Learning and Individual Differences 29 (2014) 123–130

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

The on-line assessment of metacognitive skills in a computerized learning environment

Marcel V.J. Veenman [a,b,*], Laura Bavelaar [c], Levina De Wolf [c], Marieke G.P. Van Haaren [c]

[a] Dept. of Developmental & Educational Psychology, Leiden University – Institute for Psychological Research, Wassenaarseweg 52, 2333AK Leiden, The Netherlands
[b] Institute for Metacognition Research (IMR), Hillegom, The Netherlands
[c] ICLON, Leiden University, The Netherlands

ARTICLE INFO

Article history:
Received 28 May 2012
Received in revised form 7 November 2012
Accepted 10 January 2013

Keywords:
Metacognitive skills
Assessment
Validity
Computer-based learning environment (CBLE)
Logfiles

ABSTRACT

Metacognitive skills regulate and control learning processes. For assessing metacognitive skills in learners, on-line assessment is required during actual task performance. An unobtrusive on-line method is the analysis of learner activities that are registered in logfiles of computerized tasks. As logfiles cannot reflect the learner's metacognitive considerations for enacting specific activities, logfile analysis should be validated against other on-line methods. Also, external validity of logfile measures needs to be established with related measures, such as learning performance. Fifty-two second-year students (13 years) from pre-academic education performed a computerized inductive-learning task. Traces of learner activities were stored in logfiles and automatically scored on indicators of metacognitive skills. Afterwards, participants completed learning-performance posttests. Results show high convergent validity between logfile indicators and human judgments of traced learner activities. Moreover, external validity was obtained for logfile measures in relation to learning performance (but not regarding participants' IQ scores). Implications for logfile analysis are discussed.

1. Introduction

Metacognition is a profound predictor of learning outcomes (Wang, Haertel, & Walberg, 1990). Often a distinction is made between knowledge of cognition and regulation of cognition (Brown, 1987; Schraw & Dennison, 1994). Metacognitive knowledge is declarative knowledge about the interplay between person characteristics, task characteristics, and strategy characteristics (Flavell, 1979). Having declarative metacognitive knowledge available, however, does not guarantee that this knowledge is actually used for the regulation of learning behavior (Veenman, Van Hout-Wolters, & Afflerbach, 2006; Winne, 1996). Metacognitive knowledge may be incorrect or incomplete, the learner may fail to see the usefulness or applicability of that knowledge in a particular situation, or the learner may lack the skills for doing so. Metacognitive skills refer to procedural knowledge for the actual regulation of, and control over one's learning behavior. Orientation, goal setting, planning, monitoring, evaluation and recapitulation are manifestations of those skills (Veenman, 2011a). Metacognitive skills directly shape learning behavior and, consequently, they affect learning outcomes. Veenman (2008) estimated that metacognitive skillfulness accounts for about 40% of variance in learning outcomes for a broad range of tasks.

For a decade, a debate is ongoing about how to assess metacognitive skills (Dinsmore, Alexander, & Loughlin, 2008; Veenman, 2005; Veenman et al., 2006; Winne, 2010). In general, there are two approaches to assessing metacognitive skills. Off-line methods collect students' self-reports either prior or retrospective to actual task performance. Often questionnaires (e.g., MAI of Schraw & Dennison, 1994; MSLQ of Pintrich & De Groot, 1990) and less often interviews (e.g., the SRLIS of Zimmerman & Martinez-Pons, 1990) are used for gathering self-reports of strategy use off-line. Questionnaires have the advantage that they are easy to administer to large groups. On-line methods, on the other hand, concern assessments of metacognitive strategy use during task performance. Typical on-line methods include observations (Whitebread et al., 2009), the analysis of think-aloud protocols (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995; Veenman, Elshout, & Groen, 1993), and eye-movement registration (Kinnunen & Vauras, 1995). These methods are time-consuming and labor-intensive, as they need to be individually administered. Apart from disparity in when assessments take place, another principal difference between off-line and on-line methods is that off-line measures merely rely on learner self-reports, whereas on-line measures pertain to the coding of actual learner behavior on externally defined criteria by external agencies, such as 'blind' judges and observers (Veenman, 2011a). The problem with off-line self-reports is that they need to be reconstructed from memory by the learner and, consequently, these self-reports are

* Corresponding author at: Dept. of Developmental & Educational Psychology, Leiden University – Institute for Psychological Research, Wassenaarseweg 52, 2333AK Leiden, The Netherlands. Tel.: +31 715273463.
E-mail address: Veenman@fsw.leidenuniv.nl (M.V.J. Veenman).

1041-6080/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.lindif.2013.01.003

subject to memory failure, distortion, and interpretive reconstruction (Veenman, 2011a,b).

There are three validity indices (De Groot, 1969; Nunnally & Bernstein, 1994) that are relevant to assessment of the construct of metacognitive skills (Veenman, 2007). The first one is internal consistency or reliability. Both off-line and on-line measures of metacognitive skills are usually checked for reliability, for instance, with Cronbach's alpha or with Cohen's kappa (Veenman, 2005). The second validity criterion concerns construct validity. Apart from designing meaningful assessment instruments, which refers to content validity, construct validity can be supported by establishing convergent validity (Beijk, 1977). The latter means that different assessment methods should point into the same direction, as reflected in high correlations between scores obtained with different methods in a multi-method design (Veenman et al., 2006). There is accumulating evidence that students' off-line self-reports do not converge with their actual metacognitive strategy use during task performance. Correlations between off-line and on-line measures are low ($r = .15$ on the average; Bannert & Mengelkamp, 2008; Cromley & Azevedo, 2006; Veenman, 2005, 2011a; Veenman, Prins, & Verheij, 2003) and qualitative analyses show that off-line self-reports do not converge with specific on-line behaviors (Hadwin, Nesbit, Jamieson-Noel, Code, & Winne, 2007; Winne & Jamieson-Noel, 2002). Apparently, learners do not do what they previously said they would do, nor do they accurately recollect what they have recently done. Moreover, correlations among off-line measures are often low to moderate, whereas correlations among on-line measures are moderate to high (Cromley & Azevedo, 2006; Veenman, 2005). In conclusion, off-line methods yield rather disparate results, while on-line methods converge in their assessments of metacognitive skills. Finally, the third validity criterion is external validity. An assessment instrument for metacognitive skills should behave in relation to other variables as expected by metacognition theory. For instance, most theories on metacognition claim that better metacognitive skills yield higher learning outcomes (Schraw, 2007; Veenman et al., 2006). Thus, any measure of metacognitive skills should substantially predict learning outcomes. In a review study of Veenman (2005), however, correlations with learning performance ranged from slightly negative to .36 for off-line measures, while ranging from .45 to .90 for on-line measures. Although a majority of studies rely on off-line self-reports (Dinsmore et al., 2008; Winters, Greene, & Costich, 2008), the low convergent and external validity of off-line methods would make an argument for resorting to on-line methods when assessing metacognitive skills.

With the emergence of computer-based learning environments, the on-line method of tracing metacognitive behaviors of learners in computer logfiles has become available (Kunz, Drewniak, & Schott, 1992; Veenman, Wilhelm, & Beishuizen, 2004; Veenman et al., 1993; Winne, 2010). A learning task is presented on a computer, while learner activities are recorded in a logfile and automatically coded according to a coding scheme. Obviously, the advantage is that several learners can work simultaneously on their individual computer in the classroom (or even at home), during which data are unobtrusively collected. An example of a hypermedia environment with logfile registration is *gStudy* of Winne (Hadwin et al., 2007). In *gStudy*, learners are provided with tools for making short notes, glossary entries, highlighting text, and making links across concepts during text studying. Trace data of study events in *gStudy* are registered into a log-file, which can be used to produce frequency counts, patterns of study activities, and content analysis of student notes. A limitation of logfile analysis in general, however, is that metacognitive skills need to be inferred from concrete learner activities ('events') without verbal accounts from learners for their behavior (Veenman, 2012). Nelson's model of metacognition may elucidate this limitation (Nelson, 1996; Nelson & Narens, 1990).

Nelson's model distinguishes an 'object-level' from a 'meta-level' in the cognitive system. At the object level lower-order cognitive activity takes place, while higher-order processes of evaluation and planning at the meta-level govern the object level. Two flows of information between both levels are postulated. Information about the state of the object-level is conveyed to the meta-level through monitoring processes, while instructions from the meta-level are implemented on the object-level through control processes. Thus, if an error occurs in cognitive activities on the object-level, a monitoring process will give notice of it to the meta-level and control processes will be activated to resolve the problem. Veenman (2011a) extended this model with self-instructions, that is, self-induced control processes that can be activated without prior monitoring alerts. Through learning and experience, learners may acquire a personal repertoire of self-instructions that is activated whenever they encounter a new task. These self-instructions induce cognitive activities in an orderly way on the object level, while feedback about the implemented cognitive activities is received by the meta level through the monitoring information flow. The on-line method of thinking aloud may give access to activities on the object level *and* to both information flows, whereas logfile analysis merely assesses the object level (Veenman, 2012). Thus, the cognitive activities that are registered in logfiles may either result from a metacognitive control process, or represent plain cognitive activity without metacognitive mediation. Therefore, validation research is needed to investigate the coverage of metacognitive skills by logfile analysis.

A first validation approach is to investigate the convergent validity of logfile indicators with another on-line method, such as thinking aloud or observations. For instance, in a study of Veenman et al. (1993) with a computerized environment for learning calorimetry, participants could heat blocks of different materials and weights on a burner and measure the temperature of the blocks. In a pilot study, a composite logfile measure (frequency of experiments, and frequencies of measuring the initial or final temperature of the blocks) correlated .62 with a think-aloud measure of metacognitive skillfulness (quality of orientation, goal setting, planning, monitoring, evaluation, recapitulation, and reflection). In the main study, this composite logfile measure was used to establish that thinking aloud did not affect the assessment of metacognitive skills, other than slightly slowing down the process (cf. Ericsson & Simon, 1993). In the same vein, Veenman et al. (2004) validated two logfile measures with think-aloud measures, obtained from participants performing two computerized inductive-learning tasks. Correlations between logfile and think-aloud measures ranged between .84 and .85. Next, the logfile measures were used to reveal a steep incremental development of metacognitive skills in 9, 14, 17, and 22 year old participants. In both studies, however, other potential logfile indicators had to be excluded because they were not related to think-aloud measures of metacognitive skillfulness.

Another validation approach is to compare logfile indicators with a broader judgment of the sequence of traced events in the logfiles. Human judges are more flexible and their judgments may capture elements or patterns of metacognitive activity that are not fully represented by logfile indicators (Veenman, 2012). This approach is chosen for the present study. As a first hypothesis, logfile indicators are expected to correlate substantially with human judgments of traced events.

The second research question pertains to external validity of logfile indicators. As a second hypothesis, it is expected that logfile indicators substantially predict learning outcomes, equally to human judgments of traced events.

Moreover, the intelligence of participants is assessed for reasons of external validation. Veenman et al. (2004) obtained correlations between intelligence and metacognition of .40 and .43, respectively, for 11.5 and 14 year old students who performed a less complex version of the discovery-learning task in the present study. Discovery learning is demanding because the task situation is relatively unstructured and learning processes rely heavily on inductive reasoning (De Jong & Van Joolingen, 1998). Therefore, intelligence is likely to be a relevant predictor of learning outcomes from discovery tasks, along with metacognitive skills (Snow & Lohman, 1984; Veenman & Elshout, 1995). An overview of studies by Veenman (2008) with participants of different ages, performing different tasks in different domains, has shown that intelligence correlated .45 with metacognitive skillfulness on the average. Intelligence uniquely accounted for 10% of variance in learning performance, metacognition uniquely accounted for 18% of variance, while both predictors had 22% of variance in common. Therefore, as a third hypothesis, it is expected that a composite score of logfile indicators is moderately correlated to intelligence, and that this composite score has a unique predictive value for learning performance on top of intelligence.

2. Method

2.1. Participants

In the Netherlands, three separate levels exist in secondary education with a vocational level, a general level, and a pre-academic level. Participants in this study were 52 secondary-school students from the second year of pre-academic level, 20 boys and 32 girls, all about 13 years of age (mean age was 13:2 years). Informed consent was given by their parents.

2.2. Intellectual ability

The Groninger Intelligence test for secondary education (GIVO; Van Dijk & Tellegen, 1994) was administered to all participants, one week prior to the presentation of the learning task. From this Dutch standardized intelligence test three subtests were included: Number Series, Verbal Analogies, and Unfolding Figures. These subtests represent inductive and deductive reasoning abilities, both verbal and numerical, and visuospatial ability (Carroll, 1993). Norm scores for the subtests were used to estimate overall IQ.

2.3. Learning environment

A computerized learning-by-discovery task was adapted from Veenman et al. (2004) and implemented in Authorware, an authoring environment for PC. The Otter task was originally developed by Wilhelm, Beishuizen, and Van Rijn (2001), but task difficulty was increased in order to meet with the higher level of pre-academic secondary-school students in the present study. The Otter-task required participants to experiment with five independent variables in order to discover their (combined) effects on the growth of the otter population. The five variables were habitat (either one big area, or separated small areas), environmental pollution (either natural clean, or polluted), public entrance (either no entrance, or free entrance), setting out new otter couples (none, one couple, or more couples), and feeding fish in wintertime (yes or no). Independent variables could have no effect on the otter population (public entrance), a main effect (habitat; pollution), and interact with another variable (habitat x setting out otter couples; pollution x feeding fish). For each experiment, participants could choose a value for the five variables by clicking on the pictograms on the left, and then order the computer to calculate the growth of the otter population (see Fig. 1). Results of experiments done were transferred to a storehouse on the right and participants could scroll up and down the storehouse to consult earlier results. After a minimum of 15 experiments, an exit button became available which allowed participants to leave the learning environment, although they were free to continue with further experimentation. This minimum number of experiments was set to ensure that enough data would be available for assessing all indicators of metacognitive skills from the logfiles (see below).

2.4. Metacognitive skills

All actions were logged in a text file, which logfile was automatically scored on various metacognition measures by the computer (see Table 1). The total number of experiments performed by the

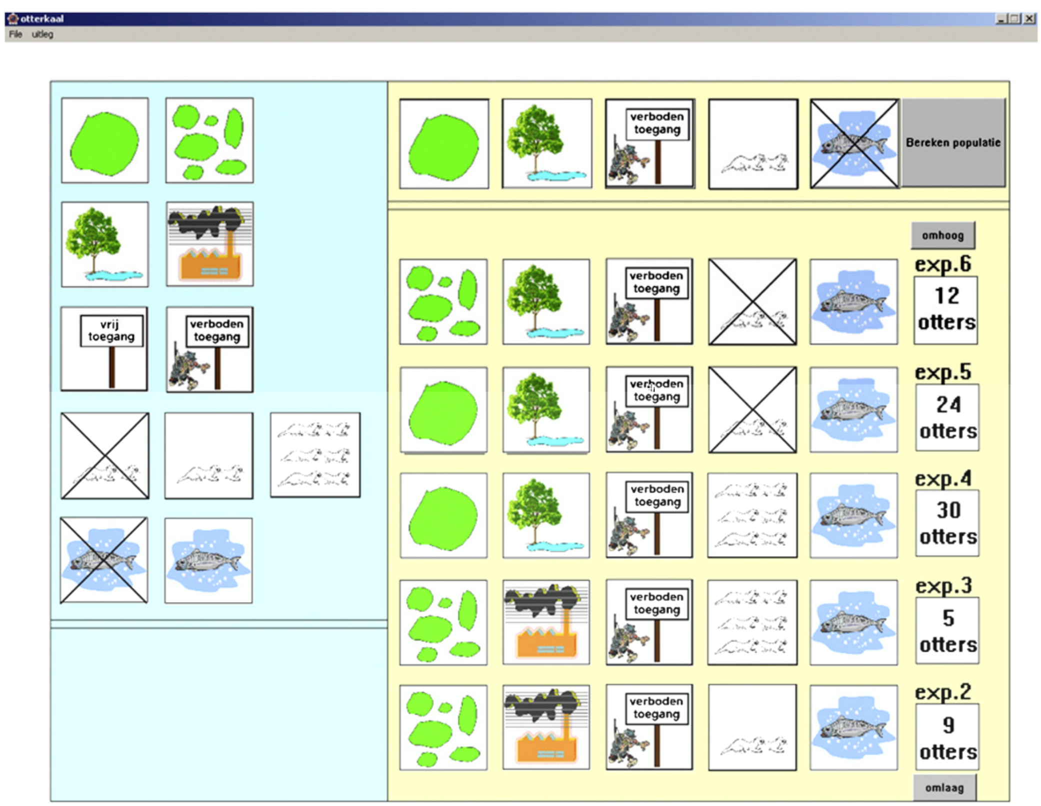

Fig. 1. Interface of the Otter task.

Table 1
Labels and descriptions for logfile and trace measures.

Label	Description
Logfiles measures	
Number.exp	Total number of experiments performed.
Thinktime	Time in sec. elapsed between receiving the outcome of a former experiment and the first move in the next experiment, accumulated over the experiments.
Scrolldown	Frequency of scrolling down to earlier experiments.
Scrollup	Frequency of scrolling back to later experiments.
Votat.pos	Number of transitions between experiments in which only one variable is altered.
Votat.neg	Mean number of variables changed in transition between experiments, minus one.
Unique.exp	Number of unique experiments performed out of 48 possible unique experiments.
Variation	Mean proportion of investigating the different levels of the five independent variables.
Trace measures	
Systematic	Score (0–4), rated from traces of learner activities, indicating the extent to which a participant revealed (idiosyncratic) systematical patterns of experimentation.
Complete	Score (0–4), rated from traces of learner activities, indicating to which extent all variables were equally varied over experiments, both singularly as well as in combination with other variables.

participant (Number.exp) was recorded as a positive indicator of metacognitive skillfulness. The higher the number of experiments, the more complete experimentation was expected to be. Evaluation and elaboration of outcomes may incite participants to do additional experiments. Indeed, earlier validation with think-aloud measures in a pilot study (Veenman, 2012) has shown that Number.exp is strongly related to orientation, evaluation, and elaboration activities (with $.74 \leq r \leq .86$). Secondly, the time elapsed between receiving the outcome of a former experiment and taking action in the next experiment (Thinktime) was registered in seconds as a positive indicator of outcome evaluation, reorientation, and planning of the next experiment. In an individual think-aloud session with the Otter task that was broadcasted on Dutch television (NTR, 2012), the participant verbally and non-verbally reacted to outcomes with surprise or confirmation, compared experiments by pointing at the screen, made notes of results and conclusions drawn, generated new ideas and hypotheses, and carefully considered the next step to be taken. Thus, participants are not idle during Thinktime, otherwise they would simply move on to the next experiment or quit the program. Thinktime was accumulated over the experiments. Thirdly, the frequency of scrolling down to earlier experiments (Scrolldown) and scrolling back up (Scrollup) were assessed as positive indicators of metacognitive skillfulness. Scrolling proved to be a valid indicator of a participant's intention to check earlier experimental configurations or to relate the outcomes of experiments (Veenman et al., 2004). In the pilot study (Veenman, 2012), Scrolling appeared to correlate .78 with think-aloud measures of evaluation. To avoid skewness in distribution of scores due to outliers, a square-root transformation was applied to Scrolldown and Scrollup. A fourth logfile indicator concerned the number of variables changed between two subsequent experiments. Varying only one variable at a time in transitions between experiments (VOTAT; Chen & Klahr, 1999; Tschirgi, 1980) proved to be a positive indicator of metacognitive skillfulness (Veenman et al., 2004), and in particular of planning and elaboration ($r \geq .91$; Veenman, 2012). Thus, the number of transitions between experiments with the value of only one variable being altered (Votat.pos) was registered as a positive indicator of metacognitive skillfulness. The mean number of variables changed per experiment minus one (Votat.neg) was also obtained for each participant, however, as a negative indicator of metacognitive skillfulness (Veenman et al., 2004). Varying more than one variable at a time represents poor planning (Veenman, Elshout, & Meijer, 1997) and a lack of experimental control (Schauble, Glaser, Raghavan, & Reiner, 1991; Shute, Glaser, & Raghavan, 1989). If more than one variable is changed between experiments, one cannot attribute the differences in outcomes to any of the changes in variables made. As a fifth indicator, the number of unique experiments performed (Unique.exp) out of 48 possible unique experiments was taken as a positive indicator for coverage of the experiment space (Klahr, Fay, & Dunbar, 1993). In the pilot study of Veenman (2012), Unique.exp correlated with all think-aloud measures of metacognition ($.78 \leq r \leq .95$). This means that metacognitively proficient learners tend to keep track of potential experiments to be done and plan their experiments accordingly. Finally, as a sixth indicator, the mean proportion of investigating the different levels of the five independent variables (Variation) was calculated over the experiments performed. Equally investigating the two or three levels of each independent variable may be regarded as a positive indicator for systematical variable variation. Even for the independent variable of public entrance without any effect on the otter population, substantial investigation of this variable is needed to exclude possible interactions with other independent variables. While based on proportions, Variation is a measure of planned and controlled behavior, independent of the number of experiments done. As said, some of these indicators (Number.exp, Scrolling, Votat, and Unique.exp) have been validated against think-aloud measures of secondary-school students from a different sample in earlier study, while a composite logfile score correlated .96 with an overall think-aloud measure of metacognition (Veenman, 2012). Unfortunately, Thinktime and Variation were not assessed in that study.

All scores on the eight measures were standardized into z-scores and the sign of Votat.neg was inverted in order to make it a positive indicator. To avoid overweighing Scrolling and Votat, mean z-scores were calculated for Scrolling over Scrolldown and Scrollup, and for Votat over Votat.pos and Votat.neg. Finally, mean z-scores were calculated over the six indicators as an overall measure of metacognitive skillfulness (MSlogfile, with Cronbach's alpha = .78).

Full traces of all learner activities were also logged and available for further inspection. These traces, which included all activities of participants on the Otter task (see the Appendix A) but not their scores on the logfile indicators, were additionally judged on two indicators of metacognition by a 'blind' judge who was unaware of the participants' scores on the logfile indicators. The first trace indicator, systematical variation (Systematic), addressed the extent to which a participant revealed systematical patterns of experimentation, even if patterns were idiosyncratic. For instance, some participants could take one specific configuration of variables as the point of departure for a series of experiments (which would subsequently violate the Votat principle; for an example, see the Appendix A). Additionally, systematical variation took into account clusters of experiments that investigated all combinations of two or three variables, while keeping the other variables constant. The second trace indicator was completeness of experimentation. Completeness (Complete) was judged on the extent to which all variables were equally varied over the experiments, both singularly as well as in combination with other variables. Both indicators were representative of controlled behavior patterns that were hard to capture with automated tallies of activities from logfiles (Veenman, 2012). Trace indicators were rated on a 5-point scale, ranging from 0 (poor) to 4 (highly systematical/complete). Scores on both indicators were standardized and a mean score (MStrace) was calculated over the two z-scores. Additionally, another 'blind' judge rated 10 randomly chosen trace protocols. Ratings of both judges fairly converged with $r = .95$ ($p < .01$) for Systematic and $r = .93$ ($p < .01$) for Complete, although the second judge tended to be slightly austere in Complete ratings.

2.5. Learning performance

In order to avoid confounding assessments of metacognition and learning performance, separate measures of learning performance were obtained afterwards, with a Multiple-choice (MC) posttest of 17 items and two Open-ended questions. For instance, an MC-item was: "In an environmental restructuring plan, separated areas become interconnected. Which other intervention from the restructuring plan would have a positive effect on the otter population as a result of that enlarged area? a) fighting pollution, b) limiting public access, c) setting out new otter couples, d) stop feeding fish." Correct answer is c. Correct MC-items could yield one point each. An Open-ended question was: "Describe exactly what the impact of the five variables is on the otter population. Also indicate which variables interact and how they interact when affecting the otter population." Open-ended questions were scored according to a strict coding scheme for correctness and completeness of the answer on a 0–15 point scale for the first question and a 0–12 point scale for the second question. Thus, the maximum total score was 44. Cronbach's alpha was .77 for total scores of learning performance.

2.6. Procedure

First the GIVO test was administered during class. Next, participants individually worked on a computer in the classroom with the Otter task. Paper and pen were provided for making notes if participants preferred to do so. After finishing the Otter task, notes were taken away and the posttest with MC-questions was handed out. When completed, the MC posttest was replaced by the open-ended questions.

3. Results

3.1. Descriptives

First, descriptives are given for IQ, for the raw scores on the eight metacognition measures obtained from the logfiles, for the raw scores on the two measures from the trace ratings, and for Learning performance (see Table 2). All measures were tested for differences between male and female participants, but no differences were found (all $p>.05$). There is one remarkable outlier of 75.67 in the IQ scores. However, IQ scores are normally distributed around the mean (with a skewness of .02). Removing the outlier would still yield a range in IQ scores from 85 to 129. We will return to this issue later.

3.2. Analyses of logfile and trace variables

Correlations among z-scores on logfile and trace measures are depicted in Table 3 (please note that the sign of Votat-neg has been inverted). All logfile measures are significantly correlated to the Systematic trace measure, while most logfile-variables are also significantly correlated to Complete (with the exclusion of Votat.neg and Variation).

Table 2
Descriptives of IQ, raw metacognition scores, raw trace scores, and Learning performance.

Variable	Mean	SD	Minimum	Maximum
IQ	103.96	11.22	75.67	129.33
Logfile measures				
Number.exp	19.98	7.88	15	52
Thinktime	265.90	212.04	61	958
Scrolldown	7.23	7.39	0	27
Scrollup	5.13	6.23	0	23
Votat.pos	4.12	4.21	0	21
Votat.neg	26.38	13.53	5	84
Unique.exp	15.44	4.24	10	33
Variation	.78	.11	.47	.95
Trace measures:				
Systematic	1.38	1.24	0	4
Complete	1.25	1.05	0	4
Learning performance	8.44	4.67	2	25

Table 3
Correlations among standardized logfile indicators (1–6) and trace measures (7–8) variables, and IQ (9).

Variable	1	2	3	4	5	6	7	8
1. Number.exp								
2. Thinktime	.37†							
3. Scroll	.48†	.27						
4. Votat	.43†	.34*	.18					
5. Unique.exp	.92†	.28*	.51†	.25				
6. Variation	.00	.12	−.17	.56†	−.20			
7. Systematic	.50†	.49†	.39†	.88†	.33*	.45†		
8. Complete	.85†	.38†	.58†	.30*	.89†	−.18	.41†	
9. IQ	.10	.16	.33*	.17	.07	−.05	.22	.03

* $p<.05$.
† $p<.01$.

Principal component analysis was performed on the six indicators of the logfile data. Results of this PCA are summarized in Table 4. Apart from Variation, all indicators show positive loadings on the first component. The second component is dominated by high loadings of Variation and Votat. Component scores were calculated and correlated with trace measures of Systematic and Complete. Scores on the first component correlated .67 ($p<.01$) with Systematic and .86 ($p<.01$) with Complete. Scores on the second component correlated .55 ($p<.01$) with Systematic and −.27 with Complete. Apparently, the first, most general component covered both systematical variation and completeness of experimentation, whereas the second component distinguished an additional aspect of systematical variation. Although Varimax rotation yielded two orthogonal components, rotation did not substantially alter the results (see Table 4). Scores on the first component were still correlated to both trace measures of Systematic ($r=.45$, $p<.01$) and Complete ($r=.90$, $p<.01$), while scores on the second component only correlated with the Systematic trace measure ($r=.74$, $p<.01$). Scores on the second component correlated .04 with Complete.

As an indicator of convergent validity, the mean logfile scores were correlated to the mean trace scores. This correlation between MSlogfile and MStrace is .92 ($p<.01$), which implies that both measures have 84% of variance in common. Removing any of the six logfile indicators from MSlogfile would reduce this correlation between MSlogfile and MStrace.

3.3. Correlations among IQ, metacognition measures, and Learning performance

Table 5 (left side) shows the correlations between IQ, MSlogfile or MStrace, and Learning performance. While both correlations of MSlogfile and MStrace with Learning performance are substantial, it appears that all correlations with IQ are rather low (in line with the correlations that were obtained between IQ on the one hand, and separate logfile and trace measures on the other; see Table 3). Earlier, an outlier of 75.67 has been detected in the IQ scores, which might distort the correlational pattern. Therefore, correlational analyses were

Table 4
Results PCA on logfile indicators.

	Unrotated		Varimax rotation	
	Component 1	Component 2	Component 1	Component 2
Eigen value	2.69	1.59		
Variance proportion	44.89	26.44		
Component loadings				
Number.exp	.93	−.10	.90	.22
Thinktime	.57	.23	.46	.41
Scroll	.67	−.30	.73	−.06
Votat	.56	.70	.29	.85
Unique.exp	.87	−.32	.93	−.01
Variation	.06	.91	−.26	.88

Table 5
Correlations among IQ, metacognition, and Learning performance (left side), and semipartial correlations (right side).

	IQ	MSlogfile	Semi-IQ	Semi-MSlogfile
Learn.perf	.17	.64†	.02	.62†
MSlogfile	.24			
	IQ	MStrace	Semi-IQ	Semi-MStrace
Learn.perf	.17	.52†	.10	.50†
MStrace	.15			

Note: Semi-IQ means semipartial correlation with either MSlogfile or MStrace partialled out; Semi-MSlogfile or Semi-MStrace refers to the semipartial correlation with IQ partialled out. Learn.perf means Learning performance.

repeated on the data of 51 participants, excluding the participant with this extremely low IQ score. Results are virtually identical (with only minor changes in correlations between IQ and Learning performance to .18, and between IQ and MStrace to .14). Apparently, the outlier was not disruptive and the data of the low-IQ participant were inserted again.

Next, the semipartial correlation (cf. Nunnally & Bernstein, 1994) with MSlogfile partialled from the correlation between IQ and Learning performance was calculated. This semipartial correlation of .02 (see Table 5, right side) reflects the unique predictive value of IQ for learning performance, independent of MSlogfile. Also, the semipartial correlation with IQ partialled from the correlation between MSlogfile en Learning performance was calculated (.62, see Table 5, right side). This semipartial correlation reflects the unique predictive value of MSlogfile for learning performance, independent of IQ. Using regression-analytic techniques for the partitioning of variance (Pedhazur, 1982), the variance accounted for in learning performance was distributed over the unique and shared sources of IQ and MSlogfile. Firstly, the squared multiple correlation of IQ and MSlogfile for predicting Learning performance was calculated from the squared correlation between IQ and Learning performance and the squared semipartial correlation of MSlogfile with Learning performance ($R^2 = (.17)^2 + (.62)^2 = .415$). The proportion of variance shared by both predictors was calculated by subtracting both squared semipartial correlations from the squared multiple correlation (shared $r^2 = .415 - (.02)^2 - (.62)^2 = .030$). Consequently, it was estimated that IQ uniquely accounted for .0% of variance in Learning Performance, MSlogfile uniquely accounted for 38.5% of variance, while both predictors had 3.0% of variance in common. In total both predictors accounted for 41.5% of variance in Learning performance. The same procedure was followed for partitioning the variance of IQ and MStrace in Learning performance. It was estimated that IQ uniquely accounted for .9% of variance in Learning Performance, MStrace uniquely accounted for 25.3% of variance, while both predictors had 2.1% of variance in common. In total both predictors accounted for 28.3% of variance in Learning performance.

4. Discussion

The first hypothesis concerned the convergent validity of logfile indicators with human judgments of traced events. Results show that separate logfile indicators were substantially correlated to either Systematic or Complete, both judged from the traced events. Overall, logfile indicators had 84% of variance in common with judgments of traced events. Together with the convergent validity of logfile indicators obtained earlier with think-aloud measures (Veenman, 2012), results endorse the notion that metacognitive skills can be adequately assessed on-line with computer logfiles from the Otter task.

Although the indicators of Number.exp and Unique.exp are highly correlated (see Table 3), removal of either indicator from MSlogfile would slightly reduce the correlation with MStrace. Including both logfile indicators not only enhances reliability of MSlogfile scores, both indicators also tap the different aspects of sheer quantity vs. scope of experimentation. Apart from commonalities in metacognitive skills shared by both logfile indicators, Unique.exp has a more profound planning component, relative to Number.exp (Veenman, 2012). In the same vein, other logfile indicators make their unique contribution to the assessment of metacognitive skills with the Otter task.

Principal component analysis of the six logfile indicators, along with the correlations of separate logfile indicators with trace measures show that logfile indicators in the Otter task capture two components. One general component represents both systematical and complete experimentation. This component may represent the planning, coordination, and evaluation of performing sufficient experiments. The other component seems to express a more specific tendency to systematically vary levels of the independent variables. In order to interpret both components more clearly, however, it would be helpful to identify the metacognitive nature of Thinktime and Variation with thinking aloud in another multi-method study.

The principle of varying one variable at the time between experiments (Votat) may be regarded as a limited approach to systematical varying independent variables across experiments. Dynamic systematical patterns of experimentation are not easily captured by relatively rigid logfile measures (Veenman, 2012). Therefore, judgmental ratings of trace protocols were used to assess various more complex patterns of systematical experimentation (e.g., see the Appendix A). Logfile measures, however, adequately appear to cover these trace assessments. Votat correlated .88 with the Systematic trace measure, indicating that varying a limited number of variables is essential to systematically investigating independent variables. Nevertheless, it would be useful to base logfile measures of Votat not only on variation between the last experiment and the current experiment, but also include a similar comparison between the before last and current experiment. It would accommodate proficient participants who take one particular experimental setting as a starting point for systematical variation.

The second hypothesis pertains to the external validity of logfile indicators as predictors of learning performance. Logfile indicators accounted for 41% of variance in learning performance on the otter task, which is even more than the 27% of variance in learning performance accounted for by judgments of traced events. Moreover, the amount of variance accounted for by logfile indicators is close to the 40% that was obtained by Veenman (2008) in an overview of studies with think-aloud and observational measures of metacognitive skillfulness. These results are in favor of the second hypothesis.

The third hypothesis about the external validity of logfiles indicators in relation to intelligence was only partially confirmed. Although logfile indicators of metacognitive skills provided for a substantial, unique source of variance in learning performance on top of intelligence, all correlations with intelligence were invariably low. These low correlations of intelligence were not due to an outlier in the intelligence scores, as removal of the outlier did not notably change the overall results. Also, low correlations cannot be attributed to a severe restriction of range in the intelligence scores. Despite the selection of pre-academic secondary students, IQ scores were normally distributed with a considerable range (see also the standard deviation in Table 2). An alternative explanation can be found in the novelty and complexity of the learning task. Elshout (1987) and Raaheim (1988) have postulated that the impact of intelligence on (learning) performance depends on the novelty and complexity of the task according to an inverted U-shaped curve. When a task is very familiar or very easy to an individual, intelligence has little impact on routine performance. On the other side, when a task is extremely unfamiliar or complex to an individual, intelligence has little to offer for advancing task performance. Despite one's intelligence, one cannot see the wood before the trees. At an intermediate level of task complexity, however, one can optimally profit from one's intellectual resources. This inverted U-curve for the impact of intelligence on learning performance has been corroborated by empirical studies (Prins, Veenman, & Elshout, 2006; Veenman & Elshout, 1999). Moreover, these studies have shown that the impact of metacognitive skillfulness on learning performance is not impaired by high levels of task complexity. In fact, with an extremely unfamiliar and complex task, the novice learner can only rely on metacognitive skills for initiating the learning process. The relatively low mean score on Learning performance in Table 2 shows that the Otter task was rather difficult to second-year pre-academic students in the present study. Consequently, high task complexity may have suppressed the correlations with intelligence, while leaving the correlations between metacognitive skills and learning performance unaffected. This explanation, however, remains to be verified in a study where the Otter task with an alleviated level of complexity is presented to second-year pre-academic students. Complexity of the task should be appropriated to the students' age and knowledge level in order to restore correlations with intelligence at a moderate level (De Jong & Van Joolingen, 1998; Prins et al., 2006). Calibration of task complexity is tough, but necessary to adequately address the full breadth of students' cognitive capacities.

What can be learned from this study with respect to the assessment of metacognitive skills through logfile analysis in general? As a first step in logfile analysis, researchers must select potentially relevant logfile indicators based on a rational analysis of the task at hand. It was argued before, however, that logfiles indicators only access the overt behavior on the object level in Nelson's model. Moreover, the metacognitive nature of activities on the object level needs to be inferred by the researcher without knowing the content of information flows between object and meta level. This makes the design of potential logfile indicators vulnerable to subjective interpretations (Veenman, 2012; Winne, 2010). For instance, the number of double experiments was initially included as a negative indicator of metacognitive skills in the logfiles of the Otter task. Repeating the same experiment over and over again was regarded as redundant. The number of double experiments, however, turned out to be slightly positively, rather than negatively correlated with other logfile indicators and with think-aloud measures of metacognitive skills (Veenman, 2012). Inspection of traced events revealed that some idle participants (but not all) used double experiments to avoid scrolling back to earlier experiments. Apparently, the participant's behavior may not always live up to the researcher's inference-based expectations. Thus, as a second step, logfile indicators should be validated against other online measures of metacognitive skills. Such convergent validation is required for all new tasks with new configurations of task-dependent logfile indicators (Veenman, 2007). It may also sift out ambiguous indicators. Finally, as a third step, one should verify whether logfile indicators uphold external validity by scrutinizing whether their relations with other, external variables fit in with theories of metacognition. A poor fit may indicate that the assessment of metacognitive skills through logfiles is inadequate, or at least that the assessment does not completely cover the construct of metacognitive skills.

In the present study, convergent validity was obtained for logfile indicators with judgments of traced events. Moreover, external validity of logfile indicators was sufficiently established in relation to learning performance. As to the low correlations between logfile indicators and intelligence, task conditions under which logfile indicators are obtained should be reconsidered. Apparently, the on-line assessment of metacognitive skills through the analysis of computer logfiles may entail an unobtrusive method for individual diagnosis of metacognitive skillfulness. A pressing question, then, is to what extent such a diagnosis can be generalized over tasks and domains. Although some researchers have argued that metacognitive skills are domain-specific by nature (Kelemen, Frost, & Weaver, 2000), there is evidence that metacognitive skills increasingly become domain surpassing with age (Veenman, 2011a). In a longitudinal design, Van der Stel and Veenman (2010) found that different measures for metacognitive skills of the same participants, obtained separately during a text-studying task in history and a problem-solving task in math, revealed a high common variance at the age of 12 years, and even more profoundly at the age of 13 years. At the age of 14 years, metacognitive skills were entirely general by nature (Van der Stel, 2011). In a cross-sectional design, Veenman and Spaans (2005) compared the metacognitive skills of 12 and 14 year old participants, who solved math problems and performed a biology task similar to the Otter task in the present study. At the age of 12 years, metacognition for math problem solving only correlated moderately with metacognition for the biology task. At the age of 14 years, however, both measures of metacognitive skills were highly correlated. Further evidence for the generality of metacognitive skills across tasks and domains has been obtained by Schraw, Dunkle, Bendixen, and Roedel (1995), Schraw and Nietfeld (1998), Veenman et al. (1997, 2004), and Veenman and Verheij (2003). It seems that learners develop a personal repertoire of metacognitive skills that is called upon whenever a learner is faced with a new task (Veenman, 2011a). In terms of Nelson's model (1996; Nelson & Narens, 1990), one can postulate that the meta level and both information flows represent general metacognitive skills. At the object level, however, general metacognitive skills must be contextualized and implemented as domain-specific activity. In the same vein, Glaser, Schauble, Raghavan, and Zeitz (1992) argued that domain-specific cognitive activities do not preclude the existence of general metacognitive strategies, such as planning and evaluation. “However, these general skills take on specific value as they are differentially useful in varying contexts” (Glaser et al., 1992, p. 370). This notion even more so stresses the relevance of validating logfile indicators with other on-line measures, preferably measures that additionally access the information flows between object and meta level. In conclusion, the individual diagnostic value of metacognitive-skill assessment in a particular task or domain setting becomes increasingly representative of a learner's general metacognitive skillfulness with age. An agenda for future research, however, is establishing the extent to which a specific instrument for assessing metacognitive skills, such as logfile analysis of the Otter task, adequately predicts GPA or learning performance in a wide range of school disciplines. For instance, we are now investigating whether logfile analysis of a similar inductive-learning task, the Aging task (Veenman et al., 2004), is suitable for selecting students in the admission to Medical College. The relation between logfile measures and prior GPA in secondary education will be examined, as well as the predictive validity of logfile measure for future study results in Medical College. This is a quest for external validity, indeed.

Appendix A

Fragment of trace data:

…

scroll-down, scroll-down, scroll-up, scroll-up, scroll-down, scroll-down, scroll-down, big-area, clean-environment, no-access, one-couples, no-fish,

10:50:17 exp.9: Thinktime = 121, big-area, clean-environment, no-access, one-couple, no-fish, growth = 26

big-area, clean-environment, correction polluted-environment, free-access, more-couples, feed-fish,

10:50:32 exp.10: Thinktime = 130, big-area, polluted-environment, free-access, more-couples, feed-fish, growth = 23

scroll-down, scroll-down, scroll-down, scroll-down, scroll-down, big-area, clean-environment, no-access, more-couples, no-fish,

10:51:02 exp.11: Thinktime = 152, big-area, clean-environment, no-access, more-couples, no-fish, growth = 30

…

Comment: As can been seen, four variables are varied between experiment 9 and 10, and three variables are varied between experiment 10 and 11. This would imply that no points are added to Votat.pos and 3 + 2 points are added to raw Votat.neg in the logfile scores. Comparison of experiment 9 with 11, however, reveals only one variable change, while scrolling is used to consult earlier results. Consequently, this would contribute to a higher score on Systematic in the scoring of trace data.

References

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.

Beijk, J. (1977). Convergerend operationalisme: Een dwingende strategie voor de gedragswetenschappen [Convergent operationalism: A compelling strategy for behavioral sciences]. *Nederlands Tijdschrift voor de Psychologie*, *32*, 173–185.

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

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098–1120.

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

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

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*, 179–201.

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.

Elshout, J. J. (1987). Problem solving and education. In E. de Corte, H. Lodewijks, R. Parmentier, & P. Span (Eds.), *Learning and instruction* (pp. 259–273). Oxford: Pergamon Books (Leuven: University Press).

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

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

Glaser, R., Schauble, L., Raghavan, K., & Zeitz, C. (1992). Scientific reasoning across different domains. In E. de Corte, M. C. Linn, H. Mandl, & L. Verschaffel (Eds.), *Computer-based learning environments and problem solving. NATO ASI series F, Vol. 84*. (pp. 345–371). Heidelberg: Springer Verlag.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, *2*, 107–124.

Kelemen, W. L., Frost, P. J., & Weaver, C. A., III (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*, 92–107.

Kinnunen, R., & Vauras, M. (1995). Comprehension monitoring and the level of comprehension in high- and low-achieving primary school childrens' reading. *Learning and Instruction*, *5*, 143–165.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111–146.

Kunz, G. C., Drewniak, U., & Schott, F. (1992). On-line and off-line assessment of self-regulation in learning from instructional text. *Learning and Instruction*, *2*, 287–301.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*, 102–116.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The Psychology of Learning and Motivation*, *26*. (pp. 125–173).

NTR (2012). Pavlov (about intelligence). http://www.wetenschap24.nl/programmas/pavlov/tv/afleveringen/seizoen-4/Dolores-Leeuwin-intelligentie.html

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

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston.

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

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale, NJ: Erlbaum.

Prins, F. J., Veenman, M. V. J., & Elshout, J. J. (2006). The impact of intellectual ability and metacognition on learning: New support for the threshold of problematicity theory. *Learning and Instruction*, *16*, 374–387.

Raaheim, K. (1988). Intelligence and task novelty. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*, *Vol. 4*. (pp. 73–97)Hillsdale: Erlbaum.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *The Journal of the Learning Sciences*, *1*, 201–238.

Schraw, G. (2007). The use of computer-based environments for understanding and improving self-regulation. *Metacognition and Learning*, *2*, 169–176.

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

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology*, *87*, 433–444.

Schraw, G., & Nietfeld, J. (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology*, *90*, 236–248.

Shute, V. J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 279–326). New York: Freeman.

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology*, *76*, 347–376.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*, 1–10.

Van der Stel, M. (2011). *Development of metacognitive skills in young adolescents*. *A bumpy ride to the high road*. Unpublished Phd thesis. Leiden: Leiden University.

Van der Stel, M., & Veenman, M. V. J. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences*, *20*, 220–224.

Van Dijk, H., & Tellegen, P. J. (1994). *GIVO. Groninger Intelligentietest Voortgezet Onderwijs.* Lisse: Swets & Zeitlinger.

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

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

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

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

Veenman, M. V. J. (2012). Assessing metacognitive skills in computerized learning environments. In R. Azevedo, & V. Aleven (Eds.), *International handbook of metacognition and learning technologies.* New York/Berlin: Springer.

Veenman, M. V. J., & Elshout, J. J. (1995). Differential effects of instructional support on learning in simulation environments. *Instructional Science*, *22*, 363–383.

Veenman, M. V. J., & Elshout, J. J. (1999). Changes in the relation between cognitive and metacognitive skills during the acquisition of expertise. *European Journal of Psychology of Education*, *14*, 509–523.

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., Elshout, J. J., & Meijer, J. (1997). The generality vs. domain-specificity of metacognitive skills in novice learning across domains. *Learning and Instruction*, *7*, 187–209.

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., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*, 159–176.

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

Veenman, M. V. J., & Verheij, J. (2003). Identifying technical students at risk: Relating general versus specific metacognitive skills to study success. *Learning and Individual Differences*, *13*, 259–272.

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

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

Whitebread, D., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., et al. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning*, *4*, 63–85.

Wilhelm, P., Beishuizen, J. J., & Van Rijn, H. (2001). Studying inquiry learning with FILE. *Computers in Human Behavior*, *21*, 933–943.

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

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, *45*, 267–276.

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

Winters, F. I., Greene, J. A., & Costich, C. M. (2008). Self-regulation of learning with computer-based learning environments: A critical analysis. *Educational Psychology Review*, *20*, 429–444.

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