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 ICLO, Leiden University, The Netherlands

A R T I C L E   I N F O
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 (CBL)
Logfiles

A B S T R A C T
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.

© 2013 Elsevier Inc. All rights reserved.

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 recaptulation 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 SRJS 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 are needed 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...
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 (Beij, 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 et al. (2004). 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. Erickson & 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 winter-time (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 logfile (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.
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 metacognitive skillfulness (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.

### 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.

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

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number.exp</td>
<td>.37</td>
<td>.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Thinktime</td>
<td>.43</td>
<td>.34</td>
<td>.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Scroll</td>
<td>.48</td>
<td>.28</td>
<td>.51</td>
<td>.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Votat.pos</td>
<td>.00</td>
<td>.12</td>
<td>.57</td>
<td>.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Unique.exp</td>
<td>.85</td>
<td>.30</td>
<td>.80</td>
<td>.18</td>
<td>.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Variation</td>
<td>.10</td>
<td>.16</td>
<td>.33</td>
<td>.17</td>
<td>.07</td>
<td>.05</td>
<td>.22</td>
<td>.03</td>
</tr>
</tbody>
</table>

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 = .54, 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 5
Correlations among IQ, metacognition, and Learning performance (left side), and partial correlations (right side).

<table>
<thead>
<tr>
<th></th>
<th>IQ</th>
<th>MSlog</th>
<th>Semi-IQ</th>
<th>Semi-MSlog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learnperf</td>
<td>.17</td>
<td>.64†</td>
<td>.02</td>
<td>.62†</td>
</tr>
<tr>
<td>MSlog</td>
<td>.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MStrace</td>
<td>.26</td>
<td>.52†</td>
<td>.10</td>
<td>.50†</td>
</tr>
<tr>
<td>Learnperf</td>
<td>.17</td>
<td>.52†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MStrace</td>
<td>.15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

The principle of varying one variable at the time between experiments (Votat) may be regarded as a limited approach to systematically varying independent variables across experiments. Dynamic systematic 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 systematic 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 experience 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 systematic 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 Raabeim (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. 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 unobstrusive 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 Niefeld (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 & Nares, 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, 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, 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

...
References


