Learning and Instruction 51 (2017) 36–46

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Effects of performance feedback valence on perceptions of invested mental effort

Steven F. Raaijmakers [a, b, *], Martine Baars [b], Lydia Schaap [b, c], Fred Paas [b, d], Tamara van Gog [a, b]

[a] Department of Education, Utrecht University, The Netherlands
[b] Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, The Netherlands
[c] Learning and Innovation Centre, Avans University of Applied Sciences, Breda, The Netherlands
[d] Early Start Research Institute, University of Wollongong, Australia

ARTICLE INFO

Article history:
Received 18 April 2016
Received in revised form
1 December 2016
Accepted 8 December 2016
Available online 16 December 2016

Keywords:
Mental effort
Feedback valence
Problem solving
Cognitive load measurement

ABSTRACT

We investigated whether the valence of performance feedback provided after a task, would affect participants' perceptions of how much mental effort they invested in that same task. In three experiments, we presented participants with problem-solving tasks and manipulated the presence and valence of feedback between conditions (no, positive, or negative feedback valence), prior to asking them to rate how much mental effort they invested in solving that problem. Across the three experiments—with different problem-solving tasks and participant populations—we found that subjective ratings of effort investment were significantly higher after negative than after positive feedback; ratings given without feedback fell in between. These findings show that feedback valence alters perceived effort investment (possibly via task perceptions or affect), which can be problematic when effort is measured as an indicator of cognitive load. Therefore, it seems advisable to measure mental effort directly after each task, before giving feedback on performance.

1. Introduction

Providing learners with feedback has proven effective for enhancing performance (Hattie & Timperley, 2007), both in classrooms (e.g., Bangert-Drowns, Kulik, Kulik, & Morgan, 1991), and in computer-based learning environments (e.g., Van der Kleij, Feskens, & Eggen, 2015). This feedback will have a certain valence for learners: when the feedback tells them their performance was incorrect, or lower than they expected, it has negative valence; when it tells them their performance was correct, or higher than they expected, the feedback has positive valence. The question addressed in the present study, is whether the valence of performance feedback on a task, has consequences for participants' perceptions of how much mental effort they invested in that task.

This question is of both theoretical and practical relevance, because subjective ratings of how much mental effort students' perceived to have invested in a task, are widely used in educational research and in (adaptive or self-regulated) computer-based learning environments as an indicator of the cognitive load that learners experienced (Paas, 1992; Paas, Van Merriënboer, & Adam, 1994; for reviews, see; Paas, Tuovinen, Tabbers, & Van Gerven, 2003; Van Gog & Paas, 2008). Theoretically, we still know very little about the "cues" that learners use when they are asked to rate how much effort they invested in a task. That is, learners' perceptions of effort investment presumably rely on multiple aspects of their experiences during task performance (i.e., cues, such as how difficult, fluent, or speedy the performance process was). Investigating whether there are external influences (such as feedback valence) on effort perceptions would be a first step towards attaining insight into which cues are probably being used. Practically, it is imperative that effort measures reliably reflect experienced cognitive load, which is no longer the case when learners' effort perceptions would be affected post-hoc by external influences (such as feedback valence). Thus, investigating whether such external influences occur, can help inform researchers and practitioners on when to best measure learners' perceptions of invested mental effort.

* Corresponding author. Department of Education, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands.
E-mail address: s.f.raaijmakers@uu.nl (S.F. Raaijmakers).

http://dx.doi.org/10.1016/j.learninstruc.2016.12.002
0959-4752/

1.1. Use of mental effort ratings in educational research and learning environments

According to Cognitive Load Theory (Sweller, Ayres, & Kalyuga, 2011), cognitive load originates from an interaction of task characteristics (e.g., the more complex the task, the higher the load it imposes) and learner characteristics (e.g., the higher a learner's knowledge, the lower the load imposed by the task). Cognitive load can be assessed in terms of processing demands, using objective measures such as dual-tasks (Brünken, Plass, & Leutner, 2003), or physiological measures (Paas, Tuovinen, Tabbers, & Gerven, 2003), or in terms of experienced cognitive load, by asking learners to rate how much effort they invested in a task. Mental effort is defined as "the aspect of cognitive load that refers to the cognitive capacity that is actually allocated to accommodate the demands imposed by the task; thus, it can be considered to reflect the actual cognitive load" (Paas et al., 2003, p. 64). Both types of measures have their strengths and weaknesses. Objective measures, are less easy to administer than subjective ratings, but have the benefit of providing moment-to-moment information regarding (fluctuations in) processing demands *during* task performance. Subjective measures, in contrast, are easy to administer and also seem sensitive to variations in cognitive load (which, as mentioned above, originates from an interaction between task and learner characteristics): subjective perceptions of effort investment have been shown to increase (or decrease) with increases (decreases) in task complexity (e.g., Paas et al., 1994; Schmeck, Opfermann, Van Gog, Paas, & Leutner, 2015), to be lower for learners with higher prior knowledge compared to learners with lower prior knowledge working on the same task (e.g., Nievelstein, Van Gog, Van Dijck, & Boshuizen, 2013), and to decrease from pretest to posttest as a consequence of knowledge acquired during a study phase (e.g., Hoogerheide, Loyens, & Van Gog, 2014). Which method is most appropriate depends on the research question being addressed.

Our present study is concerned with subjective perceptions of invested mental effort. Since it was first published, the 9-point mental effort rating scale[1] developed by Paas (1992) has become widely used in research on learning and instruction, as an indicator of learners' experienced cognitive load (for a review, see Van Gog & Paas, 2008). In combination with performance measures, subjective perceptions of how much mental effort was invested in a task are useful for obtaining information about the efficiency of instruction (Hoffman & Schraw, 2010; Paas & Van Merriënboer, 1993; Van Gog & Paas, 2008), and for guiding the selection of learning tasks in adaptive (e.g., Salden, Paas, Broers, & Van Merriënboer, 2004) or self-regulated (e.g., Kostons, Van Gog, & Paas, 2012) learning environments. For instance, Kostons et al. (2012) trained students how to select a next task based on a combination of their (self-assessed) performance and perceptions of invested mental effort. After a self-regulated learning phase, participants who had been trained in task selection showed higher knowledge gains than students who had not received training. This study exemplifies the usefulness of effort measures not only for educational research but also for educational practice (i.e., in improving self-regulated learning).

In order to effectively use effort measures in educational research or educational practice, however, it is imperative that learners' perceptions of invested mental effort reliably reflect experienced cognitive load. It is therefore important to investigate whether there are external factors (other than learners' own experiences with the task) that might affect learners' perceptions of how much effort they invested in a task, but such research is still scarce. Some recent studies have been conducted on when to best administer the effort rating scale, in which it was found that a single, overall rating of effort invested in a series of tasks, was systematically higher than the average of task-specific ratings given immediately after each task (Schmeck et al., 2015; Van Gog, Kirschner, Kester, & Paas, 2012).

These findings suggest that it is preferable to measure perceptions of invested effort directly after each task; however, it is still unclear what causes this discrepancy between overall and task-specific ratings. Which brings us to the theoretical relevance of investigating whether and how external factors affect learners' perceptions of how much effort they invested in a task: as mentioned earlier, we still know very little about what cues learners use when they are asked to rate how much effort they invested in a task. In analogy to metacognitive judgments, it is likely that learners use certain cues resulting from their experience with the task as a basis for their effort ratings. Knowing whether and which external influences affect learners' effort perceptions would constitute a first step towards attaining insight into what cues they are probably using. We start out here, by investigating whether and how the valence of performance feedback affects perceptions of invested effort.

1.2. Cue utilization, feedback, and mental effort ratings

Research on metacognitive monitoring judgments (e.g., judgments of learning; JOLs), has come a long way in past decades in uncovering which sources of information (i.e., *cues*), learners use when predicting their future memory performance (Koriat, 1997, 2015). In the cue utilization view of JOLs, Koriat distinguishes three types of cues: intrinsic, extrinsic, and mnemonic cues (Koriat, 1997). Intrinsic cues concern inherent attributes of the study material associated with the ease or difficulty of learning. For instance, learners may judge how well they will remember a certain word-pair based on the relatedness between the words, as higher relatedness is generally associated with better memory (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989). Extrinsic cues are related to how the material is presented (e.g., the number of repetitions of an item, available study time) or what strategy the learner applies when studying (e.g., level of processing, imagery). Intrinsic and extrinsic cues can affect JOLs directly, but also indirectly, by affecting the third type of cue, mnemonic cues (Koriat, 1997). Mnemonic cues can be described as internal signals from subjective experience that might indicate that information has been learned and will be remembered on a future occasion, such as fluency during encoding or retrieval (Koriat & Ma'ayan, 2005). As a learner engages with a task, the judgment shifts from being information-based (i.e., relying on intrinsic and extrinsic cues) to more experience-based (i.e., relying on mnemonic cues; Koriat & Ma'ayan, 2005).

Indicating how much effort you invested in a task that you just completed, is, of course, very different from predicting your future memory test performance by means of a JOL (indeed, effort invested in learning can in itself serve as a cue for a JOL: Koriat, Ackerman, Lockl, & Schneider, 2009; Koriat, Nussinson, & Ackerman, 2014; see also Baars, Vink, Van Gog, De Bruin, & Paas, 2014, who found a negative correlation between learners' perceptions of invested effort and their JOLs). Nevertheless, rating how much mental effort you invested in a task also constitutes a subjective, introspective judgment, that learners likely make using experience-based cues (such as their perceptions of how difficult, fluent, or speedy the task performance process was). And some of

[1] "How much mental effort did you invest in solving this problem?" with answer options ranging from (1) very very low effort, to (9) very, very high effort. Depending on the task, the question could also be phrased as "… in studying this text/animation/worked example".

these cues, such as perceived task difficulty, might be affected by external factors, such as performance feedback.

Indeed, there is substantial evidence that performance feedback can affect effort investment on a *subsequent* task, which seems to be driven by a reappraisal of *task demands* (Kahneman, 1973), or *task difficulty* (Brehm & Self, 1989; Meshkati, 1988), and serves the goal of reducing the *discrepancy* between a standard and current performance (Carver & Scheier, 1982; Kluger & DeNisi, 1996; Locke & Latham, 1990). Although the various models of the relationship between performance, feedback, and effort, all use different terms, they share the common notion that mobilization of mental effort is controlled by information about the task. When a learner receives negative feedback (i.e., performance is not up to par), this signals that task demands or task difficulty are high, and that mobilization of more effort on the next task is required in order to improve performance (i.e., bring it closer to the standard). The role of positive feedback is less pronounced in these models. However, Carver and Scheier (2000) state that positive feedback might lead people to reduce effort investment; being able to achieve a goal with less effort, frees up resources for other goals (see also Efklides, 2006). There is also evidence that when faced with competing tasks, people use the feedback that they already performed well on a certain task to allocate effort to other tasks (Larson & Callahan, 1990).

The question we are interested in, however, is not whether feedback valence affects effort investment on *subsequent* tasks, but whether it affects perceptions of how much effort was invested in the task that was just *completed*, on which the feedback was received. Research on affective influences on metacognitive judgments suggests that this might very well be the case. In her Metacognitive and Affective Model of Self-Regulated Learning, Efklides (2006, 2008, 2014) states that affective factors can influence metacognitive judgments and trigger control responses (i.e., regulation of current or subsequent study activities). In this model, estimates of effort that will be required (prospective) and how much effort was invested (retrospective) are considered as part of 'metacognitive experiences'. According to this model, the positive or negative valence of feedback, may lead to a reappraisal of task demands or task difficulty (at least partly) via affective states (Efklides, 2006). That is, positive valence (success) can be associated with positive emotions such as joy and pride (activating) or relief and contentment (deactivating), whereas negative valence (failure) can be associated, with frustration or anger (activating) or sadness and disappointment (deactivating), for instance (Pekrun, 2006; Pekrun, Frenzel, Goetz, & Perry, 2007, pp. 13–36).

There is some interesting evidence suggesting that positive and negative emotions affect task appraisal: sad stimuli (even masked) can increase experienced task difficulty and effort investment (as assessed by cardiovascular reactivity) compared to happy stimuli (Gendolla & Silvestrini, 2011). Positive emotions induced by the design of multimedia learning materials, have been found to reduce task difficulty ratings (though not effort ratings) compared to neutral material design (Plass, Heidig, Hayward, Homer, & Um, 2014; Experiment 1). Findings by Knörzer, Brünken, and Park (2016) showed that learning outcomes were reduced by positive emotions and increased by negative emotions. This seems in line with the other studies, as increased effort in response to negative emotions would lead to better learning outcomes and decreased effort in response to positive emotions would lead to lower learning outcomes. The objective load measure used seemed in line with the latter; however, the subjective effort ratings were not affected and subjective difficulty ratings were *higher* in the positive emotion condition, which contrasts with the Plass et al. study. Note though, that both types of subjective ratings were only administered once (after slide 4 out of 11); possibly the outcomes would have been different had the ratings scales been administered repeatedly after each slide (cf. Schmeck et al., 2015; Van Gog et al., 2012). Last but not least, Efklides and Dina (2004) investigated the effects of feedback on perceived difficulty and mental effort. Positive feedback *reduced* retrospectively reported feelings of difficulty and invested effort compared to prospective estimates of difficulty and effort (prior to the task), whereas negative feedback *increased* feelings of difficulty (though not effort).

In sum, even though any information provided after task performance cannot change the cognitive load that was actually experienced during task performance (as it has already been completed), and as such, should not affect perceptions of invested mental effort, the literature reviewed here suggests that it is very likely that feedback valence could affect perceptions of invested mental effort (possibly via affect), with negative feedback resulting in higher, and positive feedback in lower effort ratings compared to no feedback. The present study was designed to address this hypothesis.

1.3. The present study – main hypotheses

The present study comprises three experiments. In Experiment 1 and 2, we used five complex problem-solving tasks (so-called 'weekday problems'; cf. Van Gog et al., 2012; Experiment 3), that were so complex that it would be very hard for learners to estimate whether their answer was correct or not. This was important for the credibility of the feedback valence manipulation, which informed participants that their answer was correct (positive valence) or incorrect (negative valence), regardless of their actual performance. In Experiment 3, we used more educationally relevant problem-solving tasks (in biology; cf. Corbalan, Kester, & Van Merriënboer, 2009; Kostons et al., 2012). Because these were multistep problems (forced response on each step), we used a different feedback valence manipulation: participants were asked to assess their own performance, by indicating how many steps they thought they had performed correctly on a scale of 0–5. The positive valence condition then received feedback that their score was two points higher (i.e., two more steps correct) than they thought (bound by the maximum score of 5) and the negative valence condition received feedback that their score was two points lower (i.e., two steps less correct) than they thought.

In all experiments, participants were assigned to one of three conditions: no feedback, positive feedback valence, or negative feedback valence. After receiving feedback (depending on their assigned condition), they rated how much mental effort they invested in the task. The main hypothesis (Hypothesis 1), addressed in all three experiments, was that negative feedback would lead to higher and positive feedback would lead to lower ratings of mental effort invested in the problems, compared to the control condition that did not receive any feedback. Moreover, after the last of the five tasks participants were also asked to rate how much effort they invested in completing all tasks (overall effort rating). We hypothesized that we would replicate prior findings (Schmeck et al., 2015; Van Gog et al., 2012) that the overall effort rating would be higher than the average of task-specific ratings (Hypothesis 2), but were mainly interested in investigating whether the overall rating would be affected by the feedback in the same manner as the average of task-specific ratings (Question 1 in Experiment 1, Hypothesis 3 in Experiments 2 and 3).

2. Experiment 1

2.1. Method

2.1.1. Participants and design

Participants ($N = 186$) from the United States were recruited via

Amazon's Mechanical Turk (http://www.mturk.com). They received $0.50 for their participation (the experiment took 8 min to complete on average). Only participants with a good reputation (i.e., a Mechanical Turk approval rate of above 95%) were recruited. Four participants had to be removed due to repeated participation after a check for duplicate IP addresses. Another 32 did not finish the session and were also removed. The remaining 150 participants had a mean age of 34.63 years (SD = 12.47; range: 19–76), 41% had completed a bachelor's degree or higher, and 53 of them were male. To check whether participants attended to the questions and did not randomly click on answers, they answered several attention check questions (e.g., "Click on all the words that represent animals."; Hauser & Schwarz, 2016). All participants performed above chance.

Participants were randomly assigned to one of three conditions, in which they performed 5 tasks with either: no feedback (n = 47), predominantly positive feedback (n = 53), leading them to believe their performance was correct (not dependent on actual performance) on all tasks except for the second (PNPPP), or predominantly negative feedback (n = 50), leading them to believe their performance was incorrect (not dependent on actual performance) on all tasks except for the second (NPNNN). Feedback valence was reversed on the second task, so that the feedback manipulation would remain credible. None of the participants mentioned awareness of the feedback manipulation when we asked them afterwards what they felt the goal of the experiment was.

2.1.2. Materials

All materials were programmed and presented in Qualtrics Survey Software (http://www.qualtrics.com); participants received a link to the survey via Mechanical Turk.

2.1.2.1. Problem-solving tasks. The problem-solving tasks were five complex weekday problems (cf. Van Gog et al. (2012); Experiment 3). These 'weekday' problems were modeled after the problem used by Sweller (1993, p. 6): 'Suppose 5 days after the day before yesterday is Tuesday. What day of the week is yesterday?'. Participants were required to respond. Performance on the problem-solving tasks was rated as either correct (1 point) or incorrect (0 points). The order of tasks was the same for all participants.

2.1.2.2. Feedback. Feedback consisted of a message appearing on the screen after the task was completed, stating either "Your answer was correct." (positive feedback) or "Your answer was incorrect." (negative feedback).

2.1.2.3. Mental effort. Participants were asked to indicate how much mental effort they invested in solving the problem, on a 9-point rating scale (task-specific mental effort rating; 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*. After having performed all tasks, participants were asked to indicate how much mental effort they invested in solving all of the problems (overall mental effort rating), using the same 9-point scale.

2.1.3. Procedure

Participants were given 1 min to solve each problem, and were asked to solve the problems mentally (i.e., not using their hands to count or paper and pencil to draw out the solution steps). The Qualtrics software was programmed to give a warning when the time allotted (1 min) for solving the problem was up. One example was shown at the beginning of the experiment to familiarize participants with the type of task they would be performing.

Table 1
Mean (SD) of performance scores (range 0–1) and effort ratings (range 1–9) per condition in Experiment 1.

	Negative feedback	No feedback	Positive feedback
	n = 50	n = 47	n = 53
Mean performance	0.31 (0.22)	0.32 (0.26)	0.29 (0.22)
Mean task-specific effort	6.56 (1.27)	5.96 (1.45)	5.64 (1.50)
Mean overall effort	7.20 (1.25)	6.81 (1.48)	6.13 (1.77)

Participants then solved five problems, with each problem being immediately followed by feedback (except for participants in the control condition), after which participants rated how much mental effort they invested (task-specific mental effort rating). After all five problems, participants were asked to rate the mental effort they had invested in solving all problems (overall mental effort rating).

2.2. Results

Table 1 shows the mental effort and performance data for all conditions. Analyses were performed with analyses of variance (ANOVAs) and post hoc Tukey tests; the effect size reported for ANOVAs is partial eta-squared (η_p^2), for which 0.01 is considered to indicate a small, 0.06 a medium, and 0.14 a large effect and the effect size reported for Tukey tests is Cohen's d, for which 0.2 is considered to indicate a small, 0.5 a medium, and 0.8 a large effect (Cohen, 1988). Performance data are reported here for completeness (we did not have any hypotheses regarding performance). Mean performance on the problem-solving tasks was calculated (i.e., all scores were added and divided by the number of tasks; range of mean score = 0–1). As the tasks were complex, mean performance for all participants was low (M = 0.31, SD = 0.23; Table 1 shows the data per condition). To check for possible differences in performance between conditions an ANOVA was performed, which showed no significant differences, F (2, 147) = 0.19, p = .828, η_p^2 = 0.003.

2.2.1. Effects of feedback and timing of effort ratings

To test the effects of feedback valence on effort ratings (Hypothesis 1 and Question 1), as well as the effect of timing of the effort rating (Hypothesis 2), a repeated measures analysis was conducted, with within-subjects factor timing of effort rating (task-specific vs. overall) and between-subjects factor feedback condition (no feedback vs. Positive feedback vs. negative feedback). As can be seen in Table 1 (last rows), the overall mental effort rating after the series of five tasks, was higher than the average of task-specific mental effort ratings in each condition, in line with Hypothesis 2. The repeated measures ANOVA confirmed that this difference was significant, F (1, 147) = 110.59, p < .001, η_p^2 = 0.429. We also found a significant main effect of feedback condition, F (2, 147) = 6.34, p = .002, η_p^2 = 0.079, and no significant interaction effect of timing and feedback condition, F (2, 147) = 2.67, p = .072, η_p^2 = 0.035, meaning that the effect of feedback did not differ for task-specific and overall ratings (Question 1).[2] Tukey's post hoc tests on the main effect of condition, showed that participants in the positive

[2] In response to a reviewer comment asking whether age affected these results, we also ran the analysis with age as a covariate. The pattern of results did not change: a significant main effect of timing, F(1, 146) = 12.47, p = .001, η_p^2 = 0.079, and condition, F(2, 146) = 6.94, p = .001, η_p^2 = 0.087, but no significant interaction between condition and timing, F(2, 146) = 2.65, p = .074, η_p^2 = 0.035. There was no effect of age, F(1, 146) = 2.54, p = .113, η_p^2 = 0.017, and no collinearity was observed between age and the condition the participant was in (r = 0.10, p = .207).

feedback condition gave significantly lower effort ratings than participants in the negative feedback condition ($p = .001$, $d = 0.70$), however, neither the positive feedback condition ($p = .188$, $d = 0.35$) nor the negative feedback condition ($p = .200$, $d = 0.35$) showed significant differences relative to the no feedback condition.[3] To check if mental effort ratings increased linearly across conditions, linear contrasts were performed for both the average of task-specific effort ratings and for the overall effort ratings at the end. Both average task-specific, F (1, 147) = 10.90, $p = .001$, $\eta_p^2 = 0.069$, as well as overall effort ratings, F (1, 147) = 12.71, $p < .001$, $\eta_p^2 = 0.079$, showed a significant linear trend, indicating, in line with Hypothesis 1 (and informing on Question 1), that mental effort decreased linearly across conditions from negative feedback to positive feedback.

2.3. Discussion

Findings from Experiment 1 showed that participants' perceptions of invested mental effort were indeed influenced by feedback valence. In line with our hypothesis (Hypothesis 1), perceptions of invested effort were significantly higher after feedback that led students to believe their answers were predominantly incorrect (negative valence) than after feedback that their answers were predominantly correct (positive valence). Effort perceptions in the control condition that did not receive any feedback fell in between (not significantly different from the feedback conditions), and there was a significant linear trend from negative feedback (highest effort) to no feedback, to positive feedback (lowest effort).

We also replicated findings from prior studies (Schmeck et al., 2015; Van Gog et al., 2012) that the overall effort rating provided after all five problems, was significantly higher than the average of the task-specific effort ratings provided immediately after each problem (Hypothesis 2). Finally, it was an open question (Question 1) whether both types of effort ratings (overall and task-specific) would be influenced by feedback valence in the same manner, and our findings suggest that this was the case.

In Experiment 2, we aimed to replicate these findings from Experiment 1, and additionally explored the effects of the timing of the feedback valence reversal. That is, in Experiment 1, the feedback valence manipulation was reversed on the second task in order for the manipulation to remain credible (e.g., students who received predominantly positive feedback, got negative feedback on the second task). However, it is possible that the timing of this reversal might have an effect on mental effort ratings. With a valence reversal on the second task, participants' do not get a chance to build a consistent expectancy of their performance (or appraisal of the task) based on the feedback early on in the process, as the feedback valence changes from the first to the second, and from the second to the third task. As such, the effects of feedback valence on perceptions of invested effort might be even more pronounced, when participants do initially develop performance expectations.

For instance, Bandura (1977) stated that "After strong efficacy expectations are developed through repeated success, the negative impact of occasional failures is likely to be reduced." (p. 195). If this is true, then a feedback reversal on the fourth task in a predominantly positive feedback valence condition (PPPNP) should have less of an impact than a feedback reversal on the second task (PNPPP) would have. In other words, the effects of predominantly positive feedback on perceptions of invested effort (i.e., lower) might be stronger when the feedback valence reversal occurs later on. Although Bandura (1977) only mentioned repeated success, we speculate that a similar process is possible for repeated failure, with positive feedback after repeated failure being more likely to be ascribed to chance (NNNPN), than positive feedback early on, which might provide confidence that success is possible (NPNNN). So again, the effects of negative feedback valence on perceptions of invested effort (i.e., higher) might be stronger when the reversal occurs later. Therefore, we explored in Experiment 2 (Question 2) whether the effects of predominantly positive feedback or predominantly negative feedback on perceptions of invested effort would become even more pronounced when there is a late reversal (fourth task) than when there is an early reversal (second task).

3. Experiment 2

3.1. Method

3.1.1. Participants and design

A total of 347 participants from the United States were recruited via Amazon's Mechanical Turk (http://www.mturk.com). As in Experiment 1, only participants with a good reputation on Mechanical Turk (approval rate of above 95%) were recruited, and they were compensated with $0.50 (for ca. 8 min. Participation time on average). Fifty participants did not finish the session and were removed. No participants had to be removed for repeated participation (according to a check on IP addresses, which were inspected for duplicates). The remaining 297 participants had a mean age of 36.14 years (SD = 11.72; range: 18–76); 54% had completed a bachelor's degree or higher, and 148 were male. All participants performed above chance on the attention check.

Participants had been randomly assigned to one of five conditions, in which they performed 5 tasks with either: 1) no feedback (n = 60), 2) predominantly positive feedback with reversal on second task (PNPPP; n = 57), 3) predominantly positive feedback with reversal on fourth task (PPPNP; n = 62), 4) predominantly negative feedback with reversal on second task (NPNNN; n = 63), and 5) predominantly negative feedback with reversal on fourth task (NNNPN; n = 55). Again, none of the participants mentioned any awareness of the feedback manipulation after the experiment.

3.1.2. Materials & procedure

Materials and procedure were similar to Experiment 1.

3.2. Results

Table 2 shows the mental effort and performance data for all conditions. Again, performance data are reported for completeness. Similar to Experiment 1, mean performance for all participants was low (M = 0.28, SD = 0.28; Table 2 shows data per condition). To check for possible differences in performance between conditions an ANOVA was performed, which showed no significant differences, F (4, 292) = 0.601, $p = .662$, $\eta_p^2 = 0.008$.

3.2.1. Effects of feedback and timing of effort ratings

To test our hypotheses regarding the effects of feedback valence on effort ratings (Hypothesis 1 and 3), as well as the effect of timing of the effort rating (Hypothesis 2), a repeated measures ANOVA was conducted, with within-subjects factor timing of effort rating (average across five task-specific ratings vs. overall rating after the five tasks) and between-subjects factor feedback condition

[3] Note that this analysis involved both the task-specific and overall ratings. If we only look at average task-specific effort ratings for the last three problems (i.e., less affected by the feedback reversal), we see that the mean ratings are in the predicted direction: positive feedback (M = 5.47, SD = 1.76), no feedback (M = 6.24, SD = 1.59), and negative feedback (M = 6.59, SD = 1.41), and significantly different, F(2, 147) = 6.682, $p = .002$, $\eta_p^2 = 0.083$. Tukey's post-hoc tests showed positive feedback condition < no ($p = .043$, $d = 0.46$) and < negative feedback condition ($p = .001$, $d = 0.70$). No significant difference between negative and no feedback condition ($p = .537$, $d = 0.23$). Significant linear contrast ($p < .001$, $\eta_p^2 = 0.082$).

Table 2
Mean (SD) of performance scores (range 0–1)and effort ratings (range 1–9) per condition in Experiment 2.

	Neg. FB rev. on 2nd	Neg. FB rev. on 4th	No FB	Pos. FB rev. on 2nd	Pos. FB rev. on 4th
	$n = 63$	$n = 55$	$n = 60$	$n = 57$	$n = 62$
Mean performance	0.28 (0.26)	0.32 (0.30)	0.28 (0.27)	0.24 (0.27)	0.29 (0.31)
Mean task-specific effort	7.12 (1.10)	7.02 (1.11)	6.78 (1.60)	6.45 (1.54)	6.18 (1.51)
Mean overall effort	7.54 (1.22)	7.38 (1.15)	7.28 (1.68)	7.02 (1.51)	6.77 (1.48)

Note: FB = feedback; Neg. = negative valence; Pos. = positive valence; rev. on 2nd = feedback valence reversal on the second task; rev. on 4th = feedback valence reversal on the fourth task.

(collapsed for positive and negative, i.e., no feedback [condition 1] vs. Positive feedback [condition 2 & 3] vs. negative feedback [condition 4 & 5]). As can be seen in Table 2 (last rows), the overall mental effort rating after the series of five tasks, was higher than the average of task-specific mental effort ratings in each condition, in line with Hypothesis 2. The repeated measures ANOVA confirmed that this difference was significant, $F(1, 294) = 117.71$, $p < .001$, $\eta_p^2 = 0.286$. We also found a significant main effect of feedback condition, $F(2, 294) = 7.35$, $p < .001$, $\eta_p^2 = 0.048$, but no significant interaction effect of timing and feedback condition, $F(2, 294) = 1.84$, $p = .160$, $\eta_p^2 = 0.012$.[4] Tukey's post hoc tests on the main effect of condition showed that participants in the positive feedback condition gave significantly lower effort ratings than participants in the negative feedback condition ($p < .001$, $d = 0.49$). Similar to the results of Experiment 1, the difference between the positive and the no feedback condition was not significant ($p = .111$, $d = 0.32$), nor was the difference between the negative and no feedback condition ($p = .512$, $d = 0.17$).[5] To check if mental effort ratings increased linearly across conditions, linear contrasts were performed for both the average of task-specific effort ratings and for the overall effort ratings at the end. Both average task-specific, $F(1, 292) = 10.90$, $p = .001$, $\eta_p^2 = 0.035$, as well as overall effort ratings, $F(1, 292) = 7.15$, $p = .008$, $\eta_p^2 = 0.024$, showed a significant linear trend, indicating, in line with Hypotheses 1 and 3, that mental effort decreased linearly across conditions from negative feedback to positive feedback.

3.2.2. Exploration of effect of reversed feedback

In order to explore whether the timing of the feedback reversal had an effect on the average of task-specific effort ratings (Question 2), a 2 × 2 ANOVA was conducted (on conditions 2–5; as there was, of course, no feedback reversal present in control condition). As expected given the analyses in the previous section, there was a main effect of feedback valence, $F(1, 233) = 18.81$, $p < .001$, $\eta_p^2 = 0.075$, indicating that negative feedback resulted in significantly higher effort ratings than positive feedback (see Table 2). However, there was no main effect of feedback reversal, $F(1, 233) = 1.11$, $p = .294$, $\eta_p^2 = 0.005$, nor an interaction of feedback valence and reversal, $F(1, 233) = 0.23$, $p = .632$, $\eta_p^2 = 0.001$.

3.3. Discussion

In Experiment 2, we replicated the findings from Experiment 1. Again, perceptions of invested effort were significantly higher after feedback that led students to believe their answers were predominantly incorrect (negative valence) than after feedback that their answers were predominantly correct (positive valence), and this applied to both the average of task-specific effort ratings (Hypothesis 1) and the overall effort ratings (Hypothesis 3 in Experiment 2, this was Question 1 in Experiment 1). Perceived effort of participants in the control condition, who did not receive any feedback, fell in between (not significantly different from the feedback conditions), and there was a significant linear trend from negative feedback (highest effort) to no feedback, to positive feedback (lowest effort). We also replicated the finding from prior studies (Schmeck et al., 2015; Van Gog et al., 2012) and from Experiment 1, that the overall effort rating provided after all five problems, was significantly higher than the average of the task-specific effort ratings provided immediately after each problem (Hypothesis 2).

With regard to the open question (Question 2), we did not find any evidence that the moment of feedback reversal would affect perceptions of invested effort. If such a reversal occurs later on in the series of problems, then students have been able to build some expectation of the task and their performance on it, compared to when the reversal takes place in the beginning, which might lead to stronger effects of the feedback valence on the effort ratings. However, even though the means in Table 2 were in the expected direction for the positive valence conditions, we did not find any evidence that the average of the task-specific effort ratings was *significantly* lower (or higher, in case of negative valence) when the reversal occurred towards the end (fourth task) than at the beginning (second task).

In sum, Experiment 2 showed that the findings from Experiment 1 were replicable. The fact that these experiments were run on Mechanical Turk also means that these findings hold in a rather heterogeneous participant population. Nevertheless, one could argue that the task we used (i.e., the complex weekday problems on which it would be hard for participants to estimate whether their performance was correct or not) is not that educationally relevant, and one might question whether these findings would apply to a student population. Therefore, the purpose of Experiment 3 was to *conceptually* replicate the effect of feedback valence on effort ratings found in Experiments 1 and 2, using more educationally relevant problem-solving tasks and a different participant population of higher education students.

4. Experiment 3

To conceptually replicate the findings from Experiments 1 and 2,

[4] Again, age did not affect the results: a significant main effect of timing, $F(1, 293) = 16.35$, $p < .001$, $\eta_p^2 = 0.053$, and condition, $F(2, 293) = 7.95$, $p < .001$, $\eta_p^2 = 0.051$, but no interaction between condition and timing, $F(2, 293) = 1.91$, $p = .149$, $\eta_p^2 = 0.013$. There was no effect of age, $F(2, 293) = 3.72$, $p = .055$, $\eta_p^2 = 0.013$, and no collinearity was observed between age and the condition the participant was in ($r = 0.05$, $p = .426$).

[5] Note that this analysis of the main effect involved both task-specific and overall ratings. If we only look at effort invested in the first three problems in conditions 1, 3, and 5 (i.e., not in any way affected by the feedback reversal), we see the same pattern of results as in Experiment 1 on the last 3 tasks (see Footnote 2). Mean ratings are in the predicted direction: positive feedback ($M = 6.13$, $SD = 1.42$), no feedback ($M = 6.76$, $SD = 1.47$), and negative feedback ($M = 7.04$, $SD = 1.41$), and significantly different, $F(2, 174) = 6.125$, $p = .003$, $\eta_p^2 = 0.066$. Tukey's post-hoc tests showed: positive feedback condition < no ($p = .047$, $d = 0.44$) and < negative feedback condition ($p = .002$, $d = 0.64$). No significant difference between negative and no feedback condition ($p = .548$, $d = 0.19$). Significant linear contrast ($p = .018$, $\eta_p^2 = 0.031$).

we used a different type of problem-solving task in Experiment 3, in the domain of biology. We used monohybrid cross problems that are commonly used in teaching genetics, which could be solved in five steps. If the feedback valence manipulation would also affect mental perceptions on these learning tasks, this would not only increase the generalizability of our findings, but also their relevance for contexts in educational research and educational practice in which effort is measured.

Because a correct/incorrect manipulation of feedback valence would not be believable with these five-step problems, feedback valence was manipulated by first asking participants how many steps they felt they had performed correctly (self-assessment on a scale of 0–5) and then providing either negative feedback (indicating their performance was 2 points lower than they expected, bounded by the minimum score of 0) or positive feedback (indicating their performance was 2 points higher than they expected, bounded by the maximum score of 5).

Based on the findings from Experiments 1 and 2, we expected that negative feedback valence would result in significantly higher ratings of invested mental effort than positive feedback valence, and that there would be a significant linear trend, with perceptions of invested effort in the no feedback control condition falling in between the positive and negative feedback valence conditions (Hypothesis 1). We also expect the overall effort ratings to be higher than the average of task-specific ratings (Hypothesis 2), and that the effect of feedback valence would not differ significantly for task-specific and overall effort ratings (Hypothesis 3).

4.1. Method

4.1.1. Participants and design

Participants in Experiment 3 were 66 Dutch higher education students with a mean age of 19.70 years ($SD = 2.72$; range = 17–34; 39 male; university: $n = 20$, university of applied sciences: $n = 46$). They participated either for course credit or for a monetary reward of 5 Euro. Participants were randomly assigned to one of three conditions: 1) no feedback ($n = 23$), 2) feedback with positive valence (higher score than expected; $n = 23$), or 3) feedback with negative valence (lower score than expected; $n = 20$).

4.1.2. Materials

The materials were presented in a dedicated online learning environment developed for this experiment. Responses to all questions were required and were logged in the environment.

4.1.2.1. Genetics problem-solving tasks. The five problem-solving tasks (cf. Corbalan et al., 2009; Kostons et al., 2012) were in the domain of biology (Mendel's law of heredity) and consisted of five distinct steps (for an example, see the Appendix): (1) translating the information given in the cover story into genotypes, (2) putting this information into 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). Tasks differed in complexity by varying the number of generations, the number of unknowns, the possibility of multiple answers, and the type of reasoning used to solve the problem. This resulted in five levels of complexity (cf. Kostons et al., 2012). Performance on the problem-solving tasks was scored by assigning one point for each correct step (i.e., range per task: 0–5 points).

4.1.2.2. Self-assessment. Participants were asked to assess their own performance on a 6-point rating scale, by indicating how many steps of a problem-solving task they thought they had performed correctly, ranging from (0) no steps correct to (5) all steps correct. This self-assessment served as input for the feedback valence manipulation.

4.1.2.3. Feedback. Feedback consisted of a text message appearing on the screen after the task and self-assessment was completed. The no feedback control condition simply received the message "Your answers have been registered.". In the positive feedback valence condition, the message stated "You performed better than you expected, you performed *n* steps correctly." (where *n* was self-assessed performance + 2, bounded by the maximum score). In the negative feedback valence condition, it said "You performed worse than you expected, you performed *n* steps correctly." (where *n* was self-assessed performance - 2, bounded by the minimum score).

4.1.2.4. Mental effort rating. The task-specific and overall mental effort ratings were identical to those used in Experiments 1 and 2.

4.1.3. Procedure

Participants were provided a login code that took them to one of the three conditions in the online learning environment. In all conditions, participants were first shown an introductory video in which the correct definitions of terms were explained, along with the steps of the problem-solving procedure (without applying it to a specific problem). Then, participants observed a video modeling example, in which the problem-solving procedure was demonstrated and explained by a female model. Participants then went on to solve the five problems, which were presented in the same order for all participants and were of increasing complexity. After each task, participants were asked to self-assess their performance, received feedback (depending on assigned condition), and rated how much mental effort they invested in solving the problem (task-specific mental effort rating). After all five problems were completed, participants were asked to rate how much mental effort they had invested in solving all problems (overall mental effort rating). When the experiment was over, participants were asked if they could guess the purpose of the experiment. None of the participants indicated awareness of the feedback manipulation. In total, the experiment lasted about 40–45 min.

4.2. Results

Table 3 shows the mental effort, performance, and self-assessment data for all conditions. Data on performance and self-assessment are reported here for completeness (we did not have any hypotheses regarding performance or self-assessment). Mean performance on the problem-solving tasks (range = 0–5) was fair ($M = 3.34$, $SD = 1.00$) and did not differ significantly between conditions, $F(2, 63) = 0.25$, $p = .783$, $\eta_p^2 = 0.008$. Mean self-assessment scores did differ between conditions, $F(2, 63) = 3.95$, $p = .024$, $\eta_p^2 = 0.112$; the data from Table 3 suggest that the participants in the negative feedback valence condition gave lower self-assessment ratings than participants in the other two conditions, which is likely a consequence of the feedback consistently informing them that they scored two points lower than they

Table 3
Mean (SD) of performance scores (range 0–5), self-assessment scores (range 0–5), and effort ratings (range 1–9) per condition in Experiment 3.

	Negative feedback	No feedback	Positive feedback
	$n = 20$	$n = 23$	$n = 23$
Mean performance	3.34 (0.86)	3.44 (1.08)	3.23 (1.06)
Mean self-assessment	2.91 (1.38)	3.73 (0.99)	3.84 (1.14)
Mean task-specific effort	4.76 (1.89)	3.62 (1.54)	3.44 (1.33)
Mean overall effort	5.10 (2.29)	3.96 (1.75)	3.74 (1.84)

expected (interestingly, positive feedback did not seem to increase self-assessments compared to no feedback).

4.2.1. Effects of feedback and timing of effort rating

To test the effects of feedback valence on effort ratings, as well as the effect of timing of the effort rating, a repeated measures analysis was conducted, with within-subjects factor timing of effort rating (average across five task-specific ratings vs. overall rating after the five tasks) and between-subjects factor feedback condition (negative feedback vs. no feedback vs. Positive feedback). As can be seen in Table 3 (last rows), the overall mental effort rating after the series of five tasks was higher than the average of task-specific mental effort ratings in each condition, and the repeated measures ANOVA confirmed that this difference was significant, $F(1, 63) = 6.98$, $p = .010$, $\eta_p^2 = 0.100$. We also found a significant main effect of feedback condition, $F(2, 63) = 3.74$, $p = .029$, $\eta_p^2 = 0.106$, but no significant interaction effect of timing and feedback condition, $F(2, 63) = 0.02$, $p = .986$, $\eta_p^2 < 0.001$. Tukey's post-hoc tests on the main effect of condition showed that, similar to Experiments 1 and 2, participants in the positive feedback valence condition gave significantly lower mental effort ratings than participants in the negative feedback condition ($p = .034$, $d = 0.78$), whereas the no feedback condition fell in between and did not differ significantly from the positive ($p = .921$, $d = 0.11$), or negative ($p = .082$, $d = 0.67$) feedback valence condition. And as in Experiments 1 and 2, both the average of task-specific effort ratings, $F(1, 63) = 7.33$, $p = .009$, $\eta_p^2 = 0.103$, as well as overall effort ratings, $F(1, 63) = 5.18$, $p = .026$, $\eta_p^2 = 0.075$, showed a significant linear trend, indicating, in line with Hypotheses 1 and 3, that mental effort decreased linearly across conditions from negative feedback to positive feedback.

4.3. Discussion

In Experiment 3, we managed to conceptually replicate our findings from Experiments 1 and 2. In the next section, we will discuss the findings in more detail, and address the theoretical and practical implications.

5. General discussion

All three experiments showed significant linear trends indicating that perceptions of effort invested in the very same problem-solving tasks, were lowest when participants received positive feedback and highest when they received negative feedback. The mental effort ratings following positive and negative feedback differed significantly and substantially from each other (in Experiments 1 and 2, mental effort ratings in the negative feedback valence condition were almost one scale point higher than in the positive feedback valence condition). Effort ratings when no feedback was given, fell in between, and did not differ significantly from effort ratings following positive or negative feedback. However, in Experiment 1 and 2, where effort ratings were relatively high, there seemed to be a trend that the positive feedback condition reported having invested less effort than the no feedback condition, and this was significant when looking only at a series of 3 consecutive tasks unaffected by the feedback reversal (footnote 2 and 3). In Experiment 3, in contrast, where effort ratings were overall lower, there was hardly any difference between the positive feedback and the no feedback condition, but there seemed to be a trend in the opposite direction, with negative feedback leading to numerically higher effort ratings than no feedback (although this was not statistically significant, $p = .082$, it was a medium-sized effect).

There are at least two possible explanations for this difference in patterns of results between Experiments 1 and 2 and Experiment 3. First, because of the high complexity of the problem-solving tasks in Experiments 1 and 2, these required a high level of effort investment and therefore, there may have been no room for effort ratings to *significantly* increase compared to no feedback. Vice versa, the low complexity of the problem-solving tasks in Experiment 3 might not have allowed for effort ratings to *significantly* decrease compared to no feedback, as some effort always needs to be invested to accommodate the task demands. The linear trend shows, though, that they did numerically increase and decrease as hypothesized. Second, as will be discussed in more detail in the next section, task complexity might perhaps be a moderator of the effect of feedback on perceptions of invested effort, possibly via participants' emotional response.

5.1. Limitations and future research

This study has several limitations, which all relate to the fact that we have established that feedback valence affects perceptions of invested effort, but not the mechanisms through which this occurs. First, it is possible that students may have started investing more or less effort on *subsequent* tasks due to the feedback, in which case our results may be due to differences in actual effort investment rather than to differences in perceptions of invested effort. Our data provide no indications that this was the case; for instance, if participants actually invested more effort on a subsequent task, one would expect this to result in higher performance (especially in Experiment 3), but there were no performance differences between conditions. Nevertheless, we cannot rule out this possibility, and future research should attempt to gain more insight into the potential effect of feedback on effort invested in subsequent tasks, for instance, by including a condition in which feedback is given after the effort rating is requested, or by using objective measures (e.g., physiological, or dual task measures; Brünken et al., 2003) of effort investment during a subsequent task.

A second limitation of our study is that we did not measure participants' emotional response to the feedback. We assumed that negative feedback would result in negative emotions and positive feedback in positive emotions. However, this is a rather crude dichotomy, as positive and negative emotions can be both activating and deactivating (e.g., Pekrun et al., 2007, pp. 13–36). Moreover, these emotional responses to the feedback might differ with the complexity of the tasks. To speculate, when task complexity is high, (as in Experiments 1 and 2), positive feedback might induce relief (a positive deactivating outcome-focussed emotion) that is less likely to occur with low complexity tasks (Pekrun, Goetz, Titz, & Perry, 2002, pp. 91–105). This deactivating emotion might perhaps explain why there was a trend towards lower (perceptions of) effort investment in the positive than the no feedback condition (which was significant when looking only at the tasks unaffected by the feedback reversal) in Experiments 1 and 2 that was not present in Experiment 3 (although, as mentioned above, it might also merely reflect a bottom effect in Experiment 3). In contrast, an activating negative emotion like anger in response to negative feedback may not result in an increase in (perceptions of) effort investment on high complexity tasks, as additional effort is unlikely to make a difference, but on lower complexity tasks, increasing effort in response to negative feedback is more likely to pay off. This might explain why there was no difference between negative and no feedback in Experiments 1 and 2, while there was a trend towards higher effort in the negative compared to the no feedback condition (a medium-sized though not significant effect) in Experiment 3 (although, as mentioned above, this might also merely reflect a ceiling effect in Experiments 1 and 2). In sum, it would be interesting in future research on this topic to systematically vary task complexity within one experiment and to measure participants' emotional response to the feedback, distinguishing between

activating and deactivating positive and negative emotions, as these factors might explain effects of feedback on perceived effort invested in the task (or actual effort investment in a subsequent task).

A third, related, limitation is that our findings suggest that feedback valence affects the cues that learners use when rating invested effort, which is an interesting and important first step. However, the present study did not measure *which* cues learners use. Speculating, it seems likely that feedback alters participants' task perceptions, presumably – at least partly – via affective responses as discussed above (see also Efklides, 2008), and that this influences their perception of how much effort they invested in the task (or their actual effort investment in a subsequent task). Future research should start investigating this question of cue use in rating effort investment, and research on the cue-utilization view of metacognitive judgments (Koriat, 1997) may provide appropriate paradigms for doing so. Such studies on cue use could also point towards other external influences (e.g., mood; task design) that might affect perceptions of invested effort.

Finally, a potential limitation of our study is that we cannot rule out that other types of feedback might have had different effects; for instance, it might be worthwhile to investigate in future research whether including a form of social comparison in the feedback (e.g., relative performance feedback compared to fellow students or a norm group) would amplify affective responses (cf. Tesser, Millar, & Moore, 1988) and (thereby) affect perceptions of invested effort more strongly.

5.2. Theoretical and practical implications

The fact that we attempted and managed to directly replicate the findings from Experiment 1 in Experiment 2, and then to conceptually replicate them in Experiment 3, is considered important for educational research (Bauernfeind, 1968; Makel & Plucker, 2014), and strengthens the theoretical as well as practical relevance of this study. As for theoretical implications, our study provides a first step towards addressing the cues used by participants when rating how much effort they invested in a task. Our findings make clear that perceptions of how much effort was invested in a task, rely not only on people's direct experience with the task during performance, but can be influenced by external factors *after the fact*. As mentioned above, future research should more directly investigate the mechanisms through which this occurs in order to help researchers make informed choices on when to apply subjective rating scales.

Regarding practical implications, it is imperative that effort measures obtained in the context of research, or as a central part of task selection in adaptive or self-regulated learning environments, reliably reflect experienced cognitive load. Our results suggest that this may no longer be the case when feedback is provided before an effort rating is obtained. Although, our experiments mainly show a difference between effects of positive and negative feedback on effort ratings, there are some patterns in the data to suggest differences with the no feedback condition may arise under certain conditions. For instance, in Experiments 1 and 2 there is a difference between the positive and no feedback conditions if we look at the series of three tasks in a row that are unaffected by the feedback reversal. In Experiment 3, the difference between the no feedback and negative feedback condition was not statistically significant ($p = .082$) but did have a medium effect size. As such, it seems recommendable to measure effort directly after each task, to rule out possible external influences (e.g., of feedback valence) on retrospective effort ratings (future research should establish whether feedback valence also affects ratings on subsequent tasks, though, as measuring directly after each task will not take care of this problem in that case). Delaying effort ratings until after all tasks have been completed, does not resolve the problem that the perception of invested effort is affected by feedback, and would introduce a second problem, namely that it does not reliably reflect the average load experienced throughout the series of tasks. That is, we replicated and extended earlier findings (Schmeck et al., 2015; Van Gog et al., 2012), showing that overall mental effort ratings after a series of tasks were higher than the average of task-specific mental effort ratings.

5.3. Conclusion

To conclude, we found evidence across three experiments—-with different problem-solving tasks and participant populations—that perceptions of effort investment were affected by feedback valence. The mechanisms through which this occurs are still unclear, but might involve post-hoc changes in task perceptions (possibly via affective responses). Future research should follow-up on these findings to address those mechanisms by uncovering what cues learners use when judging how much effort they invested in a task. Although further research is needed, our findings suggest that it may be advisable in studies or learning environments in which mental effort is measured as an indicator of cognitive load, to assess mental effort investment directly after each task, and before giving feedback on performance.

Acknowledgments

This research was funded by the Netherlands Initiative for Education Research (NRO PROO; project number: 411-12-015). The authors would like to thank the Erasmus Behavioral Lab for facilitating these experiments, Guus van Schaik for facilitating Experiment 3, and the Instructional Design Pubgroup at Utrecht University for their helpful comments on a draft of this manuscript.

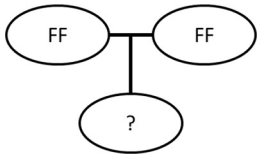
Appendix A. Example of problem-solving task (at the lowest level of complexity) used in Experiment 3.

Fur color

A guinea pig's fur color is determined by a gene that 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 the 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 needed, 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. 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

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. http://dx.doi.org/10.1016/j.learninstruc.2014.04.004.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215. http://dx.doi.org/10.1037/0033-295X.84.2.191.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238. http://dx.doi.org/10.3102/00346543061002213.

Bauernfeind, R. H. (1968). The need for replication in educational research. *The Phi Delta Kappan, 50*, 126–128.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*, 610–632. http://dx.doi.org/10.1016/0749-596X(89)90016-8.

Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology, 40*, 109–131. http://dx.doi.org/10.1146/annurev.ps.40.020189.000545.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*, 53–61. http://dx.doi.org/10.1207/S15326985EP3801_7.

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin, 92*, 111–135. http://dx.doi.org/10.1037/0033-2909.92.1.111.

Carver, C. S., & Scheier, M. F. (2000). On the structure of behavioural self-regulation. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 41–84). San Diego: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2009). Dynamix task selection: Effects of feedback and learner control on efficiency and motivation. *Learning and Instruction, 19*, 455–465. http://dx.doi.org/10.1016/j.learninstruc.2008.07.002.

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review, 1*, 3–14. http://dx.doi.org/10.1016/j.edurev.2005.11.001.

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*, 277–287. http://dx.doi.org/10.1027/1016-9040.13.4.277.

Efklides, A. (2014). How does metacognition contribute to the regulation of learning? An integrative approach. *Psychological Topics, 23*, 1–30.

Efklides, A., & Dina, F. (2004). Feedback from one's self and others: Their effect on affect. *Hellenic Journal of Psychology, 1*, 179–202.

Gendolla, G. H., & Silvestrini, N. (2011). Smiles make it easier and so do frowns: Masked affective stimuli influence mental effort. *Emotion, 11*, 320–328. http://dx.doi.org/10.1037/a0022593.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. http://dx.doi.org/10.3102/00346543029848.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: Mturk participants perform better online attention check than do subject pool participants. *Behavior Research Methods, 48*, 400–407. http://dx.doi.org/10.3758/s13428-015-0578-z.

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist, 45*, 1–14. http://dx.doi.org/10.1080/00461520903213618.

Hoogerheide, V., Loyens, S. M. M., & Van Gog, T. (2014). Comparing the effects of worked examples and modeling examples on learning. *Computers in Human Behavior, 41*, 80–91. http://dx.doi.org/10.1016/j.chb.2014.09.013.

Kahneman, D. (1973). *Attention and effort*. New Jersey: Prentice-Hall.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. http://dx.doi.org/10.1037/0033-2909.119.2.254.

Knörzer, L., Brünken, R., & Park, B. (2016). Facilitators or suppressors: Effects of experimentally induced emotions on multimedia learning. *Learning and Instruction, 44*, 97–107. http://dx.doi.org/10.1016/j.learninstruc.2016.04.002.

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. http://dx.doi.org/10.1037/0096-3445.126.4.349.

Koriat, A. (2015). Metacognition: Decision making processes in self-monitoring and self-regulation. In G. Keren, & G. Wu (Eds.), *The Wiley blackwell handbook of judgment and decision making* (pp. 356–379). Chichester, UK: John Wiley & Sons, Ltd.

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The memorizing effort heuristic in judgments of learning: A developmental perspective. *Journal of Experimental Child Psychology, 102*, 265–279. http://dx.doi.org/10.1016/j.jecp.2008.10.005.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*, 478–492. http://dx.doi.org/10.1016/j.jml.2005.01.001.

Koriat, A., Nussinson, R., & Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1624–1637. http://dx.doi.org/10.1037/xlm0000009.

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. http://dx.doi.org/10.1016/j.learninstruc.2011.08.004.

Larson, J. R., & Callahan, C. (1990). Performance monitoring: How it affects work productivity. *Journal of Applied Psychology, 75*, 530–538. http://dx.doi.org/10.1037/0021-9010.75.5.530.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. New Jersey: Prentice-Hall.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty replication in the education sciences. *Educational Researcher, 43*, 30–316. http://dx.doi.org/10.3102/0013189X14545513.

Meshkati, N. (1988). Toward development of a cohesive model of workload. *Advances in Psychology, 52*, 305–314. http://dx.doi.org/10.1016/S0166-4115(08)62394-8.

Nievelstein, F., Van Gog, T., Van Dijck, G., & Boshuizen, H. P. (2013). The worked example and expertise reversal effect in less structured tasks: Learning to reason about legal cases. *Contemporary Educational Psychology, 38*, 118–125. http://dx.doi.org/10.1016/j.cedpsych.2012.12.004.

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. http://dx.doi.org/10.1037/0022-0663.84.4.429.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63–71. http://dx.doi.org/10.1207/S15326985EP3801_8.

Paas, F., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental-effort and performance measures. *Human Factors, 35*, 737–743. http://dx.doi.org/10.1177/001872089303500412.

Paas, F., Van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*, 419–430. http://dx.doi.org/10.2466/pms.1994.79.1.419.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315–341. http://dx.doi.org/10.1007/s10648-006-9029-9.

Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. P. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz, & R. Pekrun (Eds.), *Emotion in education*. Amsterdam, The Netherlands: Academic Press.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist* (Vol. 37). http://dx.doi.org/10.1207/S15326985EP3702_4.

Plass, J. L., Heidig, S., Hayward, E. O., Homer, B. D., & Um, E. (2014). Emotional design in multimedia learning: Effects of shape and color on affect and learning. *Learning and Instruction, 29*, 128–140. http://dx.doi.org/10.1016/j.learninstruc.2013.02.006.

Salden, R. J. C. M., Paas, F., Broers, N. J., & Van Merriënboer, J. J. G. (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instructional Science, 32*, 153–172. http://dx.doi.org/10.1023/B. TRUC.0000021814.03996.ff.

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science, 43*, 93–114. http://dx.doi.org/10.1007/s11251-014-9328-3.

Sweller, J. (1993). Some cognitive processes and their consequences for the organization and presentation of information. *Australian Journal of Psychology, 45*, 1–8. http://dx.doi.org/10.1080/00049539308259112.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

Tesser, A., Millar, M., & Moore, J. (1988). Some affective consequences of social comparison and reflection processes: The pain and pleasure of being close. *Journal of Personality and Social Psychology, 54*, 49–61. http://dx.doi.org/10.1037/0022-3514.54.1.49.

Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology, 26*, 833–839. http://dx.doi.org/10.1002/acp.2883.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26. http://dx.doi.org/10.1080/00461520701756248.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes a meta-analysis. *Review of Educational Research, 85*, 475–511. http://dx.doi.org/10.3102/0034654314564881.