Metacognition Learning (2011) 6:205–211
DOI 10.1007/s11409-011-9080-x

Alternative assessment of strategy use with self-report instruments: a discussion

Marcel V. J. Veenman

Received: 9 June 2011 / Accepted: 20 June 2011 /
Published online: 29 June 2011

Strategy use is a very broad term for controlled and consciously applied procedural knowledge, in contrast with skills that are automated to a more or lesser extent (Gagné et al. 1993). Both terms, however, are used interchangeably when referring to processes that direct and support problem solving and learning. There is a gray area in between controlled strategy use and the automated performance of skills (Veenman 2011). For instance, monitoring processes may be consciously applied during reading, but often run in the background until an error or anomaly is detected. The latter is the case with more experienced readers. In this special issue, many of the strategic processes described operate in this gray area. For instance, confidence judgments or inference processes are neither fully controlled, nor fully automated. Consequently, learners may not be fully aware of ongoing processes, which may affect the verbalization of these processes in self-reports.

Strategic processes may be cognitive or metacognitive by nature. According to Nelson (1996), cognitive and metacognitive processes operate at different levels of the cognitive system that represent different functions. Higher-order metacognitive processes monitor and regulate lower-order cognitive processes that, in turn, shape behavior. Thus, drawing inferences is a cognitive activity, but the self-induced decision to initiate such activity is a metacognitive one (Veenman 2011). The authors of this special issue diverge in the perspective that they have chosen for studying strategic processes in reading. Schellings (this issue) discusses the role and assessment of learning strategies, which encompass both cognitive and metacognitive strategies, but her empirical work is entirely based on assessment methods for metacognition. Bräten and Strømsø (this issue) avoid classifying the inference processes in their study as either cognitive or metacognitive. Magliano et al. (this issue) clearly state that their RSAT does not provide a direct measure of metacognitive processes, but rather captures (cognitive) information-processing activities. On the other hand, they do acknowledge that metacognitive awareness is prerequisite to adequately

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

making inferences. Cromley and Azevedo (this issue) address both cognitive and metacognitive processes in theory, but they do not specify the nature of processes in their empirical research. Finally, Muis and Winne (this issue) discuss computational models for calibration of confidence judgments as a key metacognitive process in self-regulated learning. In conclusion, not all strategy use in this special issue is clearly defined as being cognitive or metacognitive by nature. Is that a problem? Veenman (2007) argued that the introduction of new assessment methods should be accompanied with validation of those new methods. When validating a new assessment instrument for strategy use against another (existing) assessment method, convergent validity only makes sense if both measurements represent the same construct of strategy use.

All contributions to this special issue reveal inventive designs for new assessment instruments. Schellings (this issue) converted an existing taxonomy for coding metacognitive skills from thinking-aloud protocols into a task-specific questionnaire for text studying. In addition, she used a three-point answer scale in an attempt to reduce scaling effects due to individual variation in reference points of learners (cf. Veenman et al. 2003). Questionnaire data were validated against thinking-aloud protocols of text studying. Bräten and Strømsø (this issue) developed a questionnaire for assessing shallow information accumulation vs. cross-text inferences. Both scales were used to predict within-text and across-text comprehension. Cromley and Azevedo (this issue) prompted learners with multiple-choice questions to enact strategies after reading a text fragment. MC-questions measured adequacy of strategy use, which was correlated to distinct measures of background knowledge, inferences, vocabulary, word reading, and comprehension. Magliano et al. (this issue) confronted learners with indirect questions ('what-are-you-thinking-now?') after reading a particular sentence in RSAT. Answers were analyzed by counting words borrowed from local or distal sentences, as well as new words. Word counts were compared to human judgments of the answers. Winne and Muis (this issue) addressed the *d'* statistic of signal detection theory as an alternative to gamma coefficients for assessing calibration. They compared both calibration measures across three different knowledge domains. The present discussion will critically review these new methods for assessing strategy use, their potentials, and their pitfalls.

Assessment methods: advantages and disadvantages

Generally, off-line methods are distinguished from on-line methods in the assessment of strategy use (Veenman 2005, 2011; Veenman et al. 2006). On-line methods concern measurements taken concurrent to task performance. Thinking aloud, observation, eye-movement registration, and, more recently, logfile registrations of learner activities on the computer are examples of on-line methods used for the assessment of strategy use during reading and text studying. Truly, on-line methods have their limitations. Thinking-aloud protocols may not be complete when learners do not or cannot verbalize all ongoing thoughts (the tip-of-the-iceberg phenomenon). Thinking aloud may be intrusive, especially to very poor readers who have to invest a lot of effort in executing basic reading skills. Observations and logfile registrations may only capture overt behavior, not the thoughts and motives underlying that behavior. Moreover, on-line methods are time-consuming and labor intensive as they need to be individually administered and the raw material needs to be coded according to a coding scheme. Although logfile registration can be done in large groups, the computerized coding system needs to be attuned to every new task and group of learners through validation with other on-line methods (Veenman in press). The major

strength of on-line methods, however, is that actual learner behavior is coded according to externally defined criteria. It rules out error variance due to subjective learner perceptions (Veenman 2011).

Off-line methods refer to questionnaires (e.g., MSLQ, Pintrich and De Groot 1990) and interviews (e.g., Zimmerman and Martinez-Pons 1990) that are administered either before or after task performance. In off-line methods, the learner is addressed with questions about his/her (frequency of) strategy use. Answers to these questions are based on the learner's experiences in the past, even when questions are posed immediately after task performance. The task-specific questionnaire of Schellings (this issue) and the MTSI of Bråten and Strømsø (this issue) are typical examples of retrospective questionnaires. The indirect questioning of Magliano et al. (this issue) also is an off-line assessment, although the questions are interspersed between the sentences to be read. In RSAT, the learner processes a single sentence until s/he presses the "next" button, the sentence disappears, and only then the indirect question is posed.

Off-line questionnaires have the advantage that they can be easily administered to large groups and processed accordingly. There is ample evidence, however, that learner-perceived self-reports of strategies may not correspond to actual learner behavior (for an overview, see Veenman 2005, 2011). A serious validity problem pertains to the off-line nature of the self-reports. While answering questions, learners have to consult their memory in order to reconstruct earlier processes and performance. This reconstruction process might suffer from memory failure and distortions (Ericsson and Simon 1993; Nisbett and Wilson 1977). In retrospective assessment, reconstructive interpretations may be elicited along with, or instead of correct recollections. Learners not only know more than they tell, they sometimes "tell more than [they] can know" (Nisbett and Wilson 1977, p. 247). Distortion due to memory failure is likely to increase with the interval between task performance and retrospective reports. Thus, memory reconstruction will be a serious problem for 'general' questionnaires that are administered disjointed from task performance, like the MSLQ (Veenman 2011). Retrospective self-reports that are gathered immediately after task performance may show similar distortions, albeit to a lesser extent. This memory-reconstruction problem can be mitigated by cutting back the delay between task performance and retrospective questioning to a minimum, such as Magliano et al. (this issue) do. Perhaps, memory problems can also be tempered by making the retrospective questionnaire more task-specific (see Schellings, this issue). Task-specific clues may support the reconstruction process from memory, albeit at the cost of inducing deceptive self-reports (see the prompting effect below). Although these interventions may improve reconstruction from memory, they cannot entirely resolve the memory-reconstruction problem that is inherent to off-line assessments.

Another drawback of off-line methods is that questioning interferes with the spontaneous self-report of strategy use by learners. Retrospective questions may prompt the recall of strategy use that never occurred. Not only may learners be inclined to give social-desirable answers, they are also triggered by questions to label their behavior accordingly. Incorrect labeling occurs, for instance, when the learner's declarative knowledge of strategies is poor. In fact, the learner misreads the question. Moreover, questions may evoke an illusion of familiarity with the strategies that are asked after. Learners may feel overconfident when they take, for instance, their paraphrasing for summarizing. If all learners would equally polish up their self-reports, then all self-reports would equally overestimate strategy use. However, learners who possess declarative knowledge of strategies and who have enacted those strategies likely are more critical of their strategy use than learners who do not have this declarative and procedural knowledge at their disposal. This prompting effect might

have affected answers on the retrospective questionnaires of Schellings (this issue) and Bräten and Strømsø (this issue).

Prompting may affect self-reports in a slightly different way for Magliano et al. (this issue). They argue that their indirect, open questions prompt self-reports of strategy use, "similar to those produced by thinking aloud". As these self-reports are not concurrent verbalizations, however, they are far from spontaneously generated Type-2 thinking-aloud protocols (Ericsson and Simon 1993). According to Type-2 thinking-aloud procedures, learners are instructed to think aloud concurrent to reading *before* embarking on the actual task. In case learners fall silent, they are asked to continue thinking aloud with a standard, neutral reminder. Magliano's indirect questions, posed *after* reading a sentence, are not neutral prompts for concurrently thinking aloud. The prompting effect of indirect questions could be that learners feel that they have to come up with an answer. Thus, they are inclined to *reconstruct* and *summarize* what they have done shortly *before* while reading the sentence. Moreover, they may report additional activities (paraphrasing, summarizing, and inferences) that would not have occurred without prompting questions. Finally, indirect questions run the risk of eliciting Type-3 verbalizations (interpretations and explanations; Ericsson and Simon 1993), rather than plain verbalizations of ongoing thoughts. In conclusion, questions are prompts that may distort retrospective self-reports.

Questioning learners prior or concurrent to task performance may prompt actual strategy use that otherwise would not have occurred spontaneously. Cromley and Azevedo (this issue) asked learners to enact strategies and measured the adequacy of strategies by means of Multiple-Choice questions. Although this is strictly speaking not an off-line method, strategy use is prompted by MC-questions. In order to understand what strategy use is assessed by Cromley and Azevedo, one needs to distinguish between (1) inadequate strategy use due to an availability deficiency, (2) prompted use of available strategies that may otherwise suffer from a production deficiency, and (3) spontaneously produced strategies (Veenman 2011; Veenman et al. 2000). Cromley and Azevedo's method provides valuable information about 1, but it cannot distinguish between 2 and 3. In the same vein, confidence ratings by learners (Winne & Muis, this issue) may assess inadequate calibration, but not the spontaneous use of calibration strategies. The only way to distinguish between all three levels of strategy use is to do assessments first without prompting and only then with prompting strategy use (Veenman et al. 2000, 2005). The point made here is that researchers should be aware of limitations of their assessment methods due to prompting effects.

New assessment methods: validity issues

Three validity issues should be attended to when new methods of assessment are introduced (Veenman 2007). These issues are: (1) internal consistency or reliability of a measure, (2) construct validity, and (3) external or predictive validity (De Groot 1969; Nunnally and Bernstein 1994).

Internal consistency refers to standard reliability measures, such as Cronbach's Alpha, factor analysis to determine dimensional structures, and inter-rater reliabilities of scores that are rated or judged from assessment materials. Apart from Winne and Muis (this issue), all contributions provided indicators of internal consistency. Schellings (this issue), Bräten and Strømsø (this issue), and Cromley and Azevedo (this issue) obtained adequate reliabilities for their questionnaire scales, although reliabilities for the subscales of Schellings were remarkably low. Both Schellings (this issue) and Magliano et al. (this issue) obtained a high

inter-rater reliability for judgments of verbal-aloud responses. Internal consistency is relevant to statistical interpretations, especially when statistically significant effects are lacking, but it does not provide information about *what* is being measured.

Construct validity not only refers to the content validity of meaningfully designed assessment instruments and scoring procedures (see above), but also to convergent validity of a new assessment instrument with other assessments of the same construct. Construct validity is in particular supported by convergent validity across *different* assessment instruments in a multi-method design (Veenman 2007). Schellings (this issue) found a non-significant correlation of .51 between questionnaire and thinking-aloud scores. Her conclusions, however, should be formulated with modesty for two reasons. First, results were based on $N=16$ and should be replicated in a study with more participants, which is acknowledged by Schellings. Secondly, it is doubtful whether a correlation of .51 is significantly 'higher', according to Fisher-*z* ratios (Guilford 1965), than correlations from −.07 to .42 that were found in other studies. Cromley and Azevedo (this issue) obtained correlations ranging from .23 to .74 between strategy-use MC-questions and other measures for background knowledge, inference, vocabulary, and word reading. However, validating a process measure of strategy use with non-process measures, such as vocabulary and word reading, seems like comparing apples and oranges. Convergent validity would likely increase if non-process measures were to be excluded.

Another comparison of assessments pertains to different analyses of data obtained with the *same* instruments, which are referred to as 'within-method' assessments (Veenman 2005). The construct validity rendered by within-method assessments, however, is limited, relative to multi-method assessments with separate data sources (cf. Beijk 1977). Convergence of within-method measures merely contributes to the stability of measures. Metaphorically speaking, within-method assessments are much like comparing the results of a *t*-test with ANOVA, performed on the same data. In Magliano et al. (this issue), both RSAT coding and human coding was carried out on the same answers to indirect questions. The moderate to high correlations of .44 to .70 inform us about the *stability* of measures when replacing human coding with RSAT coding, but they do not inform us about the construct validity of RSAT indirect questioning. In the vein, Winne and Muis (this issue) calculated *G* coefficients and *d'* statistics from the same confidence ratings and knowledge-test scores. Correlations between *G* and *d'*, ranging from .63 to .73, only indicate to what extent both indices overlap. They do not disclose information about the construct validity of calibration assessments. Finally, it should be noted that Bräten and Strømsø (this issue) did not include convergent measures in their study.

External validity means that an assessment instrument should behave as expected by theory in its relations to other variables. Most theories on reading processes claim that better strategy use leads to better reading comprehension (Gagné et al. 1993). Consequently, a new assessment instrument for strategy use should be an adequate predictor of reading comprehension. Unfortunately, Schellings (this issue) did not assess comprehension and, therefore, the external validity of her strategy-use questionnaire remains indeterminate. Bräten and Strømsø (this issue) obtained rather low correlations of strategy use with text comprehension, although correlations were significant for intertextual comprehension. Regression analysis with accumulation and cross-text elaboration as predictors of intertextual comprehension did not account for variance beyond that of individual predictors. Actually, the contribution of the most relevant predictor, cross-text elaboration, to the regression equation was not significant. In the study of Cromley and Azevedo (this issue), on the other hand, correlations showed that strategy use accounted for 30% of variance in comprehension on the average. Magliano et al. (this issue) revealed a mixed

pattern of low to moderate correlations of strategy use with comprehension. The predictive power of strategy use was augmented, however, by including separate strategies as predictors of comprehension in regression analyses. For Winne and Muis (this issue), external validity cannot be determined as the performance measure is incorporated in the calculation of calibration scores. In absence of a separate performance measure, one cannot conclude whether *d'* should be preferred over *G* or not.

Conclusion

New methods for assessing strategy use require thorough examination in order to gain understanding of what these methods precisely are measuring and how successful these methods are in this respect. In this discussion, the assessments of strategy use in the five contributions to this special issue were put under the microscope. It was argued that off-line methods might suffer from memory-reconstruction and prompting problems that produce flawed answers to retrospective questions. Moreover, prompting strategy use prior or concurrent to task performance restricts the scope of assessments to availability deficiencies. Researchers should be aware of these limitations as memory-reconstruction and prompting problems may pose a serious threat to the validity of assessments. Despite all efforts to design new, inventive instruments for the assessment of strategy use, not all methods in this special issue passed the test on the three validity criteria. More validation work is needed to refute the arguments raised here against off-line assessments.

References

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.

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

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

Gagné, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning* (2nd ed.). New York: HarperCollins.

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

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

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: verbal reports on mental processes. *Psychological Review, 84*, 231–259.

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

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.

Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und Metakognition: Implikationen für Forschung und Praxis* (pp. 77–99). Münster: Waxmann.

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

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

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

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

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

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.