On the role of self-assessment and task-selection in self-regulated learning

On the role of self-assessment and task-selection skills in self-regulated learning

Danny Kostons
Open University, The Netherlands

The research reported here was carried out at the

Open Universiteit
www.ou.nl

in the context of the research school

iCO

Interuniversity Center for Educational Research

ISBN: 978-90-79447-44-2

Cover design: Jeroen Berkhout
Printed by Datawyse, Maastricht, the Netherlands

On the role of self-assessment and task-selection skills in self-regulated learning

Proefschrift

ter verkrijging van de graad van doctor
aan de Open Universiteit
op gezag van de rector magnificus
prof. dr. ir F. Mulder
ten overstaan van een door het
College voor promoties ingestelde commissie
in het openbaar te verdedigen

op vrijdag 5 november 2010 te Heerlen
om 16.00 uur precies

door
Danny Daniël Nicolaas Maria Kostons
geboren op 8 februari 1980 te Maastricht

Promotor:
Prof. dr. G.W.C. Paas, Open Universiteit, Erasmus Universiteit Rotterdam

Copromotor:
Dr. T.A.J.M. van Gog, Erasmus Universiteit Rotterdam

Overige leden beoordelingscommissie:
Prof. dr. J. Elen, Katholieke Universiteit Leuven
Prof. dr. K. Scheiter, Knowledge Media Research Center
Prof. dr. J.D.H.M. Vermunt, Universiteit Utrecht
Dr. S.M.M. Loyens, Erasmus Universiteit Rotterdam
Dr. L. Kester, Open Universiteit

Voorwoord

Vier jaar geleden stond ik letterlijk te springen in de achtertuin omdat ik aangenomen was als promovendus. In die vier jaar daarna heb ik nog meerdere malen figuurlijk staan springen, dan wel van vreugde of enthousiasme, dan wel van frustratie of boosheid. Dat hoort er in het leven als promovendus nou eenmaal bij en ik had het niet willen missen. In die vier jaren van AiO-schap zijn er een aantal mensen belangrijk voor me geweest, die ik bij deze zou willen bedanken.

Fred, bedankt voor alle begeleiding, het gevoel dat ik altijd binnen kon lopen en je vermogen zaken te relativeren of voor me in perspectief te zetten. Je inzichten hebben mijn ideeën en plannen soms goed overhoop gegooid, maar dat vertaalde zich altijd in beter onderzoek en artikelen. De 'Paas'-momenten zullen me ook bijblijven; nu alleen hopen dat ik gedurende de promotie niemand 'retarded' ga noemen.

Tamara, dank je wel voor alle feedback en commentaar. Je steun en kritiek hebben me door zowel makkelijke als moeilijke tijden heen geholpen en je betrokkenheid vormde een bron van inspiratie. Je bewijst voor mij dat je heel veel kunt leren van een goed voorbeeld en wie ik nu ben als onderzoeker heb ik voor een groot deel aan jou te danken.

Mijn dank gaat ook uit naar de scholen waar ik mijn onderzoek heb mogen uitvoeren en de mensen die me bij die scholen ontvangen en ondersteund hebben: natuurlijk mijn eigen voormalige "Jeanne d'Arc college" te Maastricht (Porta Mosana; Ineke Luijckx, Martin de Cleen, & Theo Bartels); het Charlemagne college te Landgraaf (Karin Mommers); en het Bernardinus College te Heerlen (Stef Keulen).

Voor het maken van de digitale leeromgeving en het vaak binnen een dag afhebben van functionaliteiten wil ik graag Bas Jansen van Monito bedanken, net als Mat Heinen die me altijd goed technisch ondersteund heeft bij CELSTEC.

Het secretariaat heeft ook veel voor me betekend, dus bedankt Audrey, Alice, Ingrid J. en Sabine M. voor het uittikken van de verbale protocollen, bedankt Nicole voor

het financiële gedeelte, en bedankt Mieke voor het doorspitten van mijn proefschrift.

Al mijn collega's van CELSTEC wil ik graag bedanken, in het bijzonder Els, Jeroen S., Liesbeth, Maaike, Marcel v.d. K., Olga en alle (voormalige) AiO's: Amber, Amy, Bettine, Chantal, Eniko, Femke, Fleurie, Gemma, Greet, Helen, Ingrid S., Iwan, Johan, Karin, Kim, Ludo, Marjo, Milou, Monique, Pieter, Sandra en Wendy. De gezellige sfeer die jullie creëerden was uniek en ik heb dan ook met heel veel plezier de afgelopen vier jaar bij CELSTEC doorgebracht. Aan elk van jullie zou ik gemakkelijk een paragraaf kunnen wijden, maar dat zal ik jullie besparen. Op twee na dan…

Lieve 'tine, het klikte eigenlijk al meteen in het begin en sindsdien komen we elke dag wel bij elkaar langs om bij te praten over kleine en grote dingen. We hielpen elkaar nuchter te houden in turbulente tijden en ook in de laatste paar maanden van het promotietraject kon ik altijd op je rekenen om me te behoeden voor tunnelvisie. Ik er ook trots op dat Greet en jij als paranimfen bij me staan!

Lieve G., soms zeggen we wel eens dat wij elkaar een beetje te goed kennen, maar dat komt denk ik mede omdat we er de afgelopen vier jaar door dik en dun voor elkaar zijn geweest. Het was een gemis toen je naar Maastricht ging en het helpt ook niet dat ik nu naar Groningen ga, maar ik weet zeker dat onze vriendschap die afstand wel overleeft. Nu op naar jouw promotie!

Ook zou ik graag enkele mensen in mijn privé leven willen bedanken. Ad, Brigit, Cindy, Erik, Erwin, Hans, Jasper, Paul, Ramon, Rob en Robert, bedankt voor de films, spelletjes, gaan stappen, pretparken, feestjes, vakanties en al het andere! Veel van jullie ken ik al sinds de middelbare school en hoewel we steeds meer onze eigen weg gaan, hoop ik jullie zo vaak mogelijk te blijven zien.

Tenslotte wil ik mijn familie bedanken. Mijn broertje Rick, jij bent in zoveel zaken mijn tegenpool, maar als het er toe doet, weten we elkaar altijd in het midden te vinden. Mam en Pap, jullie zijn mijn rots in de branding en hebben me meer gegeven dan ik ooit terug zou kunnen doen. In alles wat ik ben, van humor tot moreel besef, herken ik altijd een beetje van jullie. Iech haw vaan uuch.

Danny Kostons,
September 2010, Heerlen

Contents

Chapter 1	Towards effective self-regulated learning: A cognitive load approach	9
Chapter 2	How do I do? Investigating effects of expertise and performance-process records on self-assessment	21
Chapter 3	Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners	33
Chapter 4	Training self-assessment and task-selection skills enhances the effectiveness of self-regulated learning	49
Chapter 5	General discussion	71
	References	81
	English summary	89
	Nederlandse samenvatting	95
	ICO dissertation series	101

Chapter 1

Towards effective self-regulated learning: A cognitive load approach

Abstract

Although self-regulated learning allows students to actively plan, monitor, and control their own learning process, these metacognitive processes require additional capacity of the limited working memory. In this chapter the theoretical framework of cognitive load is used to analyze the impact of the cognitive demands imposed by the metacognitive processes involved in self-regulated learning on students' learning outcomes. This framework also provides the theoretical foundation for the empirical studies presented in this dissertation.

Parts of this Chapter also appear in:
Van Gog, T., Kostons, D., Azevedo, R., & Paas, F. (2010). *Enhancing the effectiveness of self-regulated learning: A cognitive load perspective.* Manuscript submitted for publication.

Self-regulated learning allows students to actively plan, monitor, and control their own learning process (e.g., Boekaerts, Pintrich, & Zeidner, 2000; Pintrich, 2000a; Winne, 2001; Winne & Hadwin, 1998; Zimmerman & Schunk, 2001). Such metacognitive processes like planning, monitoring, and controlling do require additional capacity of the limited working memory. In this chapter the theoretical framework of cognitive load is used to analyze the impact of the cognitive demands imposed by the metacognitive processes involved in self-regulated learning on students' learning outcomes. This framework also provides the theoretical foundation for the studies presented in the next Chapters.

Self-regulated learning can occur at different levels, from learners controlling how long they engage in studying a given task or whether they want to restudy it (Karpicke, 2009; Metcalfe, 2009; Thiede & Dunlosky, 1999), to learners controlling what information they want to study (e.g., in a hypermedia learning environment; Azevedo, 2005; Azevedo & Cromley, 2004; Azevedo, Moos, Greene, Winters, & Cromley, 2008) or what learning tasks they want to work on (Corbalan, Kester, & Van Merriënboer, 2008; Ross, Morrison, & O'Dell, 1989). This chapter will focus primarily on metacognitive processes involved in self-regulated learning in which learners can choose their own learning tasks, which is often referred to as self-directed learning in the context of research on lifelong learning (Candy, 1991; Knowles, 1975; Loyens, Magda, & Rikers, 2008), and as learner-controlled instruction in the context of computer-assisted learning (Goforth, 1994; Merrill, 1980; Niemiec, Sikorski, & Walberg, 1996; Steinberg, 1989).

Self-regulated learning in which learners choose their own learning tasks is increasingly implemented in secondary education, because it is believed to better prepare students for tertiary education and working life. In the Netherlands, for example, a nationwide educational innovation that relies heavily on self-regulated learning (the 'study house') was implemented in the higher levels of secondary education in 1999 (see http://www.minocw.nl/english/education/293/Secondary-education.html). In addition, a heavier reliance on self-regulated learning is also seen as an opportunity to achieve personalized learning trajectories, in which instruction provides more room for each individual learner's interests and is adaptive to each learner's current level of knowledge or skill. Therefore, such personalized instruction is expected to enhance students' motivation and learning outcomes compared to non-adaptive, fixed instruction that is the same for all students (e.g., Niemiec et al., 1996). However, as we will show, there seems to be little evidence for both assumptions.

First of all, research has shown that students do not acquire self-regulation skills merely by providing them with control over their learning process, rather, they need additional training or instructional support such as prompts or tutoring (e.g., Aleven & Koedinger, 2002; Azevedo & Cromley, 2004; Azevedo, Cromley, & Seibert, 2004; Van den Boom, Paas, & Van Merriënboer, 2007; Van den Boom, Paas, Van Merriënboer, & Van Gog, 2004). For example, Azevedo and Cromley showed that providing students with a training session in which they were explained the phases and variables involved in self-regulated learning and instructed to use those during subsequent self-regulated learning in a hypermedia environment, significantly improved students' application of self-regulated learning skills during learning as well as their learning outcomes. Van den Boom et al. (2004) investigated the effects of reflection prompts and tutor feedback during self-regulated learning on students' self-regulated learning skills as measured by a difference in scores before and after the study on the regulation subscales of the Inventory of Learning Styles (see Vermunt, 1998), and found that especially tutor feedback increased students' (self-reported) self-regulation.

Secondly, the assumption that personalized instruction can foster learning compared to non-personalized instruction seems to be correct (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995; Camp, Paas, Rikers, & Van Merriënboer, 2001; Koedinger, Anderson, Hadley, & Mark, 1997; Salden, Paas, Broers, & Van Merriënboer, 2004; Shute & Zapata-Rivera, 2008). It is, however, questionable whether self-regulated learning actually results in the adaptivity to students' needs required for effective personalized instruction. When an instructional system is used to personalize instruction, it does so by a cyclical process of monitoring and assessing a student's current level of knowledge or skill to select or suggest an appropriate next learning task. The assessment can comprise several aspects of students' performance (e.g., Anderson et al., 1995; Kalyuga & Sweller, 2004; Koedinger et al., 1997) or a combination of their performance and invested mental effort (e.g., Camp et al., 2001; Corbalan et al., 2008; Kalyuga, 2006; Salden et al., 2004). For self-regulated learning to be equally adaptive and effective, students should be able to accurately monitor their own performance while working on a learning task, use the information that was thus acquired to assess their own performance after completing the task, and select an appropriate next learning task based on that assessment all by themselves. However, there is quite some evidence that students, particularly novices who lack prior knowledge of the learning tasks, are not able to do so.

We will argue here that this lack of ability to accurately monitor, self-assess, and select tasks can explain why research has shown that providing students with control over their learning process may have beneficial effects on their motivation or involvement, but often has no or even detrimental effects on their learning outcomes, especially for novice learners who lack prior knowledge of the learning tasks (see e.g., Azevedo et al., 2008; Corbalan, Kester, & Van Merriënboer, 2006; Kinzie,

1990; Kinzie & Sullivan, 1989; Lawless & Brown, 1997; Niemiec et al., 1996; Ross & Morrison, 1989; Scheiter & Gerjets, 2007; Steinberg, 1989; Uden, McGuiness, & Alderson, 2000). When neutral or positive effects on learning outcomes are found, this tends to be mostly for students with higher abilities or higher prior knowledge in the domain (e.g., Lawless & Brown, 1997; Moos & Azevedo, 2008a; Niemiec et al., 1996; Ross & Morrison, 1989; Scheiter & Gerjets, 2007; Steinberg, 1989; Uden et al., 2000). In the next sections, we will discuss how those findings can be explained by the cognitive demands that accurate monitoring, self-assessment, and task selection impose.

Monitoring

When learners are not able to monitor their performance while they are working on a learning task, they will not have a good memory representation of their performance process after completing the task. Studies applying retrospective verbal protocols suggest that learners often have a rather poor recollection of the task performance process (see e.g., Kuusela & Paul, 2000; Van Gog, Paas, Van Merriënboer, & Witte, 2005). This is problematic, because lack of an accurate representation of the performance process compromises the accuracy of self-assessment: It is considered important not only to assess the quality of the final solution or end-product, but also to assess the quality of the performance process that led up to that solution or product (e.g., Segers, Dochy, & Cascallar, 2003). Inaccurate self-assessment may in turn negatively affect task selection, in which case learners may be devoting time and effort to learning tasks that are not adaptive to their current level of knowledge and skill.

Monitoring may be difficult for learners because it requires working memory resources, which are also required for performing the learning task. Working memory capacity is limited to seven plus or minus two elements or chunks of information when holding information (Miller, 1956) and even less (about 4 elements) when processing information (Cowan, 2001; 2010). Many learning tasks are complex, that is, they are characterized by a high number of novel interacting information elements that have to be related, controlled, and kept active in working memory during task performance in order for learning to occur. Such learning tasks impose a high load on working memory, and this is referred to in cognitive load theory as intrinsic cognitive load (Chandler & Sweller, 1991; Sweller, 2010; Sweller et al., 1998). It should be mentioned that intrinsic cognitive load does not only depend on task complexity, but also on the level of expertise of the learner: As a result of learning, elements are combined into cognitive schemata stored in long-term memory that can be retrieved and handled as a single element in working memory, thereby decreasing the intrinsic load the task imposes on working memory (Sweller et al.,

1998). As a consequence, the same learning task imposes less cognitive load for a student with some prior knowledge of that task than for a novice.

Monitoring can be conceived of as a secondary task, requiring attention while performing the learning task (primary task). As such, monitoring and learning task performance are competing for the same limited working memory resources (Van Gog, Kester, & Paas, in press; Van Gog & Paas, 2009). Dual task research has shown that under high cognitive load conditions, accurate performance of the primary task, secondary task or both becomes hard to maintain (Brünken, Plass, & Leutner, 2003). That is, when the intrinsic load imposed by a learning task is high, little if any working memory resources are available for processes that impose additional capacity demands, such as concurrently monitoring performance. Learners can increase their mental effort to accommodate the dual task demands, but only to the extent that cognitive resources are still available (Paas, Tuovinen, Tabbers, & Van Gerven, 2003). When the limit of cognitive capacity has been reached, learners need to divide their resources between performing the primary and the secondary task (Kanfer & Ackerman, 1989), and as a result, monitoring, learning task performance, or both, may be hampered. Under conditions of low task complexity, or low intrinsic cognitive load, on the other hand, additional cognitive demands can be easily accommodated as ample resources are available, and the low complexity of the primary task may also make the secondary task less complex. In sum, when tasks impose high intrinsic load, monitoring performance may: a) lead to low quality monitoring (secondary task), and therefore, a poor recollection of performance on which to base self-assessment, and/or b) direct scarce cognitive resources away from the learning task (primary task), thereby hampering performance of that task at the expense of learning.

Van Gog and Paas (2009) and Van Gog et al. (in press) investigated the hypothesis that concurrent performance monitoring increases cognitive load and decreases performance on complex, high intrinsic load tasks (i.e., 9x9 Sudoku puzzles) but not on simple, low intrinsic load tasks (i.e., 4x4 Sudoku puzzles). In the pilot study reported by Van Gog and Paas, a mixed factorial design was used with task complexity as between-subjects factor and monitoring as within-subjects factor: Participants first had to work on a puzzle without the instruction to monitor their performance, then with that instruction. Cognitive load was measured using the nine-point subjective mental effort rating scale developed by Paas (1992; this scale is widely used in educational research; see Paas et al., 2003; Van Gog & Paas, 2008). The results showed that in line with the hypothesis, monitoring resulted in a trend towards higher cognitive load and significantly lower performance on the complex, but not on the simple tasks. However, these results were far from unequivocal because the pilot study had a relatively small number of participants, tasks were not counterbalanced, and the application of the monitoring instruction as within-subjects factor may have affected mental effort ratings. In the follow-up study, Van Gog et al. (in

press) therefore applied monitoring as between-subjects factor and task complexity as within-subjects factor, and included a higher number of participants. The same pattern of results was found: On the complex (i.e., high intrinsic load) task, performance was significantly lower and cognitive load ratings were significantly higher in the monitoring condition than in the no monitoring condition, whereas no significant effects of monitoring were found on the simple (i.e., low intrinsic load) task.

Although these studies did not investigate the effects on the performance of the secondary task, that is, on the quality of monitoring, and only investigated direct effects on performance, rather than learning (and performance is not always a good indicator of learning; Bjork, 1999), their findings may provide at least a partial explanation for why high prior knowledge learners seem to do better than novices in self-regulated learning (e.g., Lawless & Brown, 1997; Moos & Azevedo, 2008a; Niemiec et al., 1996; Ross & Morrison, 1989; Scheiter & Gerjets, 2007; Steinberg, 1989; Uden et al., 2000). In most of those comparative studies, the high and low prior knowledge learners were working on the same learning tasks, and because of their prior knowledge, these tasks imposed a lower intrinsic load for the high prior knowledge learners (Sweller et al., 1998). The high prior knowledge learners may therefore have had enough cognitive capacity available for performing the learning task and monitoring their performance simultaneously.

In sum, we have argued here that monitoring plays a pivotal role in the effectiveness of self-regulated learning, because the quality of monitoring can affect the accuracy of self-assessment, which in turn can affect the accuracy of task selection. The cognitive load imposed by monitoring may hamper learning when cognitive capacity is devoted to monitoring (i.e., a direct effect on learning outcomes), or it may hamper monitoring when cognitive capacity is devoted to learning, in which case the accuracy of self-assessment is likely to be negatively affected, which negatively affects the rest of the cycle of self-regulated learning (i.e., an indirect effect on learning outcomes). However, this explanation for findings regarding the effectiveness of self-regulated learning is likely to be a partial one, because there seem to be other cognitive factors that affect self-assessment accuracy.

Self-Assessment

Being able to accurately self-assess performance on a learning task is crucial for effective self-regulated learning. For example, students who are more accurate at judging their own learning, seem to be better able to decide whether or not they should restudy certain materials (e.g., Metcalfe, 2009; Thiede, Anderson, & Therriault, 2003; Thiede & Dunlosky, 1999). However, research has shown that accurate self-assessment is very difficult for learners, and especially for novices.

First of all, as mentioned in the previous section, self-assessment may be difficult when learners are not able to accurately monitor their performance, because

they will not have a good recollection of their task performance process on which to base their assessment.

Secondly, there are many types of cognitive biases that may affect the accuracy with which people assess their own performance, because these biases lead people to depend on the wrong kind of cues to assess their performance (for a review, see Bjork, 1999). For example, when people fail to solve a problem, and are subsequently provided with feedback on the correct solution, they are often inclined to overestimate the likelihood that they could have produced it themselves (i.e., hindsight bias), and when an answer comes to mind easily, it is not only more likely to be provided, but also more likely to be assumed correct (i.e., availability bias). People also often fail to distinguish their ability to perform a task from learning, that is, with the ability to reproduce that performance later on. For example, when an accurate solution to a problem is reached via a "weak strategy" such as trial-and-error or means-ends analysis, the problem may have been solved correctly, but this does not necessarily mean that a person has learned the correct solution procedure for such tasks as judged by his or her ability to solve an equivalent problem (Sweller, 1988). One of the problems in distinguishing performance from learning is that people tend to base their self-assessment on information present in the situation or in working memory at the time of the assessment, which may no longer be available in later test situations (Bjork, 1999).

Such biases may be very hard to counteract, although research on self-assessment of text comprehension and word-pair learning has shown, for example, that asking students to judge their learning at a delay rather than immediately may increase accuracy of the judgment (e.g., Nelson & Dunlosky, 1991). A possible explanation for this effect is that at a delay, students have to retrieve the information from long-term memory, which more closely resembles the situation during a test. With immediate judgments, in contrast, information can be used that is still available in working memory, but this information does not necessarily become consolidated in long-term memory, that is, it may not be available later for the test (Nelson & Dunlosky, 1991). Having students generate keywords (Thiede et al., 2003) or make summaries (Anderson & Thiede, 2008) about a text before asking them to assess their learning is another effective way to increase accuracy, especially when there is a delay between reading the text and generating keywords or summaries, presumably because this provides students with more appropriate cues of their memory for the text on which to base their assessment.

A third reason for why accurate self-assessment is especially difficult for novices is that it seems to be related to the amount of experience a person has (Boud & Falchikov, 1989; Dunning, Johnson, Erlinger, & Kruger, 2003). Accurate self-assessment requires knowledge of assessment criteria and standards, that is, about what aspects of performance should be assessed and what constitutes poor, average, good, or excellent performance on those aspects. This puts novice learners at a

disadvantage, because they not only lack knowledge about the learning tasks, but also about the assessment criteria and standards. Dunning and colleagues call this 'the double curse of incompetence' (Dunning, Heath, & Suls, 2004; Dunning et al., 2003), meaning that "… skills necessary to recognize competence are extremely close if not identical to those needed to produce competent responses" (Dunning et al., 2004, pp. 73). Another factor that might play a role is the frame of reference a learner is inclined to use during self-assessment, that is, do they compare their current performance to their own past performance (self-referenced), to their peers' performance (norm-referenced), or to an objective standard (criterion-referenced)? Lacking knowledge of the latter two, novices may have no other option than to use their own past performance. There is some evidence that this is indeed the case, at least in motor learning tasks. Ruble and Frey (1991) and Sheldon (2003) found the frame of reference to vary as a function of skill acquisition: Whereas novices tended to make more temporal comparisons to their own past performance (self-referenced), advanced individuals tended to make more social comparisons to performance of peers (norm-referenced).

These findings regarding the effects of prior knowledge on self-assessment accuracy might provide another part of the explanation for why self-regulated learning tends to be more effective for advanced learners. Not only do they have more prior knowledge of the task, due to which they have more capacity available for monitoring than novices, but they also have more knowledge of assessment criteria and standards. Both affect accuracy of self-assessment, and inaccuracies in self-assessment may negatively affect the selection of subsequent learning tasks.

Task Selection

Inaccurate task selection during self-regulated learning is problematic, because it will cause learners to work on tasks that do not fit their level of knowledge or skill (i.e., they are either too difficult or too simple), and as a result, they will not learn much. Inaccuracies in task selection can arise first of all when self-assessment is inaccurate. Data from a study by Salden et al. (2006) on personalized instruction illustrate this. In this study, learners were required to self-assess their overall task performance on a rating scale immediately after each task. This self-assessment was used in a personalized task-selection algorithm to adaptively select a next training task (i.e., learners did not self-select learning tasks in this experiment, they only self-assessed). After the experiment, participants' performance was also scored based on the log files. Comparing these objective assessment scores to the self-assessment scores, Salden et al. found that 67% of the participants tended to overestimate their performance during training. Interestingly, however, for accurate self-assessors (i.e., those whose self-assessment corresponded to the objective assessment) personalized task selection did lead to higher learning outcomes than

for the inaccurate self-assessors. For the latter group, because of the inaccuracies in their self-assessment, the task-selection algorithm that was used did not result in appropriate learning tasks that fitted their level of knowledge and skill.

In self-regulated learning, learners have to select their own learning tasks, and in this case not only their self-assessment affects what tasks they will select, but their self-efficacy beliefs, that is, their confidence in their ability to perform a task may also play a role (Bandura, 1994; Zimmerman, 2000). Ideally, a learner would seek out tasks that are challenging enough to learn from, but not too difficult. In practice, however, learners tend to select tasks that they are confident they can perform, but their confidence is not always aligned with their actual abilities. As a consequence, learners often select tasks that are too difficult (i.e., overconfidence), or tasks that are too easy (i.e., low confidence; see Pajares & Kranzler, 1995; Stone, 1994).

In addition, even when learners' self-assessment and self-efficacy beliefs would be accurate, they need to have the ability to recognize what an appropriate next task would be. As mentioned above, in system-controlled personalized instruction, task selection is based on algorithms that use assessment scores and task metadata (e.g., content, level of complexity, level of support) to determine what an appropriate next task would be. For example, consider a system that contains a task database with tasks at different levels of complexity, and different levels of support at each complexity level, such as worked examples (high support; e.g., Sweller & Cooper, 1985), completion problems (medium support; e.g., Paas, 1992) and conventional problems (no support). That system uses a task-selection algorithm based on an assessment of performance combined with mental effort scores to select tasks: When performance was high and invested mental effort was low, task complexity can be increased or support can be decreased, and when performance was low and mental effort was high, task complexity should be decreased, support increased, or both (see e.g., Camp et al., 2001; Corbalan et al., 2008; Kalyuga, 2006; Salden et al., 2004). The problem that arises during self-regulated learning is that learners do not have any knowledge about such algorithms or the important parameters underlying them. As a result, they are likely to base their selection on surface features of the task that are salient but not necessarily relevant for learning, such as a problem's cover story (Corbalan et al., 2008; Quilici & Mayer, 2002).

In sum, accuracy of monitoring, self-assessment, and task selection plays a crucial role in the effectiveness of self-regulated learning in which learners have control over what information they want to study (Azevedo, 2005; Azevedo & Cromley, 2004; Azevedo et al., 2008) or what learning tasks they want to work on (Corbalan et al., 2008; Niemiec et al., 1996). Accurate monitoring, self-assessment, and task selection seems to be especially difficult for novices, which may explain why self-regulated learning is often ineffective for novice learners. Even though learners with higher levels of prior knowledge have been found to benefit from self-regulated learning, it is important to note that in those studies, the advanced learners typi-

cally worked on the same tasks as novices. Ideally, however, students work on tasks that are at an appropriate level of difficulty for them, not on ones of which they already know a lot (unless the goal would be to automate performance, of course). In this ideal situation, all students may experience problems with monitoring due to high cognitive load, with self-assessment due to inaccurate monitoring and lack of knowledge about assessment criteria and standards, and with task selection because of inaccurate self-assessment and lack of knowledge about effective task selection heuristics.

So, are we forced to conclude that effective self-regulated learning is not feasible for novice learners? Such a conclusion might seem tempting, but would also be premature. The main aims of the research project reported in this dissertation were a) to study in detail how learners go about assessing their own performance and selecting new learning tasks, and b) to experimentally investigate whether training learners' self-assessment and task-selection skills could enhance the accuracy of those skills, and whether this would improve learning outcomes attained through self-regulated learning. An overview of the content of the remaining chapters is provided below.

Overview of the dissertation

The study presented in Chapter 2 investigated which aspects of their performance participants at different levels of expertise consider when they are asked to engage in self-assessment. To shed light on the question of whether a lack of knowledge of criteria and standards, a lack of monitoring ability, or a combination of both complicates novices' self-assessment, all participants were asked to self-assess their performance while verbalizing their thoughts. During self-assessment, half of the participants received a performance-process cue consisting of a replay of a record of their eye movements and actions performed on the computer, whereas the other half did not. The performance-process cues were expected to reduce the need for concurrent performance monitoring by allowing participants to review both physical actions (reflected in mouse/keyboard operations) and cognitive actions (reflected in eye movements) made during task performance. It was hypothesised that the cue would be helpful for novice participants to report about the performance process, whereas advanced participants do not need a cue to be able to report on the process. In addition it was hypothesized that the cue could or could not help novices evaluate their performance process (depending on which of the two explanations, lack of monitoring or lack of knowledge of criteria and standards, plays a more prominent role), but that it would help advanced learners evaluate their performance, which would show in the amount and type of assessment criteria used.

Chapter 3 describes a study that investigated differences in self-assessment and task-selection processes between novice learners who did and did not gain much knowledge from engaging in self-regulated learning, applying verbal protocol analysis and action protocol analysis to trace learners' thought processes and decisions. It was hypothesized that more effective learners (i.e., those who gained more knowledge) would be more accurate at self-assessment, would make more use of the outcomes of their self-assessment to select a new learning task, and would select tasks based more on their structural rather than superficial aspects compared to their less effective counterparts.

Chapter 4 describes two experiments that investigated whether self-assessment and task-selection accuracy could be improved by means of training and whether this would enhance the effectiveness of self-regulated learning. The first experiment investigated whether example-based learning, that is, observing a human model engaging in self-assessment, task selection, or both, could be an effective strategy for novices to acquire self-assessment and task-selection skills. It was hypothesized that observing modelling examples would improve their self-assessment and task selection accuracy. The second experiment investigated whether training self-assessment and task-selection skills, either by studying modelling examples or through practice with assessment and task selection rules, would enhance the effectiveness of subsequent self-regulated learning. It was hypothesized that learners who received training on self-assessment and task-selection skills would gain more knowledge from self-regulated learning than learners who did not. In addition, the question was explored of whether training consisting of examples or of practice would be more effective.

Finally, Chapter 5 provides an overview of the main findings of the studies presented in this dissertation, acknowledges the limitations of the studies, and discusses the findings in terms of practical and theoretical implications and directions for future research in this field.

Chapter 2

How do I do? Investigating effects of expertise and performance-process records on self-assessment

Abstract

The study described in this chapter investigated the effects of expertise and cues of the performance-process on self-assessment, using a 2 x 2 factorial design with factors 'Expertise' (Lower vs. Higher) and 'Performance-Process Cue' (Cued vs. Non-cued). The cues, consisting of replays of integrated records of participants' eye movements and actions on the computer screen, were hypothesized to help the lower expertise group to remember and the higher expertise group to evaluate their task performance, by allowing them to review both physical actions (reflected in mouse/keyboard operations) and cognitive actions (reflected in eye movements) made during task performance. The results were in line with this hypothesis. Implications of these findings for self-assessment theory and the use of eye tracking recordings as a performance-process cue are discussed.

This chapter was published as:
Kostons, D., Van Gog, T., & Paas, F. (2009). How do I do? Investigating effects of expertise and performance-process records on self-assessment. *Applied Cognitive Psychology*, *23*, 1256-1265.

Self-assessment, that is, the ability to identify strengths and weaknesses and points for improvement in one's own performance, plays an important role in self-regulated learning in school or on-the-job (Boud, 2000; Zimmerman, 2002). Unfortunately, however, many learners are not able to accurately assess their own performance (Bjork, 1994, 1999). As a consequence, they will not be able to optimally regulate their learning process, that is, select optimal future learning activities that help them improve their performance.

In order to self-assess, one also needs to be able to monitor the performance-process, because not only the end-product is important, but also the process by which it was obtained (see e.g., Segers, Dochy, & Cascallar, 2003). In addition, one needs to know the criteria and standards to which performance should be compared (Miller, 2003). The term assessment 'criteria' refers to the aspects of performance that are assessed, and 'standards' to the quality of performance on each of those aspects (see e.g., Arter & Spandel, 1992; Woolf, 2004). Research has shown that the accuracy of self-assessment improves with increasing expertise (Boud & Falchikov, 1989; Ruble & Frey, 1991). This improvement has been suggested to occur because understanding of assessment criteria and standards develops along with domain knowledge and skills. In other words: "… skills necessary to recognize competence are extremely close if not identical to those needed to produce competent responses" (Dunning, Heath, & Suls, 2004, p. 73). This has also been called the 'double curse of incompetence' (Dunning, Johnson, Erlinger, & Kruger, 2003).

The apparent close relationship between expertise and understanding of assessment criteria might imply that we have to accept that novices are simply unable to self-assess their performance. Consequently, it could be argued that it does not make sense to implement educational programs that require accurate self-assessment in order to be effective, such as learner-controlled instruction, or self-regulated learning, from the start of the learning trajectory, but only once learners have acquired some expertise.

However, in this article we present an alternative expertise-based view on novices' failure to self-assess. As mentioned before, there is another prerequisite for self-assessment, which is the ability to monitor one's performance. Novices may not be able to monitor their performance, because their learning of complex cognitive tasks by definition imposes a high cognitive load for them on their cognitive system (Sweller, Van Merriënboer, & Paas, 1998). When tasks contain a lot of new information and are complex in the sense that these information elements inter-relate, they impose a high intrinsic cognitive load (Sweller et al., 1998; Van Merriënboer &

Sweller, 2005). Monitoring can be seen as a kind of secondary task that has to compete for cognitive resources with the primary task, and, consequently, is likely to deteriorate when the load imposed by the primary task becomes too high (cf. Brünken, Plass, & Leutner, 2003). Intrinsic load interacts with expertise, however, in such a way that with increasing expertise, more elements are combined into cognitive schemata that can be retrieved from long-term memory and handled in working memory as a single element, which lowers the intrinsic load imposed by a task (Sweller, 2004). So, the more information elements of a task are combined into a schema, the lower the intrinsic load imposed on working memory by the task, and the more cognitive resources there are available for other processes, such as monitoring task performance.

When novices are unable to monitor their performance, they will only have a limited recollection of the performance process on which to base their self-assessment. In this case, providing them with a cue for self-assessment, consisting of a replay of a record of their performance process, would reduce the need to monitor their performance. For computer-based tasks, such a record can consist of a screen capture. However, this would allow the learner only to review and evaluate the problem-solving actions that were physically performed (mouse/keyboard operations), while important cognitive actions remain invisible. Therefore, a record that captures both the actions on the computer screen as well as the eye movements of the learner, could be more effective (cf. the cue used by Van Gog, Paas, Van Merriënboer, & Witte, 2005, for cued retrospective reporting; see also Hansen, 1991), because eye movements are generally believed to reflect the allocation of attention and therefore relate to cognitive processes (Duchowski, 2007; Rayner, 1998).

However, it is also plausible that the two explanations we discussed here, that is, the problems for novices with self-assessment arise because they do not know the criteria and standards, or because they are unable to monitor due to high cognitive load, interact. Even when the need to monitor is reduced by providing learners with a record of their performance process, they might not be able to self-assess because they do not know what aspects of the performance process to evaluate or what constitutes good or poor performance on those aspects.

This study, in the domain of biology, was designed to investigate these explanations, and aimed to uncover which aspects of their performance participants at different levels of expertise (i.e., novices, without prior knowledge, and advanced learners, with some prior knowledge) spontaneously consider when they are asked to engage in self-assessment, either with or without a record of their performance process. To obtain this information, learners were required to perform a task and then to assess their performance while thinking aloud (Ericsson & Simon, 1993). Half of them could self-assess based on a cue consisting of a replay of a record of

their performance process, which included their actions on the computer screen as well as their eye movements made.

It is hypothesized that: a) the cue is helpful for novice participants to report about the performance process, whereas advanced participants do not need a cue to be able to report on the process, and b) the cue may or may not help novices assess their performance process (depending on which of the two expertise explanations described earlier plays a more prominent role), but could help advanced learners evaluate their performance, which will show in the amount and type of criteria used.

Method

Participants

Forty Dutch adults (university staff) volunteered to participate in this study (26 females, 14 males; age *M* = 36.62 years, *SD* = 10.78).

Design

A 2 x 2 factorial design was used with 'Expertise' (Lower vs. Higher) and 'Performance-Process Cue' (Cued vs. Non-cued).

Apparatus and Materials

Prior knowledge test. The prior knowledge test consisted of ten multiple-choice items on the subject of heredity with four response options. Six items measured factual knowledge, and four items measured both procedural and factual knowledge. Twelve points in total could be gained; 1 for each correct factual knowledge question and 1.5 for each correct procedural-plus-factual knowledge question.

Practice exercise. The practice exercise for acquainting participants with the self-assessment procedure consisted of a puzzle, which was unrelated to the subject of the experimental tasks.

Experimental tasks. The four experimental tasks consisted of open-ended questions on the subject of heredity according to the laws of Mendel. All tasks required both knowledge of concepts, as well as procedural knowledge of how to solve such tasks. The tasks were presented on the computer screen, with text and a diagram that students had to fill in to solve the problem (See Figure 2.1 for an example).

Eye-tracking equipment. We used a 50Hz TOBII 1750 eye tracker, which is integrated into the stimulus PC monitor. Screen resolution of the stimulus PC was set at 1,024 x 768 pixels. TOBII Clearview 2.7.1 software was used for recording and re-

playing eye movements using the screen capture mode (that records at 15 frames per second), which not only captures eye-movements, but also other actions visible on the screen, such as keyboard and mouse operations. Fixations were defined as gaze-points falling within a radius of 50 pixels that together had a duration of at least 200 ms, which is appropriate for the graphical stimuli presented (Rayner, 1998).

Self-assessment cue. The cue consisted of a replay of the screen capture record, showing eye movements and mouse/keyboard operations (cf. Hansen, 1991; Van Gog et al., 2005). The recorded eye movements were played back in real time (i.e., the replay was at the same speed as the actual task performance because otherwise "time" would not be a valid self-assessment criterion), by using the replay function of the Clearview 2.7.1 software. A red dot signified the fixations. This dot became smaller or larger with decreasing or increasing fixation duration, and had a trail of 1,000 ms. (i.e., a line that showed the movement of the gaze during the last second up to the present fixation point). See Figure 2.1 for an example.

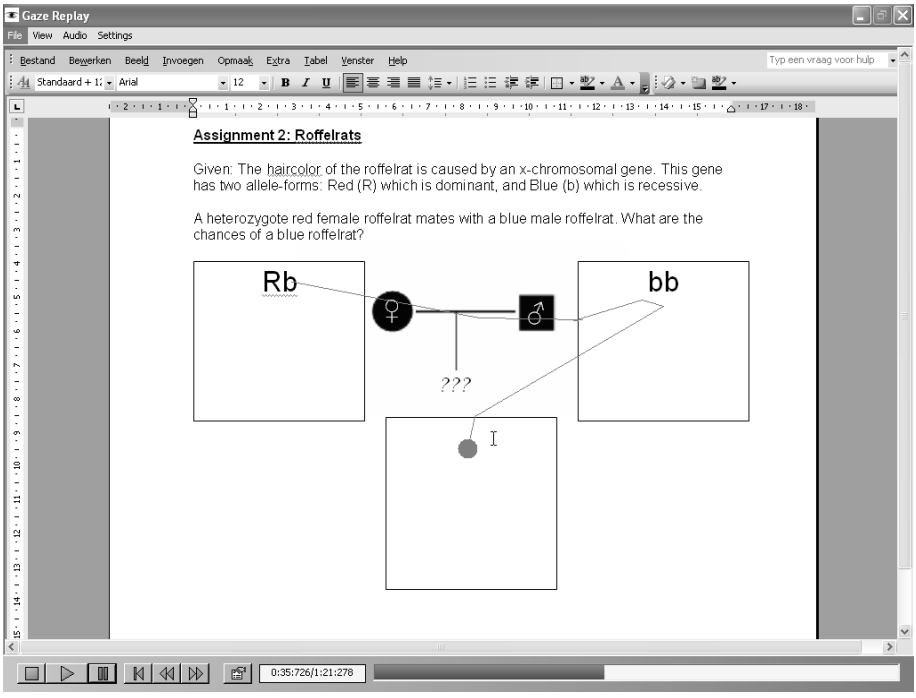

Figure 2.1. Example of the cue

Self-assessment instructions. The self-assessment instruction for the Non-cued condition was "Please evaluate your performance on this task while thinking aloud, that is, verbalize everything that comes to mind and do not mind my presence in

doing so". For the cued condition, it was "Please look at this record and evaluate your performance on this task while thinking aloud, that is, verbalize everything that comes to mind and do not mind my presence in doing so."

Audio recordings. Verbalizations during self-assessment were digitally recorded with Audacity 1.2.6 software, via a microphone attached to the stimulus PC.

Procedure

Based on their score on a prior knowledge test administered before the experimental session, participants were assigned to the higher expertise condition when they scored above 8 (n = 20, score M = 10.10, SD = 1.22, age M = 33.30 years, SD = 10.39, 5 male, 15 female) or the lower expertise condition when they scored below 5 (n = 20, score M = 2.43, SD = 2.31; age M = 39.95 years, SD = 10.36; 9 male, 11 female). Aiming at 20 novices and 20 advanced learners, we did not have to administer the prior knowledge test to more than 40 participants, as no one scored between 5 and 8.

Within the lower and higher expertise groups, participants were randomly assigned to either the Cued condition (Lower Expertise: n = 10, age M = 40.20 years, SD = 11.39, 6 male, 4 female; Higher Expertise: n = 10, age M = 32.90 years, SD = 11.21, 1 male, 9 female) or Non-cued condition (Lower Expertise: n = 10, age M = 39.70 years, SD = 9.83, 3 male, 7 female; Higher Expertise: n = 10, age M = 33.70 years, SD = 10.10, 4 male, 6 female; neither ANOVA, nor Bonferroni post-hoc tests, showed significant differences in age and gender between groups).

The study was conducted in individual sessions of approximately 60 minutes. To acquaint participants with thinking aloud during self-assessment or with thinking aloud based on the cue during self-assessment, they first received the practice exercise (approximately 3 minutes).

After the practice exercise, participants started with the four experimental tasks. Eye movements were recorded in all conditions to ensure that if knowledge of being tracked would affect task performance, this would be the same in all conditions. Participants were given a minimum of 1 minute and a maximum of 5 minutes to work on each experimental task and received a warning that their time was almost up after 4.5 minutes.

After each task, participants were required to give a self-assessment of their performance, while thinking aloud either with or without the cue depending on their assigned condition. When participants stopped thinking aloud for more than three seconds, they were prompted by the experimenter to "Keep trying to think aloud".

Data Analysis

After the experiment, the verbal protocols were transcribed and segmented based on meaningful units (i.e., partial, whole, or multiple sentences). When a subsentence within a unit had a distinct other meaning, it was made into a nested segment that could receive its own code. The segments were coded using a coding scheme with three main categories: a) Survey: comments relating to surveying information in the task and task characteristics; b) Action: comments relating to performing problem-solving steps; and c) Monitoring/Assessment: comments related to evaluating problem-solving steps, the overall performance process or product, and one's own knowledge or ability.

The category Monitoring/Assessment was further divided into four "criteria": 1. Effectiveness: evaluations of the adequacy/effectiveness of problem-solving steps, overall process, or knowledge/ability, which were further divided into being either positive or negative (e.g., "I think I did that wrong"); 2. Efficiency: evaluations including a time component (e.g., "It took me a long time to solve this"); 3. Affect: evaluations of emotional and motivational states, also divided into being either positive or negative (e.g., "This is frustrating"); and 4. Difficulty: evaluations concerning the difficulty of the task (e.g. "This was more difficult than the previous one").

Two raters with knowledge of the experimental tasks independently coded 15% of the protocols with an inter-rater reliability of .78 (Cohen's kappa). Because the inter-rater reliability was sufficient (i.e., higher than .70; Van Someren, Barnard, & Sandberg, 1994), one rater scored the remaining protocols, and this rater's scores were used in the analyses.

Results

Considering the small number of participants per condition, nonparametric tests were used for all analyses reported here. Means and standard deviations for Surveys, Actions, and Monitoring/Assessment, as well as the subdivisions Effectiveness (positive/negative), Efficiency, Affect (positive/negative), and Difficulty can be found in Table 2.1. One-tailed results are reported when an expected direction was stated in the hypotheses, otherwise two-tailed results are reported (note that one-tailed tests would still be significant when approached two-tailed). Effect size estimates for nonparametric tests were derived from z-scores and are expressed as "*r*" (see Rosenthal, 1991, p. 19), with small effects from .1 up to .3, medium effects between .3 and .5, and large effects above .5 (Field, 2005).

	Lower Expertise		Higher Expertise	
	Non-Cued	Cued	Non-Cued	Cued
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Main categories				
Surveys	3.00 (2.58)	6.30 (5.12)	6.00 (4.11)	5.40 (3.47)
Actions	13.6 (10.2)	30.2 (16.8)	22.6 (13.9)	22.5 (12.2)
Monitoring / Assessment	21.9 (5.36)	21.0 (11.4)	16.2 (7.76)	25.1 (10.6)
Sub categories Monitoring/ Assessment				
Positive effect	1.30 (1.34)	1.90 (2.28)	5.20 (2.97)	9.10 (7.91)
Negative effect	13.6 (10.2)	12.1 (5.51)	6.40 (4.77)	7.60 (5.02)
Efficiency	.40 (.52)	1.30 (1.83)	1.30 (1.16)	2.70 (1.42)
Affect Positive	.20 (.63)	.60 (.97)	1.00 (1.25)	2.50 (2.80)
Affect Negative	3.30 (4.27)	1.50 (2.51)	.10 (.32)	.80 (.63)
Difficulty	2.30 (2.21)	2.40 (4.48)	2.20 (1.69)	2.20 (1.87)

Table 2.1: Means and standard deviations for main and subcategories.

Task Scores

As a check on experimental manipulations, task performance data were compared between the lower and higher expertise participants. As one would expect, a Mann-Whitney U Test showed that higher expertise participants (M = 8.65, SD = 2.80) generally performed better on the four tasks than lower expertise participants (M = 2.10, SD = 1.74; U = 9.5, $p < .001$, r = - 0.82). The presence of a cue did not influence task performance (U = 165, *ns*).

Effects of Cue

On the categories of Survey, Action, and Monitoring/Assessment, Mann-Whitney U Tests showed that for higher expertise participants, the cue resulted in more Monitoring/Assessment comments compared to non-cued higher expertise participants (U = 25.0, p < .01, r = -.42). For lower expertise participants, the cue resulted in more Action comments compared to non-cued lower expertise participants (U = 19.0, p < .01, r = -.52), as well as a trend towards more Survey comments (U = 28.5, p = .05, r = -.37).

Because evaluations can be both positive and negative, and based on different criteria, the content of the Monitoring/Assessment statements should be considered. For higher expertise participants, we found an increase in the use of the Efficiency (U = 22.0, p < .05, r = -.48) and the Negative Affective criteria (U = 19.5, p < .01, r = -.60) when a cue was present. For lower expertise participants, we found no differences between the availability of a cue or not in any of the criteria used.

Effects of Expertise

In the non-cued condition, higher expertise participants made more Survey (U = 27.0, p < .05, r = -.39), and Action (U = 27.0, p < .05, r = -.39), and less Monitoring/Assessment (U = 25.0, p < .05, r = -.42) comments compared to lower expertise participants, whereas no such expertise effect was observable when a cue was available.

Again, the content of the Monitoring/Assessment statements should be considered. In the non-cued condition, higher expertise participants made more use of the Positive Effectiveness (U = 9.5, p < .001, r = -.69), Efficiency (U = 27.0, p < .05, r = -.42), and Positive Affective (U = 26.5, p < .05, r = -.47), but less use of the Negative Effectiveness (U = 12.0, p < .001, r = -.64) and Negative Affective (U = 17.5, p < .01, r = -.62) criteria than lower expertise participants. In the cued condition, Mann-Whitney U Test results were almost the same as in the non-cued condition (i.e., higher expertise participants used more Positive Effectiveness (U = 13.0, p < .01, r = -.63), Efficiency (U = 22.5, p < .05, r = -.47), and Positive Affective criteria (U = 23.5, p < .05, r = -.47) compared to lower expertise participants. The exception was that there was no longer a significant difference between lower and higher expertise participants in their use of the Negative Affective criterion (U = 35.5, p < *ns*, r = -.25).

Discussion

This study investigated which aspects of their performance participants at different levels of expertise considered when they were asked to engage in self-assessment

while thinking aloud, either with or without a performance-process cue consisting of a replay of a record of eye movements and actions performed on the computer. This would shed light on the question of whether lack of knowledge of standards and criteria, lack of monitoring ability, or a combination of both, hinder self-assessment.

The finding that in the non-cued condition, higher expertise participants made more Action and Survey statements than lower expertise participants, and that for lower expertise participants, the cue led to more Action and Survey statements, is in line with our hypothesis that the cue was helpful for novice participants to report about the performance process, whereas advanced participants did not need a cue to be able to report on the process. This finding also seems to suggest that lower expertise participants indeed have difficulty remembering the details of the performance process due to higher cognitive load (the higher expertise participants presumably experience less cognitive load imposed by the tasks), and that the cue allows them to review and remark on those task processes.

However, the cue did not lead to an increase in evaluations (Monitoring/Assessment) made by lower expertise participants, but, in line with our hypothesis, did seem to help higher expertise participants evaluate their performance, especially regarding Efficiency of the process and Negative Affect (which could be a possible indicator of becoming more aware of faults due to the cue compared to the non-cued condition). This finding seems to provide support for the assumption that some knowledge of standards and criteria is required in order to be able to evaluate one's own performance and that the cue can be a helpful support tool under those conditions (i.e., while higher expertise participants might not need a cue to remember the process, it seems to help them assess their performance). In other words, these findings suggest an interaction between the level of prior knowledge and self-assessment (cf. Kalyuga, Ayres, Chandler, & Sweller, 2003), in that for a cue to have a beneficial effect on self-assessment, some knowledge of the domain is required. It should again be noted that the higher expertise participants in this study were of an advanced level, but could not be considered experts. As such, whether or not true experts would benefit from the cue used here is an empirical question.

Regarding the type and use of assessment criteria, in general, higher expertise participants seem to use more criteria, and make more positive statements than lower expertise participants, who make more negative statements. However, two things should be noted. The first is that it is unclear whether the positivity of the advanced learners is warranted, since accuracy was not investigated here. The second is that in general, of the possible assessment criteria, not many were actually used, even by the higher expertise participants, but especially by the lower expertise participants. This seems to lend some support to the explanation that novices are unable to self-assess because they lack knowledge of criteria and standards. However, another possible explanation is that novices, lacking any knowledge of the

task, engage in trial and error strategies, and as such do not really have much to evaluate (i.e., cannot go beyond: "tried this, didn't work"). Indeed, Table 2.1 shows that novices produce a lot of monitoring/self-assessment comments, but split-up into subcategories, it is clear that for novices these consist mainly of negative effective and affective comments.

The performance-process cue used in this study is still rather unconventional. It has been used in a few studies as a cue for verbal reports (Schwonke, Berthold & Renkl, 2009; Van Gog et al., 2005), and was used here to cue self-assessment of task performance. Other performance-process cues, such as screen captures without eye movements or video recordings of participants, might lead to comparable results. However, as mentioned before, the added value of the eye movements is that they could trigger memory and evaluations of purely cognitive actions that did not result in physical actions (cf. Hansen, 1991; Van Gog et al., 2005). In this study we did not distinguish between reports or evaluations of cognitive or physical actions, but it might be interesting in future studies to compare whether or not different types of performance-process cues would lead to different results. It might also be interesting to investigate whether a cue could positively affect the accuracy of self-assessments.

The present study suggests that at least for somewhat advanced individuals, performance process cues could stimulate self-assessment, but not for novices. Novices did have some benefit from the cue, though, in that they were better able to recount the actions compared to when they did not have the cue. In this study, we were interested in what criteria participants would 'spontaneously' use, as providing them with lists of criteria would provide less information regarding learners' knowledge of criteria (if they are required to evaluate some aspect of their performance, they will, but whether or not they would have thought of it themselves remains unclear). Moreover, in self-regulated learning environments lists of criteria are usually not available. Self-assessment plays a crucial role in such learning environments though, as it affects subsequent learning activities. That is, if one is not aware of one's needs for performance improvement, one is unlikely to select tasks that will address those needs. Therefore, further studies on the complicated relationship between expertise, cognitive load, and the ability to evaluate one's own performance are important, as these can provide insight into how educators can support students' monitoring/assessment skills, which can ultimately contribute to improving the effectiveness of self-regulated learning.

Chapter 3

Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners

Abstract

Learner-controlled instruction is often found to be less effective for learning than fixed or adaptive system-controlled instruction. One possible reason for that finding is that students, especially novices, might not be able to accurately assess their own performance and select tasks that fit their learning needs. Therefore, this explorative study investigated the differences in self-assessment and task-selection processes between effective and ineffective learners (i.e., in terms of learning gains) studying in a learner-controlled instructional environment. Results indicated that although effective learners could more accurately assess their own performance than ineffective learners, they used the same task aspects to select learning tasks. For effective learners, who were also more accurate self-assessors, the self-assessment criteria predicted subsequent task selection. These results are discussed, particularly with regard to their potential to provide guidelines for the design of a self-assessment and task-selection training.

This chapter was published as:
Kostons, D., Van Gog, T., & Paas, F. (2010). Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education*, *54*, 932-940.

This explorative study investigated the differences in self-assessment and task-selection processes between learners who were and were not able to profit (i.e., in terms of learning gains) from studying in a learner-controlled instruction environment. Learner-controlled instruction (LCI) allows learners to regulate their own learning and to choose their own learning activities based on their personal needs and preferences (Merrill, 1980). Due to the possibility for personalization (i.e., adaptivity to learner needs and preferences), LCI was expected to be beneficial for students' learning processes and outcomes, as well as their motivation and involvement (Corbalan, Kester, & Van Merriënboer, 2008; Niemiec, Sikorski, & Walberg, 1996; Williams, 1996). However, although research on LCI has shown that it can indeed enhance learners' motivation and involvement compared to other kinds of instruction (Corbalan et al., 2008; Schnackenberg & Sullivan, 2000; Uden, McGuinness, & Alderson, 2000), most studies did not find any effects on learning (cf., Goforth, 1994; Niemiec et al., 1996; Steinberg, 1989; Uden et al., 2000; Williams, 1996), and for novices, it can have negative effects (Azevedo, Moos, Greene, Winters, & Cromley, 2008). When positive effects of LCI on learning have been reported, this is usually for learners with more prior knowledge (Lawless & Brown, 1997; Moos & Azevedo, 2008a; Niemiec et al., 1996; Scheiter & Gerjets, 2007; Schnackenberg & Sullivan, 2000; Steinberg, 1989) or higher levels of metacognitive skills (Azevedo, 2005; Scheiter & Gerjets, 2007). These findings are, however, not entirely surprising if we look at the cognitive demands imposed by LCI.

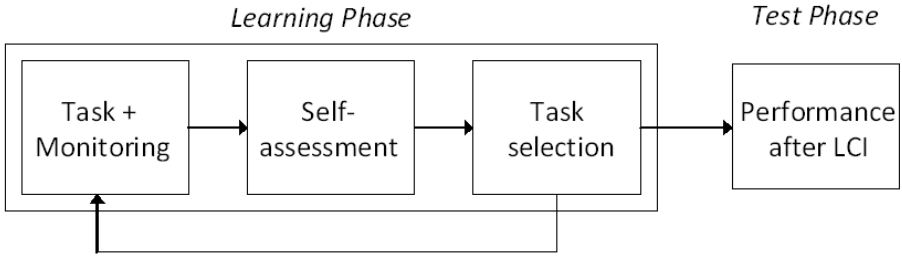

Figure 3.1: Schematic overview of LCI

For LCI to be effective for learning, students have to monitor their performance while they are working on a task, use this as input for self-assessment after completing the task, and select a new learning task based on that assessment (see Figure 3.1). Learners may experience difficulties in each of those phases. It has been shown that monitoring while engaging in a complex learning task is very difficult for

learners; often they have a poor recollection of their performance after completing a task (see Kostons, Van Gog, & Paas, 2009 - Chapter 2). In addition, monitoring may lead to lower performance on the task due to the fact that limited cognitive resources need to be shared between task performance and monitoring that performance (see Van Gog & Paas, 2009).

These difficulties with monitoring affect self-assessment as well: if learners do not have a good recollection of the task, there is not much upon which they can base their assessment. Moreover, there are other factors that complicate accurate self-assessment: They are often based on unreliable measures of performance (see Bjork, 1999, for a review). For example, learners' own judgments of how much they learned from a task (i.e., Judgments of Learning) often do not correspond with their actual learning (Koriat & Bjork, 2005; Koriat & Ma'ayan, 2005), because they are based on the wrong cues for whether information has been learned (i.e., can be recalled later). As a consequence, learners seem prone to illusions of competence (Koriat & Bjork, 2005). In addition, for self-assessment to be accurate, learners need to know the criteria and standards against which they should assess their performance (Miller, 2003). Novices usually lack knowledge of those criteria and standards and how to use them (see e.g., Dunning, Heath, & Suls, 2004). Without knowledge of 'objective' standards, learners are likely to use their own past performances as a frame of reference (Ruble & Frey, 1991; Sheldon, 2003). This tends to positively influence the interpretation of their current performance (i.e., it is better than it was before), thereby contributing to illusions of competence and overestimations of performance (Dunning et al., 2004; Topping, 2003).

Inaccurate self-assessment may in turn negatively affect the selection of a new task. If task selection is based on a wrong representation of their own performance, learners are unlikely to select a task that helps them address the weaker aspects of their performance (cf. the need to evaluate the relevancy of content and the adequacy of information for the goals to be attained in non-linear hypermedia learning; Azevedo, Guthrie, & Seibert, 2004; Moos & Azevedo, 2008b; Shapiro, 2004). However, even when self-assessment is accurate, learners may experience problems with selecting appropriate learning tasks. Particularly novice learners may find it difficult to determine which features of a task are relevant for learning and which are not (Quilici & Mayer, 2002). Rather than focussing on structural aspects of a task that are important for learning, such as the level of complexity or the amount of instructional support, they tend to focus on less relevant surface features (e.g., cover stories) when selecting a learning task (Corbalan et al., 2008). Furthermore, learners should ideally select tasks that are challenging enough to learn from (i.e., above their current performance level), but not too difficult (cf. Vygotsky's, 1978, concept of 'zone of proximal development'). In practice, however, learners tend to select tasks that they are confident they can perform (i.e., self-efficacy; Bandura, 1997), which often results in selection of too difficult tasks in case of overconfi-

dence, or too easy tasks in case of low confidence (Pajares & Kranzler, 1995; Stone, 1994).

In sum, a likely explanation for why learners, and particularly novices, do not profit from engaging in LCI, is that they experience difficulties with self-assessment and task selection. There is, however, not much research that has explored in-depth how learners go about self-assessment and task selection in LCI environments, and whether differences in these processes can explain differences in learning effectiveness. Therefore, in this explorative study, novice learners engage in LCI in an electronic learning environment on the topic of heredity, and they are asked to think aloud during self-assessment and task selection. Based on their learning outcomes, they are divided into two groups: Learners who were more effective (i.e., profited more from LCI) and less effective (i.e., did not profit much from LCI). Even though explorative, there are some expectations that can be formulated based on the theory and studies mentioned before. First, it might be expected that learners who are more effective are those who were more accurate in assessing their own performance. Second, it might be expected that effective learners consider different aspects of learning tasks during task selection than ineffective learners. Finally, effective learners might be expected to make more use of their self-assessments during task selection than ineffective learners.

Method

Participants

Thirty-two Dutch students in their 4th year of pre-university education (i.e., the highest level of secondary education in the Netherlands with a six year duration) who were enrolled in Biology courses volunteered to participate in this study (15 female, 17 male; age M = 15.9, SD = .49) and received a reward of € 7.50 for their participation. None of the participants had received any formal education on the research subject yet.

Materials and Procedure

Pre-test. Two weeks prior to the main part of the study, participants received the pre-test in a classroom at their school (15 min.). This test consisted of six paper and pencil multiple-choice problem-solving tasks on the subject of heredity: One task at each of the five complexity levels present in the learning environment (see Figure 3.2), and a novel (transfer) problem. The tasks had a multiple-choice format with three answer options and it was possible for multiple options to be correct. For each problem, full credit (1 point) was only given if all correct options were selected

and all incorrect options were left blank, and no points were awarded when any mistakes were made (i.e., a maximum of 6 points could be gained).

LCI phase. In the main phase of the study, learners engaged in LCI in individual sessions (at school) of approximately 45 minutes, using an electronic learning environment designed for this study. The learning environment consisted of a Web application with a database connected to it, which contained all learning tasks and logged participants' task solutions, task selection choices, responses on assessment scales, and time spent on each activity. On the first screen, participants received information on the procedure, that is, the three phases they would need to repeat (select, perform, assess) eight times and on how the task database was built up.

Complexity Level	Support Level	Learning Tasks				
Complexity 1 - 2 generations - 1 unknown - 1 solution - Deductive	Worked Example	Eye color	Hair structure	Shapes cat fur	Appletree blossom time	Depression
	Completion	Eye color	Hair structure	Huntington Disease	Curve chicken beak	Cleft lip
	Conventional	Eye color	Hair structure	Wolfram Syndrome	Dogtail Length	Milk allergy
Complexity 2 - 2 generations - 1 unknown - Multiple solutions - deductive	Worked Example	Eye color	Hair structure	Flower color	Fruitflies	P.R.A.
	Completion	Eye color	Hair structure	Shapes cat fur	Appletree blossom time	Tongue Curling
	Conventional	Eye color	Hair structure	Huntington Disease	Albinism	Cleft lip
Complexity 3 - 2 generations - 1 unknown - Multiple solutions - inductive	Worked Example	Eye color	Hair structure	Fruitflies	Curve Chicken beak	Wolfram syndrome
	Completion	Eye color	Hair structure	Dogtail length	Flower color	Milk allergy
	Conventional	Eye color	Hair structure	Tongue Curling	Depression	Albino
Complexity 4 - 3 generations - 1 unknown - Multiple solutions - Both ways	Worked Example	Eye color	Hair structure	Albino	Shapes cat fur	Fruitflies
	Completion	Eye color	Hair structure	Fruitflies	Tongue Curling	Flower color
	Conventional	Eye color	Hair structure	Cleft lip	P.R.A.	Depression
Complexity 5 - 3 generations - 2 unknowns - Multiple solutions - Both ways	Worked Example	Eye color	Hair structure	Milk Allergy	Depression	Huntington disease
	Completion	Eye color	Hair structure	Dogtail Length	Wolfram syndrome	Color of flowers
	Conventional	Eye color	Hair structure	Appletree blossom time	Fruiflies	Depression
	Worked example = all 5 steps worked out with rationale		Completion problem = as worked example, with 2 steps left out		Conventional problem = only problem statement.	

Figure 3.2: Overview of database structure; this was also the task selection screen

The learning task database, developed in cooperation with two biology teachers, consisted of problems in the domain of heredity (laws of Mendel), at five complexity levels tailored to the level of these learners (i.e., the highest level of complexity

matched what they should know at the end of the school year). Within each complexity level, three types of tasks were available that varied in the degree of instructional support: (a) Worked examples, which provide a fully worked-out solution procedure and an explanation of the rationale behind this procedure (cf. process-oriented worked examples; Van Gog, Paas, & Van Merriënboer, 2004) for the learner to study, (b) Completion problems (Paas, 1992), that is, partly worked-out examples, with three out of five solution steps already worked out, leaving two for the learners to complete, and (c) Conventional problems which the learners had to complete themselves (no support). Five tasks per support level per complexity level were available, resulting in a total of 75 tasks (see Figure 3.2). The heredity problems could be solved by going through the following steps: (1) translate the phenotypes (expression of genetic trait) described in the cover story into genotypes (a pair of upper and/or lower case letters representing genetic information); (2) put these genotypes into a hereditary diagram; (3) deduce missing individuals' genotypes by means of a Punnett square; (4) induce missing individuals' genotypes by means of a Punnett square; (5) compare results of step 3 and 4 and infer final solutions. Maximally five minutes were allotted for each learning task. Learners were allowed to work out the solutions on paper, and their notes were collected after each task.

After completing each task, participants had to self-assess their performance on seven criteria, implemented as nine-point rating scales, with labels at the two extremes: (1) (solution) "I think my final solution was: (1) completely wrong … (9) completely right"; (2) (approach) "I thought my approach was: (1) completely wrong … (9) completely right"; (3) (time on task) "The task took me a: (1) short time … (9) long time"; (4) (enjoyment) "I thought the task was: (1) no fun at all … (9) a lot of fun"; (5) (difficulty) "I thought the task was: (1) very easy … (9) very difficult"; (6) (mental effort) "Performing the task required: (1) very little effort … (9) very much effort"; and (7) (overall evaluation) "Altogether, I performed this task: (1) very poorly … (9) very well". These items were developed based on the study of Chapter 2 in which it was investigated which "criteria" students spontaneously mentioned during self-assessment. After this self-assessment, participants were presented with an overview of all tasks on a single screen (cf. Figure 3.2). Learners could click on the task they wanted to perform. One restriction was implemented: They needed to complete at least one conventional task within a complexity level (correctly or incorrectly) before they could proceed to a higher complexity level (cf. Corbalan et al., 2008). Participants were made aware of this rule at the beginning of their session, and were reminded of it each time they went to the task selection screen.

During self-assessment and task selection, learners were instructed by the experimenter to think aloud: "Please think aloud, that is, verbalize everything that comes to mind." (cf. Ericsson & Simon, 1993). If participants fell silent for more than three seconds, they were reminded to "Keep thinking aloud". Verbalizations were digitally recorded with Audacity 1.2.6 software (Mazzoni, 2006), via a microphone

attached to a second PC, and were transcribed afterwards. These transcriptions were entered into Multiple Episode Protocol Analysis software (MEPA; Erkens, 2005) for coding and analysis.

After completing the self-assessment on the last task, participants were asked to rank the assessment criteria according to their perceived importance by assigning a unique score to each criterion from (1) most important, to (7) least important.

Post-test. The post-test (15 min.) was completed immediately after the LCI phase and was equivalent but not identical to the pre-test (i.e., tasks had similar structural features and complexity, but different cover stories). The cover stories in the pre-test and post-test differed not only from each other, but also from those used in the learning phase.

Data Analysis

Effective and ineffective participants. Participants' post-test scores (max.: 6) were recomputed using the pre-test scores (max.: 6) as a covariate to distinguish between the more effective and less effective learners (i.e., those who did and did not learn much from engaging in LCI). A two-step cluster analysis of these corrected post-test scores revealed two groups, which we will refer to as the "Effective" group (n = 17; 7 female, 10 male; M = 4.35, SD = .33) and the "Ineffective" group (n = 15; 8 female, 7 male; M = 2.27, SD = .04) from here on. Note that their was no difference in pre-test scores between the Effective group (M = 3.06, SD = 1.68) and the Ineffective group (M = 2.64, SD = 1.34), $F(1, 30) < 1$, ns. So the difference between the groups arose due to their actions in the learning environment.

Self-assessment accuracy. For conventional tasks, scores were automatically determined by the learning environment: each multiple-choice option that was either correctly selected or correctly left blank led to assignment of one point (i.e., max. = 3 points). For completion problems, the steps participants had worked out on paper were graded (1 point for correct completion of each of two steps, plus one point if they did not make a calculation error in completing both of those steps; max. = 3 points). To be able to compute self-assessment accuracy, the scales of the performance scores and performance self-assessments were made comparable (cf. Salden, Paas, & Van Merriënboer, 2006) by recoding scores on the nine-point self-assessment scale into scores on a four point scale (1/2 became 0; 3/4 became 1; 5/6/7 became 2; 8/9 became 3). Accuracy was then determined by subtracting the self-assessed performance scores from the actual performance scores, and was expressed in terms of absolute differences.

Coding scheme. The think aloud data were analyzed using coding schemes adapted from Biggs and Collis' (1982) SOLO taxonomy. For self-assessment, verbalizations were segmented per task and per assessment criterion. These segments could be given three codes. First, the self-assessment criterion (e.g., "solution",

"approach", etc.) was indicated. Second, the quality of the statements made about that criterion, were coded using five levels: (A) participants not only explained how they arrived at a particular self-assessment score, but also drew conclusions from that assessment, (B) participants explained their self-assessment score by comparing it with assessment or performance on a different task, (C) participants only explained in general terms how they arrived at a particular score, (D) participants merely mentioned the score, and (E) participants made only vague or no comments. In the case of (A), (B) or (C), a third code could be assigned, indicating which (if any) other self-assessment criteria were used in the explanation. For task selection, segmentation was based on the smallest meaningful unit (i.e., partial or whole sentences). These segments could also get three codes: First, the segment was coded as containing merely a statement about "what" task or task feature a participant was choosing, or also explaining "why" he or she was choosing that. Then, when a "what" code was given, it could get another code, either of mentioning a task's surface feature (cover story), the complexity level, or the level of support. When a "why" code was given, it could get another code of mentioning (a) one of the self-assessment scales, (b) prior tasks, (c) aspects of the task they are about to chose, or (d) future tasks. After codes (B), (C), and (D), a third level of codes could be assigned concerning statements about: surface features, complexity level, support level, planning of a task or set of tasks, or self-efficacy with regards to a task. Because segmenting was done based on small units, cases in which multiple consecutive segments would receive the same code, were counted as a single code.

Two raters independently coded circa 15% of the transcriptions to establish inter-rater reliability (self-assessment: Cohen's κ = .86; task selection: Cohen's κ = .88). Because the inter-rater reliability was high (i.e., .70 is considered sufficient; Van Someren, Barnard, & Sandberg, 1994), one rater coded the remaining protocols and this rater's codes were used in the analyses.

Results and Discussion

For all analyses a significance level of .05 is used. Cohen's *d* is reported as a measure of effect size for the *t*-test and partial eta-squared (η_p^2) for MANOVAs.

Self-assessment

A *t*-test on self-assessment accuracy (i.e., completion and conventional problems) showed that the Effective group (M = .56, SD = .30) was more accurate (i.e., lower deviation) than the Ineffective group (M = .91, SD = .56), $t(30) = 5.93$, $p < .001$, d = .75. However, within both the Effective and Ineffective groups there were participants with higher and lower levels of self-assessment accuracy. Two-step cluster

analysis revealed that each effectiveness group consisted of two accuracy subgroups (see Table 3.1).

	Ineffective		Effective	
	n	*M (SD)*	*n*	*M (SD)*
Accurate	7	.44 (.25)	9	.33 (.13)
Inaccurate	8	1.33 (.39)	8	.81 (.21)

Table 3.1: Distribution of participants over groups, including mean inaccuracy (i.e., deviation from 0; smaller = better)

Therefore, in further analyses both effectiveness and accuracy will be included as factors. It should be noted though, that none of these groups (Effective-Accurate; Effective-Inaccurate; Ineffective-Accurate; Ineffective-Inaccurate) can be considered truly accurate: one sample *t*-tests showed that their mean accuracy differed significantly from zero (all $p < .01$). For the sake of simplicity, however, we will refer to them as Accurate and Inaccurate subgroups, rather than the "more accurate" or "less accurate" groups.

A 2 x 2 MANOVA with effectiveness and accuracy as factors of the verbal protocol data (see Table 3.2) showed a main effect of effectiveness only on the frequency of category B ("compare"), $F(1, 28) = 4.34$, $p < .05$, $\eta_p^2 = .13$, indicating that Ineffective participants ($M = 6.60$, $SD = 3.91$) made more comparisons with assessments or performance on prior tasks compared to Effective participants ($M = 3.94$, $SD = 3.54$). There was no significant main effect of accuracy, $F(1, 28) < 1$, *ns*, nor a significant interaction, $F(1, 28) = 3.57$, *ns*. As indicated in the introduction, such attribution to prior performance may lead to misinterpretation of one's performance (Miller, 2003), and although we did not find a main effect for accuracy or an interaction, Effective participants were generally more accurate self-assessors than Ineffective participants.

A 2 x 2 MANOVA with effectiveness and accuracy as factors on the self-assessment criteria that were used as part of explanations for scores on other self-assessment criteria (applicable only when self-assessment quality was category A, B or C; see Table 3.3), showed a significant main effect only for "enjoyment", $F(1, 28) = 6.46$, $p < .05$, $\eta_p^2 = .19$, indicating that Ineffective participants ($M = 1.13$, $SD = .92$) used this criterion more often while making assessments on other criteria than Effective participants did ($M = .41$, $SD = .62$). There was no significant main effect of accuracy, $F(1, 28) = 1.40$, *ns*, nor a significant interaction, $F(1, 28) < 1$, *ns*. It is interesting that "enjoyment" seems to play a role in making assessments on other criteria for Ineffective participants because this does not seem particularly relevant to the other criteria. One would expect it to play a role in task selection, but there, no such differences between Effective and Ineffective participants were found (see Task Selection section below).

	Effective		Ineffective	
	Accurate	Inaccurate	Accurate	Inaccurate
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
A	.22 (.44)	0 (0)	0 (0)	.12 (.35)
B	.22 (.44)	1.38 (1.77)	2.00 (2.08)	1.62 (1.19)
C	3.56 (3.71)	5.25 (3.62)	4.43 (3.31)	8.25 (6.54)
D	6.22 (4.15)	5.62 (5.78)	8.43 (6.93)	7.12 (5.94)
E	.67 (1.41)	.88 (1.13)	.14 (.38)	.38 (.74)

Table 3.2: Average frequency of scores of self-assessment quality categories (A = highest, E = lowest) over all tasks

	Effective		Ineffective	
	Accurate	Inaccurate	Accurate	Inaccurate
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Solution	1.33 (1.58)	1.62 (1.06)	1.14 (1.46)	1.62 (1.92)
Approach	6.22 (5.33)	4.00 (3.34)	6.29 (4.61)	4.50 (4.14)
Time on task	2.00 (1.32)	2.25 (1.39)	3.00 (2.71)	1.25 (1.75)
Enjoyment	.22 (.44)	.62 (.74)	1.00 (.82)	1.25 (1.04)
Difficulty	4.67 (2.78)	4.62 (2.56)	4.43 (2.23)	3.88 (4.36)
Mental effort	.78 (1.09)	.75 (.89)	.71 (.95)	.25 (.46)
Total evaluation	0 (0)	.13 (.35)	0 (0)	.12 (.35)

Table 3.3: Average frequency of assessment criteria used to explain scores on other assessment criteria over all tasks.

Non-parametric tests of the ranking of self-assessment criteria gave insight into differences in perceived importance of these criteria between the four groups (see Table 3.4). Both the "solution", $\chi^2(3) = 9.27$, $p < .05$, and "approach", $\chi^2(3) = 13.64$, $p < .01$ differed significantly, which seems to indicate a lower ranking of "approach" in the Effective-Inaccurate group, and of "solution" in the Ineffective-Accurate group compared to the other groups. Interestingly, all groups perceived "time on task" to be the least important assessment criterion.

A *t*-test on the time spent assessing their performance indicated that there was no significant difference between Effective participants ($M = 69.26$ sec., $SD = 21.38$) and Ineffective participants ($M = 85.41$ sec., $SD = 60.58$), $t(30) = 1.03$, *ns*, and there was no difference in the number of words verbalized, between Effective participants ($M = 121.68$, $SD = 45.83$) and Ineffective participants ($M = 123.61$, $SD = 54.05$), $t(30) < 1$, *ns*. So the finding that Effective participants were relatively more accurate in their self-assessments than Ineffective participants, cannot be explained by them thinking longer about their assessment.

	Effective		Ineffective	
	Accurate	Inaccurate	Accurate	Inaccurate
Position	Criterion *(M)*	Criterion *(M)*	Criterion *(M)*	Criterion *(M)*
1	Approach (1.89)	Solution (1.50)	Approach (1.29)	Approach (1.50)
2	Solution (1.89)	Total (3.25)	Mental Effort (4.00)	Solution (2.50)
3	Total (3.78)	Mental Effort (3.37)	Solution (4.14)	Total (3.75)
4	Mental Effort (4.44)	Approach (3.62)	Total (4.14)	Enjoyment (4.25)
5	Difficulty (4.67)	Difficulty (5.12)	Enjoyment (4.43)	Difficulty (4.75)
6	Enjoyment (5.11)	Enjoyment (5.38)	Difficulty (4.71)	Mental Effort (4.75)
7	Time (5.22)	Time (5.75)	Time (5.29)	Time (6.50)

Table 3.4: Ranking of self-assessment criteria per group. Average ranking in parentheses (Lower = Higher ranking)

Task Selection

A 2 x 2 MANOVA with effectiveness and accuracy as factors of the verbal protocol data on task selection considerations showed no significant differences between conditions, $F(1, 28) < 1$, *ns*.

Self-assessment as Predictor of Task Selection

To investigate relationships between self-assessment and subsequent task selection, predictive models were built using linear regression analysis. Scores on the seven assessment criteria on conventional problems were used as parameters and the task selection step-size as the dependent variable. Step-size indicates the increase (or decrease) in task difficulty. Each level of support (see Figure 3.2) was numbered over complexity levels, with complexity level one, worked example being (1) and complexity level five, conventional problem being (15). So a step-size of +1 for example, would mean selection of a task with less support (from worked example to completion problem or from completion to conventional problem) within a complexity level, or when the previous task was a conventional problem, the selection of the first type of task (worked example) at the next complexity level.

No predictive models could be created for the Accurate- or Inaccurate-Ineffective participants (all $F < 1$, *ns*). A predictive model could be created for the Inaccurate-Effective participants, $F(3, 19) = 7.94$, $p < .01$, $R^2 = .56$, which included "solution" ($\beta = -.78$, $p < .05$), "time on task" ($\beta = -.61$, $p < .01$) and "total evaluation"

(β = .72, p < .05) as significant parameters (β describes the strength and direction of the parameter's prediction, and R^2 the strength of the model). This model indicates that a larger positive step-size (i.e., selecting a more difficult task) is predicted when assessment on the "solution" was lower, assessment on "time on task" was lower, and/or assessment on "total evaluation" was higher. For the Effective-Accurate participants, a predictive model could also be created, $F(3, 22) = 4.69$, $p < .05$, R^2 = .39, with "time on task" (β = -.50, p < .01), "difficulty" (β = -.82, p < .05) and "mental effort" (β = -.86, p < .05) as significant parameters. This model indicates that a larger positive step-size (i.e., selecting a more difficult task) is predicted when assessment of "time on task" was lower, when assessment of "difficulty" was lower, and/or assessment of "mental effort" was lower. Note though, that the fact that these assessment criteria predict the task selection step size, does not mean that participants are consciously using them in selecting tasks. Indeed the verbal protocol data show that the frequency of mentioning self-assessment criteria during task selection was very low (ranging from not at all for "solution", "mental effort" and "total evaluation" to a maximum of three times over all eight tasks for "time on task"). Moreover, assessment of time spent on the task was a significant predictor in both models, even though students ranked this criterion as least important (cf. Table 3.4).

Conclusion

This explorative study investigated differences in self-assessment and task-selection processes between novices who showed learning gains in LCI (effective learners) and those who did not (ineffective learners). It is important to stress here that there were no differences in pre-test scores between those groups, so their differential effectiveness resulted from their actions in the learning environment. It may come as a surprise that the effective group was rather large (the distribution was about 50-50), given that previous research has shown very little effects of LCI on learning for novice students. It is important to note though, that the task database as well as the tests used in this study were appropriate for the level of those novice learners, that is, even the highest complexity level was in principle attainable. This might not have been the case in previous studies which usually find learning effects of LCI only for high prior knowledge learners. If there still was something to learn for high prior knowledge learners in those studies, it can safely be assumed that the learning environment as well as the tests must have contained tasks at a complexity level that was way out of novices' reach.

Our first expectation that effective learners would be more accurate in their self-assessments of task performance than ineffective learners was confirmed. This finding seems to support the position that accurate self-assessment is beneficial for

making effective use of LCI, although it should be noted that our results also showed that: (a) more accurate does not mean absolutely (i.e., 100% accurate), and (b) within both effective and ineffective groups there were more and less accurate participants, which supports the notion that self-assessment accuracy is not the only factor influencing the effective use of LCI. An interesting question, however, is what makes the effective learners more accurate. It seems from the verbal protocol data that the ineffective learners assessed their performance more often by comparing it to their prior performance than effective learners, which could lead to misinterpretation of their performance (Miller, 2003). But our data do not explain what the effective learners compared their performance to. Given that they were all novices and did not differ in prior knowledge, it is unlikely that those participants knew more about assessment criteria and standards.

We did not find support for our second expectation that effective learners would consider different aspects of tasks during task selection than ineffective learners. Possibly, this is a result of the fact that all learners were novices, despite the fact that some learned more effectively in this LCI environment than others. Future research might consider the use of interview techniques or cued retrospective verbal reports (Van Gog, Paas, Van Merriënboer, & Witte, 2005) instead of or in addition to think-aloud procedures to gain further insight into potential differences in task-selection processes between effective and ineffective learners. Such techniques might provide the opportunity to uncover more information because they are used after task performance, rather than concurrent with task performance, which is less demanding in terms of cognitive resources and gives learners more time to verbalize their thoughts and motives.

Our third expectation, that effective learners would make more use of their self-assessment for task selection than ineffective learners, seems to be confirmed when looking at the predictive models, but no evidence for this was found in the verbal protocols. For ineffective learners, no predictive models could be created. For the effective but inaccurate learners, time on task, total (overall) evaluation, and quality of their solution were significant predictors. However, whereas the direction seems to make sense for “time on task” and “total evaluation” (lower score on time or higher score on total evaluation leads to larger step size, i.e., selection of a more difficult task), it is rather puzzling for “solution”. Why would someone who rates the quality of his or her solution as low, then go on to select a more difficult task? It is possible that self-efficacy beliefs played a role here. If those participants would have high self-efficacy, they might have been willing to select (and persist on) challenging tasks even if they rate the quality of their solution as low. However, this possible explanation requires further investigation, as the verbal protocols do not shed light on this question; they showed no differences between conditions. For the effective learners who were accurate self-assessors, a larger positive step size (i.e., selecting a more difficult task) was predicted when assessment of “time on task”

was lower, when assessment of "difficulty" was lower, and/or assessment of "mental effort" was lower. This seems logical, as increases in expertise are associated not only with better performance, but also with a decrease in the amount of mental effort and the time required to complete a task (see e.g., Van Gog, Paas, & Van Merriënboer, 2005). These variables are also often taken into account in adaptive system-controlled assessment and task selection algorithms (see e.g., Camp, Paas, Rikers, & Van Merriënboer, 2001; Corbalan et al., 2008; Kalyuga, 2006; Kalyuga & Sweller, 2004, 2005; Salden et al., 2006). However, it should be noted that performance of the effective inaccurate participants also improved, even though mental effort did not predict task selection for them. Moreover, it is very important to keep in mind that for both groups of learners, the verbal protocol data did not show that learners deliberately used these criteria. To investigate the extent to which learners may or may not be aware of their use of assessment criteria during task selection, future research might again consider the use of cued retrospective reporting or interview techniques instead of or in addition to think-aloud procedures.

Limitations

There are some limitations to this explorative study. First of all, the findings from this study have to be interpreted with some caution because the number of participants per subgroup was rather low. Post-hoc power analyses using G*Power (Erdfelder, Faul, & Buchner, 1996) showed that power on the accuracy analyses and regression models was relatively high despite the low number of participants, but on the verbal protocol data analyses power was rather low (which was likely due to the fact that frequencies were very low). Secondly, we did not measure students' self-efficacy beliefs or learning approaches, and it might be that these factors play a key role in the distinction between more and less effective learners or more accurate and more inaccurate self-assessors. Finally, we asked learners to think aloud only during self-assessment and task selection, which were the processes of primary interest in this study, as asking them to think aloud during the entire procedure might have overburdened them. However, it is possible that the effective and ineffective learners also engaged differently in the learning tasks they chose, for example, planning or monitoring their actions differently (cf. Azevedo et al., 2004).

Despite those limitations, this study did uncover some interesting differences in self-assessment and task-selection processes between novice learners who showed learning gains when having control over their instruction and those learners who did not. Future research could use these findings as input for the design of training to enhance students' self-assessment and task-selection abilities, which may in turn enhance the effectiveness of LCI for novice learners.

Future Research

First of all, it would be an interesting question for future research whether learners become more accurate self-assessors if they are trained in the use of relevant criteria and objective standards for performance. Our results suggest that such an increase in self-assessment accuracy might lead to more effective learning in LCI. Secondly, the effects of training learners on how to use certain assessment criteria during task selection could be investigated. Making learners aware of what assessment criteria are used by effective learners or by effective adaptive instructional systems, and how these are used, might enhance the quality of task selection and foster learning. A third interesting question is whether training in self-assessment and task selection would positively affect the ability to monitor performance while working on the learning tasks, as cognitive schemata are built that may guide attention to appropriate task aspects already during performance. This may or may not lower the overall cognitive load experienced, but at least it may direct cognitive load away from ineffective processes for performing the primary and secondary task (i.e., decrease extraneous load) to processes that contribute to performing those tasks (i.e., increase germane load; Sweller, Van Merriënboer, & Paas, 1998). Last but not least, when self-assessment and task-selection skills can be enhanced by training, it should be investigated whether this indeed positively affects learning in LCI.

Another interesting question in terms of cognitive load, is at what time such a training should be provided. It is possible to do this before learners engage in LCI, for example by means of examples that demonstrate how good self-assessors evaluate their performance and use this as input to select a next task. It may also be provided while they are engaging in LCI. For example, research on self-regulated learning in hypermedia environments has been concerned with influencing students' self-regulatory behaviour during study, for example by using prompts (Azevedo & Cromley, 2004; Bannert, 2004) or feedback (Narciss, Proske, & Koerndle, 2007). These procedures not only seem to positively affect self-regulation processes, but also the primary process of learning. Such guidance for task selection and self-assessment while engaging in LCI should, however, probably be reduced as expertise increases. From research on instructional guidance, it is known that more experienced learners no longer benefit from elaborate guidance and might even be hindered by it (Kalyuga, Ayres, Chandler, & Sweller, 2003).

To conclude, this study provided insight into differences in self-assessment and task-selection processes between effective and ineffective novice learners, which are not only interesting from a theoretical point of view, but also from a practical point of view, as they can be used in the design of a training for self-assessment and task-selection skills, which might improve the effectiveness of LCI.

Chapter 4

Training self-assessment and task-selection skills enhances the effectiveness of self-regulated learning

Abstract

For self-regulated learning to be effective, students need to be able to accurately assess their own performance on a learning task and use this assessment for the selection of a next learning task. Evidence suggests, however, that students have difficulties with accurate self-assessment and task selection, which may explain the poor learning outcomes often found with self-regulated learning. Experiment 1 investigated and confirmed the hypothesis that observing a human model engaging in self-assessment, task selection, or both could be effective for the acquisition of self-assessment and task-selection skills. Experiment 2 investigated and confirmed the hypothesis that acquisition of self-assessment and task–selection skills, either through examples or through practice, would enhance the effectiveness of self-regulated learning.

This chapter is submitted for publication as:
Kostons, D., Van Gog, T., & Paas, F. (2010). *Training self-assessment and task-selection skills enhances the effectiveness of self-regulated learning*. Manuscript submitted for publication.
Part of the data from Experiment 1 are also reported in:
Van Gog, T., Kostons, D., & Paas, F. (in press). Teaching students self-assessment and task-selection skills with video-based modeling examples. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Self-regulated learning is an active, constructive process in which learners plan, monitor, and control their own learning process (e.g., Pintrich, 2000a; Winne, 2001; Winne & Hadwin, 1998; Zimmerman & Schunk, 2001). Self-regulated learning can occur at different levels, from learners controlling how long they engage in studying a given task or whether they want to restudy it (Karpicke, 2009; Metcalfe, 2009; Thiede & Dunlosky, 1999), to learners controlling what information they want to study (e.g., in a hypermedia learning environment; Azevedo, 2005; Azevedo & Cromley, 2004; Azevedo, Moos, Greene, Winters, & Cromley, 2008) or what learning tasks they want to work on (Corbalan, Kester, & Van Merriënboer, 2008; Ross, Morrison, & O'Dell, 1989). This article focuses on self-regulated learning in which learners can choose their own learning tasks, which is often referred to as self-directed learning in the context of research on lifelong learning (Candy, 1991; Knowles, 1975; Loyens, Magda, & Rikers, 2008), and as learner-controlled instruction in the context of computer-assisted learning (Goforth, 1994; Merrill, 1980; Niemiec, Sikorski, & Walberg, 1996; Steinberg, 1989). Research has shown that having control over what information to study or what tasks to work on is not effective for novices' self-regulated learning (Azevedo et al., 2008; Lawless & Brown, 1997; Niemiec et al., 1996; Williams, 1996). We assume that this may be due to novices' lack of self-assessment and task-selection skills, which play a crucial role in this kind of self-regulated learning. To verify this assumption, we investigate whether training these skills results in higher self-assessment and task-selection accuracy (Experiment 1) and whether this in turn leads to higher learning outcomes attained through self-regulated learning (Experiment 2).

Providing learners with control over the learning tasks they work on, is believed to foster their self-regulated learning skills, and to result in personalized learning trajectories. Such personalized instruction is expected to enhance students' motivation and learning outcomes more than non-personalized instruction that is the same for all students. However, there is little evidence for both assumptions. First of all, research has shown that students do not apply and acquire self-regulation skills merely by engaging in self-regulated learning, rather, they need additional training or instructional support such as prompts or tutoring (e.g., Aleven & Koedinger, 2002; Azevedo & Cromley, 2004; Van den Boom, Paas, & Van Merriënboer, 2007; Van den Boom, Paas, Van Merriënboer, & Van Gog, 2004). Secondly, although the assumption that personalized instruction can foster learning compared to non-personalized instruction seems to be correct (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995; Camp, Paas, Rikers, & Van Merriënboer, 2001; Koedinger, Anderson,

Hadley, & Mark, 1997; Salden, Paas, Broers, & Van Merriënboer, 2004), it is questionable whether self-regulated learning actually results in the adaptivity to students' needs that is required for effective personalized instruction.

When an instructional system is used to personalize instruction, it does so by monitoring and assessing a student's current level of knowledge and skill to select or suggest an appropriate next learning task. The assessment can comprise several aspects of students' performance (e.g., Anderson et al., 1995; Kalyuga & Sweller, 2004; Koedinger et al., 1997) or a combination of their performance and invested mental effort (e.g., Camp et al., 2001; Corbalan et al., 2008; Kalyuga, 2006; Salden et al., 2004). For self-regulated learning to be equally adaptive and effective, students should be able to accurately monitor and assess their own performance and recognize what an appropriate next task would be. However, there is quite some evidence that students, particularly novices who lack prior knowledge of the learning tasks, are not very accurate at monitoring, self-assessment, and task selection.

By monitoring one's own performance while working on a task, a student can construct a mental representation of the task performance process, which is a prerequisite for accurate self-assessment. However, both activities compete for limited working memory resources, which might become problematic under conditions of high cognitive load, as monitoring, task performance, or both may be negatively affected by a lack of resources (Van Gog, Kester, & Paas, in press). Most learning tasks impose a high cognitive load, especially for novice learners (Sweller, Van Merriënboer, & Paas, 1998). When concurrent monitoring is hampered, learners will have a poor recollection of their performance, which may hamper their self-assessment of that performance after the task (see also Kostons, Van Gog, & Paas, 2009 –Chapter 2).

But even with a good recollection of performance, accurate self-assessment is not guaranteed. Self-assessment may also be hampered by several biases that may cause learners to depend on the wrong kind of cues to assess their performance (for a review, see Bjork, 1999), such as hindsight bias (i.e., once an answer or solution procedure is known, e.g., after feedback, students are more likely to think that they could have produced it themselves), or availability bias (i.e., answers that come to mind easily are not only more likely to be provided but are also more likely to be assumed correct). Moreover, accurate self-assessment also seems to require some domain expertise (Dunning, Heath, & Suls, 2004; Dunning, Johnson, Erlinger, & Kruger, 2003). Individuals with higher levels of prior knowledge have been found to be more accurate self-assessors. This may be because their experience lowers the cognitive load imposed by the learning task (see Sweller et al., 1998), allowing them to devote more cognitive resources to monitoring their task performance, which likely provides them with a more accurate memory representation on which to base their assessment (Van Gog et al., in press). In addition, individuals with more prior knowledge might be more accurate self-assessors because their experience pro-

vides them with more knowledge of the criteria and standards that good performance should meet (Dunning et al., 2003, 2004).

Inaccurate self-assessment, in turn, may negatively affect selection of an appropriate new learning task. For example, if students overestimate their performance, they may choose a task that is too difficult for them (cf. Azevedo, Guthrie, & Seibert, 2004; Moos & Azevedo, 2008b; Shapiro, 2004). Learners should ideally select tasks that are challenging, but not too difficult (cf. Vygotsky's, 1978, concept of 'zone of proximal development'). In practice, however, learners tend to select tasks that they are confident they can perform (i.e., self-efficacy, Bandura, 1997), which often results in selection of tasks that are too difficult in case of overconfidence, or tasks that are too easy in case of low confidence (Pajares & Kranzler, 1995; Stone, 1994). Moreover, even when self-assessment is accurate, novices may still experience problems in selecting appropriate learning tasks. When selecting a task, it is important to discern which aspects of a task are relevant for learning, such as the structural features of the task (e.g., type of task, complexity level, amount of support provided), from aspects that are less relevant, such as superficial cover stories. Research has shown that novices may experience difficulties in discerning between these aspects (Chi, Glaser, & Rees, 1982; Quilici & Mayer, 2002; Ross, 1989) and tend to choose tasks based on irrelevant aspects (Ross & Morrison, 1989). When task selection is inaccurate, the chosen tasks are unlikely to be adaptive to the learner's level of prior knowledge or skills, so learners will end up working on tasks that are not aligned with their learning needs.

In sum, inaccuracies in self-assessment and task selection may lead to ineffective self-regulated learning. Support for this assumption comes from studies that have shown that providing novice learners with control over their learning process may have beneficial effects on their motivation or involvement (e.g., Corbalan et al., 2008; Schnackenberg & Sullivan, 2000), but has detrimental effects on learning outcomes when compared to teacher or computer controlled fixed or personalized instruction (see e.g., Azevedo et al., 2008; Lawless & Brown, 1997; Niemiec et al., 1996; Williams, 1996). Beneficial effects on learning outcomes attained through self-regulated learning have been found mainly for learners with higher levels of prior knowledge (Lawless & Brown, 1997; Moos & Azevedo, 2008a, b; Niemiec et al., 1996; Scheiter & Gerjets, 2007; Schnackenberg & Sullivan, 2000; Steinberg, 1989), who, as mentioned above, are better able to monitor and assess their own performance than novices. In addition, Kostons, Van Gog, and Paas (2010 –Chapter 3) investigated whether secondary education students who differed in the amount of knowledge gained from studying in a self-regulated learning environment, also differed in self-assessment and task-selection skills. The results indicated that students who gained more knowledge, were also more accurate self-assessors, and made better use of that assessment when selecting new tasks.

Given the important role that self-assessment and task-selection skills play in the effectiveness of self-regulated learning, an important question is whether novice learners can be trained to become more accurate self-assessors and task selectors (Experiment 1), and even more importantly, whether this higher accuracy improves the learning outcomes they attain through self-regulated learning (Experiment 2).

Experiment 1

Experiment 1 aimed to investigate whether an example-based training of self-assessment and task-selection skills leads to higher accuracy of self-assessment and task selection. Research on *worked examples* inspired by cognitive theories such as ACT-R (Anderson, 1993) and cognitive load theory (Sweller, 1988; Sweller et al., 1998) and research on *modelling examples* inspired by social learning theory (Bandura, 1986) and cognitive apprenticeship (Collins, Brown, & Newman, 1989) has shown that both of these types of example-based learning are highly effective during the initial stages of skill acquisition (for reviews, see Atkinson, Derry, Renkl, & Wortham, 2000; Van Gog & Rummel, 2010). These types of example-based learning differ in the format of the examples. Whereas worked examples are primarily based on a written account of a model's problem-solving procedure, modelling examples involve observing a model performing the task, which can take a variety of forms, not only live observation, but also watching a video in which the model is visible (e.g., Braaksma, Rijlaarsdam, & Van den Bergh, 2002), a video consisting of a screen capture of the model's computer screen in which the model is not visible (though s/he can be heard when a spoken explanation of what s/he is doing is provided; e.g., McLaren, Lim, & Koedinger, 2008; Van Gog, Jarodzka, Scheiter, Gerjets, & Paas, 2009), or an animation in which the model is represented by a pedagogical agent (e.g., Atkinson, 2002; Wouters, Paas, & Van Merriënboer, 2009). They also differ in the kinds of tasks for which they are used. Worked examples have until recently been used mainly to teach procedures for solving highly structured problems in domains such as algebra (e.g., Cooper & Sweller, 1987; Kalyuga & Sweller, 2004; Sweller & Cooper, 1985), probability (e.g., Berthold & Renkl, 2009; Catrambone, 1995, 1996), statistics (e.g., Atkinson, Catrambone, & Merrill, 2003; Paas, 1992; Quilici & Mayer, 1996), geometry (e.g., Kalyuga & Sweller, 2004; Paas & Van Merriënboer, 1994; Schwonke, Renkl, Krieg, Wittwer, Aleven, & Salden, 2009), physics (e.g., Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Van Gog, Paas, & Van Merriënboer, 2006), or learning to use a database program (Tuovinen & Sweller, 1999). Several recent studies have applied worked examples with equal success in teaching less structured tasks such as argumentation (Schworm & Renkl, 2007), or learning to recognize designer styles in art education (Rourke & Sweller, 2009). Modelling examples have been used more often than worked examples with less structured

tasks such as writing (e.g., Braaksma et al., 2002; Zimmerman & Kitsantas, 2002), assertive communication (e.g., Baldwin, 1992), or collaboration (e.g., Rummel & Spada, 2005), and to a lesser extent for teaching problem-solving skills in highly structured domains such as mathematics (e.g., Schunk, 1981; Schunk & Hanson, 1985) or chemistry (e.g., McLaren et al., 2008). Interesting from the perspective of the goal of our study, is that modelling examples have been used for teaching *meta*-cognitive skills as well, for example in trying to improve dart-throwing skills (Kitsantas, Zimmerman, & Cleary, 2000) or writing skills (Zimmerman & Kitsantas, 2002). Therefore, we used modelling examples consisting of computer screen-recordings with spoken text to investigate our hypothesis that training secondary education students' self-assessment and task-selection skills would enhance the accuracy of those skills. By providing students with either no self-assessment and task-selection examples, only self-assessment examples, only task-selection examples, or both self-assessment and task-selection examples, we were not only able to investigate whether self-assessment and task selection can be trained, but also to determine whether an increase in accuracy in one skill would transfer to the other, for example, whether an increase in self-assessment accuracy would also lead to an increase in task-selection accuracy even if the latter was not modelled.

Method

Participants and design

Participants were 80 Dutch students (44 female, 36 male; age M = 15.23, SD = .53) in their fourth year of pre-university education (i.e., the highest level of secondary education in the Netherlands with a six year duration). A 2 x 2 factorial design was used with factors Self-Assessment Modelling Examples (Yes vs. No) and Task-Selection Modelling Examples (Yes vs. No). Participants were randomly assigned to one of the four conditions: (1) self-assessment and task-selection modelling examples (SA + TS; n = 20), (2) only self-assessment modelling examples (SA; n = 20), (3) only task-selection modelling examples (TS; n = 19), or (4) no self-assessment or task-selection examples (NO; n = 21).

Materials

Pre-test and post-test. The pre-test and post-test consisted of five paper and pencil heredity problems on the laws of Mendel, at five levels of complexity. The problems were presented in random order (i.e., not in order of complexity) and the order differed between the pre-test and post-test. The post-test contained problems that were equivalent but not identical to the pre-test problems: they had simi-

lar structural features but the surface features (cover stories) differed. The problems could be solved in five steps: (1) translating the phenotypes (i.e., expressions of genetic traits) described in the cover story into genotypes (a pair of upper and/or lower case letters representing genetic information); (2) putting these genotypes into a hereditary diagram; (3) determining the number of required Punnett Squares by looking at the direction of reasoning (deductive/inductive); (4) filling in the Punnett Square(s); and (5) extracting the final solution(s) from the Punnett Square(s). On both tests, participants were instructed to not only provide the final answer, but also write down the steps they took to reach the solution.

Mental effort rating. After each problem in the pre-test and post-test, participants rated how much mental effort they had invested in solving the problem on the nine-point rating scale developed by Paas (1992) which ranges from (1) "very, very little effort" to (9) "very, very much effort".

Self-assessment. After completing the mental effort scale, participants self-assessed their performance on tasks in the pre-test and post-test on a six point rating scale ranging from 0 to 5. After the experiment, participants' performance was scored by the experimenter on the same scale, assigning one point for each correct step, that is, (0) indicated "none of the steps correct" and (5) indicated "all steps correct" (i.e., max. score per problem: 5; max. score on the test: 25).

Task selection. After self-assessment, participants indicated on an overview of the task database (see Figure 4.1) what problem they would have selected next. Note, though, that they did not get to work on that problem because the tasks in the pre-test and post-test were the same for all students. The task database contained tasks at five levels of complexity (left column) and at each level, there were tasks that contained three levels of support: 1) high support: completion problems (i.e., partially worked-out examples that the learner has to complete, see Paas, 1992) with many steps already worked out and few for the learner to complete (white row); 2) low support: completion problems with few steps already worked out and many for the learner to complete (light gray row); and 3) no support: conventional problems which participants had to complete entirely (dark gray row). At each level of support within each complexity level, there were five tasks to choose from, which consisted of different cover stories. Before selecting the task, participants were informed what the complexity level was of the problem they had just worked on.

Complexity Level	Support Level	Learning Tasks				
Complexity 1 - 2 generations - 1 unknown - 1 solution - Deductive	Completion High support	Eye color	Hair structure	Shapes cat fur	Japanese Apple tree	Depression
	Completion Low support	Eye color	Hair structure	Sickle cell Anemia	Curve chicken beak	Guinea Pigs
	Conventional No support	Eye color	Hair structure	Huntington	Milk Allergy	Cleft Lip
Complexity 2 - 2 generations - 1 unknown - Multiple solutions - Deductive	Completion High support	Eye color	Hair structure	Flower color	Widow's peak	P.R.A.
	Completion Low support	Eye color	Hair structure	Shapes cat fur	Albinism	Pea plant
	Conventional No support	Eye color	Hair structure	Tongue Curling	Japanese Apple tree	Fruit flies
Complexity 3 - 2 generations - 1 unknown - Multiple solutions - Inductive	Completion High support	Eye color	Hair structure	Fruit flies	Curve Chicken beak	Wolfram syndrome
	Completion Low support	Eye color	Hair structure	Dog tail length	Japanese Apple tree	Milk allergy
	Conventional No support	Eye color	Hair structure	Freckles	Flower Color	Earlobes
Complexity 4 - 3 generations - 1 unknown - Multiple solutions - Both ways	Completion High support	Eye color	Hair structure	Albino	Shapes cat fur	Fruit flies
	Completion Low support	Eye color	Hair structure	Fruit flies	Tongue Curling	Flower color
	Conventional No support	Eye color	Hair structure	Pea plant	Dimples	Depression
Complexity 5 - 3 generations - 2 unknowns - Multiple solutions - Both ways	Completion High support	Eye color	Hair structure	Milk Allergy	Depression	Huntington disease
	Completion Low support	Eye color	Hair structure	Dog tail Length	Wolfram syndrome	Flower color
	Conventional No support	Eye color	Hair structure	Cystic Fibrosis	Fruit flies	Rat fur

Figure 4.1: Task selection database

Modelling examples. Participants were given four modelling examples consisting of digital videos of the model's computer screen recorded with Camtasia Studio, along with a spoken explanation by the model. The gender of the models was varied: two examples were by two different male models, and the other two examples were by two different female models (because the model's gender might possibly influence students' learning by affecting self-efficacy; Schunk, 1987). The four examples either showed the model solving a heredity problem (NO), the model solving a heredity problem and assessing his or her own performance (SA), the model solving a heredity problem and selecting a new task based on a performance score that was presented as a given and not further explained (TS), or the model solving a heredity problem, assessing his or her own performance, and selecting a new task (SA+TS), depending on the assigned condition. The content of the examples was as follows:

(1) Problem solving (all conditions). The model performed the problem solving task. Two models worked on problems of complexity level 1, and two models

worked on problems of complexity level 2 (i.e., of the five complexity levels present in the task database and in the pre-test and post-test; see Table 4.1). The quality of the models' performance varied between the examples: the first example showed a model accurately solving the problem, but in the other three examples the models made one or more errors (see Table 4.1). This was done to create variability in phases 2 and 3 of the examples, that is, in the model's self-assessment scores and task selections (i.e., if the model would not make any errors or would detect and correct them immediately, they would always have the highest possible self-assessment score. Following task performance, the models rated their invested mental effort on the nine point rating scale.

Example	Model	Problem-Solving Performance	Complexity Level
1	Male 1	0 errors	Level 1
2	Female 1	2 errors	Level 1
3	Male 2	4 errors	Level 2
4	Female 2	1 error	Level 2

Table 4.1: Modelling example characteristics

(2) Self-assessment (SA and SA+TS conditions): The models assessed their performance on the 6-point rating scale, assigning themselves one point for each correct step. The models' self-assessment was always accurate.

(3) Task selection (TS and SA+TS conditions): Then, the model selected a new task based on a combination of the performance score and the mental effort score. The models used a table (see Figure 4.2) in which the relationship between performance and mental effort scores was depicted. This table could be used to infer a recommended 'step size' for task selection. For example, a performance score of three, and a mental effort score of two, would result in a step size of +2, which essentially indicates the number of rows one is recommended to go back or progress to in the second column from the left in Figure 4.1. A positive step size means a recommendation to select a more challenging task (i.e., less support or higher complexity level), a step size of zero means repeating a comparable task (i.e., same level of support and same complexity level), and a negative step size means a recommendation to select a simpler task (i.e., higher level of support or lower level of complexity). This kind of task-selection algorithm, based on performance and mental effort scores, has proven to lead to an effective learning path in studies on adaptive, personalized task selection (e.g., Camp et al., 2001; Corbalan et al., 2008; Kalyuga, 2006; Salden et al., 2004). The models' task selection was always accurate.

Performance / *4-5*	+2	+1	0
2-3	+1	0	-1
0-1	0	-1	-2
	1-3	*4-6*	*7-9* / *Effort*

Figure 4.2: Determining task-selection step-size using performance and mental effort scores

Participants in the NO condition observed the model's only performing the problem-solving task. In the time in which the participants in the other conditions observed the model's self-assessment and/or task selection, participants in the NO condition were instructed to indicate whether the model made any errors during task performance, and if so, what the errors were and what the correct step would have been. Finding and fixing errors in examples may foster the acquisition of problem solving skills (see Große & Renkl, 2007), and it can also be expected to direct students' attention towards assessment of performance (of the model) to some extent.

Procedure

The experiment was conducted in a computer room at the participants' school in sessions of approximately 70 minutes duration, with 10 to 24 students per session. Prior to the experiment, participants had been randomly assigned to conditions. First, all participants completed the pre-test on paper. They were given four minutes to complete each problem, followed by one minute for assessing their performance and selecting a next learning task (Chapter 3 had shown this to be sufficient time for solving conventional problems). Participants were not allowed to proceed to the next problem before the time was up and time was kept by the experimenter using a stopwatch. After completing the pre-test, participants studied the modelling examples on the computer which varied according to their assigned condition (see materials section). Each participant had a headset for listening to the

model's explanations. Finally, all participants completed the post-test on paper, according to a procedure similar to the pre-test.

Data analysis

Self-assessment accuracy was determined by computing the absolute difference between participants' objective performance score and their self-assessed performance score. The lower this difference, the more accurate participants' self-assessment was (i.e., 0 = 100% accurate). Each participant's mean self-assessment accuracy was computed for the post-test, not including those problems on which both the objective and the subjective assessment were zero. This was done because we did not want to overestimate self-assessment accuracy; it is logical that students can state quite accurately that they were not able to solve a problem at all, and this is probably not very indicative of their self-assessment skill. This is also the reason why we will not analyze mean self-assessment and task-selection accuracy on the pre-test: in both conditions there were too many problems that had not even been partially solved. We included self-assessment and the task-selection ratings in the pre-test so that participants would get acquainted with them, as the models in the experimental conditions also used these and participants had to use them again during the post-test.

Task-selection accuracy was determined by first recoding subjective (i.e., self-assessed) performance scores and the mental effort scores indicated by the participants according to the table that was also used for the task-selection by the model (see Figure 4.2), and then reading off the recommended step-size from this table. Recommended step size essentially indicates the number of rows one is recommended to go back or progress to in the second column from the left in Figure 4.1. For example, if a participant had completed the conventional problem at complexity level 2 and was recommended to progress two steps, this recommendation meant to choose a task with low support at complexity level 3. The absolute difference between the recommended step size and the step size chosen by the participant was then computed to indicate task-selection accuracy based on participants' self-assessed performance score. For five participants, task-selection accuracy could not be computed due to missing values in mental effort or task-selection data. One could also compute task-selection accuracy based on the objective performance scores, but we preferred to use the self-assessed performance score because this does not penalize participants for inaccuracies in their self-assessment, while using the objective performance score would.

Results

Data were analyzed using 2 x 2 ANOVAs with Self-Assessment Modelling Examples and Task-Selection Modelling Examples as factors, and the significance level set at .05. Partial eta squared (η_p^2) is reported as a measure of effect size, 0.01, 0.06, and 0.14 corresponding to small, medium, and large effect sizes, respectively.

Learning gains

Pre-test data from two participants were missing. Participants' mean performance on the pre-test problems was 3.42 (*SD* = 5.01), and their mean performance on the post-test problems was 15.74 (*SD* = 6.08), so they acquired procedural skills for solving heredity problems from the modelling examples. A 2 x 2 ANOVA on the pre-test to post-test knowledge gain did not show any significant differences between conditions, $F(3, 74) < 1$, *ns*.

Self-assessment accuracy

In line with our hypothesis, a 2 x 2 ANOVA on self-assessment accuracy showed a significant main effect of the factor Self-Assessment Modelling Examples, indicating that participants who had studied self-assessment modelling examples were more accurate (*M* = .81, *SD* = .54) than participants who had not studied those examples (*M* = 1.23, *SD* = .75), $F(1, 76) = 8.04$, *MSE* = 33.08, $p = .006$, $\eta_p^2 = .10$. No main effect of Task-Selection Modelling Examples nor an interaction effect was found, both $F(1, 76) < 1$, *ns*.

Task-selection accuracy

In line with our hypothesis, a 2 x 2 ANOVA on task-selection accuracy showed a significant main effect of Task-Selection Modelling Examples, indicating that participants who had studied task-selection modelling examples were more accurate (*M* = 2.31, *SD* = 2.18) than participants who had not studied those examples (*M* = 3.96, *SD* = 2.06), $F(1, 71) = 11.57$, *MSE* = 323.27, $p = .001$, $\eta_p^2 = .14$. No main effect of Self-Assessment Modelling Examples, nor an interaction effect was found, both $F(1, 71) < 1$, *ns*.

Discussion

Results of this experiment showed that students can acquire self-assessment and task-selection skills, which are considered to play a pivotal role in the effectiveness

of self-regulated learning, from studying modelling examples. The results suggest that both skills need to be explicitly trained, as we found no indications (i.e., no interaction effects) that an increase in for example self-assessment accuracy also led to an increase in task-selection accuracy when the latter was not modelled. For reasons of experimental control, all students received the same tasks on the post-test, so in this experiment they did not actually get to work on the tasks they had selected. A very important question therefore, which we will address in Experiment 2, is whether students can apply the self-assessment and task-selection rules they acquired from studying modelling examples in a self-regulated learning environment in which they are allowed to select which problems to work on. If they are able to do so, we would expect this to enhance the learning outcomes attained through self-regulated learning.

In this experiment, we chose modelling examples as a means to train self-assessment and task-selection skills, because research has shown that example-based learning is a powerful instructional strategy. Thus far, in educational settings, examples have mostly been used for teaching cognitive skills, and this study adds further evidence that they are useful for teaching metacognitive skills as well (see also Kitsantas et al., 2000; Zimmerman & Kitsantas, 2002). A lot of research, especially on learning from worked examples, has demonstrated that for the acquisition of problem-solving skills instruction consisting of studying examples is more effective for novices than instruction consisting of practicing problem solving (see Atkinson et al., 2000; Sweller et al., 1998; Van Gog & Rummel, 2010). In this study, we did not investigate whether training self-assessment and task-selection skills via modelling examples was more effective than training those skills in some other way, for example via practice after having been explained the assessment and selection rules (i.e., how to come to a performance assessment score and how to combine performance and mental effort scores to select a new task). Therefore, Experiment 2 also investigated the effectiveness of examples compared to practice with self-assessment and task-selection rules.

Experiment 2

This experiment investigated the effects of teaching self-assessment and task-selection skills on the effectiveness of self-regulated learning. Learning outcomes attained after engaging in self-regulated learning will be compared for students who were not taught those skills, students who were taught those skills via modelling examples as in Experiment 1, and students who were taught those skills by explaining them the self-assessment and task-selection 'rules' and then allowing them to practice application of these rules by having them assess the model's performance and subsequently select a new task for the model based on that assessment. This

third condition thus involves a kind of 'peer-assessment'. Peer-assessment is often implemented not only as a grading procedure, but also as a means to foster the development of both content knowledge and assessment skills (Dochy, Segers, & Sluijsmans, 1999). It has been suggested that engaging in peer-assessment activities might also develop self-assessment skills (Dochy et al., 1999; Somervell, 1993). However, only a few studies have tried to empirically demonstrate that assessing a peer's performance may subsequently improve self-assessment skills (see Oldfield & Macalpine, 1995; Searby & Ewers, 1997).

It is hypothesized that training self-assessment and task-selection skills leads to higher learning outcomes attained after self-regulated learning. Whether training consisting of examples or of practice is more effective, is an open question. On the one hand, based on findings concerning the acquisition of problem-solving skills, we might expect example-based learning to be more effective. On the other hand, in studies comparing learning from worked examples and learning from practicing with problem solving, students are not generally provided with any information concerning how to solve the problem. In this study, they are first explained the rules before they have to apply them to the model's performance.

Method

Participants and design

90 Dutch students (50 female, 40 male; age M = 14.66, SD = .71) in their third year of Higher General Secondary Education (the second highest level of secondary education in the Netherlands, with a 5 year duration) participated in this experiment. Participants were randomly assigned to one of three conditions: (1) self-assessment and task-selection skills taught via modelling examples (Modelling; n = 32); (2) self-assessment and task-selection skills taught via practice (Practice; n = 25), and no teaching of self-assessment and task-selection skills (Control; n = 33).

Materials

Pre-test and post-test. The pre-test and post-test were the same as in Experiment 1.

Mental effort rating. The same 9-point rating scale was used as in Experiment 1.

Self-assessment/peer-assessment. The same 6-point rating scale was used as in Experiment 1.

Task selection. The same procedure was used as in Experiment 1. During the self-regulated learning phase, students could click on the task they wanted to per-

form in the overview of the database (see Figure 4.1) that was visible on their computer screen and then received that task to work on.

Training conditions. The Control (i.e., no training) condition was the same as in Experiment 1: participants observed the video of the model performing the problem-solving task and then had to find and fix errors in the model's performance. The Modelling condition was the same as the SA+TS condition in Experiment 1. In the Practice condition, participants were first explained the self-assessment and task-selection rules that were also used by the modelling examples, but without a concrete example. That is, concerning assessment, it was explained that each of the five steps in the problem-solving process could be right or wrong and contributed equally to the total assessment score. Concerning task selection, it was explained how mental effort and performance scores could be combined to infer a recommended step size, which could then be used to determine the next task by going back or progressing the recommended number of rows in the database overview. Then, they observed the video of the model performing the problem-solving task, assessed the model's performance, and selected a new task for the model. They were reminded of the rules immediately before assessment and selection was required.

Self-regulated learning. An electronic learning environment was used for this study (adapted from Kostons et al., 2010 - Chapter 3) consisting of a Web application with a database connected to it that contained all learning tasks (see Figure 4.1). The environment logged participants' answers to the problems, responses on mental effort and self-assessment rating scales, and task-selection choices. Participants had to go through the cycle of selecting a task, performing the task, rating their mental effort, and assessing their performance eight times. A maximum of five minutes was allotted per task (a previous study showed this to be more than enough time; see Chapter 3). A visual and auditory warning was given when only one minute was left, and the system automatically continued to the self-assessment phase once time was up. One restriction on task selection was implemented of which participants were informed beforehand and were reminded of each time they went to the task-selection screen: They needed to complete at least one conventional task within a complexity level (correctly or incorrectly) before they could proceed to a higher complexity level (cf. Corbalan et al., 2008; see also Chapter 3). This rule had been implemented for two reasons. First, this prevented students from choosing only tasks with high levels of support, allowing them to test whether they could perform the task without any support, as they would have to do on the test. Second, this led to at least some conventional tasks being performed during the SRL-phase, allowing for analyses on self-assessment and task-selection accuracy during SRL (see data analysis section). This rule had also been explained in the task-selection training (i.e., in the modelling examples and practice conditions).

Procedure

The experiment was run at the participants' school in sessions of approximately 110 minutes duration, with 9 to 20 students per session. Participants had been randomly assigned to conditions prior to the experiment. They first completed the pre-test, according to the same procedure used in Experiment 1. After the pre-test, participants in the Modelling condition observed the four problem solving, self-assessment and task-selection modelling examples. Participants in the Practice condition received the explanation of the rules, observed the model solving the problem and then assessed the model's performance and selected a next task for the model four times. Participants in the Control condition observed the model solving the problem and then engaged in finding and fixing the mistakes in the model's demonstrated performance four times. Then, all participants engaged in self-regulated learning in the electronic learning environment, selecting learning tasks, performing those tasks, rating their mental effort, and self-assessing their performance eight times. Finally, participants proceeded to the post-test, completed according to the same procedure as in Experiment 1.

Data analysis

Self-assessment and task-selection accuracy on the post-test and during the self-regulated learning phase were determined according to the same procedure as used in Experiment 1. Mean task-selection accuracy could not be computed for 31 participants due to missing values in either mental effort or task selection (these participants were rather equally distributed across conditions: Modelling: n = 10, Practice: n = 9, Control: n = 12). During the self-regulated learning phase, self-assessment and task-selection accuracy was only computed for conventional problems, because the completion problems provided support consisting of worked-out steps, which had consequences for self-assessment (i.e., assigning oneself a point for those steps, is not a judgment of whether one has correctly performed it, but of whether one is confident that one could correctly perform it) and because of the task-selection rule, restrictions in task selection applied as well. Note that not all participants performed conventional problems, and that the mean accuracy may be based on different numbers of conventional problems for different participants.

Results

Means and standard deviations of pre-test and post-test performance, learning gains, self-assessment, and task-selection accuracy per condition are provided in Table 4.2. ANOVAs with planned contrasts (significance level of .05) were used to

test our hypotheses. Cohen's *d* is provided as a measure of effect size, with 0.25, 0.50, and 0.80 corresponding to small, medium, and large effect sizes, respectively.

Learning gains

In line with our expectation, the Control condition gained less knowledge than both the Modelling condition ($t(87)= 2.68$, $p = .005$, one-tailed, $d = 0.64$) and the Practice condition ($t(87)= 2.27$, $p = .013$, one-tailed, $d = 0.61$). The Modelling and the Practice conditions did not differ from each other ($t(87) < 1$, *ns*).

	Modelling	Practice	Control
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
Pre-test score (max. = 25)	3.69 (3.73)	3.00 (2.92)	3.91 (3.36)
Post-test score (max. = 25)	15.91 (6.42)	14.84 (5.92)	12.06 (7.07)
Learning gain (max. = 25)	12.22 (6.03)	11.84 (5.44)	8.15 (6.66)
Self-assessment accuracy post-test (lower = better)	1.41 (.78)	1.55 (.48)	1.71 (.68)
Self-assessment accuracy self-regulated learning (lower = better)	1.52 (1.23)	1.22 (1.15)	1.47 (0.89)
Task-selection accuracy post-test (lower = better)	4.20 (1.49)	4.87 (2.13)	4.28 (1.87)
Task-selection accuracy self-regulated learning (lower = better)	2.28 (1.60)	2.25 (1.60)	3.90 (2.67)

Table 4.2: Means (and *SD*) of pre-test and post-test performance and accuracy data per condition

Self-assessment accuracy

In line with our expectation, self-assessment accuracy on the post-test was higher in the Modelling condition than in the Control condition ($t(87)= 1.81$, $p = .037$, one-tailed, $d = 0.41$). In contrast to our hypothesis, there was no significant difference between the Control condition and the Practice condition ($t(87) < 1$, ns). In addition, the Modelling and the Practice conditions did not differ from each other ($t(87) < 1$, *ns*). There were no significant differences between conditions in self-assessment accuracy on the conventional problems completed during the self-regulated learning phase (all $t(64) < 1$, *ns*).

Task-selection accuracy

In contrast to our expectation, there were no significant differences between conditions in task-selection accuracy on the post-test (all $t(56) < 1$, *ns*). However, there

were significant differences between conditions in task-selection accuracy on the conventional problems completed during the self-regulated learning phase. In line with our expectation, accuracy in the Modelling condition was higher than in the Control condition ($t(63) = 2.74$, $p = .004$, one-tailed, $d = 0.74$) and accuracy in the Practice condition was higher than in the Control condition ($t(63) = 2.52$, $p = .007$, one-tailed, $d = 0.75$). The Modelling and the Practice conditions did not differ from each other ($t(63) < 1$, *ns*).

Discussion

This experiment showed that in line with our hypothesis, training students' self-assessment and task-selection skills enhanced the effectiveness of self-regulated learning in terms of learning gains. Training via modelling and practice seemed to be equally effective. One would expect the increase in learning gains in these conditions compared to the control condition to be due to increased self-assessment and task-selection accuracy. The results partially support this assumption, but are not unequivocal. Regarding self-assessment accuracy, the modelling condition indeed outperformed the control condition on the post-test, but the practice condition did not. Moreover, no differences between conditions in self-assessment accuracy during self-regulated learning were found. Regarding task-selection accuracy, no significant differences were found on the post-test, but the modelling and practice conditions significantly outperformed the control condition during self-regulated learning.

General Discussion

Our main aim was to investigate whether self-assessment and task-selection accuracy can be increased through training and whether such training enhances the effectiveness of self-regulated learning. Experiment 1 demonstrated that training consisting of observing human models who engage in self-assessment and task selection improved students' self-assessment and task-selection skills. Experiment 2 showed that training self-assessment and task-selection skills, either through modelling as in Experiment 1 or by being explained the rules and then practicing those skills by assessing the model's performance and selecting a new task for the model, indeed enhanced the effectiveness of self-regulated learning.

Previous research on improving self-regulated learning in hypermedia environments has shown that training can improve students' application of self-regulation activities such as monitoring or planning during task performance and that this can increase their learning outcomes (e.g., Azevedo & Cromley, 2004). Our study ex-

tends that research by focusing on the development of training for self-regulated learning situations in which learners have full control over the learning tasks they engage in. The training we implemented was relatively simple and focused primarily on teaching students the kind of rules for self-assessment and task selection that are also implemented in e-learning systems to personalize instruction (e.g., Camp et al., 2001; Corbalan et al., 2008; Kalyuga, 2006; Salden et al., 2004). Even though the exact content of the rules might differ between tasks and domains, especially for self-assessment (i.e., the scale on which performance is assessed, or how performance is assessed), the underlying principles that the task-selection rules convey, such as "when selecting a task, do not only regard your performance, but also the amount of effort you invested" and "if your performance was high and your invested mental effort was low, you can select a more complex task" are likely to be effective across tasks and domains. An interesting question for future research therefore, is whether acquired self-assessment and task-selection skills can transfer to other tasks in the same domain or even to other domains.

Another interesting question is whether training of self-assessment skills positively affects monitoring during task performance. That is, if students know what is important for assessing their performance, their monitoring might become more focused. It will probably still require some cognitive capacity that cannot be devoted to the learning task, but when monitoring is more focused this may be less detrimental for learning. It might even foster learning, as students might be able to adjust their performance in response to evaluations of certain steps made during task performance.

Finally, given the reported relationship between prior knowledge and accuracy of self-assessment, an interesting question would be whether students with higher levels of prior knowledge than the novices who participated in our studies would still benefit from training self-assessment and task-selection skills, or whether such training would be unnecessary or even harmful for them (i.e., an 'expertise reversal effect' might occur; Kalyuga, 2007; Kalyuga, Ayres, Chandler, & Sweller, 2003).

A limitation of the studies presented here, is that in Experiment 2, we did not manage to fully replicate the effects on accuracy that we found in Experiment 1. Had we used a think-aloud methodology during self-regulated learning (e.g., Azevedo & Cromley, 2004; Kostons et al., 2010 - Chapter 3), this could not only have provided more detailed insight into how students applied the self-assessment and task-selection rules they had learned in the training, but it could perhaps also have helped explain this difference between the results concerning accuracy obtained in Experiments 1 and 2. Another limitation is that this study focused solely on cognitive factors. It should be noted that the standard deviations on learning gains were quite large in all conditions. So even though conditions that were trained had higher learning gains on average than the control condition, differences in learning gains within conditions could potentially be explained by differences in students' motiva-

tion or goal orientation (e.g., Pintrich, 2000a). Combining measures of cognitive and affective variables in future studies might shed light on this issue. In addition, a limitation of this study is that we did not explicitly address effects on conceptual knowledge in measuring learning gains, although conceptual knowledge was required for successful problem solving (e.g., to be able to write down the required genotypes, a student would have to know what a genotype is and how specific givens determine what the genotype looks like). Finally, this study did not assess longer-term effects of training self-assessment and task-selection skills on the effectiveness of self-regulated learning. Future research should investigate whether the effects of training fade over time and if so, at what intervals the training should be repeated.

Despite these limitations, our studies resulted in an important finding for educational practice, showing that a relatively simple training intervention can significantly increase the amount of knowledge or skills students gain from self-regulated learning.

Chapter 5
General discussion

Abstract

This final chapter first describes the main findings and conclusions of the studies presented in this dissertation. Then, limitations of the studies are acknowledged, and findings are discussed in terms of practical and theoretical implications. Finally, suggestions for directions for future research in this field are provided.

Parts of this Chapter also appear in:
Van Gog, T., Kostons, D., Azevedo, R., & Paas, F. (2010). *Enhancing the effectiveness of self-regulated learning: A cognitive load perspective.* Manuscript submitted for publication

This dissertation started with a chapter providing an analysis of the cognitive demands that self-regulated learning in which students can choose their own learning tasks impose, especially for novice learners. This analysis suggested that performing learning tasks that impose a high cognitive load, while concurrently monitoring that performance, may hamper learning directly as well as complicate self-assessment, which may hamper learning indirectly. In addition to the negative effects that problems with monitoring might cause, accurate self-assessment may be complicated by novices' lack of knowledge of assessment criteria and standards. Inaccuracies in self-assessment, in turn, can affect the accuracy of task selection. Additionally, novices lack knowledge of how to use their self-assessments to select a next learning task, and what features of the task are important to consider during task selection. At the end of the first chapter, the main aims of the research reported in this dissertation were presented. The first aim was to study in detail how learners go about assessing their own performance and selecting next learning tasks. The second aim was to experimentally investigate whether training students' self-assessment and task-selection skills could enhance the accuracy of those skills, and whether this would improve learning outcomes attained through self-regulated learning.

Regarding the first aim, the study in Chapter 2 suggested that lower expertise participants (for whom the task imposed higher cognitive load) indeed had more difficulties reporting about their performance than higher expertise participants (for whom the task imposed less cognitive load). When lower expertise participants were provided with a performance-process cue consisting of a replay of a record of eye movements and actions performed on the computer, their remarks about the process increased, suggesting that the cue was helpful for recalling the performance process for those participants. However, higher expertise participants made more remarks about the task performance process than lower expertise participants even when they did not receive a this cue, suggesting that the higher expertise participants indeed had a better memory representation of their performance process. This study also showed that in general, the amount of assessment criteria that were spontaneously used was rather low. For higher expertise participants, the evaluative remarks increased when they received a cue, but this was not the case for lower expertise participants, which suggests that at least some knowledge of criteria and standards was required in order to be able to evaluate one's own performance and that the cue was a helpful support tool under those conditions.

The study in Chapter 3 provided evidence that self-assessment accuracy indeed plays an important role in the effectiveness of self-regulated learning: effective (i.e.,

in terms of learning gains) learners more accurately assessed their own performance compared to ineffective learners, and effective learners' self-assessment scores predicted subsequent task selection. Note that the verbal protocols suggested that this does not imply a conscious use of those criteria during task selection. Both self-reported mental effort and time-on-task scores were, in the case of effective learners, predictors of the tasks they selected next; however, when students had to rank the importance of the assessment criteria, mental effort and time on task were ranked only intermediate and very low, respectively. Taken together, the studies of Chapters 2 and 3 underline the need for explicit training of self-assessment and task-selection skills in order to enhance the effectiveness of self-regulated learning.

The second aim of the research reported in this dissertation was to experimentally investigate whether training students' self-assessment and task-selection skills could enhance the accuracy of those skills, and whether this would improve learning outcomes attained through self-regulated learning. This aim was addressed in the two experiments reported in Chapter 4. The first experiment showed that training with video-based modelling examples could increase self-assessment and task-selection accuracy. The results also suggested that both skills need to be trained, as we found no indications that an increase in self-assessment accuracy also led to an increase in task-selection accuracy when the latter was not modelled, or vice versa. The second experiment showed that training self-assessment and task-selection skills, either via modelling examples as in the first experiment, or via practice, enhanced the effectiveness of self-regulated learning in terms of learning gains. Training via modelling and practice seemed to be equally effective. The results of this second experiment partially, but not unequivocally, supported the assumption that the increase in learning gains was due to enhanced self-assessment and task-selection skills.

Limitations

There are several limitations to the research reported in this dissertation. First of all, the studies presented in Chapters 2 and 3 used relatively small number of participants. Even though this is not uncommon in research applying detailed process measures, larger sample sizes would have provided more statistical power. Another limitation to Chapter 2 was that self-assessment accuracy could not be determined. Participants were not provided with a predetermined set of assessment criteria, because we intended to examine which criteria would be considered spontaneously. But this also meant that there was no set of subjective scores to compare with a set of objective scores on those criteria and the lack of such data made it impossible to determine whether the performance-process cue affected self-assessment accuracy, which would have been interesting to determine in the con-

text of the theoretical framework presented in Chapter 1. A third limitation to Chapter 2 was that mental effort (being a possible assessment criterion) was not measured in this study, and as such, based on the verbal report data, only indirect conclusions could be drawn regarding the assumed beneficial effects of cueing on cognitive load (see Chapter 1).

Next to the relatively low number of participants, a second limitation of the study reported in Chapter 3 was that learners were asked to think aloud only during self-assessment and task selection. This was done because these were the processes of primary interest in this study, and asking students to think aloud during the entire procedure could have been exhausting and could therefore have reduced the quality of the data towards the end of the procedure. However, the lack of think aloud data from the learning tasks made it impossible to investigate whether effective and ineffective learners engaged in different processes during task performance (e.g., planning or monitoring; cf. Azevedo, Cromley, & Seibert, 2004), which could have influenced self-assessment and task selection.

As for Chapter 4, these experiments involved much higher numbers of participants, and the results clearly demonstrated the benefits of training on the accuracy of self-assessment and task-selection skills and subsequent learning gains with self-regulated learning. However, because no verbal reports were collected in this study, there was no information on whether and how students used what they had learned in the training during self-regulated learning. Such data could have helped explain why the increases in self-assessment and task-selection accuracy we found in Experiment 1 of this chapter could not be fully replicated. The application of a test immediately after the training, as in Experiment 1, might also have provided different results regarding accuracy. However, such a test was not taken, because this would also have provided a learning opportunity (Roediger & Karpicke, 2006). Therefore, such a test immediately after training could have provided different learning opportunities to students who did and did not receive training, and as a consequence, it would have been less clear whether differences found on the final test after self-regulated learning could be attributed solely to students' behaviour during self-regulated learning.

Theoretical and practical implications

Despite these limitations, the research presented in this dissertation has some interesting theoretical and practical implications. As mentioned in Chapter 1, self-regulated learning is increasingly implemented in secondary education, because it is believed to better prepare students for tertiary education and working life (e.g., in the Netherlands, a nationwide educational innovation that relies heavily on self-regulated learning was implemented in the higher levels of secondary education in 1999; http://www.minocw.nl/english/education/293/secondary-education.html).

With educational research showing that self-regulated learning often results in higher motivation or involvement (e.g., Corbalan, Kester, & Van Merriënboer, 2008; Schnackenberg & Sullivan, 2000), but in lower learning outcomes (see e.g., Azevedo, Moos, Greene, Winters, & Cromley, 2008; Lawless & Brown, 1997; Niemiec, Sikorsky, & Walberg, 1996; Williams, 1996), it is not surprising that the question of how to improve the effectiveness of self-regulated learning occupies the minds of educational researchers, as well as educational practitioners and policy makers (e.g., Dijsselbloem, 2008).

Theoretical implications

Concerning theoretical implications, this dissertation provided a new perspective on self-regulated learning in which learners are free to choose their own learning tasks. Cognitive load theory was used as a theoretical framework to explain the cognitive demands imposed by self-regulated learning, particularly with regard to accurate monitoring. Although the notion that monitoring plays an important role in the effectiveness of self-regulated learning is not new (e.g., Butler & Winne, 1995; Moos & Azevedo, 2008b; Pintrich, 2000b; Zimmerman, 1990), the cognitive load perspective seems to have added value because it has the power to explain why monitoring is often difficult for learners, especially for novices, and why monitoring can have adverse effects on learning. This cognitive load perspective can help identify new research questions concerning techniques to foster monitoring, such as the performance process cues studied in Chapter 2. Moreover, this research contributed to research on self-regulated learning by focussing specifically on the role of self-assessment and task-selection processes and investigating a new instructional technique of enhancing the effectiveness of self-regulated learning by decreasing the load imposed by self-assessment and task selection skills through a training of those skills. Even though not all research on self-regulated allows students to choose their learning tasks, self-assessment (or self-evaluation) is considered important in many studies on self-regulated learning and usually some kind of information selection usually is required (e.g., in hypermedia learning environments; Azevedo, 2005; Azevedo & Cromley, 2004; Scheiter & Gerjets, 2007).

This dissertation also extends the research on self-assessment by: a) highlighting the role that cognitive load plays in the ability to monitor performance, which subsequently influences accuracy of self-assessment; b) providing a detailed analysis of how students go about assessing their own performance when assessment criteria are not provided (Chapter 2); c) calling attention to individual differences in assessment accuracy between learners (Chapters 2 and 3); d) underlining the role that knowledge of assessment criteria and standards plays in accurate self-assessment (e.g., Dunning, Heath, & Suls, 2004; Miller, 2003); e) investigating new ways of training students to enhance their self-assessment accuracy (Chapter 4); and f) teaching them how to use self-assessment as input for task selection (Chapter 4). Last but not

least, an important contribution to the research on self-assessment is that this research was not solely concerned with investigating ways to increase self-assessment accuracy (as e.g., Miller, 2003; Oldfield & Macalpine, 1995; Searby & Ewers, 1997), but that it also directly investigated the consequences self-assessment (in)accuracy has for learning.

Finally, this dissertation contributes a new perspective to research on (e-learning) systems for personalized instruction, by showing that the kind of algorithms used in such systems can also be used by students themselves (after training) to increase adaptivity of the tasks to their learning needs, thereby improving learning outcomes.

Practical implications

As for practical implications, some researchers have proposed that it might be better not to have novices engage in full-scale self-regulated learning, but rather, provide a more gradual transition from teacher-controlled to learner-controlled instruction, for example via shared control over instruction (e.g., Corbalan et al., 2006). Corbalan et al. (2008) had a system for personalized instruction assess the learner's performance and select a set of appropriate tasks based on that assessment. The learner then got to choose which task s/he wanted to work on. This was still motivating for learners, presumably because it gave them a sense of control, and prevented negative effects on learning. Even though this seems to be a good solution, as it is more effective in terms of learning outcomes, one drawback is that such a system of shared control does not reduce the problems learners will encounter once they have to assess their own performance and select learning tasks all by themselves. To reduce those problems, instructional interventions that focus directly on increasing the accuracy of monitoring, self-assessment, and task selection are required. Several interventions were suggested in this dissertation.

Monitoring. To reduce monitoring difficulties that arise when tasks are high in intrinsic load, educators could for example implement access to a performance process cue, such as a video or screen-recording of task performance (Chapter 2). Including eye movements as we did might be complicated in practice, but easy-to-use screen recording software is becoming more and more available and affordable. Such a performance-process cue will take up some additional study time, because learners need to review their entire task performance process, but it might cancel out the negative influences that unsuccessful monitoring might have on self-assessment. Moreover, there are indications that reviewing ones own performance can directly contribute to learning (e.g., Fireman, Kose, & Solomon, 2003), making this extra time investment worthwhile.

Self-assessment and task selection. To foster self-assessment and task selection accuracy, we used training consisting of modelling examples or practice (Chapter 4). Thus far, examples are mostly used in education to teach cognitive skills, and this

study adds further evidence that they are useful for teaching metacognitive skills as well (see also Kitsantas et al., 2000; Zimmerman & Kitsantas, 2002). Interestingly for educational practice, the training intervention we used was relatively simple, as it was based on conveying the procedures and algorithms used in instructional systems for personalized instruction (Camp et al., 2001; Corbalan et al., 2008; Kalyuga, 2006; Salden et al., 2004). Our experiment showed that even such a simple intervention can have a strong impact on the effectiveness of self-regulated learning. As a result of such training, learners can fully deploy the possibilities offered by self-regulated learning to personalize instruction: they become capable of making task selection choices that fit their level of knowledge or skill development.

Future research

The studies presented in this dissertation also gave rise to new questions that future research might address, primarily concerning training. Firstly, from our theoretical perspective, it would be interesting to investigate whether training self-assessment would also have positive effects on monitoring performance while working on the learning tasks. If students know what is important for assessing their performance, it may direct their monitoring activities, which in this case may still require some cognitive capacity that cannot be devoted to the learning task, but may be less detrimental for learning.

Secondly, as easy as the training we used might be to implement in practice, some further research is advisable before doing so. For example, the results from our training study were very positive, but were obtained in an experiment of relatively short duration. Future research should investigate whether the effects of training would fade over time and if so, at what intervals the training should be repeated. If the training has to be repeated, would it have to be in full, or would it be sufficient to, for example, provide prompts to apply the principles that were learned (e.g., Azevedo & Cromley, 2004; Bannert, 2004) or provide feedback concerning the (in)correct application of those principles (e.g., Kicken, Brand-Gruwel, Van Merriënboer, & Slot, 2009; Narciss, Proske, & Koerndle, 2007; Taminiau, Corbalan, Kester, & Van Merriënboer, 2008)? When prompts and feedback are provided during task performance, cognitive load should again be taken into account; prompts and/or feedback again provides another 'task' to attend to, which may increase cognitive load. The timing of such prompts or feedback may therefore be crucial (i.e., just-in-time information, Kester, Van Merriënboer, & Kirschner, 2006), regardless of whether it is static (i.e., provided at fixed points in time), or dynamic (i.e., provided depending on the learner's task progress; e.g., Azevedo et al., 2004). For example, providing prompts during the task may increase cognitive load, though this might differ depending on whether the prompt is proactive or retrospective. Proactive prompts direct learners' attention to what is to come, which requires

them to keep the prompt in mind (high cognitive load). Retrospective prompts on the other hand, prompt learners to think about their last actions which are still fresh in their memory (see e.g., Helsdingen, Van Gog, & Van Merriënboer, 2010).

A related question, given that learners will gain knowledge during self-regulated learning – at least when they have received training -, is whether learners with higher levels of expertise would still benefit from training, or whether it would be unnecessary or even harmful for them (cf. 'expertise reversal effect'; Kalyuga, 2007; Kalyuga, Ayres, Chandler, & Sweller, 2003). Moreover, it is important to establish whether a training would have to be developed for each domain and for each type of task within that domain, or whether the effects of training will transfer to other tasks within or across domains.

Last but not least, the research presented in this dissertation focused mostly on cognitive factors. However, it is important to keep in mind that training monitoring, self-assessment, or task selection strategies does not guarantee that learners will actually use them later on. For example, in Chapter 4, participants in the condition in which self-assessment and task-selection skills were trained gained much more knowledge than those in the non-training condition, but there were also large differences in knowledge gains between students within the trained condition. Applying the techniques learned from instructional interventions to increase monitoring, self-assessment, and task selection requires continuous effort from learners, and whether or not they are willing to invest this effort will be determined by affective and motivational processes such as their sense of usefulness of the techniques, and their motivation and self-efficacy beliefs (Bandura, 1997; Kanfer & Ackerman, 1989; Kinzie & Sullivan, 1989; Paas, Tuovinen, Van Merriënboer, & Darabi, 2005; Pintrich, 2000; Zimmerman, 2000). Combining measures of cognitive and affective variables in future studies might shed more light on this issue.

References

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, *26*, 147-179.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*, 167-207.

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, *128*, 110-118.

Arter, J. A., & Spandel, V. (1992). An NCME instructional module on: Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice*, *11*, 36-45.

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, *94*, 416-427.

Atkinson, R. K., Catrambone, R., & Merrill, M. M. (2003). Aiding transfer in statistics: Examining the use of conceptually oriented equations and elaborations during subgoal learning. *Journal of Educational Psychology*, *95*, 762-773.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*, 181-214.

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

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*, 523-535.

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology*, *29*, 344-370.

Azevedo, R., Guthrie, G. T., & Seibert, D. (2004). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. *Journal of Educational Computing Research*, *30*, 87-111.

Azevedo, R., Moos, D., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development*, *56*, 45-72.

Baldwin, T. T. (1992). Effects of alternative modelling strategies on outcomes of interpersonal-skills training. *Journal of Applied Psychology*, *77*, 147-154.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behaviour* (Vol 4, pp 71-81). New York: Academic Press. (Reprinted in H. Friedman (Ed.). *Encyclopedia of mental health*, San Diego: Academic Press, 1998).

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Bannert, M. (2004). Designing metacognitive support for hypermedia learning. In H. M. Niegemann, D. Leutner, & R. Brünken (Eds.), *Instructional design for multimedia learning* (pp. 19–30). Münster: Waxmann.

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology*, *101*, 70-87.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. E Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435-459). Cambridge, MA: MITPress.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (Eds.) (2000). *Handbook of self-regulation*. San Diego: Academic Press.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*, 151-167.

Boud, D., & Falchicov, N. (1989). Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *Higher Education*, *18*, 529-549.

Braaksma, M. A. H., Rijlaarsdam, G., & Van den Bergh, H. (2002). Observational learning and the effects of model-observer similarity. *Journal of Educational Psychology*, *94*, 405-415.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*, 53-61.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*, 245-281.

Camp, G., Paas, F., Rikers, R., & Van Merriënboer, J. J. G. (2001). Dynamic problem selection in air traffic control training: A comparison between performance, mental effort and mental efficiency. *Computers in Human Behavior*, *17*, 575-595.

Camtasia Studio (Version 6)[Computer software]. Okemos, Michigan: Techsmith.

Candy, P. C. (1991). *Self-direction for lifelong learning*. San Francisco: Jossey-Bass.

Catrambone, R. (1995). Aiding subgoal learning: Effects on transfer. *Journal of Educational Psychology*, *87*, 5-17.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1020-1031.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293-332.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*, (Vol. 1, pp 7-75). Hillsdale, NJ: Erlbaum.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp 453-494). Hillsdale, NJ: Erlbaum.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*, 347-362.

Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2006). Towards a personalized task selection model with shared instructional control. *Instructional Science*, *34*, 399-422.

Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2008). Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary Educational Psychology*, *33*, 733-756.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-114.

Cowan, N. (2010). The magical number four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*, 51-57.

Dijsselbloem, J. (2008). *Tijd voor onderwijs. Parlementair onderzoek der Tweede Kamer: Commissie Dijsselbloem* [Time for education. Parliamentary research of the Second Chamber: Committee Dijsselbloem]. The Hague, The Netherlands: SDU.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in Higher Education: A review. *Studies in Higher Education*, *22*, 233-239.

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice* (2nd ed.). London: Springer.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*, 69-106.

Dunning, D., Johnson, K., Erlinger, J., & Kruger, J. (2003) Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83-87.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers*, *28*, 1-11.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: MIT.

Erkens, G. (2005). *Multiple Episode Protocol Analysis* (Version 4.10 [Computer software]. Utrecht: Utrecht University.

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

Fireman, G., Kose, G., & Solomon, M. J. (2003). Self-observation and learning: The effects of watching oneself on problem solving. *Cognitive Development*, *18*, 339-354.

Goforth, D. (1994). Learner control = decision making + information: A model and meta-analysis. *Journal of Educational Computing Research*, *11*, 1-26.

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, *17*, 612-634.

Hansen, J. P. (1991). The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica, 76,* 31-49.

Helsdingen, A. S., Van Gog, T., & Van Merriënboer, J. J. G. (2010). *The effects of practice schedule and critical thinking instruction on learning and transfer of a complex judgment task*. Manuscript submitted for publication.

Kalyuga, S. (2006). Assessment of learners' organized knowledge structures in adaptive learning environments. *Applied Cognitive Psychology, 20*, 333-342.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*, 509-539.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*, 23-31.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology*, *93*, 579-588.

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology*, *96*, 558-568.

Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to optimize the efficiency of e-learning. *Educational Technology, Research and Development*, *53*, 83-93.

Kanfer, R. & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatments interaction approach to skill acquisition. *Journal of Applied Psychology*, *74*, 657-690.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469-486.

Kester, L., Kirschner, P. A., & Van Merriënboer, J. J. G. (2006). Just-in-time information presentation: Improving learning a troubleshooting skill. *Contemporary Educational Psychology*, *31*, 167-185.

Kicken, W., Brand-Gruwel, S., Van Merriënboer, J. J. G., & Slot, W. (2009). The effects of portfolio-based advice on the development of self-directed learning skills in secondary vocational education. *Educational Technology Research and Development*, *57*, 439-460.

Kinzie, M. B. (1990) Requirements and benefits of effective interactive instruction: Learner control, self-regulation, and continuing motivation. *Educational Technology Research and Development*, *38*(1), 5-21.

Kinzie, M. B., & Sullivan, H. J. (1989). Continuing motivation, learner control, and CAI. *Educational Technology Research and Development*, *37*(2), 5-14.

Kitsantas, A., Zimmerman, B. J., & Cleary, T. (2000). The role of observation and emulation in the development of athletic self-regulation. *Journal of Educational Psychology*, *92*, 811-817.

Knowles, M. S. (1975). *Self-directed learning: A guide for learners and teachers*. New York: Association Press.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology*, *31*, 187-194.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*, 478-492

Kostons, D., Van Gog, T., & Paas, F. (2009). How do I do? Investigating effects of expertise and performance-process records on self-assessment. *Applied Cognitive Psychology*, *23*, 1256-1265.

Kostons, D., Van Gog, T., & Paas, F. (2010). Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education*, *54*, 932-940.

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology*, *113*, 387-404.

Lawless, K. A., & Brown, S. W. (1997). Multimedia learning environments: Issues of learner control and navigation. *Instructional Science*, *25*, 117-131.

Loyens, S. M. M., Magda, J., & Rikers, R. M. J. P. (2008). Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational Psychology Review*, *20*, 411-427.

Mazzoni, D. (2006). *Audacity* (Version 1.2.6) [Computer software].

McLaren, B. M., Lim, S., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? New results and a summary of the currents state of research. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2176-2181). Austin, TX: Cognitive Science Society.

Merrill, M. D. (1980). Learner control in computer based learning. *Computers and Education*, *4*, 77-95.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*, 159-163.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Miller, P. J. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education*, *28*, 383-394.

Moos, D. C., & Azevedo, R. (2008a). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Educational Psychology*, *33*, 270-298.

Moos, D. C., & Azevedo, R. (2008b). Monitoring, planning, and self-efficacy during learning with hypermedia: The impact of conceptual scaffolds. *Computers in Human Behavior, 24*, 1686-1706.

Narciss, S., Proske, A., & Koerndle, H. (2007). Promoting self-regulated learning in web-based learning environment. *Computers in Human Behavior*, *23*, 1126-1144.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*, 267-270.

Niemiec, R. P., Sikorski, C., & Walberg, H. J. (1996). Learner-control effects: A review of reviews and a meta-analysis. *Journal of Educational Computing Research*, *15*, 157-174.

Oldfield, K. A., & Macalpine, J. M. K. (1995). Peer and self-assessment at tertiary level: An experiential report. *Assessment in Higher Education*, *20*, 125-132.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429-434.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*, 63-71.

Paas, F., Tuovinen, J. E., Van Merriënboer, J. J. G., & Darabi, A., (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Educational Technology, Research & Development*, *53*(3), 25-33.

Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*, 122-133.

Pajares, F., & Kranzler, J. (1995). Self-efficacy and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology*, *20*, 426-443.

Pintrich, P. R. (2000a). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, *92*, 544-555.

Pintrich, P. R. (2000b). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, and M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 451-502). San Diego, CA: Academic Press.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*, 144-161.

Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology*, *16*, 325-342.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372-422.

Roediger, H. L., III., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181-210.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (revised edition). Newbury Park, CA: Sage.

Ross, B. H. (1989). Remindings in learning and instruction. In S. Vosniadou & A. Rotony (Eds.), *Similarity and analogical reasoning* (pp. 438-469). Cambridge, MA: Cambridge University Press.

Ross, S. M., & Morrison, G. R. (1989). In search of a happy medium in instructional technology research: Issues concerning external validity, media replications, and learner control. *Educational Technology Research and Development*, *37*(1), 19-33.

Ross, S. M., Morrison, G. R., & O'Dell, J. K. (1989). Uses and effects of learner control of context and instructional support in computer-based instruction. *Educational Technology Research and Development*, *37*(4), 29-39.

Rourke, A., & Sweller, J. (2009). The worked-example effect using ill-defined problems: Learning to recognize designers' styles. *Learning and Instruction*, *19*, 185-199.

Ruble, D. N., & Frey, K. S. (1991). Changing patterns of comparative behavior as skills are acquired: A functional model of self-evaluation. In J. Suls (Ed.), *Social comparison: Contemporary theory and research* (pp. 79-113). Hillsdale, NJ: Erlbaum.

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem-solving in computer-mediated settings. *Journal of the Learning Sciences*, *14*, 201-241.

Salden, R. J. C. M., Paas, F., Broers, N. J., & Van Merriënboer, J. J. G. (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instructional Science*, *32*, 153-172.

Salden, R. J. C. M., Paas, F., Van der Pal, J., & Van Merriënboer, J. J. G. (2006). Dynamic task selection in flight management system training. *The International Journal of Aviation Psychology*, *16*, 157-174.

Salden, R. J. C. M., Paas, F., & Van Merriënboer, J. J. G. (2006). Personalised adaptive task selection in Air Traffic Control: Effects on training efficiency and transfer. *Learning and Instruction, 16*, 350-362.

Scheiter, K., & Gerjets, P. (2007). Learner control in hypermedia environments. *Educational Psychology Review, 19*, 285-307.

Schnackenberg, H. L., & Sullivan, H. J. (2000). Learner control over full and lean computer-based instruction under differing ability levels. *Educational Technology Research and Development*, *48*(2), 19-35.

Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of Educational Psychology*, *73*, 93-105.

Schunk, D. H. (1987). Peer models and children's behavioural change. *Review of Educational Research*, *57*, 149-174.

Schunk, D. H., & Hanson, A. R. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology*, *77*, 313-322.

Schwonke, R., Berthold, K., & Renkl, A. (2009). How multiple representations are used and how they can be made more useful. *Applied Cognitive Psychology*, *23*, 1227-1243.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. J. C. M. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, *25*, 258-266.

Schworm, S. & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology*, *99*, 285-296.

Searby, M., & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education. *Assessment & Evaluation in Higher Education*, *22*, 371-383.

Segers, M., Dochy, F., & Cascallar, E. (Eds.). (2003). *Optimising new modes of assessment: In search of qualities and standards.* Boston: Kluwer Academic Publishers.

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal*, *41*, 159-189.

Sheldon, J. P. (2003). Self-evaluation of competence by adult athletes: Its relation to skill level and personal importance. *The Sport Psychologist*, *17*, 426-443.

Shute, V. J., & Zapata-Rivera, D. (2008). Adaptive Technologies. In J. M. Spector, M. D. Merrill, J. J. G. Van Merriënboer, & M. P. Driscoll (Eds.) *Handbook of research on educational communications and technology* (pp 277-294). New York, Lawrence Erlbaum Associates.

Somervell, H. (1993). Issues in assessment, enterprise and higher education: The case for self-peer and collaborative assessment. *Assessment & Evaluation in Higher Education*, *18*, 221-233.

Steinberg, E. R. (1989). Cognition and learner control: A literature review, 1977-88*. Journal of Computer-Based Instruction, 16*, 117-124.

Stone, D. (1994). Overconfidence in initial self-efficacy judgments: Effects on decision processes and performance. *Organizational Behavior and Human Decision Processes*, *59*, 452-474.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257-285.

Sweller, J. (2004). Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture. *Instructional Science, 32*, 9-31.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*, 123-138.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*, 59-89.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251-296.

Taminiau, B., Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2008). Procedural advisory models and the acquisition of domain-specific skills. In Kinshuk, D. G. Sampsion, J. M. Spector, P. Isaias, & D. Ifenthaler (Eds.), *Cognition and exploratory learning in digital age (CELDA). Proceedings of the conference of the international association for development of the information society* (pp 420-421). International Association for Development of the Information Society (IADIS).

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*, 66-73.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1024-1037.

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity, and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.) *Optimising new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology*, *91*, 334-341.

Uden, L., McGuinness, V. M., & Alderson, A. (2000). A comparative study of learner control and system control in computer-aided learning. In D. Benzie & D. Passey (Eds.), *Proceedings of the International Conference on Educational Uses of Information and Communication Technologies* (pp. 367-370). Beijing, China: IFIP.

Van den Boom, G., Paas, F., & Van Merriënboer, J. J. G. (2007). Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. *Learning and Instruction*, *17*, 532-548.

Van den Boom, G., Paas, F., Van Merriënboer, J. J. G., & Van Gog, T. (2004). Reflection prompts and tutor feedback in a web-based learning environment: Effects on students' self-regulated learning competence. *Computers in Human Behavior*, *20*, 551-567.

Van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., & Paas, F. (2009). Attention guidance during example study via the model's eye movements. *Computers in Human Behavior*, *25*, 785-791.

Van Gog, T., Kester, L., & Paas, F. (in press). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology*.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*, 16-26.

Van Gog, T., & Paas, F. (2009). Effects of concurrent performance monitoring on cognitive load as a function of task complexity. In N. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1605-1608). Austin, TX: Cognitive Science Society.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2004). Process-oriented worked examples: Improved transfer performance through enhanced understanding. *Instructional Science*, *32*, 83-98.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2005). Uncovering expertise-related differences in troubleshooting performance: Combining eye movement and concurrent verbal protocol data. *Applied Cognitive Psychology, 19*, 205-221.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction*, *16*, 154-164.

Van Gog, T., Paas, F., Van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*, 237-244.

Van Gog, T., & Rummel, N. (2010). Example-based learning. Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, *22*, 155-174.

Van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*, 147-177.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes.* London: Academic Press.

Vermunt, J. D. (1998). The regulation of constructive learning processes. *British Journal of Educational Psychology*, *68*, 149-171.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Williams, M. (1996). Learner control and instructional technologies. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 957–983). New York: Scholastic.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk, *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153-189). Mahwah, NJ: Lawrence Erlbaum.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser, *Metacognition in education and practice. The educational psychology series* (pp. 277-304). Mahwah, NJ: Lawrence Erlbaum.

Woolf, H. (2004). Assessment criteria: Reflections on current practices. *Assessment & Evaluation in Higher Education*, *29*, 479-493.

Wouters, P., Paas, F., & Van Merrienboer, J. J. G. (2009). Observational learning from animated models: Effects of modality and reflection on transfer. *Contemporary Educational Psychology*, *34*, 1-8.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*, 3-17.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*, 82-91.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*, 64-70.

Zimmerman, B. J., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology*, *94*, 660-668.

Zimmerman, B. J., & Schunk, D. H. (Eds.) (2001). *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

English summary

Self-regulated learning is an active, constructive process in which learners plan, monitor, and control their own learning process. Self-regulated learning can occur at different levels, from learners controlling how long they engage in studying a given task or whether they want to restudy it, to learners controlling what information they want to study or what learning tasks they want to work on. This dissertation will focus primarily on self-regulated learning in which learners can choose their own learning tasks, which is often referred to as self-directed learning in the context of research on lifelong learning, and as learner-controlled instruction in the context of computer-assisted learning. Because learners get to choose their own learning tasks, self-regulated learning is assumed to result in personalized learning trajectories in which instruction is adaptive to the individual learner's needs and preferences. Such personalized instruction is expected to enhance learning and motivation compared to non-personalized instruction that is the same for all learners. Self-regulated learning in which students have some choice over what tasks they work on is increasingly implemented in secondary education, because it is believed to better prepare students for tertiary education and working life. For example, in the Netherlands a nationwide educational innovation that relies heavily on self-regulated learning (the 'study house') was implemented in the higher levels of secondary education in 1999 (e.g. http://www.minocw.nl/english/education/293/-secondary-education.html).

Studies on the effects of self-regulated learning have shown that providing learners with the opportunity to regulate their learning may indeed have beneficial effects on their motivation and involvement. However, learning outcomes with self-regulated learning are often similar or, in the case of novices, inferior to those achieved with non-personalized instruction. When an instructional system is used to personalize instruction, it does so by a cyclical process of monitoring and assessing a learner's current level of knowledge and/or skill to select or suggest an appropriate next learning task, and such adaptation has been shown to be beneficial for learning. For learners to self-regulate their learning just as effectively, they should be

able to accurately monitor and assess their own performance and determine what an appropriate next learning task would be themselves.

The main aims of the research project reported in this dissertation were to investigate how learners assess their own performance and select next learning tasks, and whether training learners' self-assessment and task-selection skills could enhance the accuracy of those skills, and subsequently improve the effectiveness of their self-regulated learning.

The theoretical analysis in Chapter 1 suggests that accurate monitoring of performance, self-assessment and task selection plays a crucial role in the effectiveness of self-regulated learning in which learners control what learning task they want to work on. However, this chapter also provides evidence that learners, particularly novices, may experience difficulties with accurate monitoring, self-assessment and task selection.

First, performing a task and monitoring that performance both require resources from a limited working memory capacity. Cognitive resources devoted to performing the learning task are unavailable for monitoring that performance, and vice versa. The more resources the activity of task performance requires, the fewer resources are available for monitoring that performance. Consequently, it can be expected that the competition of the processes of performance and monitoring for limited resources, can result in poor performance, poor monitoring, or both, especially with complex tasks.

Second, when monitoring is hampered, learners will likely have a poor recollection of their performance, inhibiting accurate self-assessment. This may be why learners with more expertise are also more accurate self-assessors. Their experience lowers the cognitive load imposed by the learning task, allowing them to devote more cognitive resources to monitoring their task performance, which likely provides them with a more accurate memory representation on which to base their assessment. But even with a good recollection of performance, accurate self-assessment is not guaranteed. Self-assessment may also be hampered by several biases that may cause learners to depend on the wrong kind of cues to assess their performance. Moreover, accurate self-assessment also seems to require knowledge of the criteria and standards that good performance should meet, which learners with little experience in the domain are unlikely to have.

Third, inaccurate self-assessment, when used as input for task selection, may negatively affect selection of an appropriate new learning task. For example, if learners overestimate their performance, they are likely to choose a next task that is too difficult for them. But even when self-assessment is accurate, learners may still experience problems in selecting appropriate learning tasks. When selecting a task, it is important to discern which aspects of a task are relevant for learning, such as the structural features of the task (e.g., type of task, complexity level, amount of support provided), from aspects that are less relevant, such as cover stories. Re-

search has shown that novices may experience difficulties in discerning between these aspects and tend to choose tasks based on irrelevant aspects. When task selection is inaccurate, the chosen tasks are unlikely to be adaptive to the learner's level of prior knowledge or skills, so learners will end up working on tasks that are not aligned with their learning needs.

The first chapter thus discusses evidence that learners, particularly novices, have difficulties with accurate monitoring, self-assessment and task selection, which are required for effective self-regulated learning. However, it also explains why learners experience these difficulties, which may help in the development of interventions to deal with these difficulties and stimulate effective self-regulated learning.

The study presented in Chapter 2 investigated which aspects of their performance participants at different levels of expertise consider when they are asked to engage in self-assessment. To shed light on the question whether a lack of knowledge of criteria and standards, a lack of resources for monitoring, or a combination of both complicate novices' self-assessment, half of the participants received a performance-process cue during self-assessment, whereas the other half only received a blank screen. The performance-process cues consisted of a replay of a record of participants' eye movements and actions performed on the computer, and were expected to reduce the need for concurrent performance monitoring by allowing participants to review both physical actions (reflected in mouse/keyboard operations) and cognitive actions (reflected in eye movements) made during task performance. It was hypothesized that the cue would be helpful for novice participants to report about the performance process, whereas advanced participants would not need a cue to be able to report on the process, because the task was less complex for them so they had more cognitive capacity available for monitoring. In addition it was hypothesized that the cue could or could not help novices evaluate their performance process (depending on which of the two explanations, lack of monitoring or lack of knowledge of criteria and standards, played a more prominent role), but that it would help advanced learners evaluate their performance, which would show in the amount and type of assessment criteria used. Analyses of verbal protocols suggest an interaction between the level of prior knowledge and self-assessment. Novices indeed seemed to have difficulty remembering the details of the performance process, presumably due to problems with monitoring their performance, and the cue allowed them to better review and remark on those task performance processes. The cue had no effect on the advanced learners' ability to report on their performance-process. In contrast, whereas the cues did not help novices with evaluating their performance, presumably because they lack knowledge of criteria and standards, advanced learners did benefit from the cue, as shown by enhanced use of self-assessment criteria. This study sheds more light on the complicated relationship between monitoring, self-assessment, and level of expertise. The results seem to indicate that offering a cue to novices may be beneficial for the way they

generate performance comments, but that they likely need some knowledge of criteria and standards in order to make good use of this cue.

Chapter 3 describes a study which investigated the differences in self-assessment and task-selection processes between effective and ineffective learners studying in a self-regulated e-learning environment. Fourth year pre-university students engaged in a cyclic process of choosing their own learning task from a database, performing that task, assessing their performance using a provided list of self-assessment criteria, then selecting a new task, et cetera; they repeated this cycle eight times in total. During self-assessment and task selection, students were asked to verbalize what they were thinking. After data-collection, students were assigned to either an effective or ineffective student-group, based on their performance gains between pre-test and post-test. Analyses based on those groups indicated that effective students more accurately assessed their own performance compared to ineffective learners. There were no differences between the groups in task-aspects considered when choosing a task, but whereas for the effective students their self-assessment outcomes did seem to predict task selection, this was not the case for the ineffective students. However, this does not mean students are making these choices deliberately on their assessments; whereas for effective learners both mental effort and time-on-task were significant predictors of task selection, these ranked intermediate and very low respectively on students' perceived importance of assessment criteria. This study provided insight into differences in self-assessment and task-selection processes between effective and ineffective self-regulated novice learners, and underlines the potential of training self-assessment and task-selection skills as a key to improving the effectiveness of self-regulated learning.

Such training was the object of study in the two experiments described in Chapter 4, which investigated whether self-assessment and task-selection accuracy could be improved by means of training and whether this would enhance the effectiveness of self-regulated learning. The first experiment investigated whether example-based learning, that is, learning by observing a human model engaging in self-assessment, task selection, or both, could be an effective strategy for training novices' (fourth year pre-university students) self-assessment and task-selection skills. In the experimental conditions, the models demonstrated task performance and, dependent on condition, additionally self-assessment, task selection, or both. In the control condition, participants only observed the models' task performance and were subsequently instructed to detect and correct errors made by the models. Results indicate that students presented with modelling examples of self-assessment and/or task-selection skills were more accurate on those skills at a post-test than students who did not receive those examples. The second experiment investigated whether training novices' (third year higher general secondary education students) self-assessment and task-selection skills, either through modelling or

practicing, could improve the effectiveness of subsequent self-regulated learning. This experiment therefore also investigated the effectiveness of another training method. Students in the Modelling condition viewed performance, self-assessment and task-selection modelling examples as in Experiment 1. Students in the Practice condition were explained the self-assessment and task-selection rules which they had to apply to the model's performance they had observed (i.e., they assessed the model's performance and selected a next task for the model). The Control condition was similar to Experiment 1. After the training (or in case of the control condition, error detection and correction), all participants engaged in self-regulated learning in an e-learning-environment (comparable to that of Chapter 3), again performing the cycle of choosing a learning task, performing that learning task, and assessing their performance eight times. The results showed that students who received self-assessment and task-selection training showed higher pre-test to post-test learning gains than students in the control group. Training via modelling and practice seemed to be equally effective. These two experiments resulted in an important finding for educational practice, showing that a relatively simple training intervention to improve self-assessment and task-selection skills can enhance accuracy of those skills and significantly increase learning outcomes students attain through self-regulated learning.

Finally, Chapter 5 provides an overview of the main findings of the studies presented in this dissertation, discusses the limitations of those studies as well as their theoretical and practical implications, and the questions they raise for future research. Taken together, the studies presented in this dissertation provide evidence that learners need accurate self-assessment and task-selection skills in order to be effective self-regulated learners and that training those skills can result in more effective self-regulated learning as evidenced by higher learning outcomes.

Nederlandse samenvatting

Zelfregulerend leren is een actief, constructief proces, waarin de lerende het eigen leerproces moet plannen, monitoren en aansturen. Zelfregulerend leren kan op meerdere niveaus plaatsvinden, variërend van lerenden de controle geven over hoe lang of hoe vaak ze iets willen bestuderen, tot aan welke informatie ze willen bestuderen of welke leertaken ze willen maken. In dit proefschrift wordt vooral gekeken naar zelfregulerend leren waarbij lerenden hun eigen leertaken kunnen kiezen, wat in de context van 'leven-lang-leren' ook wel zelfgestuurd leren en in de context van computerondersteund leren ook wel leerling-gestuurde instructie wordt genoemd. Omdat de lerende zelf het eigen leerpad kan bepalen, zou zelfregulerend leren waarin de lerenden zelf taken kiezen tot meer gepersonaliseerde instructie moeten leiden, die beter aansluit bij de behoeften en doelen van de individuele lerende. Dergelijke gepersonaliseerde instructie zou tot zowel hogere leerresultaten als motivatie moeten leiden ten opzichte van niet-adaptieve instructie. Zelfregulerend leren waarin lerenden zelf taken kiezen wordt steeds meer toegepast in het voortgezet onderwijs, mede omdat het gezien wordt als een manier om leerlingen beter voor te bereiden op hoger onderwijs en het werkende leven. Zo is bijvoorbeeld het studiehuis in 1999 ingevoerd als onderwijsinnovatie binnen de hogere niveaus van het voortgezet onderwijs, waarin een sterke nadruk gelegd wordt op het zelfregulerend leren (http://www.rijksoverheid.nl/documenten-en-publicaties/vragen-en-antwoorden/wat-houdt-de-tweede-fase-in-het-voortgezet-onderwijs-vo-in.html).

Hoewel onderzoek laat zien dat motivatie inderdaad toeneemt door zelfregulerend leren, is er vaak geen verschil in leeruitkomsten, of zijn deze voor novieten soms zelfs slechter, in vergelijking met niet-adaptieve instructiesystemen. Wanneer een adaptief instructiesysteem wordt gebruikt voor het selecteren van leertaken, hanteert dit systeem gewoonlijk een cyclisch proces waarin het huidige kennis- en/of vaardigheidniveau van de lerende geregistreerd en beoordeeld wordt en op basis daarvan een gepaste leertaak gekozen of voorgesteld wordt. Dergelijke adaptieve instructie blijkt een gunstig effect te hebben op de leeruitkomsten. Maar wanneer de lerende de controle over het leerproces krijgt, zal hij of zij zelf het monito-

ren en beoordelen van de prestatie en het kiezen van een volgende leertaak op zich moeten nemen.

Het onderzoek dat in dit proefschrift beschreven wordt, heeft tot doel te bestuderen hoe lerenden hun prestaties beoordelen en volgende leertaken kiezen en te onderzoeken of het trainen van zelfbeoordeling- en taakselectievaardigheden gunstige effecten heeft op zowel de accuratesse waarmee die vaardigheden toegepast worden alsmede op de leeruitkomsten behaald door middel van zelfregulerend leren.

De theoretische analyse in hoofdstuk 1 stelt dat nauwkeurig monitoren, zelfbeoordelen en taken selecteren een cruciale rol speelt in de effectiviteit van zelfregulerend leren waarin lerenden zelf taken kiezen. Echter, het hoofdstuk beschrijft ook bewijs uit de literatuur dat lerenden, vooral novieten, moeilijkheden kunnen ervaren met nauwkeurig monitoren, zelfbeoordelen en taken selecteren.

Ten eerste kosten zowel het uitvoeren van een taak alsmede die taakuitvoering gelijktijdig monitoren werkgeheugen capaciteit - en die capaciteit is beperkt. Capaciteit die besteed wordt aan het uitvoeren van de leertaak is niet beschikbaar voor het monitoren van die uitvoering en vice versa. Hoe meer capaciteit nodig is voor het uitvoeren van de taak, hoe minder er beschikbaar zal zijn voor het monitoren van die taak uitvoering. Deze competitie om werkgeheugen capaciteit tussen het uitvoeren van de taak en het monitoren daarvan kan leiden tot slechte taakprestatie, slecht monitoren, of beide, vooral wanneer de leertaken complex zijn.

Ten tweede zal inaccuraat monitoren waarschijnlijk leiden tot een gebrekkige herinnering van het proces van taakuitvoering, wat accuraat zelfbeoordelen kan hinderen. Dit is mogelijk een verklaring waarom lerenden met meer expertise zichzelf meer accuraat kunnen beoordelen. Door hun ervaring zijn leertaken minder belastend voor het werkgeheugen van lerenden met meer expertise dan ze voor novieten zijn, waardoor meer capaciteit overblijft voor monitoren, waardoor een meer accurate herinnering ontstaat die tot meer accurate zelfbeoordeling zou moeten leiden. Maar zelfs met een correcte herinnering aan de taakuitvoering kan het accuraat beoordelen van de eigen prestatie gehinderd worden. Lerenden baseren hun beoordeling vaak op verkeerde informatie (‘biases’). Tevens speelt kennis van beoordelingcriteria en -standaarden een rol, maar de ontwikkeling van dergelijke kennis lijkt hand in hand te gaan met ervaring in het domein.

Ten derde leiden inaccurate zelfbeoordelingen, wanneer deze gebruikt worden als input voor taakselectie, mogelijk tot de selectie van taken die niet passen bij het voorkennisniveau van de lerende. Als een lerende zijn/haar prestatie bijvoorbeeld overschat, kiest hij/zij waarschijnlijk een volgende taak gekozen die te moeilijk is. Maar zelfs als de zelfbeoordeling wel accuraat is, kunnen lerenden nog moeilijkheden ervaren in het kiezen van een geschikte taak. Bij het kiezen van een leertaak is het belangrijk onderscheid te kunnen maken tussen de aspecten van de taak die er voor het leren toe doen, zoals de structurele aspecten van de taak (bijv. niveau van

complexiteit, hoeveelheid ondersteuning), en aspecten van de taak die minder of niet belangrijk zijn voor het leren (bijv. de verhaaltjes die de leertaak inkleden). Onderzoek heeft aangetoond dat vooral novieten problemen hebben met het maken van dit onderscheid tussen relevante en irrelevante taakaspecten voor leren en de neiging hebben om taken te kiezen op aspecten die irrelevant zijn voor het leren. Wanneer taakselectie niet nauwkeurig is, zullen de gekozen taken niet passen bij het kennisniveau van de lerende en zal de lerende dus niet werken aan een taak die past bij zijn/haar leerbehoeftes.

Dit eerste hoofdstuk beschrijft dus bewijs dat lerenden, vooral novieten, problemen ervaren met accuraat monitoren, zelfbeoordelen en taken selecteren en beargumenteert dat deze processen een belangrijke rol spelen in de effectiviteit van zelfregulerend leren. Echter, het geeft ook verklaringen voor waarom lerenden deze problemen ervaren en biedt daarom een aanknopingspunt voor de ontwikkeling van interventies om deze problemen tegen te gaan en effectief zelfregulerend leren te stimuleren.

De studie in hoofdstuk 2 onderzocht welke aspecten van hun prestatie proefpersonen met verschillende expertiseniveaus in acht namen wanneer ze gevraagd werden hun prestatie te beoordelen. Om de vraag te beantwoorden of problemen met zelfbeoordeling te wijten zijn aan capaciteitsproblemen met monitoren, gebrekkige kennis van beoordelingscriteria en -standaarden, of beide, kreeg de helft van de proefpersonen een hulpmiddel (‘cue’) van het taakuitvoeringsproces te zien gedurende de zelfbeoordelingen, terwijl de andere helft van de deelnemers een blanco scherm te zien kreeg wanneer ze zichzelf moesten beoordelen. Door deze ‘cues’, die bestonden uit het afspelen van en opname van de oogbewegingen van de proefpersoon gecombineerd met de acties die de proefpersonen op de computer uitvoerden, zouden proefpersonen hun prestatie niet gelijktijdig met de taakuitvoering hoeven te monitoren, omdat deze ‘cues’ hen toestaan achteraf zowel fysieke handelingen (te zien aan hun bewerkingen met muis en toetsenbord) als mentale handelingen (te zien aan hun oogbewegingen) te bekijken. De verwachting was dat deelnemers met minder voorkennis (novieten) de ‘cues’ nodig hadden om te rapporteren over het taakuitvoeringsproces, terwijl deelnemers met meer voorkennis (de gevorderden) ze daarvoor niet nodig zouden hebben. Ook werd verwacht dat de ‘cues’ de novieten of wel, of niet konden helpen bij het beoordelen van de prestatie (afhankelijk van of de problematiek lag bij het monitoren of gebrekkige kennis van criteria en standaarden), maar dat de ‘cues’ in ieder geval de gevorderden zouden helpen bij het beoordelen van de prestatie, wat zich zou uiten in het gebruik van meer of andersoortige criteria. De resultaten lieten een interactie zien tussen de ‘cues’ en het niveau van voorkennis. Novieten hadden problemen met het zich herinneren van taakuitvoering, waarschijnlijk door hoge cognitieve belasting, en de cues hielpen hen zich deze beter te herinneren. De ‘cues’ hadden geen positief effect op het terugrapporteren van de taakuitvoering voor de gevorderden. Maar

waar de 'cues' de novieten niet leken te helpen met hun zelfbeoordelingen, waarschijnlijk vanwege een gebrek aan kennis van beoordelingcriteria en –standaarden, werden de gevorderden wel door de 'cues' geholpen bij hun zelfbeoordelingen. Deze studie brengt de ingewikkelde relatie tussen monitoren, zelfbeoordelen, en het niveau van expertise in zicht.

In hoofdstuk 3 wordt een exploratieve studie beschreven, waarin de verschillen in zelfbeoordeling- en taakselectieprocessen werden onderzocht tussen lerenden die meer en lerenden die minder leerden in een elektronische leeromgeving waarin ze zelfregulerend leerden. VWO-4 leerlingen werd gevraagd acht maal de volgende cyclus te doorlopen: een taak te kiezen, deze te maken, de prestatie op deze taak te beoordelen aan de hand van een lijst van criteria, een nieuwe taak te kiezen, et cetera. Gedurende het kiezen van leertaken en het beoordelen van prestaties werd leerlingen gevraagd hardop te denken. Op basis van de mate van toename tussen pre- en posttest scores werden de leerlingen achteraf toegewezen aan een effectieve (veel toename) of ineffectieve (nauwelijks toename) groep. De analyses op deze groepen laten zien dat de effectieve leerlingen meer accuraat hun prestatie beoordeelden vergeleken met de ineffectieve leerlingen. Tevens leken de uitkomsten van de zelfbeoordelingen de selectie van een volgende taak te voorspellen voor effectieve leerlingen, maar niet voor de ineffectieve leerlingen. Dit betekent echter niet dat leerlingen bewust gebruik maakten van de uitkomsten van zelfbeoordelingen in hun taakselecties; waar voor effectieve leerlingen zowel de hoeveelheid moeite als de tijdsinvestering in taakuitvoering voorspellers waren voor volgende taakselectie, werden deze door leerlingen respectievelijk gemiddeld en zeer laag in een ordening van relevantie geplaatst. Verder werden er geen verschillen gevonden tussen de groepen in de taakaspecten op basis waarvan de leerlingen een keuze voor een taak maakten. Deze studie laat zien dat er verschillen zijn in zelfbeoordeling- en taakselectieprocessen tussen effectieve en ineffectieve leerlingen. Deze bevindingen onderstrepen de noodzaak van een zelfbeoordeling- en taaksectie training als sleutel tot het verhogen van de effectiviteit van zelfregulerend leren.

Een dergelijke training werd onderzocht in de twee experimenten gerapporteerd in hoofdstuk 4. Het eerste experiment in dit hoofdstuk onderzocht of een training bestaande uit het observeren van voorbeelden, dat wil zeggen, modellen die een taak uitvoerden, de eigen prestatie beoordeelden en een volgende taak kozen, tot meer accurate zelfbeoordeling- en taakselectie kon leiden bij VWO-4 leerlingen. In de experimentele condities kreeg men vier maal een voorbeeld te zien van taakuitvoering door het model en, afhankelijk van de toebedeelde conditie, vervolgens ook hoe het model die prestatie beoordeelt, een nieuwe taak kiest op basis van deze beoordeling, of beide. In de controleconditie kreeg men ook de voorbeelden van taakuitvoering door het model te zien en moest men de fouten in deze voorbeelden proberen te vinden en te verbeteren. De resultaten laten zien dat leerlingen die

voorbeelden van zelfbeoordeling of taakselectie kregen op de posttest accurater waren in deze vaardigheden.

Het tweede experiment in hoofdstuk 4 onderzocht bij HAVO-3 leerlingen of het trainen van zelfbeoordeling- en taakselectievaardigheden, ofwel door voorbeelden ofwel door oefenen, de effectiviteit van zelfregulerend leren verhoogde. In dit tweede experiment werd dus ook gekeken of een andere training ook effectief zou zijn. Leerlingen in de voorbeeldconditie keken naar voorbeelden van prestatie, zelfbeoordeling en taakselectie, zoals in Experiment 1. Leerlingen in de oefenconditie kregen eerst de criteria en regels uitgelegd en moesten konden na elk voorbeeld van taakuitvoering door het model de prestatie van het model beoordelen en een taak kiezen voor het model. Leerlingen in de controleconditie moesten wederom fouten zoeken en verbeteren in de taakuitvoering van het model zoals in Experiment 1. Na de training (of het zoeken van fouten) werd er zelfregulerend geleerd in een elektronische leeromgeving (vergelijkbaar met de leeromgeving in hoofdstuk 3) waarin leerlingen wederom acht maal de cyclus doorliepen van een taak kiezen, deze maken, en de prestatie beoordelen, om dan weer een volgende taak te kiezen, etc. De resultaten lieten zien dat leerlingen in beide trainingcondities hogere toenames in leerresultaten behaalden in vergelijking met de controleconditie. Beide trainingscondities verschilden niet van elkaar in behaalde resultaten. Samengevat resulteerden de twee experimenten in hoofdstuk 4 in een belangrijke bevinding voor de onderwijspraktijk door te laten zien dat met een relatief kleine interventie gericht op het verbeteren van zelfbeoordeling- en taakselectievaardigheden deze meer accuraat worden en een significante toename bewerkstelligd wordt in de hoeveelheid kennis die opgedaan wordt door middel van zelfregulerend leren.

Tenslotte wordt in hoofdstuk 5 een samenvatting gegeven van de bevindingen van dit proefschrift, worden de beperkingen van het onderzoek evenals theoretische en praktische implicaties uiteengezet en worden suggesties voor toekomstig onderzoek gegeven. Gezamenlijk laten de hoofdstukken in dit proefschrift zien dat nauwkeurige zelfbeoordeling en taakselectie nodig is voor effectief zelfregulerend leren en dat training van zelfbeoordeling- en taakselectievaardigheden tot effectiever zelfregulerend leren leidt, zoals blijkt uit de hogere leeruitkomsten.

ICO dissertation series

ico

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the following universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Groningen, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University (and formerly Radboud University Nijmegen and Tilburg University).

139. Gijlers, A.H. (23-09-2005). *Confrontation and co-construction: Exploring and supporting collaborative scientific discovery learning with computer simulations.* Enschede: University of Twente.
140. Stevenson, M.M.C. (27-09-2005). *Reading and writing in a foreign language: A comparison of conceptual and linguistic processes in Dutch and English.* Amsterdam: University of Amsterdam.
141. Saab, N. (14-10-2005). *Chat and explore: The role of support and motivation in collaborative scientific discovery learning.* Amsterdam: University of Amsterdam.
142. Löhner, S. (11-11-2005). *Computer-based modeling tasks: The role of external representation.* Amsterdam: University of Amsterdam.
143. Beers, P.J. (25-11-2005). *Negotiating common ground: Tools for multidisciplinary teams.* Heerlen: Open University of the Netherlands.
144. Tigelaar, E.H. (07-12-2005). *Design and evaluation of a teaching portfolio.* Maastricht: Maastricht University.
145. Van Drie, J.P., (20-12-2005). *Learning about the past with new technologies. Fostering historical reasoning in computer-supported collaborative learning*. Utrecht: Utrecht University.
146. Walrecht, E.S. (09-01-2006). *Brede innovatie, passende strategie?: De Groninger Vensterschool als casus van onderzoek naar strategie en invoering.* Groningen: University of Groningen.
147. De Laat, M. (03-02-2006). *Networked learning*. Utrecht: Utrecht University.
148. Prince, C.J.A.H. (21-04-2006). *Problem-based learning as a preparation for professional practice.* Maastricht: Maastricht University.
149. Van Gog, T. (28-04-2006). *Uncovering the problem-solving process to design effective worked examples.* Heerlen: Open University of the Netherlands.
150. Sins, P.H.M. (18-05-2006). *Students' reasoning during computer-based scientific modeling.* Amsterdam: University of Amsterdam.
151. Mathijsen, I.C.H. (24-05-2006). *Denken en handelen van docenten.* Utrecht: Utrecht University.
152. Akkerman, S.F. (23-06-2006). *Strangers in dialogue: Academic collaboration across organizational boundaries.* Utrecht: Utrecht University.
153. Willemse, T.M. (21-08-2006). *Waardenvol opleiden: Een onderzoek naar de voorbereiding van aanstaande leraren op hun pedagogische opdracht.* Amsterdam: VU University Amsterdam.

154. Kieft, M. (19-09-2006). *The effects of adapting writing instruction to students' writing strategies.* Amsterdam: University of Amsterdam.
155. Vreman-de Olde, G.C. (27-09-2006). *Look experiment design: Learning by designing instruction.* Enschede: University of Twente.
156. Van Amelsvoort, M. (13-10-2006). *A space for debate: How diagrams support collaborative argumentation-based learning.* Utrecht: Utrecht University.
157. Oolbekking-Marchand, H. (9-11-2006). *Teachers' perspectives on self-regulated learning: An exploratory study in secondary and university education.* Leiden: Leiden University.
158. Gulikers, J. (10-11-2006). *Authenticity is in the eye of the beholder: Beliefs and perceptions of authentic assessment and the influence on student learning*. Heerlen: Open University of the Netherlands.
159. Henze, I. (21-11-2006). *Science teachers' knowledge development in the context of educational innovation.* Leiden: Leiden University.
160. Van den Bossche, P. (29-11-2006). *Minds in teams: The influence of social and cognitive factors on team learning.* Maastricht: Maastricht University.
161. Mansvelder-Longayroux, D.D. (06-12-2006). *The learning portfolio as a tool for stimulating reflection by student teachers.* Leiden: Leiden University.
162. Visschers-Pleijers, A.J.S.F. (19-01-2007). *Tutorial group discussion in problem-based learning: Studies on the measurement and nature of learning-oriented student interactions.* Maastricht: Maastricht University.
163. Poortman, C.L. (16-02-2007). *Workplace learning processes in senior secondary vocational education.* Enschede: University of Twente.
164. Schildkamp. K.A. (15-03-2007). *The utilisation of a self-evaluation instrument for primary education.* Enschede: University of Twente.
165. Karbasioun, M. (20-04-2007). *Towards a competency profile for the role of instruction of agricultural extension professionals in Asfahan.* Wageningen: Wageningen University.
166. Van der Sande, R.A.W. (04-06-2007). *Competentiegerichtheid en scheikunde leren: Over metacognitieve opvattingen, leerresultaten en leeractiviteiten.* Eindhoven: Eindhoven University of Technology.
167. Pijls, M. (13-06-2007). *Collaborative mathematical investigations with the computer: Learning materials and teacher help.* Amsterdam: University of Amsterdam.
168. Könings, K. (15-06-2007). *Student perspectives on education: Implications for instructional design.* Heerlen: Open University of the Netherlands.
169. Prangsma, M.E. (20-06-2007). *Multimodal representations in collaborative history learning.* Utrecht: Utrecht University.
170. Niemantsverdriet, S. (26-06-2007). *Learning from international internships: A reconstruction in the medical domain.* Maastricht: Maastricht University.
171. Van der Pol, J. (03-07-2007). *Facilitating online learning conversations: Exploring tool affordances in higher education.* Utrecht: Utrecht University.
172. Korobko, O.B. (07-09-2007). *Comparison of examination grades using item response theory: A case study*. Enschede: University of Twente.
173. Madih-Zadeh, H. (14-09-2007). *Knowledge construction and participation in an asynchronous computer-supported collaborative learning environment in higher education.* Wageningen: Wageningen University.
174. Budé, L.M. (05-10-2007). *On the improvement of students' conceptual understanding in statistics education.* Maastricht: Maastricht University.
175. Meirink, J.A. (15-11-2007). *Individual teacher learning in a context of collaboration in teams.* Leiden: Leiden University.
176. Niessen, T.J.H. (30-11-2007). *Emerging epistemologies: Making sense of teaching practices.* Maastricht: Maastricht University.
177. Wouters, P. (07-12-2007). *How to optimize cognitive load for learning from animated models.* Heerlen: Open University of the Netherlands.
178. Hoekstra, A. (19-12-2007). *Experienced teachers' informal learning in the workplace.* Utrecht: Utrecht University.
179. Munneke-de Vries, E.L. (11-01-2008). *Arguing to learn: Supporting interactive argumentation through computer-supported collaborative learning*. Utrecht: Utrecht University.

180. Nijveldt, M.J. (16-01-2008). *Validity in teacher assessment. An exploration of the judgement processes of assessors*. Leiden: Leiden University.
181. Jonker, H.G. (14-02-2008). *Concrete elaboration during knowledge acquisition*. Amsterdam: VU University Amsterdam.
182. Schuitema, J.A. (14-02-2008). *Talking about values. A dialogue approach to citizenship education as an integral part of history classes*. Amsterdam: University of Amsterdam.
183. Janssen, J.J.H.M. (14-03-2008). *Using visualizations to support collaboration and coordination during computer-supported collaborative learning*. Utrecht: Utrecht University.
184. Honingh, M.E. (17-04-2008). *Beroepsonderwijs tussen publiek en privaat: Een studie naar opvattingen en gedrag van docenten en middenmanagers in bekostigde en niet-bekostigde onderwijsinstellingen in het middelbaar beroepsonderwijs.* Amsterdam: University of Amsterdam.
185. Baartman, L.K.J. (24-04-2008). *Assessing the assessment: Development and use of quality criteria for competence assessment programmes*. Utrecht: Utrecht University.
186. Corbalan Perez, G. (25-04-2008). *Shared control over task selection: Helping students to select their own learning tasks*. Heerlen: Open University of the Netherlands.
187. Hendrikse, H.P. (22-05-2008). *Wiskundig actief: Het ondersteunen van onderzoekend leren in het wiskunde onderwijs*. Enschede: University of Twente.
188. Moonen, M.L.I. (26-09-2008). *Testing the multi-feature hypothesis: Tasks, mental actions and second language acquisition*. Utrecht: Utrecht University.
189. Hooreman, R.W. (18-11-2008). *Synchronous coaching of the trainee teacher: An experimental approach*. Eindhoven: Eindhoven University of Technology.
190. Bakker, M.E.J. (02-12-2008). *Design and evaluation of video portfolios: Reliability, generalizability, and validity of an authentic performance assessment for teachers.* Leiden: Leiden University.
191. Kicken, W. (12-12-2008). *Portfolio use in vocational education: Helping students to direct their learning.* Heerlen: Open University of the Netherlands.
192. Kollöffel, B.J. (18-12-2008). *Getting the picture: The role of external representations in simulation-based inquiry learning*. Enschede: University of Twente.
193. Walraven, A. (19-12-2008). *Becoming a critical websearcher: Effects of instruction to foster transfer.* Heerlen: Open University of the Netherlands.
194. Radstake, H. (14-05-2009).*Teaching in diversity: Teachers and pupils about tense situations in ethnically heterogeneous classes*. Amsterdam: University of Amsterdam.
195. Du Chatenier, E. (09-09-2009). *Open innovation competence: Towards a competence profile for inter-organizational collaboration in innovation teams*. Wageningen: Wageningen University.
196. Van Borkulo, S.P. (26-06-2009). *The assessment of learning outcomes of computer modelling in secondary science education*. Enschede: University of Twente.
197. Handelzalts, A. (17-09-2009). *Collaborative curriculum development in teacher design teams*. Enschede: University of Twente.
198. Nievelstein, F.E.R.M. (18-09-2009). *Learning law: Expertise differences and the effect of instructional support*. Heerlen: Open University of the Netherlands.
199. Visser-Wijnveen, G.J. (23-09-2009). *The research-teaching nexus in the humanities: Variations among academics*. Leiden: Leiden University.
200. Van der Rijst, R.M. (23-09-2009). *The research-teaching nexus in the sciences: Scientific research dispositions and teaching practice*. Leiden: Leiden University.
201. Mainhard, M.T. (25-09-2009). *Time consistency in teacher-class relationships*. Utrecht: Utrecht University.
202. Van Ewijk, R. (20-10-2009). *Empirical essays on education and health*. Amsterdam: University of Amsterdam.
203. Seezink, A. (18-11-2009). *Continuing teacher development for competence-based teaching*. Tilburg: Tilburg University.
204. Rohaan, E.J. (09-12-2009). *Testing teacher knowledge for technology teaching in primary schools*. Eindhoven: Eindhoven University of Technology.
205. Kirschner, F.C. (11-12-2009). *United brains for complex learning*. Heerlen: Open University of the Netherlands.
206. Wetzels, S.A.J. (18-12-2009). *Individualized strategies for prior knowledge activation*. Heerlen: Open University of the Netherlands.