the answer scale. The answer scale ("very often") may direct respondents to think about regular learning, whereas the scale "Did you execute this activity very often on this task?" triggers the respondent to keep explicitly thinking about the task at hand.

As mentioned in the introduction, Messick (1995) sees construct validity to be essential for all educational and psychological measurements and two major threats may hamper the construct validity of assessment: the construct-irrelevant variance and construct underrepresentation. By tailoring the questionnaire to the task at hand, we believe we have reduced the "construct-irrelevant-information" which is very likely resulting from questionnaires (see Introduction). By having the students complete the questionnaire while thinking aloud, we have also checked whether the questionnaire's items were intelligible for the respondents. We did find a fair correlation between the questionnaire and the think-aloud method, although the correlation might be higher still in new research. However, we did not examine the consequences of the narrowness of the think-aloud method. With the think-aloud method used, we do not know whether the learners have verbalized all their thoughts and "construct-under-representation" may have been present. According to Messick, validity is a property of an interpretation of data, and to do this properly we need more data. At this moment, we have compared one questionnaire with one type of the think-aloud method concerning one task (although within three different studies). In order to be able to validly interpret the data, much more research is needed including different types of the think-aloud method and different tasks. Since in different tasks, or with different texts, both the construct-irrelevance variance and the construct underrepresentation of both the questionnaire's items and think-aloud categories might vary. For practical purposes, we strive for instruments that are adequate alternatives for the labor-intensive think-aloud methods, but, both instruments should probably be continuously used in mixed-method designs for theoretical considerations. Given the relatively low correlations and internal consistencies at on subscale level (except for the Elaboration and evaluation scale), our research seems to indicate that a task-specific questionnaire may not tap *particular* study activities. However, this critical note may also be made about the think-aloud method (cf. Magliano et al. 2011).

To conclude, based on the comparison with the think-aloud protocols, a task-specific questionnaire may be a more promising instrument than sometimes thought, yet some further improvements should be examined in more depth.

Overall conclusion

This article discusses measuring learning strategies with the help of questionnaires. In 'multi-method' research, in which think-aloud measures are compared with questionnaires, low or moderate correlations are found. A conclusion often drawn is that learners are not able to verbally report on their learning activities. However, we presented some alternative explanations for these regularly-found low correlations which consider the possibility that the methods compared may be aimed at different strategies and different tasks. For a fair comparison, both the questionnaire and the think-aloud method should be aimed at measuring the *same learning strategies* that are performed on the *same learning task.* In

examining the possibilities and problems by applying questionnaires, we used two ways in comparing the questionnaire to the think-aloud method. First, we constructed a questionnaire straightforwardly based on a taxonomy for coding think-aloud protocols in text studying. Second, the think-aloud method was used during the administration of the questionnaire. The correlation found between the questionnaire and the think-aloud protocols was higher than regularly reported and the case-study sheds some valuable light on how to improve the questionnaire. We cautiously conclude that if task-specific questionnaires are meticulously constructed and examined in new ways, they might become reasonably adequate alternatives for the labor-intensive think-aloud method in measuring learners' learning strategies.

Appendix 1: questionnaires for the measurement of learning strategies

– Approaches to Learning and Studying Inventory (ALSI, Entwistle and McCune 2004)
– Approaches to Studying Inventory (ASI; Entwistle & Ramsden 1983)
– ASSIST: A reconceptualisation of the Approaches to Studying Inventory (Tait et al. 1998)
– Awareness of Independent Learning Inventory (AILI, translation of a Dutch questionnaire; Meijer et al. 2006b)
– Dynamic and Active Learning Inventory (DALI, Iran-Nejad & Chissom 1992).
– Effective Study Test (Brown 1986)
– Index of Learning Styles (ILS, Felder & Spurlin 2005)
– The revised Inventory of Learning processes (ILP-R, Schmeck et al. 1991)
– Inventaris Leerstijlen (ILS-HO, Inventory of learning styles- Higher Education, Vermunt 1992)
– Learning and Study Strategies Inventory (LASSI, Weinstein & Palmer 2002)
– Learning Strategies Survey (Kardash & Amlund 1991)
– Learning Style Inventory (LSI, Dunn et al. 1987)
– Learning-Style Inventory – version 3 (LSI, Kolb 1999)
– Metacognitive Awareness Inventory (MAI, Schraw & Dennison 1994)
– Metacognitive Awareness of Reading Strategies Inventory, Version 1.0 (MARSI; Mokhtari and Reichard 2002)
– Motivated Strategies for Learning Questionnaire (MSLQ, Pintrich et al. 1993)
– Revised Approaches to Studying Inventory (RASI, Entwistle & Tait 1995)
– Self-Efficacy and Metacognition Learning Inventory-Science (SEMLI-S, Thomas et al. 2008)
– State Metacognitive Inventory (SMI, O'Neil & Abeli 1996)
– Study Attitudes Methods Survey (Michael et al. 1988)
– The Revised two-factor Study Process Questionnaire: R-SPQ-2 F (Biggs et al. 2001)

References

Admiraal, W., Hoeksma, M., Duin, G., & Janssen, J. (2011). *Assessment evidence of teacher competences in video portfolios: Reliability, construct validity and consequential validity.*
Alexander, P. A. (2006). *Psychology in learning and instruction*. Upper Saddle River: Pearson Prentice Hall.

Anderson, D., Nashon, S., & Thomas, G. P. (2009). Evolution of research methods for probing and understanding metacognition. *Research in Science Education, 39*, 181–195.

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect the learning? *Metacognition and Learning, 3*(1), 39–58.

Biggs, J., Kember, D., & Leung, D. Y. P. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology, 71*, 133–149.

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*(1), 41–58.

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

Broekkamp, H., & Van Hout-Wolters, B. H. A. M. (2007). Students' adaptation of study strategies when preparing for classroom tests. *Educational Psychology Review, 19*(4), 401–428.

Brown, W. F. (1986). *Effective study test*. Iowa City: ACT.

Butler, D. L., & Cartier, S. C. (2004). Promoting effective task interpretation as an important work habit: A key to successful teaching and learning. *Teachers College Record, 106*, 1729–1758.

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

Cromley, J. G., & Azevedo, R. (2006). Self-report of reading comprehension strategies: What are we measuring? *Metacognition and Learning, 1*(3), 229–247.

Dunn, R., Dunn, K., & Price, G. E. (1987). *Learning style inventory*. Lawrence: Price Systems.

Entwistle, N., & McCune, V. (2004). The conceptual bases of study strategies inventories. *Educational Psychology Review, 16*(4), 325–346.

Entwistle, N., & Ramsden, P. (1983). *Understanding student learning*. London: Croom Helm.

Entwistle, N. J., & Tait, H. (1995). *The revised approaches to studying inventory. Centre for research on learning and instruction*. Edinburgh: University of Edinburgh.

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

Felder, R. M., & Spurlin, J. (2005). Applications, reliability and validity of the index of learning styles. *International Journal of Engineering Education, 21*(1), 103–112.

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.

Hopfenbeck, T. N. (2009). *How can we know that they think what we mean? Investigating students' self-reports of self-regulated learning*. Paper presented at the 13th Biennial Conference for European Research on Learning and Instruction, EARLI, Amsterdam, August 25–29th.

Iran-Nejad, A., & Chissom, B. S. (1992). Contributions of active and dynamic selfregulation to learning. *Innovative Higher Education, 17*, 125–136.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., et al. (2007). Cognitive processing of self-report items in educational research: do they think what we mean? *Educational Psychologist, 42*(3), 139–151.

Kardash, C. M., & Amlund, J. T. (1991). Self-reported learning strategies and learning from expository text. *Contemporary Educational Psychology, 16*, 117–138.

Kolb, D. (1999). *Learning-style inventory—version 3*. Boston: Hay/McBer Training Resources Group.

Magliano, J. P., Millis, K. K., The RSAT Development Team, Levinstein, I., & Boonthum, C. (2011). *Assessing comprehension during reading with the reading strategy assessment tool (RSAT)*.

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

Meijer, J., Elshout-Mohr, M., Van Daalen-Kapteijns, M., Meeus, W., & Tempelaar, D. (2006b). *Construction and validation of a questionnaire on metacognition*. Paper presented at the Second bi-annual conference of the special interest group on Metacognition of EARLI, Cambridge.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Michael, W. B., Michael, J. J., & Zimmerman, W. S. (1988). *Study attitudes and methods survey*. San Diego: EDITS.

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.

Mokhtari, K., & Reichard, C. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology, 94*(2), 249–259.

Molenaar, I., Van Boxtel, C. A. M., & Sleegers, P. J. C. (2011). *The effects of scaffolding metacognitive activities in small groups.*

O'Neil, H. F., Jr., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *Journal of Educational Research, 89*(4), 234–245.

Perry, N. E., & Winne, P. H. (2006). Learning from learning kits: Study traces of students' self-regulated engagements with computerized content. *Educational Psychology Review, 18,* 211–228.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385–408.

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

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Predictive validity and reliability of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement, 53,* 801–813.

Rachal, K. C., Daigle, S., & Rachal, W. S. (2007). Learning problems reported by college students: are they using learning strategies? *Journal of Instructional Psychology, 34*(4), 191–200.

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.

Schmeck, R., Geisler-Brenstein, E., & Cercy, S. P. (1991). Self-concept and learning: The revised inventory of learning processes. *Educational Psychology, 11,* 343–362.

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

Sins, P., van Joolingen, W., Savelsbergh, E., & van Hout-Wolters, B. (2008). Motivation and performance within a collaborative computer-based modelling task: Relations between students' achievement, goal orientation, self-efficacy, cognitive processing, and achievement. *Contemporary Educational Psychology, 33,* 58–77.

Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2004). Metacognition and self-regulated learning constructs. *Educational Research and Evaluation, 10*(2), 117–139.

Tait, H., Entwistle, N. J., & McCune, V. (1998). ASSIST: A reconceptualisation of the approaches to studying inventory. In C. Rust (Ed.), *Improving student learning: Improving students as learners.* Oxford: Oxford Centre for Staff and Learning Development.

Thomas, G. P., Anderson, D., & Nashon, S. (2008). Development of an instrument designed to investigate elements of science students' metacognition, self-efficacy and learning processes: the SEMLI-S. *International Journal of Science Education, 30,* 1701–1724.

Van Hout-Wolters, B. H. A. M. (1992). *Cognitieve strategieën als onderwijsdoel. [Cognitive Strategies as Educational Goal.].* Groningen: Wolters-Noordhoff.

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 Hout-Wolters, B. H. A. M., Simons, P. R. J., & Volet, S. (2000). Active learning: Self-directed learning and independent work. In P. R. J. Simons, J. L. van der Linden, & T. M. Duffy (Eds.), *New learning* (pp. 21–37). Dordrecht: Kluwer.

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

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*(1), 3–15.

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

Vermunt, J. D. (1992). *Leerstijlen en sturing van leerprocessen in het hoger onderwijs [Learning styles and regulation of learning processes in higher education].* Lisse: Swets & Zeitlinger.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Hillsdale: Erlbaum.

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.

Weinstein, C. E., & Hume, L. M. (1998). *Study strategies for lifelong learning*. Washington: American Psychological Association.

Weinstein, C. E., & Palmer, D. R. (2002). *LASSI user's manual: For those administering the learning and study strategies inventory*. Clearwater: H&H Pub.

Zimmerman, B. J. (1994). Dimensions of academic self regulation: A conceptual framework for education. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance: Issues and educational applications* (pp. 3–21). Hilsdale: Erlbaum.