Metacognition and Learning (2019) 14:21–42
https://doi.org/10.1007/s11409-019-09189-5

Effects of self-assessment feedback on self-assessment and task-selection accuracy

Steven F. Raaijmakers[1] · Martine Baars[2] · Fred Paas[2,3] · Jeroen J. G. van Merriënboer[4] · Tamara van Gog[1]

Received: 22 September 2017 / Accepted: 1 February 2019 / Published online: 8 March 2019

Abstract

Effective self-regulated learning in settings in which students can decide what tasks to work on, requires accurate self-assessment (i.e., a judgment of own level of performance) as well as accurate task selection (i.e., choosing a subsequent task that fits the current level of performance). Because self-assessment accuracy is often low, task-selection accuracy suffers as well and, consequently, self-regulated learning can lead to suboptimal learning outcomes. Recent studies have shown that a training with video modeling examples enhanced self-assessment accuracy on problem-solving tasks, but the training was not equally effective for every student and, overall, there was room for further improvement in self-assessment accuracy. Therefore, we investigated whether training with video examples followed by feedback focused on self-assessment accuracy would improve subsequent self-assessment and task-selection accuracy in the absence of the feedback. Experiment 1 showed, contrary to our hypothesis, that self-assessment feedback led to *less* accurate future self-assessments. In Experiment 2, we provided students with feedback focused on self-assessment accuracy plus information on the correct answers, or feedback focused on self-assessment accuracy, plus the correct answers and the opportunity to contrast those with their own answers. Again, however, we found no beneficial effect of feedback on subsequent self-assessment accuracy. In sum, we found no evidence that feedback on self-assessment accuracy improves subsequent accuracy. Therefore, future research should address other ways improving accuracy, for instance by taking into account the cues upon which students base their self-assessments.

Keywords Self-regulated learning · Problem solving · Self-assessment · Task-selection · Feedback

Students in secondary education are increasingly required to learn how to self-regulate their learning, because self-regulating your learning is considered imperative for learning in higher education and workplace settings (Bransford et al. 2000). Most models assume self-regulated learning to be a cyclical process proceeding "...from some kind of a preparatory or preliminary

✉ Steven F. Raaijmakers
s.f.raaijmakers@uu.nl

Extended author information available on the last page of the article

phase, through the actual performance or task completion phase, to an appraisal or adaption phase" (Puustinen and Pulkinen 2001 p. 280). Key components of self-regulated learning are *monitoring* (i.e., keeping track of your learning/performance), evaluation of performance, and *control* (i.e., regulating your study behavior in response to that evaluation; e.g., Nelson and Narens 1990; Winne and Hadwin 1998; Zimmerman 1990). For self-regulated learning to be effective, both monitoring judgments and control decisions need to be accurate (e.g., Bjork et al. 2013; Panadero et al. 2017). Unfortunately, students' monitoring accuracy is often low (Brown and Harris 2013; Winne and Jamieson-Noel 2002). Because accurate control depends on accurate monitoring (Thiede et al. 2003), the control of study behavior can suffer as well (e.g., Winne and Jamieson-Noel 2002), and consequently, self-regulated learning often leads to suboptimal learning outcomes (Kornell and Bjork 2008). Hence, it is important to improve learners' monitoring accuracy.

Training self-regulated learning skills has been found to be effective for improving learners' monitoring and control accuracy (e.g., Azevedo and Cromley 2004; Bol et al. 2016; Kostons et al. 2012). For example, during a 30-min training session, Azevedo and Cromley (2004) trained college students on how to regulate their learning in a hypermedia environment. During the training the experimenter explained the different phases and areas of regulation with the help of a table, described a model of self-regulated learning with a diagram, and gave a definition of each self-regulated learning variable on a list of 17 variables. After the training the students were allowed to learn about the human cardiovascular system using the hypermedia environment. Students in the training condition verbalized more about monitoring and learned more about the topic than students in the control condition. The training used by Bol et al. (2016) consisted of self-regulated learning exercises given to community college students during a 3-week mathematics course. Students had to perform four exercises each week and each of these exercises corresponded to a particular self-regulated learning component. For instance, one of the exercises was to set a weekly academic goal (corresponding with the component 'goal setting'). Students in the training condition reported using more metacognitive self-regulation (actual behavior was not measured) and their achievement was higher than that of students in the control condition.

Another type of self-regulated learning training focused on improving self-assessment (i.e., monitoring) and task-selection (i.e., control) skills in a learning environment in which students could choose their own problem-solving tasks. Previous research on adaptive instructional systems has already shown that selecting students' subsequent tasks based on their performance, possibly combined with mental effort invested to attain such performance, yields effective learning paths (Corbalan et al. 2008). However, in a self-regulated learning environment tasks are not selected by the system, but by the students themselves. Allowing students to have control over the selection of tasks makes effective learning conditional upon the accuracy of both students' self-assessments, as well as their task selections. In essence, a cyclical model emerges proceeding from performance, through self-assessment, to task selection.

Kostons et al. (2012) used video modeling examples to train self-assessment and task selection. These video modeling examples are based on principles from both social learning theory (Bandura 1977) and example-based learning (Renkl 2014). In a video modeling example a model (either visible or not) demonstrates a problem-solving process or procedure to a learner (Van Gog and Rumme 2010). Video modeling examples have proven effective for the acquisition of various domain-specific skills (e.g., probability calculations; Hoogerheide et al. 2016) as well as self-regulated learning skills (Kostons et al. 2012; Zimmerman and

Kitsantas 2002). Kostons et al. (2012) used four such video modeling examples to train self-assessment and task-selection skills. The examples consisted of screen recordings with voice overs, showing models demonstrating how they performed a problem-solving task, rating how much mental effort they invested, self-assessing their performance, and selecting a next task based on a combination of self-assessed performance and invested effort (e.g., when performance was high and effort was low, a more complex task was selected). In their first experiment, Kostons et al. (2012) found that videos in which self-assessment was modeled led to more accurate self-assessment and that videos in which task selection was modeled led to more accurate task selection. In their second experiment, a pre-training with videos in which both self-assessment and task selection were modeled led to improved performance after a self-regulated learning phase. However, even though these results were promising, there was room for further improvement in self-assessment accuracy. Moreover, standard deviations were large, meaning there were substantial individual differences in the effectiveness of the training (Kostons et al. 2012).

This might suggest that some students had learned more from the video modeling examples training than others or that some found it difficult to apply the self-assessment skills they had learned from the video modeling examples during the self-regulated learning phase. These students might benefit from additional support after the self-regulated learning skills training. Additional support could be provided by giving feedback on self-assessments in a way similar to which feedback is usually given on performance. This may allow students to adjust their self-assessments on subsequent tasks, which is the main aim of this study (Butler and Winne 1995).

Self-assessment feedback

When asked to self-assess their performance, students have to compare their answer to an internal standard (i.e., knowledge of what constitutes poor, average, or good performance) unless an external standard is available (i.e., criteria are given). Butler and Winne (1995) presented this idea in their influential model of self-regulated learning and it still plays a key role in more recent models (for a review, see Panadero 2017). In the Butler and Winne model, during monitoring, performance is compared to an internal standard (i.e., internal model of what represents a correct answer). Students can decide, based on the evaluation, if they have learned enough (to pass a test), if learning went well, or if they have performed well. In the case of a multistep problem-solving task (which we use in this study) a student uses standards that include knowledge of the problem-solving procedure to assess his or her performance. If such a standard is of poor quality, self-assessment accuracy would be negatively affected. For example, students could mistakenly identify incorrectly performed steps as correctly performed steps, or students could use cues that are not necessarily predictive of their performance, such as fluency or effort, that might bias their self-assessments (see Koriat 1997). When novices start learning how to solve a problem, they lack proper internal standards (i.e., they have no internal model of what constitutes a correct answer) and, consequently, cannot evaluate their performance properly. The conjunction of both incompetence and the recognition thereof has been called ‘the double curse of incompetence’ (Kruger and Dunning 1999). The lack of internal standards might explain why some learners’ self-assessments are inaccurate—even after training such as given in Kostons et a. (2012).

Providing students with external standards through feedback in the form of correct answers could support self-assessment. They can then contrast their own answers with the correct answers and, hopefully, make a more accurate assessment of their learning. Indeed, when learners are provided with correct answers when assessing their own learning, their self-assessment accuracy has been shown to improve (Baars et al. 2014; Dunlosky et al. 2010; Lipko et al. 2009; Rawson and Dunlosky 2007).

However, because in all of these studies the standards (correct answers) were provided during self-assessment, we cannot know for sure whether the quality of students' *internal* standards improved, that is, whether they would be able to make more accurate self-assessments in the absence of the standards. That would require measuring on subsequent tasks, in the absence of the feedback, whether accuracy improved from the feedback (which is done in the present study). Moreover, when provided with the correct answers, learners are able to study those answers and learn from them. This could lead to higher test performance for the learners who had access to the correct answers (cf. Rawson and Dunlosky 2007), which may lead to two problems. First, when such (re)study opportunities lead to higher performance, it is more difficult to conclude that self-assessment has improved independently of performance because – as mentioned above – higher performance is associated with higher self-assessment accuracy (Dunning et al. 2004; Kruger and Dunning 1999). Second, because learners typically overestimate their performance, self-assessment accuracy may improve not so much because learners became more accurate self-assessors (i.e., acquired better standards to assess performance), but because their performance improved (i.e., if learners overestimate their performance and do not take into account the increase in performance from learning phase to test; cf. Kornell and Bjork 2009).

In order to avoid the issues that standards in the form of correct answers bring along, yet still provide learners with support to better calibrate their internal standards, researchers have suggested that feedback could be provided with a focus on the accuracy of their self-assessments instead of their performance (self-assessment feedback; Panadero et al. 2015; Winne and Nesbit 2009). By providing learners with the information that their self-assessment was (in)accurate they could deduce that their internal standards need to be adjusted, without having access to the correct answers. In a study by Nietfeld et al. (2006), participants in the treatment group were provided with self-assessment exercises after each class. These exercises asked students to self-rate their understanding of the day's class, identify difficult content, write down how they would improve their understanding, and, finally, answer review questions with confidence judgments (i.e., 0–100%) after each question. After these exercises, the answers to the review questions were discussed in class. Students were encouraged to compare their confidence judgments with their performance. On each test (there were four tests in total) all students were asked to provide confidence judgments for each item on the test. Calibration (i.e., the difference between confidence judgment and performance) improved over time for students who received self-assessment exercises (Nietfeld et al. 2006). Although the feedback was still mostly focused on performance, participants were prompted to generate self-assessment feedback themselves by comparing their confidence rating to the performance feedback.

The current study

The aim of the study presented here was to investigate whether – after an initial self-regulated learning training – self-assessment accuracy can further improve from

feedback that is focused on the accuracy of self-assessments, instead of focused on students' problem-solving task performance. In two experiments, participants first received self-assessment and task-selection training with video modeling examples (cf. Kostons et al. 2012; Raaijmakers et al. 2018), then engaged in problem-solving and self-assessment during a learning phase in which they received self-assessment feedback, followed by problem-solving and self-assessment during a test phase in which the feedback was no longer present. The primary outcome measures were self-assessment accuracy during the test phase (i.e., with no feedback present), and task-selection accuracy (because self-assessment accuracy is considered a necessary condition for task-selection accuracy, task-selection accuracy would be expected to improve from improved self-assessment accuracy).

In Experiment 1, we compared the effectiveness of two types of self-assessment feedback to a no-feedback control condition (which only received the self-assessment and task-selection training): general feedback, which indicated whether a self-assessment was correct or incorrect and *how many* steps had actually been performed correctly, versus more specific feedback, which -in addition to the general feedback- also indicated *which* steps had actually been performed correctly or incorrectly (by flagging the steps as correct/incorrect, but not providing the correct answer to the step). In a study by Miller and Geraci (2011), when learners were given more specific performance feedback, self-assessment accuracy improved, but not when learners were given more general feedback. Similarly, specific self-assessment feedback would allow learners to adjust their internal standards at the level of the problem-solving steps (e.g., "I thought I understood how to solve step 1, 2, 3, and 4, but not step 5, but from the self-assessment feedback I learned that I don't understand step 4 either"). General self-assessment feedback, in contrast, would only allow the learner to determine whether they overestimated or underestimated their overall performance at the level of the task (e.g., "I thought I performed 3 out of the 5 steps correctly, but apparently I only performed 1 step correctly").

We hypothesized that during the learning phase, participants in both feedback conditions would be able to make more accurate task-selection decisions than participants in the control condition (Hypothesis 1), as these could be directly based on the actual performance indicated in the feedback. Regarding the test phase (i.e., in the absence of the feedback), we hypothesized that participants in the feedback conditions would demonstrate more accurate self-assessment than participants in the control condition (Hypothesis 2a), and that the specific self-assessment feedback condition would demonstrate more accurate self-assessment than participants in the general self-assessment condition (Hypothesis 2b; i.e., Hypothesis 2: specific > general > `no). If hypothesis 2 would be confirmed, we would expect a similar pattern of results for task-selection accuracy during the test phase (Hypothesis 3). **Experiment 1.**

Method

Participants and design A total of 204 adolescents in their third year of Dutch secondary education participated in this study. They were in senior general secondary education (the second highest level with a five year duration; $n = 154$) or pre-university education (the highest level with a six year duration; $n = 50$). Three participants were excluded due to too much prior knowledge of the learning tasks (i.e., scoring 60% or higher). Another 93 participants had to be

excluded because they did not manage to finish the experiment within the lesson period.[1] The remaining sample of 108 participants had a mean age of 14.35 years ($SD = 0.67$), and contained 49 boys and 59 girls. Participants were randomly assigned to one of the three conditions: no feedback ($n = 37$), general self-assessment feedback ($n = 34$), or specific self-assessment feedback ($n = 37$).

Materials and procedure The experiment was conducted in computer rooms at students' schools with ca. 20–30 students per session. Sessions were restricted by the length of a lesson period (~50 min). All materials were presented in a dedicated web-based learning environment created for this study. Participants were provided with a personal login and password (which handled the random assignment) and were asked to fill out some demographic information on paper before logging in (i.e., age, gender, prior education). Before starting with the experiment, participants were instructed to perform all the tasks by themselves and in private. Participants first completed the pretest, then received the video-modeling examples training (~15 min), after which they went on to the learning phase and finally, to the test phase. The pretest, learning phase, and test phase were self-paced.

Problem-solving tasks The problem solving tasks (see Appendix 1) were in the domain of biology (monohybrid cross problems) and the problem solving procedure consisted of five distinct steps (cf. Corbalan et al. 2009; Kostons et al. 2012): (1) translating the information given in the cover story into genotypes, (2) putting this information in a family tree, (3) determining the number of required Punnett squares, (4) filling in the Punnett square(s), (5) finding the answer(s) in the Punnett square(s). The tasks used in this experiment were selected from a database (cf. Kostons et al. 2012) with 75 tasks at five complexity levels and three support levels within each complexity level (Fig. 1). Tasks increased in complexity across levels by an increase in the number of generations, an increase in the number of unknowns, the possibility of multiple correct answers, and the type of reasoning needed to solve the problem. Tasks with high support had 4 steps already worked-out, leaving 1 for the student to complete, tasks with low support had 2 steps already worked-out, leaving 3 for the student to complete, and tasks with no support required students to solve the entire problem on their own (cf. completion problems: Paas 1992; and fading strategy: Renkl and Atkinson 2003).

Pretest The pretest was used to check students' prior knowledge. It consisted of three problem-solving tasks without support from the first three complexity levels (of the task database in Fig. 1).

Self-assessment and task-selection training The self-assessment and task-selection training, which all participants received, consisted of an introductory video, in which the main concepts of the problem-solving tasks were explained (i.e., dominant/recessive, homozygous/heterozygous), and four video modeling examples. The video modeling examples were screen recordings created with Camtasia Studio (cf. Kostons et al. 2012; Raaijmakers et al. 2018). The video modeling examples showed a computer screen recording with voice over of the model

[1] For this reason, additional participants were recruited as we were aiming for at least 30 participants per condition, hence the high overall sample size. Exclusion was equal across conditions, $\chi^2(2) = 0.281$, $p = .869$. Excluded participants did not perform significantly different from non-excluded participants during the learning phase, $t(187) = 1.24$, $p = .217$, and did not self-assess significantly different, $t(187) = 0.50$, $p = .621$.

Complexity level 1			Complexity level 2			Complexity level 3			Complexity level 4			Complexity level 5		
High support	Low support	No support	High support	Low support	No support	High support	Low support	No support	High support	Low support	No support	High support	Low support	No support
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 structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure	Hair structure
Milk allergy	Dog tail length	Cystic Fibrosis	Albinism	Fruit flies	Pea plant	Fruit flies	Dog tail length	Freckles	Flower color	Cat fur shape	Tongue curling	Cat fur shape	Sickle cell	Fruit flies
Dimples	Wolfram	Fruit flies	Cat fur shape	Tongue curling	Dimples	Chicken beak	Apple tree	Flower color	Widow's peak	Albinism	Apple tree	Apple tree	Chicken beak	Milk allergy
Earlobes	Flower color	Rat fur	Fruit flies	Flower color	Depres-sion	Wolfram	Milk allergy	Earlobes	P.R.A.	Pea plant	Fruit flies	Depres-sion	Guinea pigs	Cleft lip

Fig. 1 Task database containing the 75 problem-solving tasks showing the different levels of complexity, different levels of support, and the different surface features of the learning tasks

(see Table 1) performing a problem-solving task (at the first or second level of complexity, see Table 1) by writing out the solution to each step that s/he was able to complete. The model then rated how much effort s/he invested in solving that problem, by circling the answer on a scale of 1 to 9 (Paas 1992). The scale was presented horizontally, with labels at the uneven numbers: (1) very, very little mental effort, (3) little mental effort, (5) neither little nor much mental effort, (7) much mental effort, and (9) very, very much mental effort. The model then rated how many steps s/he thought s/he had performed correctly (i.e., self-assessment) on a scale ranging from (0) no steps correct to (5) all steps correct by circling the answer and selected an appropriate subsequent task from the task database (according to the selection algorithm shown in Fig. 2), by circling that task. For example, one model had performed a task with a self-assessment of 3 steps formed correctly and invested very much (8) mental effort into that performance. Combining those two numbers resulted in a task-selection advice of minus 1, which meant going one step to the left in the task database (Fig. 1). While solving the problem, rating effort, self-assessing, and selecting a next task, the model was thinking aloud.

Learning phase and self-assessment feedback The learning phase consisted of three problems without support, one from each of the first three complexity levels (of the task database in Fig. 1). In the learning phase, participants engaged in the same activities as they had observed in the modeling examples: they first solved a problem, then rated how much effort they invested on

Table 1 Different features of the video modeling examples

Video modeling example	Gender of the model	Complexity level	Problem-solving performance	Invested mental effort	Task-selection advice
1	Female	Level 1	5 steps	2	+2
2	Male	Level 1	5 steps	5	+1
3	Female	Level 2	4 steps	7	0
4	Male	Level 2	3 steps	8	-1

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

Fig. 2 Algorithm used for task-selection advice showing the jump size and direction for each of the combinations of self-assessed performance and mental effort

the effort rating scale of 1–9, assessed their performance on a scale ranging from 0 to 5, and then selected an appropriate subsequent task from the database (Fig. 1) using the algorithm they had seen the model use (note though that the information on the algorithm shown in Fig. 2 was no longer available to participants). Participants did not actually receive the problem they selected; as the tasks in the learning (and test phase) were fixed (participants were made aware of this).

On the learning phase problems, participants received self-assessment feedback. In the general self-assessment feedback condition, a message appeared on the screen after a self-assessment was made, stating whether or not the self-assessment was correct and how many steps had actually been performed correctly (e.g., "Your self-assessment was incorrect. You performed 2 steps correctly."). The specific self-assessment feedback condition additionally received a list of the five steps with information on which steps were performed correctly and which incorrectly in the form of either a green check mark or a red cross mark (see Fig. 3). In the no feedback control condition, participants only saw a message stating that their answer had been registered.

Test phase The three problems in the test phase were isomorphic to the learning phase problems (i.e., same structural features but different surface features). Again, participants engaged in problem solving, effort rating, self-assessment, and task selection (but they did not receive the selected task). Self-assessment feedback was no longer provided on these tasks.

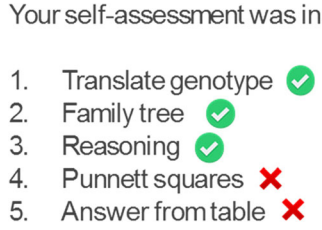


Fig. 3 Feedback provided during the learning phase to participants in the specific self-assessment feedback condition in Experiment 1. The general self-assessment feedback condition was only provided with the top sentence

Data analysis Performance on the problem-solving tasks from the pre-test, learning phase, and test phase, was scored by assigning one point for each correct step (i.e., range per problem: 0–5 points). Self-assessment accuracy was calculated by taking the absolute difference between the self-assessed and actual performance score for each problem-solving task and then averaging it over the problems (i.e., range: 0–5; Schraw 2009). To calculate task-selection accuracy, first the task-selection advice was derived with the algorithm (i.e., invested effort and actual performance on the task were combined into the task-selection advice), and then the absolute difference was taken between the task chosen and the advised task and averaged over the problems (i.e., range: 0–14; Kostons et al. 2012).

Results

Table 2 shows an overview of the results. Data were analyzed with (repeated measures) ANOVAs, and the effect size reported is partial eta-squared (η_p^2), for which .01 is considered a small, .06 a medium, and .14 a large effect size (Cohen 1988).

Preliminary checks Before conducting the analyses to test our hypotheses, we performed some checks. First, we checked whether prior knowledge did not differ between conditions (randomization check). An ANOVA on pretest performance showed no significant difference between conditions, $F(2, 105) = 0.27$, $p = .766$, $\eta_p^2 = .005$.

Table 2 Descriptive statistics from Experiment 1: Mean Performance during Pretest, Learning Phase, and Test Phase (range: 0–5); Mean Mental Effort during Pretest, Learning Phase, and Test Phase (range: 1–9); Mean Self-Assessments and the Accuracy of those Assessments during the Learning and Test phase (range: 0–5); Mean Task Selection Level and Accuracy of Task Selection during the Learning and Test Phase (range: 0–14)

Dependent variables	Condition					
	Control ($n = 37$)		General self-assessment feedback ($n = 34$)		Specific self-assessment feedback ($n = 37$)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Pretest						
Performance	0.53	0.47	0.62	0.48	0.55	0.59
Self-assessment	1.21	1.35	1.20	1.41	1.52	1.30
Mental effort	5.81	2.61	6.04	2.46	6.45	2.06
Learning phase						
Performance	3.15	1.51	2.88	1.59	3.10	1.58
Mental effort	3.56	1.20	4.10	1.83	3.77	1.57
Self-assessment	3.79	0.98	3.30	1.23	3.72	1.26
SA accuracy [a]	1.40	1.16	1.68	0.74	1.31	0.99
TS level	7.02	1.55	6.17	2.53	6.28	2.65
TS accuracy [a]	1.28	1.19	2.17	1.72	1.85	1.65
Test phase						
Performance	3.71	1.46	3.32	1.23	3.47	1.66
Mental effort	3.08	1.37	3.77	1.80	3.72	1.91
Self-assessment	3.78	1.31	3.25	1.53	3.59	1.40
SA accuracy [a]	0.96	0.85	1.59	0.91	1.13	1.04
TS accuracy [a]	2.30	2.17	2.77	2.37	2.36	1.94

[a] lower, better; SA, self-assessment; TS, task selection

Second, we checked whether problem-solving performance (both actual performance and self-assessed performance) increased over time and effort decreased over time, which is what one would expect to occur as a result of training and practice with the problems. Repeated measures ANOVAs with test moment (pre-test, learning phase, and test phase) as within-subjects factor and condition as between-subjects factor were conducted. On actual performance, this analysis showed a main effect of test moment on problem-solving performance, $F(1.92, 201.62) = 271.79$, $p < .001$, $\eta_p^2 = .721$.[2] Repeated contrasts showed that performance increased significantly from the pretest to the learning phase ($p < .001$) and from the learning phase to the test phase ($p < .001$), showing that participants performance improved as a result of training as well as from engaging in problem solving. Given that all students received the same tasks, there was no main effect of condition, $F(2, 105) = 0.34$, $p = .714$, $\eta_p^2 = .006$, and no interaction effect of test moment and condition, $F(3.84, 201.62) = 0.61$, $p = .653$, $\eta_p^2 = .011$.

On self-assessed performance there was a main effect of test moment, $F(1.42, 148.85) = 199.25$, $p < .001$, $\eta_p^2 = .655$.[3] Repeated contrasts showed that self-assessments increased significantly from the pretest to the learning phase ($p < .001$), but stayed at the same level from the learning phase to the test phase ($p = .428$) showing that participants perceived that their performance improved from training, but not from engaging in problem solving. There was no main effect of condition, $F(2, 105) = 1.31$, $p = .273$, $\eta_p^2 = .024$, and no interaction effect of test moment and condition, $F(2.84, 148.85) = 1.04$, $p = .374$, $\eta_p^2 = .019$.

On mental effort, there was a main effect of test moment, $F(1.32, 138.71) = 103.40$, $p < .001$, $\eta_p^2 = .496$.[4] Repeated contrasts showed that mental effort decreased significantly from the pretest to the learning phase ($p < .001$) and from the learning phase to the test phase ($p = .009$). There was no main effect of condition, $F(2, 105) = 1.29$, $p = .280$, $\eta_p^2 = .024$, and no interaction effect of test moment and condition, $F(2.64, 138.71) = 0.74$, $p = .513$, $\eta_p^2 = .014$.

Hypothesis 1: Task-selection accuracy during the learning phase We hypothesized that the self-assessment feedback provided to the experimental conditions would lead to more accurate task selections during the learning phase (where self-assessment feedback was provided). To test this hypothesis, an ANOVA with a planned contrast (general self-assessment condition and specific self-assessment condition vs. control condition) on task-selection accuracy during the learning phase was conducted, which showed that the difference between the experimental conditions and the control condition during the learning phase was significant, $t(105) = 2.34$, $p = .021$, $d = 0.46$. However, in contrast to our hypothesis, task selections were *less* accurate instead of more accurate in the experimental conditions compared to the control condition (see Table 2; higher deviation = less accurate). In order to explain these surprising findings, we additionally investigated which percentage of the self-assessment feedback had been negative (i.e., the self-assessment feedback indicated that the self-assessment was inaccurate). Of all instances of feedback, 73% were negative during the learning phase. Finally, we explored if the experimental conditions systematically chose easier tasks (after receiving self-assessment

[2] Mauchly's Test indicated that the assumption of sphericity was violated ($p = .013$). Because the Greenhouse-Geisser estimate of sphericity (ε) was greater than .75 ($\varepsilon = 0.926$), a Huynh-Feldt correction was applied.
[3] Mauchly's Test indicated that the assumption of sphericity was violated ($p < .001$). Because the Greenhouse-Geisser estimate of sphericity (ε) was less than .75 ($\varepsilon = 0.709$), a Greenhouse-Geisser correction was applied.
[4] Mauchly's Test indicated that the assumption of sphericity was violated ($p < .001$). Because the Greenhouse-Geisser estimate of sphericity (ε) was less than .75 ($\varepsilon = 0.661$), a Greenhouse-Geisser correction was applied.

feedback) than the control condition, $t(105) = 1.71$, $p = .090$, indicating a slight tendency to choose easier tasks.

Hypothesis 2: Self-assessment accuracy during the test phase We hypothesized that in the test phase (i.e., in the absence of feedback) participants who received self-assessment feedback during the learning phase would demonstrate more accurate self-assessment than participants in the control condition (Hypothesis 2a), and that the specific self-assessment feedback condition would demonstrate more accurate self-assessment than participants in the general self-assessment feedback condition (Hypothesis 2b). An ANOVA on self-assessment accuracy during the test phase showed a main effect of condition, $F(2, 105) = 4.29$, $p = .016$, $\eta_p^2 = .076$. Tukey's post-hoc tests showed that, in contrast to our hypotheses, the general self-assessment feedback condition was significantly *less* accurate than the no feedback control condition ($p = .015$). The specific self-assessment feedback condition fell in between, the means show that it was somewhat more accurate than the general feedback condition, but not significantly so ($p = .712$), and somewhat less accurate than the no feedback control condition, but not significantly so ($p = .100$).

Hypothesis 3: Task-selection accuracy during the test phase Finally, we hypothesized that the expected increase in self-assessment accuracy in the feedback conditions in the test phase (Hypothesis 2) would also positively affect task-selection accuracy during the test phase in those conditions. Because expected improvement in self-assessment accuracy did not occur, it was unlikely that task-selection accuracy would be affected, and indeed, the ANOVA on task-selection accuracy in the test phase showed no significant differences between conditions, $F(2, 105) = 0.50$, $p = .606$, $\eta_p^2 = .009$.

Discussion

The findings from Experiment 1 showed that participants who received self-assessment feedback did not make more accurate task selections than participants in the control condition during the learning phase (Hypothesis 1); they actually made *less* accurate task selections. Participants in the feedback conditions did not self-assess their performance more accurately than participants in the control condition during the test phase (Hypothesis 2a) and participants who received specific self-assessment feedback did not self-assess their performance more accurately than participants who received general self-assessment feedback during the test phase (Hypothesis 2b). Receiving general self-assessment feedback even seemed to result in *less* accurate self-assessments than receiving no feedback. However, these differences were not reflected in the accuracy of task selection during the test phase (Hypothesis 3).

The finding that self-assessment feedback led to *lower* task-selection accuracy during the learning phase, even though participants could directly rely on the feedback (which stated how many steps they performed correctly) to make their task-selection, suggests that receiving feedback may have biased task selection in a systematic way. Possibly, participants have reacted to the negative self-assessment feedback in a similar way as people commonly react to negative performance feedback (Ilgen and Davis 2000). If people attribute the feedback to their ability, they might select easier tasks in order to reach self-enhancement (Strube and Roemmele 1985). The data indeed show a slight tendency for the experimental conditions (who received self-assessment feedback) to choose easier tasks during the learning phase (see Table 2).

One possible explanation for the lack of beneficial effects of self-assessment feedback is that students might not have been able to adjust their internal standards adequately because they were not provided with information on the correct answers, which would make it difficult for them to assess performance on subsequent (test phase) tasks. That is, they would know either that they overestimated or underestimated their overall performance or which steps they performed correctly/incorrectly (specific feedback), but they would not know what they did wrong exactly. Therefore, Experiment 2 investigated whether adding correct answer feedback to the general self-assessment feedback and having students contrast their own answers with the correct answers would improve self-assessment accuracy in the absence of feedback.

Experiment 2

In Experiment 2, we investigated the combination of general self-assessment feedback and correct answer feedback. As mentioned in the introduction, being presented with the correct answers during self-assessment is known to improve accuracy (Baars et al. 2014; Baker et al. 2010; Dunlosky et al. 2010; Dunlosky and Rawson 2012; Lipko et al. 2009; Rawson and Dunlosky 2007; Rawson et al. 2011), but it is unclear whether participants would still show better accuracy in the absence of the answers. Therefore, Experiment 2 used a similar design as Experiment 1, distinguishing a learning phase (with feedback) and test phase (without feedback), to shed light on this issue. To control for the possibility that students would learn from studying the correct answers, thereby improving their performance (which could in turn lead to improved self-assessment accuracy; Dunning et al. 2003; Kruger and Dunning 1999), we added a condition in which students could study the correct answers, but could not compare them to their own answers. This enabled us to isolate the added effect of contrasting your own answers to correct answers. If restudy of the correct answers would lead to an increase of performance (and improved self-assessment accuracy), this could be detected in the difference between the correct answer condition and the control condition.

We hypothesized that during the learning phase, participants in both feedback conditions would be able to make more accurate task-selection decisions than participants in the control condition (Hypothesis 4), as these could be directly based on the actual performance indicated in the feedback. Regarding the test phase (i.e., in the absence of the feedback), we hypothesized that participants in the feedback conditions would demonstrate more accurate self-assessment than participants in the control condition (Hypothesis 5a), and that the self-assessment feedback + contrasting own answers with correct answers condition would demonstrate more accurate self-assessment than participants in the self-assessment feedback + correct answers condition (Hypothesis 5b; i.e., Hypothesis 5: contrast > correct > no). If Hypothesis 5 would be confirmed, we would expect a similar pattern of results for task-selection accuracy during the test phase (Hypothesis 6).

Method

Participants and design A total of 136 Dutch students in their second year of secondary education (second highest and highest level of secondary education) participated in this study. Five of those participants possessed too much prior knowledge (i.e., scoring 60% or higher on the pretest) and had to be excluded. Another 15 participants were excluded because they did not manage to finish the experiment within the class period. Six participants had to be removed

due to missing data on outcome variables.[5] The remaining sample of 110 participants had a mean age of 13.72 ($SD = 0.54$), and contained 54 boys and 56 girls. Participants were randomly assigned to one of the three conditions: (1) no feedback ($n = 36$), (2) self-assessment feedback + correct answers ($n = 38$), or (3) self-assessment feedback + contrasting own and correct answers ($n = 36$).

Materials, procedure and data analysis The materials, procedure and data analysis were identical to Experiment 1 except for the feedback intervention during the learning phase. In the self-assessment feedback + correct answers condition, a message appeared on the screen after a self-assessment was made, stating whether or not the self-assessment was correct and how many steps had actually been performed correctly (cf. the general self-assessment feedback in Experiment 1, e.g., "Your self-assessment was incorrect. You performed 2 steps correctly."), and what the correct answers were (see Fig. 4). The self-assessment feedback + contrasting condition additionally saw their own answers next to the correct answers and were instructed to compare and contrast their answers with the correct answers. In the no feedback condition, participants only saw a message stating that their answer had been registered (cf. Experiment 1).

Results

Table 3 shows an overview of the results.

Preliminary checks Before conducting the analyses that test our hypotheses, we checked whether the data met the following demands: prior knowledge should not differ between conditions (check on success of randomization procedure), problem-solving performance (both absolute and self-assessed) should increase over time and effort should decrease over time (check that learning occurred). An ANOVA on pretest performance showed no significant difference between conditions, $F(2, 107) = 0.42$, $p = .658$, $\eta_p^2 = .008$.

A repeated measures ANOVA with test moment (pretest, learning phase, and test phase) as within-subjects factor and condition as between-subjects factor, showed a main effect of test moment on problem-solving performance, $F(2, 214) = 510.64$, $p < .001$, $\eta_p^2 = .827$. Repeated contrasts showed that performance increased significantly from the pretest to the learning phase ($p < .001$) and from the learning phase to the test phase ($p < .001$). Although students in the experimental condition had the opportunity to study the correct answers, there was no main effect of condition, $F(2, 107) = 0.22$, $p = .801$, $\eta_p^2 = .004$, and no interaction effect of test moment and condition, $F(4, 214) = 0.76$, $p = .554$, $\eta_p^2 = .014$.

A similar repeated measures ANOVA on self-assessments of performance showed a main effect of test moment, $F(1.44, 153.79) = 449.75$, $p < .001$, $\eta_p^2 = .808$.[6] Repeated contrasts showed that self-assessed performance increased significantly from the pretest to the learning phase ($p < .001$), but stayed at the same level from the learning phase to the test phase

[5] Exclusion was equal across conditions, $\chi^2(2) = 0.141$, $p = .932$. Excluded participants did not perform worse than non-excluded participants during the learning phase, $t(128) = 1.727$, $p = .087$, but did assess their performance more accurately, $t(128) = 2.00$, $p = .048$.

[6] Mauchly's Test indicated that the assumption of sphericity was violated ($p < .001$). Because the Greenhouse-Geisser estimate of sphericity (ε) was less than .75 ($\varepsilon = 0.719$), a Greenhouse-Geisser correction was applied.

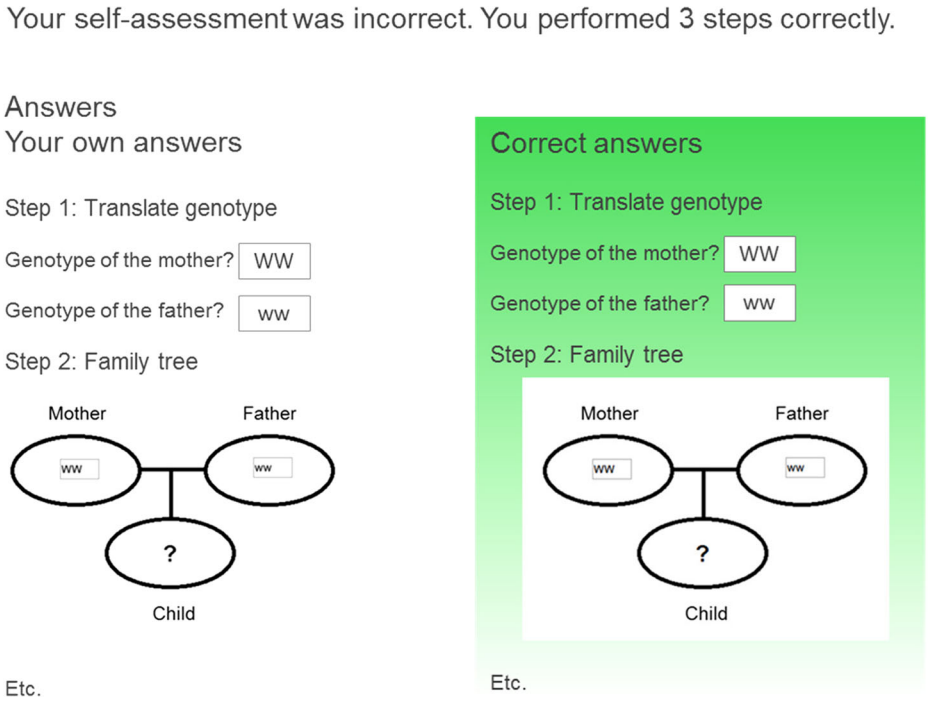


Fig. 4 Feedback provided during the learning phase to participants in the self-assessment feedback + contrast condition in Experiment 2. The self-assessment feedback + answers condition was only provided with the top sentence and the correct answers (the right part)

($p = .798$). There was no main effect of condition, $F(2, 107) = 0.33$, $p = .722$, $\eta_p^2 = .006$, and no interaction effect of test moment and condition, $F(2.88, 153.79) = 0.64$, $p = .585$, $\eta_p^2 = .012$.

A third repeated measures ANOVA on mental effort showed a main effect of test moment, $F(1.30, 138.65) = 95.99$, $p < .001$, $\eta_p^2 = .473$.[7] Repeated contrasts showed that mental effort decreased significantly from the pretest to the learning phase ($p < .001$) and from the learning phase to the test phase ($p = .043$). Similar to Experiment 1, there was no main effect of condition, $F(2, 107) = 1.14$, $p = .323$, $\eta_p^2 = .021$. However, there was an interaction effect of test moment and condition, $F(2.59, 138.65) = 3.60$, $p = .020$, $\eta_p^2 = .063$, indicating that the decrease of mental effort over time differed between the conditions with the self-assessment feedback + contrast condition showing an increase in mental effort from the learning phase to the test phase, while the mental effort in other conditions decreased from learning phase to test phase.

Hypothesis 4: Task-selection accuracy during the learning phase We hypothesized that the self-assessment feedback provided to the experimental conditions would lead to more accurate task-selection decisions during the learning phase (where self-assessment feedback was provided). To test this hypothesis, an ANOVA with a planned contrast (self-assessment feedback + correct answers condition and self-assessment feedback + contrast condition vs.

[7] Mauchly's Test indicated that the assumption of sphericity was violated ($p < .001$). Because the Greenhouse-Geisser estimate of sphericity (ε) was less than .75 ($\varepsilon = 0.648$), a Greenhouse-Geisser correction was applied.

Table 3 Descriptive statistics from Experiment 2: Mean Performance during Pretest, Learning Phase, and Test Phase (range: 0–5); Mean Mental Effort during Pretest, Learning Phase and Test Phase (range: 1–9); Mean Self-Assessments and the Accuracy of those Assessments during Learning and Test Phase (range: 0–5); Mean Task Selection Level and Accuracy of Task Selection during the Learning and Test Phase (range: 0–14)

Dependent variables	Condition					
	Control (n = 36)		Self-assessment feedback + answer (n = 38)		Self-assessment feedback + contrast (n = 36)	
	M	SD	M	SD	M	SD
Pretest						
Performance	0.89	0.80	0.98	0.89	0.81	0.65
Self-assessment	1.01	1.21	1.08	1.04	1.04	1.34
Mental effort	6.72	2.13	5.92	2.24	5.19	2.45
Learning phase						
Performance	3.31	1.06	3.46	1.23	3.50	1.08
Self-assessment	4.07	0.88	4.05	0.93	3.90	0.86
Mental effort	3.69	1.63	3.50	1.57	3.38	1.44
SA accuracy*	1.19	0.78	1.21	0.85	1.18	0.78
TS level	6.77	1.97	6.71	1.54	6.73	1.50
TS accuracy*	1.43	1.29	1.20	1.70	1.23	1.12
Test phase						
Performance	3.94	0.89	3.93	1.09	4.01	0.91
Self-assessment	4.20	1.01	3.91	1.29	3.99	1.22
Mental effort	3.16	1.69	3.17	1.88	3.30	1.69
SA accuracy*	0.87	0.60	0.75	0.64	0.88	0.79
TS accuracy*	1.64	1.66	1.42	1.54	1.63	1.74

*lower, better; SA, self-assessment; TS, task selection

control condition) on task-selection accuracy during the learning phase was conducted, which showed that the difference between the experimental conditions and the control condition during the learning phase was not significant, $t(107) = 0.29$, $p = .773$, $d = 0.06$. Similar to Experiment 1, we investigated the percentage of negative self-assessment feedback. During the learning phase 75% of the feedback messages were negative. Finally, we explored if the experimental conditions systematically chose easier tasks (after receiving self-assessment feedback) than the control condition. This was not the case, $t(107) = -0.14$, $p = .890$.

Hypothesis 5: Self-assessment accuracy during the test phase We hypothesized that in the test phase (i.e., in absence of feedback) participants who received self-assessment feedback during the learning phase would demonstrate more accurate self-assessment than participants in the control condition (Hypothesis 5a), and that the self-assessment feedback + contrast condition would demonstrate more accurate self-assessment than participants in the self-assessment + correct answers condition (Hypothesis 5b). An ANOVA on self-assessment accuracy during the test phase showed no significant differences between conditions, $F(2, 107) = 1.18$, $p = .310$, $\eta_p^2 = .022$.

Hypothesis 6: Task-selection accuracy during the test phase Finally, if the participants in the self-assessment feedback conditions indeed improved on self-assessment accuracy during the test phase, then this could be expected to result in more accurate task-selection decisions during the test phase (Hypothesis 5). Because we did not see the expected improvement in self-

assessment accuracy during the test phase though, it was unlikely that task-selection accuracy would be affected, and, indeed, the ANOVA on task-selection accuracy in the test phase showed no significant differences between conditions, $F(2, 107) = 0.54$, $p = .586$, $\eta_p^2 = .010$.

Discussion

Participants who received self-assessment feedback did not make more accurate task selections than participants in the control condition during the learning phase (Hypothesis 4). Participants in the self-assessment feedback + contrast condition did not make more accurate self-assessments during the test phase than participants in the self-assessment feedback + correct answers condition, who did not make more accurate self-assessments during the test phase than participants in the control condition (Hypothesis 5a and 5b). Not surprisingly, therefore, participants who received self-assessment feedback did not show more accurate task selections in the test phase (Hypothesis 6).

Contrary to our findings, previous studies using correct answers to improve self-assessment accuracy did find beneficial effects on self-assessment accuracy (Baars et al. 2014; Dunlosky et al. 2010; Lipko et al. 2009; Rawson and Dunlosky 2007). However, the methods used in those studies differ critically from the methods used in the present study. While in the previous studies the correct answers were provided *during* self-assessment (or monitoring) of performance (i.e., students could base their self-assessment on the comparison of their answer to the correct answer), in the present study the correct answer feedback was provided *after* self-assessment. We expected that this would allow participants to refine their internal standards based on the feedback, as a consequence of which subsequent self-assessments without feedback should have been more accurate. However, we did not find any indications that self-assessment accuracy improved from self-assessment feedback.

General discussion

The main objective of the current study was to investigate whether – after an initial self-regulated learning training – the accuracy of self-assessments would improve from feedback focused on self-assessment accuracy (i.e., self-assessment feedback). Based on the SRL model of Butler and Winne (1995) we predicted that the self-assessment feedback received during the learning phase, would allow learners to refine their internal standards, leading to higher self-assessment accuracy during the test phase, in the absence of this feedback. In Experiment 1 we investigated the potential benefit of more specific self-assessment feedback (i.e., feedback information on the level of distinct problem-solving steps) instead of more general self-assessment feedback (i.e., feedback information on the level of the task). In Experiment 2 we investigated the effects of adding correct answers to the self-assessment feedback and having learners contrast those with their own answers (with an extra condition to check if learning from the correct answers could lead to improved self-assessment accuracy). In both experiments, self-assessment feedback failed to improve self-assessment accuracy.

We will now discuss possible explanations for our null findings, provided by our data, the limitations of our study, or other possible (speculative) mechanisms. As for our data, we could investigate the possibility that the self-assessment feedback could (inadvertently) have caused learners to simply correct their self-assessments downwards without considering their actual performance (cf. Roelle et al. 2017). That is, the self-assessment feedback that the learners received indicated that their self-assessment was *incorrect* most of the time, and thus had

mostly a negative connotation. Consequently, learners might have adjusted their self-assessments downwards, and we have found some evidence that this was the case. Self-assessments were lower in the general self-assessment feedback condition compared to the other conditions in Experiment 1 (although not significantly; see Table 2). If learners simply gave themselves lower self-assessment ratings without considering their actual performance, they would not take into account the fact that their performance would improve from the learning phase to the test phase due to practice, and thus, their self-assessments would remain inaccurate. Indeed, the analysis of self-assessed performance across the different phases suggests that learners perceived that their performance improved as a consequence of training (i.e., from pretest to learning phase), but not from engaging in problem solving (i.e., from the learning to the test phase) even though their actual performance continued to improve (see Table 2). These explanations are not mutually exclusive and, interestingly, both involve a mechanism in which the learner directs less rather than more attention to monitoring.

A potential reason (similar to the explanation for the lower self-assessment accuracy in Experiment 1) for the lack of support for our hypotheses regarding task-selection accuracy could also be explored using our data. That is, participants may have interpreted the self-assessment feedback negatively and might have chosen easier tasks (which learners are likely to do after negative feedback, see Ilgen and Davis 2000). This might be mediated through self-efficacy, as a recent meta-analysis showed a relationship between self-assessment interventions and self-efficacy (Panadero et al. 2017). Indeed, our data show that learners in the experimental conditions chose slightly easier tasks during the learning phase (see Table 2), but this was not the case in Experiment 2 (see Table 3). Given that the feedback in Experiment 1 did not provide the correct answer, but feedback in Experiment 2 did, it is possible that this effect is mediated by attributions (i.e., how students explain their successes or failures; cf. Weiner 1986). For instance, when students attribute the (negative) self-assessment feedback to ability, their self-efficacy might go down and they might reflexively adjust their self-assessments downwards ("I must not be very good at solving these problems") and, consequently, would be more inclined to select easy tasks after feedback. The reason this did not occur in Experiment 2, might be because the knowledge of the correct answers helped them interpret the assessment feedback in a different manner that did not lower self-efficacy, or because it lead to an expectation that they would be able to solve similar problems in the future (cf. hindsight bias; Kornell 2015). However, this is highly speculative, as we did not assess students' considerations for self-assessments.

This study also had some limitations that need to be considered as potential explanations for our findings. As all experiments in this study were performed in the classroom (in order to increase the ecological validity) there was limited time available for the intervention. As the length of an intervention is often related to the effect of the intervention (e.g., Bangert-Drowns et al. 2004), this might have had a diminishing effect on our intervention. Future studies might include interventions extended over multiple lessons. This could also provide the opportunity to investigate student motivation, which could provide insight into the students' motivational explanations concerning task selection. Another limitation of this study concerns the size, rather than the length, of the interventions. The experiments in this study use small interventions that are easily implemented into electronic learning environments. However, using small interventions could have decreased the power to detect differences between the conditions, which might be an additional explanation for the null findings. Lastly, we do not know if these results would be the same using different tasks, different age groups, or other domains. Future research could explore such generalization of the results.

Finally, we could speculate about possible mechanisms that might explain our results. We do not know precisely on which cues students base their self-assessments (see De Bruin and Van Merriënboer 2017; Koriat 1997). An endeavor to uncover which cues are used, would require process measures such as interviews, think-aloud protocols, et cetera (Azevedo 2009; De Bruin and Van Gog 2012). However, we could speculate that attributions might eventually trigger self-protective strategies (i.e., choosing easier tasks) if self-assessment feedback is attributed to ability. In such cases self-enhancement is more important than accurate self-assessment (Strube and Roemmele 1985). In this case, it might be more effective if the intervention would redirect students' attention from the invalid cues they used (i.e., cues that are not predictive of their actual performance) towards more valid (i.e., predictive) cues, which might better enable students to adjust their internal standards (i.e., what constitutes good performance; Winne and Hadwin 1998), resulting in more accurate self-assessments in the absence of feedback. This would, ultimately, allow learners to more accurately control their learning process and improve their learning outcomes.

Another possible mechanism is that participants might have started to expect similar feedback on subsequent tasks and might, consequently, have payed less attention to their subsequent performance (i.e., monitoring *during* problem solving) and self-assessments (Salmoni et al. 1984). Thus, counterintuitively, the self-assessment feedback might have led to paying *less*, instead of *more*, attention to self-assessment (Shute 2008). The field of self-assessment research might benefit from future research into which specific types of self-assessment feedback produce higher learning gains.

Acknowledgments The authors would like to thank all participating schools and students.

Funding This research was funded by the Netherlands Initiative for Education Research (NRO PROO; project number: 411–12-015).

Compliance with ethical standards

Conflict of interest The authors declare they have no conflict of interest.

Appendix 1

Example of problem-solving task used in pretest and posttest (first level of complexity)

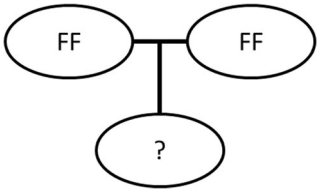
Fur color

A guinea pig's fur color is determined by a gene, which expresses itself as black in its dominant form (F) and white in its recessive form (f). Two guinea pigs, who are both black and homozygote for that trait, produce offspring. What are the possible genotypes for this offspring?

Step 1. Translate information from text into genotypes.

– Both guinea pigs are homozygote for the dominant allele, so both genotypes are FF.

Step 2. Fill in a family tree.



Step 3. Determine number of Punnett squares by deciding if problem is to be solved deductively or inductively.

– Both parents are given, so we can solve the problem deductively. Solving problems deductively only requires one Punnett square.

Step 4. Step 4. Fill in the Punnett square.

	F	F
F	FF	FF
F	FF	FF

Step 5. Find the answer in the Punnett square.

– The only possible genotype for the offspring is FF.

References

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning, 4*, 87–95. https://doi.org/10.1007/s11409-009-9035-7.

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*, 523–535. https://doi.org/10.1037/0022-0663.96.3.523.

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.

Baker, J. M. C., Dunlosky, J., & Hertzog, C. (2010). How accurately can older adults evaluate the quality of their text recall? The effect of providing standards on judgment accuracy. *Applied Cognitive Psychology, 24*, 134–147. https://doi.org/10.1002/acp.1553.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs: Prentice Hall.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*, 29–58. https://doi.org/10.3102/00346543074001029.

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-1130011-143823.

Bol, L., Campbell, K. D. Y., Perez, T., & Yen, C. J. (2016). The effects of self-regulated learning training on community college students' metacognition and achievement in developmental math courses. *Community College Journal of Research and Practice, 40*, 480–495. https://doi.org/10.1080/10668926.2015.1068718.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school*. Washington: National Academy Press.

Brown, G. T., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 367–393). Thousand Oaks: Sage Publications.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. Review of Educational Research, 65(3), 245-281. https://doi.org/10.3102/00346543065003245.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum.

Corbalan, G., Kester, L., & Van Merriënboer, J. J. (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. (2009). 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.

De Bruin, A. B. H., & 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.

De Bruin, A. B., & Van Merriënboer, J. J. G. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction, 51*, 1–9. https://doi.org/10.1016/j.learninstruc.2017.06.001.

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.

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2010). Improving college students' evaluation of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology, 64*, 467–484. https://doi.org/10.1080/17470218.2010.502239.

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. https://doi.org/10.1111/j.1529-1006.2004.00018.x.

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.

Hoogerheide, V., Van Wermeskerken, M., Loyens, S. M., & Van Gog, T. (2016). Learning from video modeling examples: Content kept equal, adults are more effective models than peers. *Learning and Instruction, 44*, 22–30. https://doi.org/10.1016/j.learninstruc.2016.02.004.

Ilgen, D., & Davis, C. (2000). Bearing bad news: Reactions to negative performance feedback. *Applied Psychology, 49*, 550–565. https://doi.org/10.1111/1464-0597.00031.

Koriat, A. (1997). Monitoring one's 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.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—And costs—Of dropping flashcards. *Memory, 16*, 125–136. https://doi.org/10.1080/09658210701763899.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449–468. https://doi.org/10.1037/a0017350.

Kornell, N. (2015). If it is stored in my memory I will surely retrieve it: Anatomy of a metacognitive belief. *Metacognition and Learning, 10*, 279–292. https://doi.org/10.1007/s11409-014-9125-z.

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.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15*, 307–318. https://doi.org/10.1037/a0017599.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*, 303–314. https://doi.org/10.1007/s11409-011-9083-7.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 125–173). New York: Academic Press.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1*, 159–179. https://doi.org/10.1007/s10409-006-9595-6.

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.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422.

Panadero, E., Brown, G. T., & Strijbos, J. W. (2015). 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.

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004.

Puustinen, M., & Pulkkinen, L. (2001). Models of self-regulated learning: A review. Scandinavian Journal of Educational Research, 45(3), 269-286. https://doi.org/10.1080/00313830120074206

Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., Van Merriënboer, J., & Van Gog, T. (2018). Training self-regulated learning skills with video modeling examples: Do task-selection skills transfer?. *Instructional Science*, *46*(2), 273–290. https://doi.org/10.1007/s11251-017-9434-0

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*, 559–579. https://doi.org/10.1080/09541440701326022.

Rawson, K. A., O'Neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied, 17*, 288–302. https://doi.org/10.1037/a0024749.

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*, 1–37. https://doi.org/10.1111/cogs.12086.

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.

Roelle, J., Schmidt, E. M., Buchau, A., & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology, 109*, 99–117. https://doi.org/10.1037/edu0000132.

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin, 95*, 355–386. https://doi.org/10.1037/0033-2909.95.3.355.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 33–45. https://doi.org/10.1007/s11409-008-9031-3.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. https://doi.org/10.3102/0034654307313795.

Strube, M. J., & Roemmele, L. A. (1985). Self-enhancement, self-assessment, and self-evaluative task choice. *Journal of Personality and Social Psychology, 49*, 981–993. https://doi.org/10.1037/0022-3514.49.4.981.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*, 155–174. https://doi.org/10.1007/s10648-010-9134-7.

Weiner, B. (1986). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*, 548–573. https://doi.org/10.1007/978-1-4612-4948-1_6.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 279–306). Hillsdale: Erlbaum.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology, 27*, 551–572. https://doi.org/10.1016/S0361-476X(02)00006-1.

Winne, P. H., & Nesbit, J. C. (2009). Supporting self-regulated learning with cognitive tools. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 259–277). New York: Routledge.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*, 3–17. https://doi.org/10.1207/s15326985ep2501_2.

Zimmerman, B. J., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology, 94*, 660–668. https://doi.org/10.1037/0022-0663.94.4.660.

Affiliations

Steven F. Raaijmakers [1] · Martine Baars [2] · Fred Paas [2,3] · Jeroen J. G. van Merriënboer [4] · Tamara van Gog [1]

[1] Department of Education, Utrecht University, P.O. Box 80140, 3508 TCUtrecht, The Netherlands

[2] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands

[3] School of Psychology, University of Wollongong, Wollongong, Australia

[4] School of Health Professions Education, Maastricht University, Maastricht, The Netherlands