


Self-regulation of secondary school students: self-assessments are inaccurate and insufficiently used for learning-task selection

Michelle L. Nugteren^{1,3}  · Halszka Jarodzka^{1,2} · Liesbeth Kester^{1,3} · Jeroen J. G. Van Merriënboer^{1,4}

Received: 29 September 2016 / Accepted: 15 January 2018 / Published online: 23 January 2018
© The Author(s) 2018. This article is an open access publication

Abstract Self-assessment and task selection are important self-regulated learning skills for secondary school students. More specifically, selecting new tasks based on self-assessments is very important for them, because teachers are not always present or able to select tasks for them individually. However, little is known about the processes underlying these self-regulated learning skills, and thus no guidelines exist for teaching self-assessment and the selection of subsequent learning-tasks. We propose a model for self-regulated learning-task selection (SRLTS) which represents a possible pathway for the task-selection process, and which students could use as a norm when making task selections. The model could help students to decide what possible new tasks might be suitable for their current skill level, based on self-assessments. The aim of this study is to evaluate to what extent secondary school students select learning tasks according to this model, and whether they use self-assessments to this end. Secondary school students ($N = 15$) selected learning tasks in the domain of genetics from a structured task database. The tasks varied in difficulty and amount of support provided (i.e., completion problems vs. traditional problems). We used eye tracking, performance estimates, estimates of mental effort, judgments of learning, and open questions to gain more insight in what students focus on and think about when selecting a task. Results suggest that students roughly follow the SRLTS model, but they base their decisions on inaccurate self-assessments. This implies that students might benefit from self-assessment and task-selection advice, which could

✉ Michelle L. Nugteren
M.L.Nugteren@uu.nl

¹ Faculty of Psychology and Educational Sciences, Welten Institute Research Center for Learning, Teaching and Technology, Open University of the Netherlands, P.O. Box 2960, 6401 DL Heerlen, The Netherlands

² Humanities Lab, Lund University, Box 201, SE-221 00 Lund, Sweden

³ Department of Education and Pedagogy – Education, Faculty of Social and Behavioral Sciences, Utrecht University, P.O. Box 80127, 3508 TC Utrecht, The Netherlands

⁴ Faculty of Health, Medicine and Life Sciences, School of Health Professions Education (SHE), Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

provide feedback on self-assessments and stimulate students to use self-assessment information as input for task selection in the way the model prescribes to optimize their learning.

Keywords Self-assessment · Task selection · Eye tracking · Judgments of learning · Mental effort

Introduction

Self-regulated learning is becoming more and more important in secondary schools (Istance and Dumont 2010). Insight into self-regulated learning processes might shed light on how to improve current instructional methods. Therefore, this article focuses on the processes taking place during self-regulated learning. Specifically, it focuses on selecting appropriate next learning tasks by using self-assessments of one's own prior performance. This is an important skill for secondary school students, because they often need to practice many different tasks when learning a new skill. For instance, when learning how to solve Mendelian genetics problems, students need to study many example solutions and practice many different problems with different difficulty levels and different contexts before mastering this skill (e.g., Bruner 1960; Van Merriënboer and Kirschner 2013). Teachers often only select examples and problems for an entire class of students, which means students usually do not practice at a level that is most appropriate for them individually. In addition, when students are practicing at home for a test, their teachers are not present to help them select practice tasks. Hence, it is useful when students are able to select appropriate tasks themselves.

In this article, we will describe a model that shows one way in which students could select learning tasks by using self-assessments (note that we define tasks as both examples and problems). This model is adapted from Zimmerman's self-regulated learning model (Zimmerman 2002; Zimmerman and Campillo 2003) and Van Merriënboer and Kirschner's (2013) model for dynamic task selection. We explored to what extent secondary school students select tasks according to this adapted model. In addition, we will describe the important role of self-assessments for task selection, and why it is important for self-assessments to be accurate.

The self-regulated learning-task selection (SRLTS) model

Most self-regulated learning models include an interaction between monitoring processes and controlling actions (Bjork et al. 2013; Efklides 2011; Nelson and Narens 1990; Panadero 2017; Winne 2001; Winne and Hadwin 1998; Zimmerman 2002; Zimmerman and Campillo 2003). When students practice one or more learning tasks, they can *monitor* (i.e., self-assess) their performance and use this self-assessment information to *control* their next learning step. A next learning step could focus on the selection of new learning tasks, aimed at practicing a new skill. However, in most models of self-regulated learning, the focus is on how students regulate their learning *within* a task (Loyens et al. 2008). Another line of research that deals with a similar topic is self-directed learning (SDL). In these models, the focus is on how students regulate their learning *between* tasks (Loyens et al. 2008).

Currently existing SRL models often provide analyses of task aspects and problem-solving strategies for specific tasks (Winne and Hadwin 1998; Zimmerman and Campillo

2003). In addition, the Conditions, Operations, Products, Evaluations, and Standards (COPEs) model (Winne 2001; Winne and Hadwin 1998) includes planning for future learning in one of its phases. However, these models do not focus specifically on selection of the task itself. Current SDL models focus more on students choosing their own goals, but these models do not specifically focus on the process of task selection itself either (Loyens et al. 2008). This lack of specificity about the task-selection process in both research lines makes it difficult for secondary school students to use SRL/SDL models for this goal, because they do not specify *how* self-assessments might be used for task selection. Thus, there are no practical guidelines on how students could select tasks during practice. Studies have shown that students do not seem very good at selecting learning tasks themselves without training or guidance. For instance, even more accurate self-assessors select the same learning tasks as less-accurate students (Kostons et al. 2010).

To overcome the gap between these two research lines and provide students with a norm for effective task selection, we propose an adapted model that focuses on the task-selection process from a self-regulated learning perspective. We present a model for self-regulated learning-task selection (SRLTS; see Fig. 1) describing a systematic way in which students could select tasks to develop their knowledge and skills. The SRLTS model could help both students when selecting tasks, as well as advance current theories on task selection by providing insight in the highly specified way in which the task-selection process can occur. This can then be used for adaptive task selections, which have been shown to lead to more efficient learning than non-adaptive task selections (e.g., Corbalan et al. 2008).

This model shows one possible theory-based elaboration of the task-selection process. Students can *self-assess* their performance on previous tasks (arrow a), which includes making performance estimates, mental effort ratings, and Judgments of Learning (JOLs; e.g., Koriat 1997). They can compare these assessments to *standards* (arrows c, e, and g, which are described in Sect. [The SRLTS model in action](#)).

Note that the definitions of self-assessment vary on what measures are included in self-assessments (see Panadero et al. 2016, for an overview of different ways to categorize self-assessments; and De Bruin and Van Gog 2012, for different measures that can be used for self-assessments). We have decided to include assessment measures of the process of learning, namely performance estimates, mental effort ratings and JOLs. Performance estimates are a common component of self-assessments, and mental effort ratings have

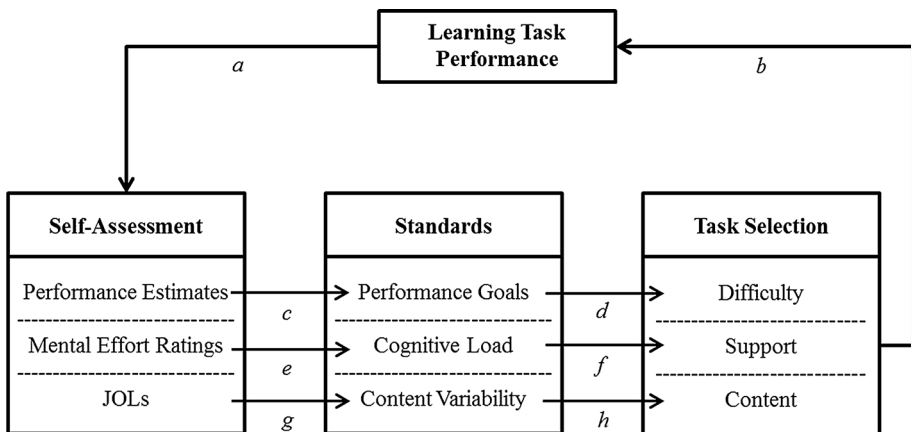


Fig. 1 Model for self-regulated learning-task selection

been shown to be a good indicator of *cognitive load* (Paas 1992). These measures are qualified as retrospective monitoring. JOLs are estimates of future performance, and are qualified as prospective monitoring (Baars, Vink, Van Gog et al. 2014). Together, these three assessment measures represent different elements of the learning process, they are all task-specific measures, and they are all aimed at providing input for future improvements. This is similar to the definition of self-assessments by Andrade (2010). We have defined self-assessment as the process by which students make judgments, based on a set of criteria, about their performance on previous learning tasks and the resulting outcomes (Boud and Falchikov 1989). In this definition, self-assessment is regarded as a part of the monitoring process (Bjork et al. 2013; Nelson and Narens 1990), although it can be argued that the two terms are different. For instance, self-assessment can also be considered as an overall statement about a certain skill level in general, in which case self-monitoring is considered as an assessment after individual tasks (Eva and Regehr 2011). Regardless of the name, a moment-to-moment assessment is considered to be more accurate than a global judgement (Eva and Regehr 2011). In the current study, we decided to stay in line with most other studies on self-assessment and task selection by using the term self-assessment for the assessment of individual tasks (e.g., Kicken et al. 2009; Kostons et al. 2010, 2012; Taminau et al. 2013).

Self-assessments are a crucial part of self-regulated learning (Andrade 2010; Panadero and Alonso-Tapia 2013). It is therefore important for self-assessments to be accurate (Brown et al. 2015; Brown and Harris 2014; Panadero et al. 2016). When self-assessment are inaccurate, this leads to errors in the next steps of the self-regulated learning process (Brown and Harris 2014). Both over- and underconfidence can negatively impact the accuracy of self-assessments (Dunlosky and Rawson 2012). However, many studies have shown that accurately assessing themselves is often difficult for students (e.g., Brown et al. 2015; Panadero et al. 2016). The accuracy of self-assessments depends on many factors, including expertise (novices are less accurate than experts; e.g., Dunning et al. 2003) and having clear self-assessment criteria (Panadero et al. 2016). These criteria can be externally imposed, for instance, by a teacher, or they can be internally imposed by the students themselves in the form of internal standards (Andrade and Du 2007; Panadero and Alonso-Tapia 2013). Note that these internal standards can be different from externally imposed standards (Andrade and Du 2007). By comparing self-assessment information and standards, students can *control task selection* (arrow b) by determining the difficulty and support levels as well as the contents of suitable tasks. Gradually increasing difficulty levels is necessary for acquiring the desired skill in a so-called 'spiral curriculum' (Bruner 1960). Within a difficulty level, tasks with varying levels of support can provide different amounts of cognitive load. Tasks with different levels of support include worked examples, in which both a problem statement and all the worked-out solution steps are provided to the learner, and completion problems, in which a problem statement is provided but only part of the solution is worked out. Traditional problems only provide problem statements and, thus, offer hardly any support (Sweller et al. 1998; Van Merriënboer and Kirschner 2013).

At each level of difficulty, beginning students should benefit most from studying a worked example, because worked examples should direct their attention to the problem states and solution steps, and thus should result in low cognitive load (also known as the 'worked example effect'; Sweller et al. 1998). More advanced students should benefit more from completion problems, which should stimulate students to study the solution steps more carefully than with worked examples. Students cannot complete the solution if they do not understand the given steps (known as the 'completion problem effect'; Sweller et al. 1998). Finally, when the students are experienced, traditional problems should be most

suitable as they allow students to perform all solution steps themselves (Newell and Simon 1972; Van Merriënboer and Kirschner 2013). This fading of task support through problem types at one level of task difficulty is also known as ‘scaffolding’ (Renkl and Atkinson 2003). It reflects the idea that instructional methods that work well for novice learners are different from methods that work well for more experienced learners. Specifically, students who just start to work on a particular difficulty level should benefit from high support such as given by worked examples, while this can be detrimental for more experienced students who profit more from traditional problems (known as the ‘expertise reversal effect’; e.g., Kalyuga et al. 2003).

Development of the SRLTS model

The basis for the design of the SRLTS model were Zimmerman’s self-regulated learning model (Zimmerman 2002; Zimmerman and Campillo 2003) and Van Merriënboer and Kirschner’s (2013) model for dynamic task selection. We adapted the cyclical path from Zimmerman’s model into the SRLTS model, because it provides a highly adaptable model base for self-regulated learning. We made an adaptation into a specific model for learning-task selection, because Zimmermann’s basic cycle does not specify which factors could be important for learning-task selection. Similarly, we based the SRLTS model on the model for dynamic task selection (Van Merriënboer and Kirschner 2013), because it accounts for many factors that are important for learning-task selection. However, this model was originally designed as a selection model to be used by teachers. Thus, we transformed it into a self-regulated learning model, which represents the selection process when students make task selections themselves.

Two factors that we added to the SRLTS model that are not explicitly specified in Zimmerman’s self-regulated learning model, and Van Merriënboer and Kirschner’s model for dynamic task selection are standards, and a link between JOLs and content selection. Standards resemble a central part of the COPEs model (Winne 2001; Winne and Hadwin 1998). We added this to stress, just like the COPEs model does, that calibration should be optimal to ensure that the self-regulated learning process runs effectively (Winne 2004). This means that self-assessments need to be accurate to ensure that the task selections are based on correct information, because only then can the task selections be effective. Finally, we added content selection based on an evaluation of variability over previous tasks, because learning-task selection often includes content options that make it possible to work on a highly varied set of tasks. These facilitate generalization and abstraction, and thus transfer of learning (Sweller et al. 1998). We linked content selection to JOLs, because JOLs affect task selections (Bjork et al. 2013; Metcalfe and Finn 2008) and can be phrased to focus on similar content on subsequent tests (e.g., Baars et al. 2013).

The SRLTS model in action

Students can use the SRLTS model as a normative model to decide what might be a suitable next task, based on their self-assessments. The model shows them the path they could follow to select a task that matches their current skill level. Students can follow this cycle each time they want to select a next task.

To choose the *difficulty level* of a next task in the SRLTS model, students can first *estimate their performance* from the last task they performed. Performance estimates are then compared to the *internal performance goals* set for the task (arrow c), which indicate the desired learning outcome. Students could use the distance between the current

performance state and the goal state to choose a difficulty level for the next task (e.g., Locke and Latham 1990; Zimmerman and Campillo 2003). When the performance estimates are equal to or higher than the goal, the student could select a task with a higher difficulty level. On the other hand, if performance is lower than the goal, the student can select tasks with the same or an easier difficulty level. Students might go to an easier level when the goal is still very far away. On the other hand, if they are very close to the goal, they could choose to repeat the same level (arrow d).

Similarly, students can use *mental effort ratings* to choose the *support level* of the new learning task (Van Merriënboer and Kirschner 2013). They can compare their current level of invested mental effort to the acceptable internal standard for cognitive load (arrow e). Students can reduce high load by selecting a task with more support, such as by moving from a completion assignment to a worked example or from a traditional task to a completion problem. Students can increase the cognitive load by selecting a task with less support, such as by moving from a completion assignment to a traditional problem or from a worked example to a completion assignment. This way, students can regulate support to match their current needs (arrow f; Paas et al. 2003; Renkl and Atkinson 2003; Sweller et al. 1998). Similar to choosing a difficulty level, students could decide to repeat a support level when they have experienced moderate mental effort and would like practice more with the same support level. When combining the levels, students might also choose to increase or decrease a level and keep another level constant.

Finally, students might also choose between tasks with different *contents*. It is important to practice with a high variability of contents to cover as much as possible of the learning domain (arrow g). This is known as the ‘variability principle’ (Sweller et al. 1998). Practicing varied tasks in different contexts increases the ability to recognize essential task elements in new contexts (Paas and Van Merriënboer 1994; Sweller et al. 1998; Van Merriënboer and Kirschner 2013) which benefits the transfer of learning. Students can use JOLs to determine content variability, and use this for content selection (arrow h). Variability can be low if the content variability of the tasks students can select from is low, or when students do not utilize the full range of available content due to limited personal interests. JOLs can indicate this by asking students if they could perform similar tasks on a future test. If they feel they cannot, they should increase content variability to increase transfer.

Since most students are not very good at assessing themselves (e.g., Brown et al. 2015; Panadero et al. 2016), the model could also be used for the design of self-assessment and task-selection advice. When students do not master the skills that are required to follow the pathways described in the model, they might benefit from guidance to help them acquire those skills. Studies on advice for task selection are described below.

Supporting the task-selection process

There are two options for students to acquire the self-assessment and task-selection skills as described in the SRLTS model. First, it is possible to provide guidance to support the acquisition of these skills. Previous studies on guidance for self-assessment and task-selection skill development have shown mixed results. Specifically, modeling by a teacher of self-assessment and task-selection skills can lead to higher performance, but not always more than mere practice (Kostons et al. 2012; Van Gog et al. 2010). Also, providing task-selection advice might even hamper the learning of task-selection skills (Taminiau et al. 2013). On the other hand, results from other studies do suggest that self-assessment skills can be improved through guidance (e.g., Andrade 2010; Panadero and Alonso-Tapia 2013), and that task-selection advice could improve task selections (Kicken et al. 2009). Together,

these results suggest that guidance for self-assessment and task selection could be helpful, but might not always be necessary for the acquisition of these skills. According to the SRLTS model, students would not need guidance if they already base their task selection on all three relevant task characteristics in the model, which are difficulty, support, and content. Second, students could improve through practice (such as distributed practice, e.g., Cepeda et al. 2009). Students might master the cyclical path in the SRLTS model by simply repeating it several times. If this were the case, students would not need guidance to learn task-selection skills, which could also explain the mixed results from task-selection guidance studies mentioned above.

This study

The goal of the current study is to evaluate the SRLTS model by investigating whether students follow the cyclical path described in the model, and whether their self-assessments are accurate or not. Earlier studies have uttered a need to yield objective information on which factors students take into account when selecting tasks, for instance by studying the task-selection process with process measures such as eye tracking (Corbalan et al. 2008, 2009b). The current study is the first to do this and to investigate task selection and self-assessment with eye tracking as a process measure.

The first research question is (RQ1): Do secondary school students take difficulty, support and content into account when selecting learning tasks without guidance? The answer to this question will help to evaluate the cyclical path from the SRLTS model. Furthermore, results could indicate whether students need guidance for learning-task selection.

As mentioned above, the accuracy of self-assessments is important for adequate self-regulated learning. Inaccurate self-assessments can lead to suboptimal task selections. For example, students might select a task that is too difficult for them if they have incorrectly self-assessed their performance as high. Previous studies have shown that students often do not make accurate self-assessments (e.g., Brown et al. 2015; Dunning et al. 2003; Panadero et al. 2016). Hence, the second research question is (RQ2): How do inaccurate self-assessments influence task selections?

Finally, it might be possible for self-assessment accuracy and task selections to improve through mere practice. Boud and Falchikov (1989) suggest that more research is needed before statements can be made about the improvement of self-assessments over time; especially in the context of task selection, such improvements have not been studied before. However, the possibility of self-assessment accuracy and task-selection improvement over time could have implications for potential task-selection and self-assessment guidance. If students improve over time through practice alone, they might not need guidance for this process. Therefore, the third research question is (RQ3): Does the quality of self-assessments and unsupported task selections improve over time?

Method

Participants

Participants were 15 students from Dutch higher general and pre-university secondary schools ($M_{age} = 13.93$, $SD = 1.49$ years; 8 females), recruited through local

advertisements. Participation in this experiment was voluntarily and extra-curricular. Participants gave informed consent. Each participant received a gift certificate of €10 at the end of the experiment.

Measures and materials

Genetics tasks

Participants studied tasks in the domain of genetics and had to determine genotypes (i.e., the expression of a particular trait needed to be determined from the hereditary characteristics of the parents). The tasks focused on various parent characteristics, such as eye color, hair type, and several diseases and syndromes. Each task required participants to determine the genotype of one or two descendants from the task description by following five steps: (1) determine the genotypes, (2) make a family tree, (3) determine the number and direction of the contingency tables, (4) fill in the contingency tables, and (5) choose the right answer from the contingency tables. These steps had to be completed for each task in this order, thus providing a clear start and end for each task. See the Appendix for an example of a task. Each correctly performed step received one point, with a maximum of five points per task.

Task database

The table in Fig. 2 displays 75 genetics tasks. Participants selected tasks from this table on a computer screen. The 75 tasks were divided over five difficulty levels and, within each difficulty level, three support levels, as indicated by different columns and colors. Difficulty varied on four factors. First, the number of generations present in a task varied with the different levels. Tasks on the higher difficulty levels involved more generations. Second, participants had to determine one genotype on the lower four levels, and two

Level 1 - 2 generations - 1 unknown - 1 solution - deductive			Level 2 - 2 generations - 1 unknown - several solutions - deductive			Level 3 - 2 generations - 1 unknown - several solutions - inductive			Level 4 - 3 generations - 1 unknown - several solutions - both ways			Level 5 - 3 generations - 2 unknown - several solutions - both ways		
1/2 done 1/2 self	1/4 done 1/4 self	Everything yourself	1/2 done 1/2 self	1/4 done 1/4 self	Everything yourself	1/2 done 1/2 self	1/4 done 1/4 self	Everything yourself	1/2 done 1/2 self	1/4 done 1/4 self	Everything yourself	1/2 done 1/2 self	1/4 done 1/4 self	Everything yourself
Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color	Eye Color
Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape	Hair Shape
Cat Fur	Huntington (disease)	Wolman Disease	Apert Syndrome	Flower Color	Huntington (disease)	Fruit Flies	Tail Length Chimpanzee	Tongue Rolling	Albinism	Fruit Flies	Cleft Lip	Milk Allergy	Tail Length Chimpanzee	Flowering Period Apple Tree
Flowering Period Apple Tree	Beak Chicken	Tail Length Chimpanzee	Fruit Flies	Flowering Period Apple Tree	Albinism	Beak Chicken	P.R.A. (disease dogs)	Depression	Cat Fur	Flower Color	P.R.A. (disease dogs)	Apert Syndrome	Wolman Disease	Fruit Flies
Depression	Cleft Lip	Milk Allergy	P.R.A. (disease dogs)	Tongue Rolling	Breast Cancer	Wolman Disease	Milk Allergy	Apert Syndrome	Depression	Tongue Rolling	Breast Cancer	Huntington (disease)	Flower Color	Beak Chicken

Fig. 2 Visual representation of the set of 75 tasks. Participants used this table to select eight tasks. The top row indicated the five difficulty levels and the different columns represented the different support levels (cf. Corbalan et al. 2008, 2009a, 2009b, 2011; Kostons et al. 2010, 2012; Van Gog et al. 2010)

genotypes on the highest difficulty level. Third, the easiest tasks only had one right answer, while the more difficult tasks had multiple possible answers. Participants needed to find all possible answers to score the maximum amount of points on those tasks. Finally, the easiest tasks only required deductive reasoning to find the right solution. Tasks on the third level required inductive reasoning, and the most difficult tasks required both types of reasoning.

The top row in Fig. 2 indicates the five difficulty levels. Three different types of support were present within each difficulty level. The most left columns (which are labeled “½ done, ½ yourself”) represented completion problems for which half of the solution was already worked out. Steps 1 and 3 were fully completed, and steps 2 and 4 were half completed; thus, participants themselves needed to perform half of steps 2 and 4 as well as step 5. The tasks in the middle columns represented completion problems where a quarter of each task’s solution was worked out. This means participants had to perform half of steps 1, 2 and 3, as well as steps 4 and 5. The tasks in the most right columns (which are labeled “Everything yourself”) represented traditional problems in which participants had to perform all five steps. For each difficulty and support level, participants could choose from different contents. Each participant selected and performed eight of these tasks. All tasks and the task database were derived from Corbalan et al. (2008, 2009a, 2009b, 2011), Kostons et al. (2010, 2012), and Van Gog et al. (2010).

Reasons for task selection

Participants wrote down their reasons for selecting each task. Their answers were categorized by determining whether a selection was based on content (i.e., the varying parent characteristics), difficulty level, or support level. For some answers, more than one category applied. One example of an answer from the category of *content* was “[I want] to do as many different topics as possible”. An example answer from the category of *difficulty* was “I found the previous task rather difficult, so I thought it would be wiser to do a task that is a little bit easier again”. An example answer that was coded as *support* was “I chose the color blue [= low support] because I thought I could do some of it without a lot of help, but might still need some help”. Actual task selections were also considered when coding the answers. For instance, when a participant wrote down that s/he selected a more difficult task than the previous one, but s/he switched from a task with low support to a task with no support, this was coded as a reason referring to support. So when the written answer and the actual selection were different, the answer was coded based on the actual selection of a new task.

Performance estimate

Participants answered the question “Indicate how many steps you think you performed correctly on the task” on a scale ranging from “zero steps” to “five steps”. The answers indicated how participants estimated their own performance.

Mental effort rating scale

Mental effort was measured with the question: “How much effort did this task cost?” Participants answered this question on a 9-point scale, ranging from “very, very low

mental effort” to “very, very high mental effort” (Paas 1992). The answers indicated participants’ perceived mental effort for each task.

Judgments of learning (JOLs)

Participants answered the following question: “How many steps do you think you could solve on a similar task on a test?” on a scale ranging from “zero steps” to “five steps”. Note that this question focused on solving similar tasks, and not the exact same task that students just finished. The answers to this question were compared to the estimates of current performance to check if they would be different.

Self-assessment accuracy

This measure indicated on a scale from -5 to 5 how accurate participants were at estimating their own performance. Median self-assessment accuracy was calculated by first subtracting the *actual performance score* from the *performance estimate* for each participant. A median score of 0 indicated a correct estimate; a negative score indicated underestimation, and a positive score indicated overestimation.

Estimated performance change

Estimated performance change indicated on a scale from -5 to 5 if participants believed they would do better or worse on a future test, compared to their current performance. Median estimated performance change was calculated by subtracting the *performance estimates* from the *JOLs* for each participant. A change score of 0 indicated that participants thought their future performance would be the same as their current performance, a positive score that their future performance would be higher than their current performance, and a negative score that their future performance would be lower than their current performance.

Eye tracker

A Tobii 1750, with a screen size of 17 inch and a 50 Hz rate, measured participants’ eye movements while they were selecting a task from the database as depicted in Fig. 2. The eye movements indicated whether participants took difficulty levels or support levels into account when selecting a task (Holmqvist et al. 2011). Fixation durations and times to first fixation were calculated for each separate difficulty level and for each separate support level. Fixation durations indicated how long a participant fixated on a certain difficulty or support level. Times to first fixation indicated the time until the first fixation in a difficulty or support level.

Differences over Time

Data from all measures were split into the first four task selections (early selections) and the final four selections (final selections) to gain more insight in how the task-selection process develops. We split the tasks in two groups of four to have an even distribution. The differences between early selections and final selections were analyzed with Wilcoxon-signed rank tests.

Procedure

Participants completed the experiment individually. First, they listened to a short verbal explanation on the topic of the learning environment (i.e., genetic problem solving tasks) and they were told that they would have to select and perform eight tasks. Next, they selected the first task from the database presented on the computer screen and gave their reason for selecting that task by writing down their answer on the answer sheet. After performing the task on paper, they answered questions about their estimated performance, invested mental effort, and JOLs on the computer. Participants repeated this procedure eight times without a time limit. Mean time spent on the experiment was 1 h and 4 min ($M = 64.50$ min; $SD = 14.08$).

Results

Research question 1: do secondary school students take difficulty, support and content into account when selecting learning tasks without guidance?

To answer the first research question, we asked participants to give their main reason for selecting a particular task. Table 1 shows the categories for different reasons for task selections. All 15 participants made eight task selections, but one participant did not answer this question twice, leading to a total response number of 118. All answers were coded by two raters. The single measure intraclass correlation coefficients were .98 for content, .95 for difficulty level, and .84 for support level. Disputes were solved through discussion.

Participants indicated that they based a median of 6.00 selections out of 8 on content, 2.00 selections out of 8 on difficulty level, and 1.00 selections out of 8 on support level, with some overlap in the different categories (i.e., one selection could be assigned to multiple categories). Moreover, out of all 118 selections that were made, only 2 selections were based these on all three categories (i.e., content, difficulty and support).

Table 1 Reasons for task selection, categorized by content, difficulty level and support level

Category	Median number of selections per participant out of a total of 8 selections
Content	6.00
Difficulty level	2.00
Support level	1.00
Random selection	0

More than one category could apply to a given reason

Research question 2: how do inaccurate self-assessments influence task selections?

To answer the second research question, participants' mean performance scores, mean performance estimates, and mean JOLs were calculated per task. The median performance score was 2.63 steps correct (range is 0 to 5 steps). The median performance estimate was 3.63 steps correct (possible range is 0 to 5 steps). The median JOL was 3.75 steps (possible range is 0 to 5 steps). Median self-assessment accuracy was $(2.63 - 3.63) = -1$ step. This suggests that participants on average overestimated their own performance by one step out of five. This difference between the actual performance scores and the performance estimates was significant, $z = -2.70$, $p = .01$, $r = -0.49$. Median estimated performance change was $(3.75 - 3.63) = 0.12$ steps. This suggests that participants on average did not indicate that their performance would increase or decrease in the future. This difference between the JOLs and the performance estimates was not significant, $z = -0.98$, $p = .33$.

Furthermore, we compared participants' actual selections to the SRLTS model-based selection, which were calculated with participants' actual performance scores (note that participants could only use their own performance estimates). The SRLTS model-based selection was calculated in the following way: A low actual performance score (0 or 1 points out of 5) meant that participants should have chosen a lower difficulty level. A medium score (2 or 3 points) meant that participants should have remained at the same difficulty level. A high score (4 or 5 points) meant that participants should proceed to a higher difficulty level. Similarly, the mental effort ratings were used to calculate the support level participants could have chosen if they had followed the SRLTS model. According to the SRLTS model, a low rating (1, 2 or 3 points) would prescribe a task with less support, a medium rating (4, 5 or 6 points) would prescribe that participants should remain at their current support level, and a high rating (7, 8 or 9 points) would prescribe that participants should choose a task with more support (see Table 2). Results from this comparison suggest that participants deviated more than half a difficulty level ($Mdn = 0.57$) and more than half a support level ($Mdn = 0.57$) from the SRLTS model-based selections. This means that participants selected difficulty and support levels that, on average, were different by about half a level from the levels that the model prescribed.

Table 2 Coding scheme for calculations of scores into selections based on the SRLTS model

Score	SRLTS model-based selection
0 or 1 steps performed correctly on a learning task	Go to a lower difficulty level
2 or 3 steps performed correctly on a learning task	Remain at the same difficulty level
4 or 5 steps performed correctly on a learning task	Go to a more difficult level
Mental effort rating of 1, 2 or 3 points	Go to a level with less support
Mental effort rating of 4, 5 or 6 points	Remain at the same support level
Mental effort rating of 7, 8 or 9 points	Go to a level with more support

Research question 3: does the quality of self-assessments and unsupported task selections improve over time?

Performance score

Median performance score for the first four tasks (early selections) is 2.50 steps correct. Median performance score for the final four tasks (final selections) is 2.75 steps correct. There was no significant difference on performance score between early and final selections, $z = -1.51$, $p = .13$.

Performance estimate

There was a significant difference between early selections ($Mdn = 3.50$ steps) and final selections ($Mdn = 3.75$ steps) for the performance estimates, $z = -2.49$, $p = .01$, $r = -0.45$. This suggests that participants estimated their performance significantly higher on the final four tasks than on the first four tasks.

Judgments of learning (JOLs)

Median JOL for early tasks is 3.50 steps. Median JOL for final tasks is 3.75 steps. There was no significant difference between early and final selections for JOLs, $z = -1.59$, $p = .11$.

Mental effort

There was a significant difference in mental effort scores between early task selections ($Mdn = 4.75$) and final task selections ($Mdn = 4.25$), $z = -2.24$, $p = .03$, $r = -0.41$. This suggests that participants rated their mental effort scores as significantly higher for the first four tasks than for the final four tasks.

Self-assessment accuracy

Median self-assessment accuracy for early tasks is $(2.50-3.50) = -1$ step, $z = -2.68$, $p = .01$, $r = -0.49$. Median self-assessment accuracy for final tasks is $(2.75-3.75) = -1$ step, $z = -2.40$, $p = .02$, $r = -0.44$. There was no significant difference between the overall self-assessment accuracy scores for early and final selections, $z = -0.91$, $p = .36$.

Estimated performance change

Median estimated performance change for early tasks is $(3.50-3.50) = 0$ steps, $z = -1.77$, $p = .08$, $r = -0.32$. Median estimated performance change for final tasks is $(3.75-3.75) = 0$ steps, $z = -0.18$, $p = .86$. There was no significant difference between early and final task selections for estimated performance change, $z = -1.19$, $p = .23$.

Fixation duration

Figure 3 shows the fixation durations for the different difficulty levels. Fixation duration was significantly different between early task selections ($Mdn = 6.34$ s) and final task

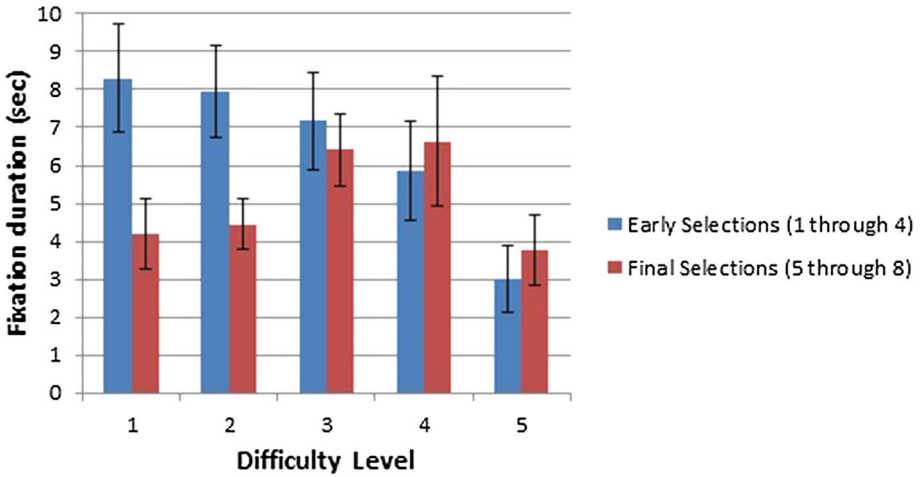


Fig. 3 Mean fixation duration in seconds for difficulty levels, split for early and final selections. Error bars represent one standard error of the mean (SEM)

selections ($Mdn = 3.24$ s) for difficulty level one, $z = -2.50$, $p = .01$, $r = -0.46$. Fixation duration was also significantly different between early task selections ($Mdn = 7.57$ s) and final task selections ($Mdn = 4.53$ s) for difficulty level two, $z = -2.33$, $p = .02$, $r = -0.43$. The difference between early and final selections was not significant for difficulty level three, $z = -0.68$, $p = .50$, difficulty level four, $z = -0.40$, $p = .69$, and difficulty level five, $z = -0.63$, $p = .53$. This suggests participants fixated longer on difficulty levels one and two during the early selections than during the final selections.

Figure 4 shows the fixation durations for the three support levels. Fixation duration was significantly different between the early task selections ($Mdn = 2.01$ s) and the final task

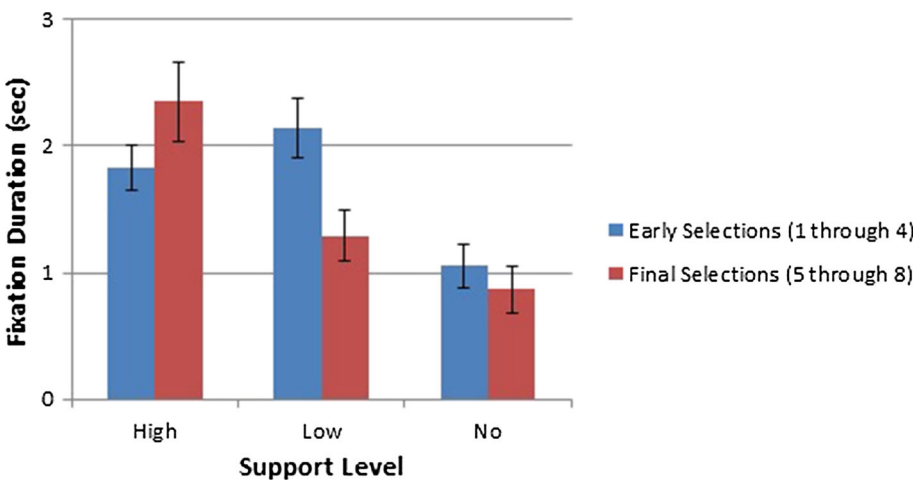


Fig. 4 Mean fixation duration in seconds for the support levels, split for early and final selections. Error bars represent one SEM

selections ($Mdn = 1.12$ s) for the low support level, $z = -2.61$, $p = .01$, $r = -0.48$. Fixation duration was not significantly different for the high support level, $z = -1.59$, $p = .11$, and the no support level, $z = -1.36$, $p = .17$. This suggests participants fixated longer on the low support levels during the early task selections than during the final task selections.

Time to first fixation

Figure 5 shows the times to first fixation for the five difficulty levels. Time to first fixation was significantly different between the early task selections ($Mdn = 17.20$ s) and the final task selections ($Mdn = 5.13$ s) for difficulty level four, $z = -2.33$, $p = .02$, $r = -0.43$. The differences in time to first fixation between early task selections ($Mdn = 23.93$ s) and final task selections ($Mdn = 12.04$ s) was marginally significant for difficulty level five, $z = -1.76$, $p = .08$, $r = -0.32$, and not significant for difficulty level one, $z = -0.41$, $p = .68$, difficulty level two, $z = -0.97$, $p = .33$, and difficulty level three, $z = -1.19$, $p = .23$. This suggests participants fixated faster on the higher difficulty levels (four and five) during the final task selections than during the early task selections.

Figure 6 shows the times to first fixation for the three support levels. The differences between early task selections and final task selections were not significant for the high support level, $z = -1.08$, $p = .28$, the low support level, $z = -0.40$, $p = .69$, and the no support level, $z = -1.42$, $p = .16$.

Discussion

The current study investigated in an explorative experiment the task selection behavior of secondary school students, and the role of self-assessments in this process. With regard to the first research question (RQ1), our findings suggest that only a very small amount of task selections took the difficulty level, support level *and* contents of the tasks into account. Students based most task selections on content only, and sometimes include difficulty and/

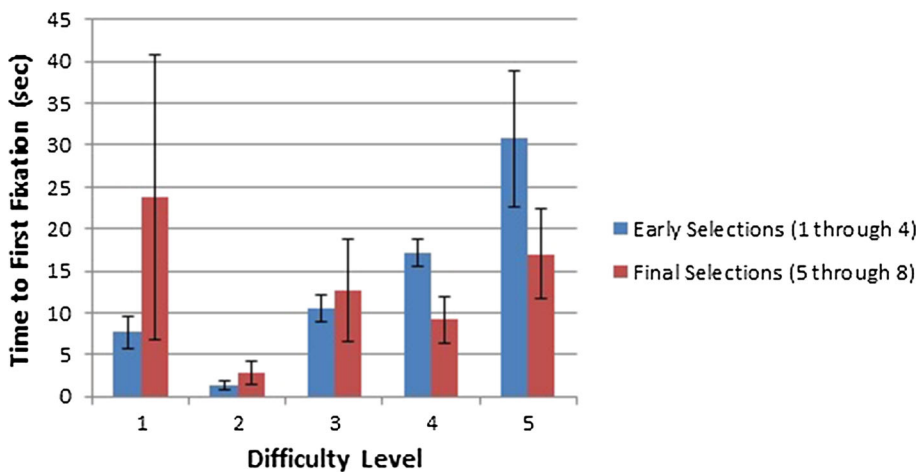


Fig. 5 Mean time to first fixation for difficulty level, split by early and final selections. Error bars represent one SEM

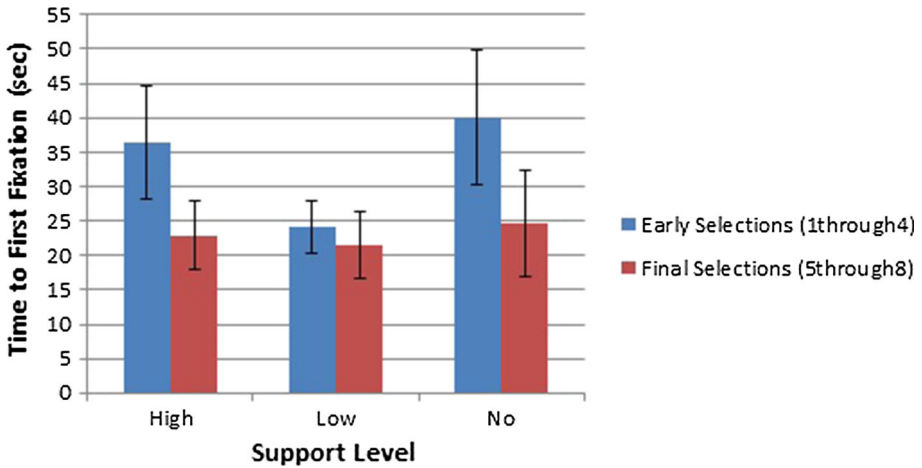


Fig. 6 Mean time to first fixation for support level, split by early and final task selections. Error bars represent one SEM

or support level in their decision. Furthermore, the reasons for making content selections were almost always unrelated to content variability. Only one participant indicated that she wanted to practice with as many different topics as possible. All other reasons for content selections were based on personal interest, mostly by indicating that a certain topic seemed fun or interesting.

These results support other studies that based task-selection advice on similar ideas (Kicken et al. 2009; Kostons et al. 2010; Taminiu et al. 2013; Van Gog et al. 2010). In these studies, the four-component instructional design (4C/ID) model (Van Merriënboer and Kirschner 2013) was used to design the structure of the task databases and to design advice on task selection, and the results were mixed. Specifically, studies in which modeling examples provided the advice were successful in helping students to acquire these skills (Kostons et al. 2012; Van Gog et al. 2010). However, other studies in which the advice was provided on-demand in textual form (Taminiu et al. 2013) or through supervision meetings and electronic portfolios (Kicken et al. 2009) were less successful.

Our findings regarding the second research question (RQ2) show that students in this sample are inaccurate in self-assessing their performance. They overestimated their current performance, and they did not seem to expect changes in their future performance compared to their current performance. Actually, our results suggest that asking for JOLs does not have additional benefits to asking for performance ratings, because the difference between these measures was 0 in 74.17% of all cases, which means participants gave the same estimate for actual performance and JOL on nearly 75% of all tasks. Our findings are in agreement with other studies reporting that (unguided) self-assessments are of low quality (e.g., Boud and Falchikov 1989; Dunning et al. 2003; Kostons et al. 2010; Panadero and Alonso-Tapia 2013; Taminiu et al. 2013), and that the process responsible for poor self-assessments might also be responsible for poor learning-task selection. That is, according to the cue-utilization approach, different cues have different effects on the accuracy of self-assessments (Koriat 1997). It is possible that the use of inaccurate cues is also responsible for poor learning-task selections. For instance, when students wrongly believe they are highly skilled at performing genetic tasks, they will estimate their performance too high and thus make inaccurate self-assessments. At the same time, they will

be more likely to choose difficulty levels that are too high for their current skill level. The current results confirm this idea, because they suggest that participants in this sample tended to choose suboptimal difficulty and support levels compared to what the SRLTS model would have prescribed. Possibly, they based these selections on their inaccurate self-assessments, thus leading to inaccurate learning-task selections.

Findings regarding the third research question (RQ3) are mixed. Not many changes occurred from early to later self-assessments. Only performance estimates seemed to increase slightly, while mental effort ratings tended to decrease. This suggests that participants in this sample found the tasks easier after some practice, although the actual performance scores did not reflect this. Surprisingly, self-assessment accuracy and estimated changes in performance did not increase over time. This is in line with the finding for research question 2 that participants overestimated their current performance: They are also not accurate in predicting their future performance. However, this was a short study. It is possible that with more training and/or guidance, participants could become more accurate in their self-assessments of current and future performance. For eye tracking measures, there were some notable differences between early and final task selections. For early task selections, participants in this sample fixated longer on easier tasks with low support; for final selections, participants focused faster on tasks that are more difficult. Thus, participants seemed to show an orientation towards the easier tasks for the first selections and towards the more difficult tasks for the final selections.

The longer fixations during the early selections might be explained by participants trying to familiarize themselves with the task selection table (Fig. 2). The faster fixations during the final task selections could indicate that participants have become familiar with the table and therefore fixate faster on the difficulty levels and support levels they want to choose next. The low performance scores (participants only performed about half of the five steps correctly) suggest that participants focused on the more difficult tasks, even though they were not yet ready to perform these tasks. Guidance could accommodate this by adapting the advice at different points in time. For example, advice on selecting early tasks could incorporate more explanation and guidance about the structure of the task-selection table. Advice on selecting later tasks could incorporate performance scores on earlier tasks and suggest to students to retry easier tasks when necessary, instead of progressing to tasks that are too difficult.

The shift in attentional focus from easy tasks during the early selections to more difficult tasks during the final selections is in a sense surprising, because the self-assessments and the reasons given for task selections do not designate such a shift. In other words, participants seem to shift their attention to higher difficulty levels, while their self-assessments and answers to the open question about their reasons for task selection do not indicate why. However, this shift is congruent with the other results, which suggested that in some instances, participants did follow the selection rules described by the SRLTS model without receiving guidance. This suggests that students might be unaware that they base their selections on these rules. This is in accordance with the idea that self-regulated learning is sometimes based on unknown and unexpected processes (Butler and Winne 1995; Schraw et al. 2006).

Implications

The results from this study have four implications for the SRLTS model. First, students might not consciously consider all factors as specified in the SRLTS model. The paths in the SRLTS model for choosing difficulty levels (arrows a, b, c and d in Fig. 1) and support

levels (arrows a, b, e and f in Fig. 1) did seem to be followed in some of the selections, but students did not seem to decide this consciously. Students in this sample usually did take content into account when selecting a task, but there were no signs that they followed the content path (arrows a, b, g and h in Fig. 1). Instead of checking for variability, they based content selections on personal interests. Thus, our results have provided a first indication that students can sometimes successfully follow the theoretical paths to task selection that the SRLTS model describes, but they might benefit from study materials that are more adapted to these specific factors and their paths. Study materials could be adapted to show a more clear structure to students (i.e., to have clear difficulty and support levels, and provide different contents), and to provide more guidance for the selection of these materials. For instance, study materials could include task databases for common topics that require practice, such as inheritance tasks, algebra or the laws of physics. These databases would need a clear distinction between different contents, difficulty levels and support levels; and instructions for the students on how to select these three factors through self-assessments. That way, students could select their own tasks with content and levels that match their individual competence levels and needs.

Second, self-assessments were inaccurate, which suggests that students made suboptimal learning-task selections compared to the selections they could have made with correctly estimated performance scores and the SRLTS model. This illustrates the importance of making accurate self-assessments, because they are needed to make optimal selections about difficulty and support levels of selected tasks. Although inaccurate self-assessments might have a positive effect on motivation, realistic self-assessments are preferred over a falsely-positive view on one's own abilities (Panadero et al. 2016). Thus, students might benefit from self-assessment guidance, which could, for example, provide assessment criteria and feedback (Brown et al. 2015; Panadero and Alonso-Tapia 2013). Additionally, self-assessment accuracy increases with domain-specific expertise (Panadero et al. 2016), which could also be stimulated through guidance.

Third, JOLs do not seem to add any new self-assessment information compared to the performance estimates. Students usually gave the same answer to both these questions (respectively, the number of steps they believed they currently performed correctly, and the number of steps they believed they could perform correctly on a similar task on a future test). This illustrates other notions about the similarity between JOLs and performance estimates (Baars et al. 2014). Therefore, JOLs might either be removed from the SRLTS model, or broadened by asking students about their expected future performance on a series of tasks as an additional measure next to the single-task performance estimates. Though, these results seem to suggest for educational practice that performance estimates for single tasks could be sufficient without additional JOLs. Thus, it seems that it might be better to focus on improving the accuracy of performance estimates without adding other time-consuming measures.

Fourth, our findings seem to suggest that students could benefit from task-selection advice and feedback on their self-assessments. Self-assessment feedback could help students to obtain correct self-assessment information, which they could then use to select learning tasks. In addition, students could receive task-selection advice, which could focus on using self-assessments for choosing difficulty levels, support levels, and content respectively. Thus, the SRLTS model could be a guide when designing task-selection advice, by using the paths from the model as a basis for the advice.

Limitations and future research

A limitation of our study is that the task environment used in the current study was very well structured. There were 75 tasks with different difficulty levels, support levels and contents. When students practice learning tasks at school or at home, they mostly have to work from textbooks and materials provided by their teacher, which are often less well structured. Many textbooks and materials do not contain worked examples or completion problems, and even if they do, students would need to invest more effort than in the current study to unravel the different difficulty and support levels during practice. Moreover, we used genetics tasks; hence, it is unknown whether these results can be generalized to other domains or other types of tasks. Furthermore, the lab setting limited the ecological validity of this study. To maintain some level of ecological validity, we only tested secondary school students who were able to come to our lab, which unfortunately resulted in a low number of participants. Even though small sample sizes are common in studies with process measures (e.g., Goldberg and Wichansky 2002; Jarodzka et al. 2010; Tsai et al. 2012; Wiebe et al. 2009), further research is urgently needed to replicate these findings with a larger sample size and in a real-life setting.

Another limitation is that the participants did not receive feedback on their performance, which might have affected both the actual performance scores and the performance estimates. Future studies could therefore incorporate feedback to improve the students' self-assessments and thus investigate how correcting inaccurate self-assessments affects task selections. In addition, students in the current experiment were not externally motivated to score as high as possible, nor did we investigate whether participants' standards were congruent to our theorized standards. Both of these factors could have influenced the results and could be investigated in future studies. Moreover, the eye tracking data for the first task selection might have been influenced by the participants first trying to familiarize themselves with the selection table. Additionally, participants' viewing strategy might become more efficient over time, thus also affecting the data for the last selection. Future studies could prevent this by familiarizing participants with the selection table before the start of the experiment with unrelated material.

The current study does not yet establish that selecting tasks according to the SRLTS model actually leads to higher achievement. Previous studies that used similar set-ups suggest clear benefits from this theory-driven combination of factors (e.g., Camp et al. 2001; Corbalan et al. 2009b, 2011), thus giving the model a solid theoretical and empirical foundation. However, evidence for this specific model is necessary to warrant further conclusions about this. Future studies could investigate how beneficial selecting tasks according to the SRLTS model is for learning. In addition, these studies could investigate whether advice on self-assessment and task selection, such as corrections for inaccurate self-assessments, and guidance on the use of self-assessments for learning-task selection, improve this.

The students in this experiment seemed to have based most of their task selections on content only, which suggests that they might benefit from guidance on self-assessment and task selection. Future studies could investigate this by designing a type of guidance based on the SRLTS model, and measure whether this influences task selections and/or domain-specific learning (indirectly through different practice tasks). Moreover, future studies could also investigate whether changes over time might need to be taken into account when providing guidance. For instance, early guidance might need to include more explicit instructions, whereas later guidance might phase these out and focus instead on alerting

students to incorporate self-assessment information in their selections. This might help students to better understand on which factors they can base task selections and how to do this, and to prevent them from progressing to fast to higher levels when they might not be ready for this.

Finally, future research could use process measures like eye tracking and think-aloud protocols (Ericsson and Simon 1993) for further investigating the unknown processes that underlay self-regulated learning. The combination of eye tracking and think-aloud protocols (specifically, cued retrospective reporting) has been shown to be an effective method in uncovering metacognitive processes (Van Gog et al. 2005).

Conclusion

In conclusion, this exploratory study suggests that secondary school students are not good at self-assessing their performance, and provides an indication that secondary school students therefore might benefit from guidance, advice and feedback for reaching better self-assessments. Furthermore, they could benefit from guidance to incorporate self-assessment information in their task selections, since they do not seem to do this on their own. Future studies could explore different types of external feedback and guidance. Furthermore, these results suggests that the SRLTS model could provide rules and structure for task selections that students might not exert themselves. The results also provide leads on ways to improve the SRLTS model, most notably by investigating the addition of guidance (in the form of advice and feedback on self-assessments), and motivation as important factors for task selection. Future research will be needed to investigate these potential additions. Finally, this study showed differences in task selections over time, despite the short duration of the study. Guidance might be adapted to these differences, for example, by first giving students advice on how to interpret the database of available tasks and then giving them advice on how to self-assess and select tasks from this database. Future research should look more deeply into this.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Funding This study was funded by the Netherlands Organization for Scientific Research (NWO, file number 411-12-015).

Appendix

Example of a high support task from difficulty level 1, translated from Dutch.

Level 1. Cat fur

Cats can have nice fur. A gen causes the shape of the stripes on the cat, for which rings (C) are dominant over stripes (c). Two cats have a kitten. Both cats are homozygote for the characteristics and have rings in their fur.

What is the possible genotype of the kitten?

Step 1: Determine the genotypes

A genotype is the information that lies in the genes and consists of 2 alleles (letters). The phenotype is how this information is expressed. Two things should be taken into account for this: Is it dominant/recessive and is it homozygote/heterozygote? Dominant means that if at least one dominant allele is present in the genotype, you will always see this characteristic, despite of what else is in the genotype. At least 1 of the 2 letters will always be the capital letter. Recessive means that there cannot be a dominant allele in the genotype and that the genotype will always consist of two small letters. Homozygote means that both alleles are the same (so both have the dominant letter, or both are recessive). Heterozygote means that both the dominant and recessive alleles are in the genotype.

Parents: Both parents show the dominant characteristic, because they both have rings in their fur. Therefore, there must be at least one dominant allele in the genotype. Furthermore, it is known that both parents are homozygote, which means that both alleles must be the same. Therefore, the genotype for each parent is CC.

Step 2: Make a family tree

A family tree (see Fig. 7) is a graphic representation of the genotypes. The parents are all the way at the top, below that is the next generation of the children, and below that the generation of grandchildren, etc.

Step 3: Determine the number and direction of the contingency tables

You can reason in two different ways. Deductive means that you know both parents and should reason towards the unknown child (from top to bottom in the tree). For this, one contingency table is enough. Inductive means that you only know one of the parents and the child, but you do not know the other parents. For this, you need to reason from bottom to top and you need a separate contingency table for each possible combination of the unknown parent (so a maximum of 3).

In this case, you know both parents, so your reasoning is deductive and you only need to make 1 contingency table.

Step 4: Fill in the contingency tables

First, draw the tables (see Fig. 8), fill in the parents at the edges and sum up the alleles to get all combinations. The first two steps are already done, so now make the combinations yourselves

Fig. 7 An example of a family tree, which was provided with this task. Participants could copy this family tree on their answer sheet and fill it in

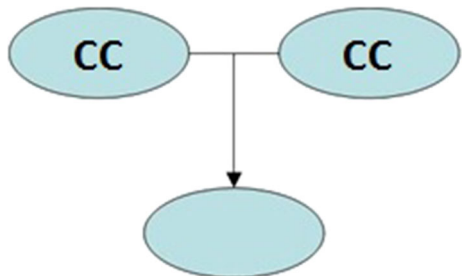


Fig. 8 An example of a contingency table, which was provided with this task. Participants could copy this table on their answer sheet and fill it in

	C	C
C		
C		

Step 5. Choose the right answer from the contingency tables

Compare the answers to the ones you get from the contingency tables and retrieve your final answer from there. Mark the correct answer.

What is the possible genotype of the unknown child?

- A. CC
- B. Cc
- C. cc
- D. CC and Cc
- E. CC and cc
- F. Cc and cc
- G. CC, Cc and cc

References

- Andrade, H. L. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. *NERA Conference Proceedings, 2010*, 25.
- Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education, 32*, 159–181. <https://doi.org/10.1080/02602930600801928>.
- Baars, M., Vink, S., Van Gog, T., De Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92–107. <https://doi.org/10.1016/j.learninstruc.2014.04.004>.
- Baars, M., Visser, S., Van Gog, T., De Bruin, A., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology, 38*, 395–406. <https://doi.org/10.1016/j.cedpsych.2013.09.001>.
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. <https://doi.org/10.1146/annurev-psych-113011-143823>.
- Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*, 529–549. <https://doi.org/10.1007/BF00138746>.
- Brown, G. T. L., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice, 22*, 444–457. <https://doi.org/10.1080/0969594X.2014.996523>.
- Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research, 2*, 22–30. <https://doi.org/10.14786/flr.v2i1.24>.
- Bruner, J. S. (1960). *The process of education*. Cambridge: Harvard University Press.
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281. <https://doi.org/10.3102/00346543065003245>.

- Camp, G., Paas, F., Rikers, R., & Van Merriënboer, J. (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. [https://doi.org/10.1016/S0747-5632\(01\)00028-0](https://doi.org/10.1016/S0747-5632(01)00028-0).
- Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236–246.
- 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. <https://doi.org/10.1016/j.cedpsych.2008.02.003>.
- Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2009a). Combining shared control with variability over surface features: Effects on transfer test performance and task involvement. *Computers in Human Behavior*, *25*, 290–298. <https://doi.org/10.1016/j.chb.2008.12.009>.
- Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2009b). Dynamic task selection: Effects of feedback and learner control on efficiency and motivation. *Learning and Instruction*, *19*, 455–465. <https://doi.org/10.1016/j.learninstruc.2008.07.002>.
- Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2011). Learner-controlled selection of tasks with different surface and structural features: Effects on transfer and efficiency. *Computers in Human Behavior*, *27*, 76–81. <https://doi.org/10.1016/j.chb.2010.05.026>.
- De Bruin, A., & Van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction*, *22*, 245–252. <https://doi.org/10.1016/j.learninstruc.2012.01.003>.
- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271–280. <https://doi.org/10.1016/j.learninstruc.2011.08.003>.
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83–87. <https://doi.org/10.1111/1467-8721.01235>.
- Efklikes, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*, 6–25. <https://doi.org/10.1080/00461520.2011.538645>.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev ed.). Cambridge, MA: MIT Press.
- Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education*, *16*, 311–329. <https://doi.org/10.1007/s10459-010-9263-2>.
- Goldberg, J. H., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In Hyönä, J., Radach, R., & Deubel, H. (Eds.), *The mind's eyes: Cognitive and applied aspects of eye movement research* (pp. 493–516). Oxford: Elsevier Science.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.
- Istance, D., & Dumont, H. (2010). Future directions for learning environments in the 21st century. In H. Dumont, D. Istance, & F. Benavides (Eds.), *The nature of learning: Using research to inspire practice* (pp. 317–338). Paris, France: OECD, Centre for Research and Educational Innovation. <https://doi.org/10.1787/9789264086487-15-en>.
- Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*, 146–154. <https://doi.org/10.1016/j.learninstruc.2009.02.019>.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*, 23–31. https://doi.org/10.1207/S15326985EP3801_4.
- 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. <https://doi.org/10.1007/s11423-009-9111-3>.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>.
- 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. <https://doi.org/10.1016/j.compedu.2009.09.025>.
- Kostons, D., Van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction*, *22*, 121–132. <https://doi.org/10.1016/j.learninstruc.2011.08.004>.

- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Englewood Cliffs, NJ: Prentice-Hall.
- Loyens, S. M., Magda, J., & Rikers, R. M. (2008). Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational Psychology Review*, *20*, 411–427. <https://doi.org/10.1007/s10648-008-9082-7>.
- Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*, 174–179. <https://doi.org/10.3758/PBR.15.1.174>.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). San Diego, CA: Academic Press.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs: Englewood Cliffs, NJ.
- 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. <https://doi.org/10.1037/0022-0663.84.4.429>.
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*, 1–4. https://doi.org/10.1207/S15326985EP3801_1.
- 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. <https://doi.org/10.1037/0022-0663.86.1.122>.
- Panadero, E. (2017). A Review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 1–28. <https://doi.org/10.3389/fpsyg.2017.00422>.
- Panadero, E., & Alonso-Tapia, J. (2013). Self-assessment: Theoretical and practical connotations. When it happens, how is it acquired and what to do to develop it in our students. *Electronic Journal of Research in Educational Psychology*, *11*, 551–576. <https://doi.org/10.14204/ejrep.30.12200>.
- Panadero, E., Brown, G. T. L., & Strijbos, J. W. (2016). The future of student self-assessment: a review of known unknowns and potential directions. *Educational Psychology Review*, *28*, 803–830. <https://doi.org/10.1007/s10648-015-9350-2>.
- Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist*, *38*, 15–22. https://doi.org/10.1207/S15326985EP3801_3.
- Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36*, 111–139. <https://doi.org/10.1007/s11165-005-3917-8>.
- Sweller, J., Van Merriënboer, J. J. G., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296. <https://doi.org/10.1023/A:1022193728205>.
- Taminiau, E. M., Kester, L., Corbalan, G., Alessi, S. M., Moxnes, E., Gijsselaers, W. H., et al. (2013). Why advice on task selection may hamper learning in on-demand education. *Computers in Human Behavior*, *29*, 145–154. <https://doi.org/10.1016/j.chb.2012.07.028>.
- Tsai, M. J., Hou, H. T., Lai, M. L., Liu, W. Y., & Yang, F. Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education*, *58*, 375–385. <https://doi.org/10.1016/j.compedu.2011.07.012>.
- Van Gog, T., Kostons, D., & Paas, F. (2010). 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* (pp. 296–301). Austin, TX: Cognitive Science Society.
- 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*, *11*, 237–244. <https://doi.org/10.1037/1076-898X.11.4.237>.
- Van Merriënboer, J. J. G., & Kirschner, P. A. (2013). *Ten steps to complex learning: A systematic approach to four-component instructional design* (2nd (Rev ed.)). New York: Routledge.
- Wiebe, E. N., Minogue, J., Jones, M. G., Cowley, J., & Krebs, D. (2009). Haptic feedback and students' learning about levers: Unraveling the effect of simulated touch. *Computers & Education*, *53*, 667–676. <https://doi.org/10.1016/j.compedu.2009.04.004>.
- Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153–189). Mahwah, NJ: Lawrence Erlbaum Associates.
- Winne, P. H. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research*, *41*, 466–488. <https://doi.org/10.1016/j.ijer.2005.08.012>.

- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. *Metacognition in Educational Theory and Practice*, 93, 27–30.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, 41, 64–70. https://doi.org/10.1207/s15430421tip4102_2.
- Zimmerman, B. J., & Campillo, M. (2003). Motivating self-regulated problem solvers. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 233–262). New York: Cambridge University Press.