Effects of Feedback on Self-Evaluations and Self-Regulation in Elementary School

van Loon, M. H., Roebers, C. M.

(2017)

Applied Cognitive Psychology, 31, 508-519.

DOI: 10.1002/acp.3347

Abstract

Elementary school learners are typically highly confident when judging accuracy of their test responses, relatively independent of whether these are correct. While feedback has been shown to improve accuracy of adults' and adolescents' self-evaluations and subsequent self-regulation, little is known about beneficial effects for elementary school children. We investigated effects of fine-grained feedback on 4th and 6th graders self-evaluations and restudy selections by presenting them the ideas they were meant to bring up in their test responses. One group received full-definition feedback standards, whereas the other group received idea-unit feedback standards. The two types of feedback strongly improved 4th and 6th graders' self-evaluations for commission errors and for partially correct responses. While restudy selections before feedback were more adaptive for 6th than 4th graders, age differences disappeared after receiving feedback. Findings imply that feedback standards are a suitable tool to calibrate elementary school learners and to support effective self-regulation.

Keywords: Children; Development; Feedback; Self-Evaluations; Self-Regulation

Effects of Feedback on Self-Evaluations and Self-Regulation in Elementary School

Children, as well as adults, need to self-evaluate learning in order to identify discrepancies between what has already been learned and what is yet to be learned (Dunlosky & Rawson, 2012). Based on self-evaluations, learners can decide how to allocate their study time (Metcalfe & Finn, 2013). For adults as well as for children, a strong relation has been found between accurate self-evaluations, adaptive regulation of learning, and academic achievement (Dunlosky & Rawson, 2012; Rinne & Mazzocco, 2014; Thiede, Anderson, & Therriault, 2003). However, research shows that the vast majority of children in elementary school still lack the metacognitive skills necessary to accurately evaluate their performance. Hence, most children are overconfident (Bol & Hacker, 2001; Schneider & Löffer, 2016), and this is often resulting in maladaptive regulation and low performance (Bol, Hacker, O'Shea, & Allen, 2005). In the present approach, we attempted to increase the accuracy of elementary school children`s self-evaluations.

Out of the different means to assess learners` accuracy of their self-evaluations, item-specific Self-Score Judgments (SSJs) seem to suit our purposes best. SSJs give detailed insights into students' self-assessment of their knowledge and their discrimination ability, and can provide item-specific information about how evaluations are related to item-specific restudy decisions (Nietfeld, Cao, & Osborne, 2005). When investigating item-specific SSJs, miscalibration (i.e., biased self-evaluations) most often appears in terms of overconfidence for incorrect (commission errors) and incomplete (partially correct) test responses (Lipko, Dunlosky, Hartwig, Rawson, Swan, & Cook, 2009). Such overconfidence for SSJs has been explained by the accessibility theory (Koriat, 1993). That is, students often base evaluations on the accessibility of information they were able to retrieve from memory, while neglecting to accurately evaluate its quality. Consequently, students may give themselves credit for any

test response, regardless of whether or not the information they provided is objectively correct and complete.

Due to overconfidence, incorrectly and incompletely learned items are typically not selected for further study (Lockl & Schneider, 2002; Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013). To improve efficient self-regulation, it is thus important that children are supported to accurately self-evaluate their learning.

Improving self-evaluations and self-regulation through feedback

A possible way to help children improve the accuracy of their self-evaluations is to provide feedback; which is among the most critical influences on students' learning process (Hattie & Timperley, 2007). Effective feedback illustrates details about the standards of a fully correct response, in order to support students to reduce discrepancies between current self-evaluations and actual performance (Hattie &Timperley, 2007). Importantly, item-based feedback seems more beneficial than global feedback to improve self-evaluations, regulation, and academic achievement (Miller & Geraci, 2011). The learning process and outcomes are not easily improved by mere outcome feedback, probably because it does not give concrete insights into students' actual learning progress in relation to standards. To make students aware of their learning progress, they rather need fine-grained, detailed feedback. That is, feedback should address performance on individual items and precisely elaborate on reasons why performance is correct or incorrect, rather than just addressing correctness (Miller & Geraci, 2011; Nietfeld, Cao, & Osborne, 2006; Renner & Renner, 2001; Van der Kleij et al., 2015). Item-specific feedback has been demonstrated to improve adolescents and adults' self-evaluations and self-regulation (Callender, Franco-Watkins, & Roberts, 2015; Labuhn, Zimmerman, & Hasselhorn, 2010; Miller & Geraci, 2011). When feedback on responses to test items is fine-grained and sufficiently elaborates on the relation between actual performance and learning standards, learners are supported to identify and correct errors, to

self-regulate learning more effectively, and thus to improve learning outcomes (Butler & Winne, 1995; Kitsantas & Zimmerman, 2006; Van der Kleij, Feskens, and Eggen, 2015). Also for primary school children, item-based feedback on test responses is beneficial for their performance (Lipko-Speed, Dunlosky, & Rawson, 2014). The one study that provided primary school children (5th graders) with item-specific feedback showed that this improved achievement (Lipko-Speed et al., 2014). However, it is unclear to what extent feedback can also improve children's self-evaluations and self-regulation.

Although young learners would need detailed feedback, teachers in typical classroom setting usually do not have the time and resources to give their students item-specific feedback. Hence, they mainly provide learners with global outcome feedback (e.g., a grade). One time-effective, yet understudied technique to provide students in the classroom with fine-grained feedback is giving feedback standards that show the correct answer for each test question (Rawson & Dunlosky, 2007). Feedback standards have beneficial effects on the accuracy of adults' (Rawson & Dunlosky, 2007) as well as adolescents' self-evaluations (Lipko, Dunlosky, Hartwig, et al., 2009). In studies investigating effects of feedback standards, participants were asked to learn definitions of difficult concepts. After studying, they were first asked to complete an open-ended test of concept understanding, and then to self-score the quality of their test-responses by indicating whether these were incorrect, partially correct, or fully correct. Based on random allocation, half of the students were assigned to a feedback standard group, while the other half did not receive feedback. For the feedback group, the feedback standard showed the correct definition of the concept for every studied and tested concept. Feedback standards improved self-evaluations; that is, the external standards helped students to compare their own responses to the correct definition and better judge the quality of their recall (Lipko, Dunlosky, Hartwig, et al., 2009; Rawson & Dunlosky, 2007). The standards were especially suitable to counteract overconfidence for commission

errors. In fact, investigations of adults' overconfidence for commission errors show that the degree of overconfidence of SSJs was 55% when using no feedback, and only 25% when using feedback standards showing the full definition (Rawson & Dunlosky, 2007). A similar pattern was found for the adolescents (13-year-olds), who's degree of overconfidence was 54% for commission errors without feedback, and only 34% with full definition standards (Lipko, Dunlosky, Hartwig, et al., 2009).

Follow-up research (Dunlosky, Hartwig, Rawson, & Lipko, 2010) further addressed how the format of the feedback standard can be most beneficial to improve learners' SSJs. As outlined above, full definition feedback reduced overconfidence in adults and adolescents in comparison to no feedback (Lipko, Dunlosky, Hartwig, et al., 2009; Rawson & Dunlosky, 2007). However, some overconfidence remained, possibly because full definition standards do not provide sufficient feedback concerning the specific ideas required in a test response. Fine-grained idea-unit standards that highlight the specific ideas by parsing the definition into smaller idea units, may be more suitable to reduce overconfidence (Dunlosky et al., 2010; Lipko, Dunlosky, Hartwig, et al., 2009). Experimental studies comparing effects of full-definition and idea-unit standards show that idea-unit standards are most beneficial to improve self-evaluations for adults (Dunlosky et al., 2010) as well as for adolescents (Lipko, Dunlosky, Hartwig, et al., 2009). After receiving idea-unit feedback, overconfidence rates reduced to 15-22% for adults and 28% for adolescents, respectively (Dunlosky et al., 2010). Similarly, for partially correct responses, overconfidence reduced to 20% for both adults (Dunlosky et al., 2010) and adolescents (Lipko, Dunlosky, Hartwig, et al., 2009).

However, the few studies addressing effects of feedback standards have mainly been conducted with adults and adolescents. The one study that provided primary school children (5^{th} graders) with feedback standards showed that feedback improved achievement, but not children's self-regulated learning (Lipko-Speed et al., 2014). Presumably, the children only

processed the feedback superficially, because they were not stimulated to do so in a more active manner. Even though a key issue in children's education is ensuring that feedback is actively processed and targeted at the appropriate level, little is known about how feedback affects self-evaluations and self-regulation in elementary school children, and also to what extent beneficial effects of feedback depend on learners' age.

Children's Use of Feedback

As outlined, it is yet unclear under which circumstances feedback can lead to improvements in children's item-specific self-evaluations and the differential regulation of their study time. Around the age of 6 to 8 years, children increasingly acquire insights into their own learning and remembering; they begin to differentiate between items or questions for which they are able and for which they are unable to provide a response (Destan, Hembacher, Roebers, & Ghetti, 2014; Roebers, Moga, & Schneider, 2001). Children's overconfidence declines slowly but steadily, that is, over the course of elementary school, children's self-evaluations of their responses (e.g., in tests) become more accurate and realistic (Krebs & Roebers, 2010; for a review on metacognitive development see Roebers, 2014). Studies using tasks in which children have to learn and remember textual information, do not show strong developmental differences in self-evaluation accuracy in the late elementary school years (Flavell, Friedrichs, & Hoyt, 1970; Schneider & Löffer, 2016).

Moreover, it is in the late elementary school years (4th to 6th grade) that children also become capable of regulating their learning. That is, from about an age of 9/10 years on, children can differentially withdraw previously given responses when they are unsure about the correctness of that response (Krebs & Roebers, 2012). Because self-regulatory skills are still developing in the late elementary school years, learners become increasingly able to regulate their learning (Schneider & Löffer, 2016). While adult students' study choices are directly influenced by their confidence about what they believe they know (Metcalfe & Finn,

2008), for young children this relation between self-evaluations (i.e., self-monitoring) and self-regulation (i.e., control) is not yet pronounced (Roebers, Krebs, & Roderer, 2014). Rather, children seem to have difficulty in interpreting their awareness of learning deficits as a need for increased restudy (Cao & Nietfeld, 2007). During elementary school years, children begin to base study decisions more strongly on their previous self-evaluations (Metcalfe & Finn, 2013)

The few existing studies that confronted children of various ages with feedback on performance revealed pronounced developmental progression in the ability to incorporate feedback. While preschoolers and first graders have strong difficulties in using feedback at all (Lipko, Dunlosky, & Merriman, 2009), older elementary school students seem to start implementing feedback (Labuhn et al., 2010). However, even when presented with feedback, children in the later elementary school years still appear to be at least partially deficient in detecting and correcting errors (Hacker, 1997, Lipko-Speed et al., 2014; Salles, Ais, Semelman, Sigman, & Calero, 2016).

Two separate, not mutually exclusive explanations have been suggested for children's overconfidence despite feedback: Sensitivity to a wishful thinking bias and limitations in cognitive resources. Firstly, children have the desire to perform well, and base their self-evaluations on their desires (i.e., their wishful thinking) rather than on their actual performance (Lipowski, Merriman, & Dunlosky, 2013; Schneider, 1998). Because they see negative feedback as self-threatening information, they may process it in a shallow manner and recall negative feedback less than positive feedback (Sedikides & Green, 2009). As a consequence, children may fail to learn from their errors. Although children as well as adults are sensitive to the wishful thinking bias, this bias is much stronger in children than in adults (Bjorklund & Green, 1992). Therefore, it may be more difficult for children to learn from

feedback on errors as they may remain overly optimistic even after receiving negative feedback (Salles et al., 2016).

Secondly, feedback standards may not have similar effects in children as in adolescents and adults because of limitations in children's cognitive resources (Case & Griffin, 1990; Krebs & Roebers, 2012). Self-evaluating performance with use of feedback standards may be demanding for Working Memory Capacity (WMC). The working memory system processes and maintains information (Baddeley, 2003), but its capacity is limited and fatigue-sensitive. WMC predicts performance in a wide range of tasks, ranging from reading (Daneman & Carpenter, 1980) and cognitive reasoning (Kane, Hambrick, Tuholski, Wilhelm, Payne, & Engle, 2004) to making accurate self-evaluations (Lewandowsky, 2011). In elementary school years, WMC is still developing. Hence, older elementary school children typically show better operational efficiency, a higher processing speed, and they can more effectively deal with cognitive load compared to their younger peers (Bayliss, Jarrold, Gunn, & Baddeley, 2003; Case, Kurland, & Goldberg, 1982; Fry & Hale, 2000). Younger children's limitations in processing capacity make it more difficult to compare own responses with feedback standards. However, comparing a given test response to an externally provided standard, one idea at a time, may reduce cognitive load and may be easier to process than more complex full definition standards (Leahy & Sweller, 2011). For this reason, we presume that especially for children, idea-unit feedback may be more suitable than full-definition feedback.

An important novelty of the present research is that we do not only investigate self-evaluations, but also subsequent restudy selections. Research with adults shows that feedback has beneficial effects on the self-regulated learning process. Feedback can indicate to learners that the studied information is yet inadequately understood. When these insights are linked to subsequent study strategies, students' altered self-evaluations improve subsequent regulatory

decisions and effective task persistence (Butler & Winne, 1995; Van der Kleij et al., 2015). Research with adolescents shows that feedback standards supported identification of items that were not yet fully learned and improved subsequent restudy selections for these items (Baars, Vink, Van Gog, De Bruin, & Paas, 2014). However, for elementary school children it is yet unclear whether and how feedback benefits self-regulation, and how this is affected by children's age. Research by Lipko-Speed et al. (2014) showed that the beneficial effect of feedback on 5th graders' regulation was very limited. However, the children were not asked to actively process the feedback; possibly, self-evaluating each test item by comparing responses to standards may support active processing and make feedback more beneficial for children's regulation.

Research showing developmental differences in effective regulation of learning in late elementary school years suggests that the control component (e.g., restudy selections, allocation of study time) of self-regulated learning develops later than the self-evaluation component (Lockl & Schneider, 2002; Schneider & Löffer, 2016). Therefore, feedback may have similar effects on self-evaluations for both 4th and 6th graders, but when selecting items for further study, younger children are presumably less effective in translating their acquired insights in performance into effective regulation.

Present Study

With this study, we investigate effects of two different types of feedback (i.e., full definition and idea-unit feedback) on 4th and 6th grade children's self-evaluations and restudy selections when studying concepts. We aimed to give children the most supportive circumstances to make self-evaluations as accurate as possible. This was done by asking them to make SSJs and restudy selections after test-taking. Self-evaluating and regulating learning after test-taking allows children to refer to their experience with the task. Hence, postdictive self-evaluations are typically more accurate than predictive self-evaluations that are made

before test-taking (Maki, 1998; Pieschl, 2009). Further, because judgment format can affect accuracy of self-scoring (Lipowski et al., 2013; Metcalfe & Finn, 2013), we attempted to support accurate self-scoring by making the self-scoring scale comparable to objective scoring. That is, children were asked to score the number of correct ideas in their test responses.

Children made self-evaluations and restudy selections, and then received feedback (either full definition or idea-unit feedback). Following the feedback, the children had to once again make self-evaluations and restudy selections. We assume that feedback standards improve SSJs and global performance judgments.

In the present approach, both item-specific and global judgments were elicited. It is clear that there are conceptual and methodological differences between self-evaluation types (SSJs and GJs) and judgment timing (before and after feedback). SSJs are made per item; a representation of the given response on each item is available to the child when indicating the number of correct ideas in the response. When making GJs, participants self-evaluate their overall number of correct responses; the evaluation is based on the aggregate likelihood of correct performance without having direct access to the test responses. In most studies, either one of these measures is used, and little is known about the consistency of self-evaluations across measures. Research with adults assumes that different self-evaluation measures may be correlated, indicating consistency in self-evaluations (Kleitman & Stankov, 2007). In this study, we exploratively investigate the relation between children's SSJs and GJs, and in addition, we address the consistency of SSJs and GJs, both before and after receiving feedback.

Specifically, the aim of this research was to answer the following three questions: 1) How well can children evaluate the quality of their responses, and is judgment accuracy improved by feedback? We hypothesize that children are overconfident, especially for

commission errors and partially correct responses due to the accessibility of information, but that feedback reduces overconfidence. Moreover, we predict that overconfidence will be lower after receiving idea-unit feedback than after receiving full definition feedback because of the resource-demanding nature of the full definition feedback.

2) To what extent are improvements in self-evaluations translated into improved regulation? We hypothesize that restudy selections are improved and more focused on correction of erroneous and partially correct responses after children have received feedback than before, and that idea-unit feedback leads to better restudy selections than full-definition feedback. 3) Do developmental differences affect accuracy of self-evaluations, restudy selections, and effects of feedback? Although we do not expect differences between age groups in accuracy of SSJs, we hypothesize that restudy selections are more closely linked to self-evaluations in 6th than in 4th graders, both before and after receiving feedback.

Methods

Participants

The sample consisted of 100 children; 49 4th grade children (*M* age = 10.09 years, *SD* = 6.4 months, 25 females) and 51 6th grade children (*M* age = 11.89 years, *SD* = 6.0 months, 28 females). Participants were recruited from three different primary schools in the German speaking part of Switzerland and they possessed sufficient German language skills to understand and complete the study tasks. Informed consent from parents/caretakers was acquired prior to testing.

Materials

The materials consisted of 16 concepts for 4th graders and 18 concepts for 6th graders. Selection of these concepts was based on children`s curriculum, and on a pilot study in which children in the same age groups participated (these were 31 4th graders and 21 6th graders). For this pilot study, both age groups completed a pretest, a study phase, and a posttest for 24

concepts. Selection of these concepts was based on teaching materials and the learning objectives outlined in the curriculum description for teachers. The pretest of the pilot study showed that 4th graders had 4.3% prior knowledge for the tested materials; the 6th graders had 10% correct prior knowledge. The posttest pilot results showed that 4th graders had 18.7% correct test performance, 6th graders` test performance was 33.3%. Based on the pretest and posttest results in this pilot study, concepts were selected which were unknown for most of the children before learning; that is, the pretest indicated that children had a low level of prior knowledge. Further, concepts were selected that were likely to have a higher posttest score than a pretest score, indicating that the concepts were not too difficult for these age groups, and that children had the possibility to at least partially learn these in the study phase.

The concept task consisted of 10 phases:

1) Pretest. This pretest consisted of the to-be-studied concepts and empty lines on which children could write their responses.

2) Study phase, for which concepts were presented with its definition given in a sentence; each definition consisted of three idea units. Further, each concept was presented in an example sentence, to clarify use of the concept.

3) Concept test. Similar to the pretest, the studied concepts were presented with empty lines on which children could write down the meaning of the concepts.

4) Item-specific self-evaluations pre-feedback. For these self-score judgments (SSJs), a judgment scale with four squares was printed next to the lines on which children could write test responses. These squares ranged from 0 – 3 points, children could mark one of the four squares to indicated how many idea units they thought they had correct.

5) Global self-evaluation pre-feedback. The global judgments (GJs) were made on a horizontal continuous line with end points labelled "nothing" and "everything". Children

could indicate how many of their test responses they thought to be correct, by marking the corresponding point of their choice.

6) Restudy selections pre-feedback. When making restudy selections, the concepts were presented on one page in two columns, children could mark the concepts they liked to restudy.

7) Feedback phase. Children were randomly assigned to the full-definition feedback standards group or the idea-unit standard feedback group. The full-definition feedback group received feedback containing the concepts and the definitions as given when studying, and there was no separation of idea units. Similarly, the idea-unit feedback contained the concepts and the definitions as given in the definition when studying, but, in addition to the full-definition feedback, the three idea units in the meaning were separated. Similar to when making SSJs before feedback, they could mark one of the four squares indicating the number of correct idea units, ranging from 0 – 3.

8) Item-specific self-evaluations post-feedback. Similar as when making SSJs pre-feedback, these SSJs were made on a scale ranging from 0 – 3 points, children could mark one of the four squares to indicated how many idea units they thought they had correct

9) Global self-evaluation post-feedback. Similar as when making pre-feedback GJs, after receiving the feedback, children could indicate how many of their test responses they thought to be correct. These post-feedback GJs were made by marking the corresponding point on a line with end points labelled "nothing" and "everything".

10) Restudy selections post-feedback. Similar to when making pre-feedback restudy selections, in this phase, restudy selections were made by marking concepts for restudy.

There were three different versions of the concept task for each age group, with the order of items being different. Further, in each subtask, the order of the items was changed. All self-evaluative judgments were postdictions and were delayed until after completion of

the full test. These design decisions were made to provide children with the most beneficial circumstances to make judgments; postdictions are typically more accurate than predictions of performance (Hacker, Bol, Horgan, & Rakow, 2000), and delaying judgments until after completing the test leads to higher judgment accuracy than immediately making judgments when studying (Rhodes & Tauber, 2011).

Procedure

Prior to starting, children were shown the folders and the content, they were instructed that they would study concepts with their definitions. They were shown that each concept definition consisted of 3 separate idea units, and instructed that they should try to remember all three ideas for the test. Children then received a folder with all the subtasks, and were instructed to start with the pretest. Children were told to write down the meaning if they thought they knew it, and leave the lines blank if they did not know the concept. Then, the study phase started. Study was self-paced, but after 12 minutes, the experimenter told the children to continue with the next task. After studying, children completed the posttest. After completing the posttest, children were asked to give item-specific SSJ for each of the concepts, indicating how many points (0-3) they thought they would receive for each individual item. Then children were asked to make a GJ, indicating how many test responses they thought would be correct. After making the GJ children made restudy selections; they were instructed to mark the concepts they would like to study again if they had a chance.

Then, the feedback phase started, children received feedback on a separate piece of paper. Half of the children received full-definition feedback standards, the other group received idea-unit standard feedback. Children were instructed to inspect the feedback and use it to again self-score their test responses by making item-specific SSJs. Then, children were again to make a global evaluation judgment (GJ), and finally, the children were again asked to indicate which items they would like to select for restudy.

Scoring

Each definition of a studied concept consisted of 3 idea units. Pretest and posttest responses were scored as omission (no response given), commission error (an entirely incorrect response given, response does not contain any idea unit of the definition of the concept), one idea correct (response contains one idea unit of the studied definition of the concept), two idea units (containing two idea units of the definition of the concept) and three idea units (response contains all correct idea units of the studied definition). In line with the scoring used by Rawson and Dunlosky (2007), ideas were scored correct when either provided verbatim or as a paraphrase of the original idea unit. Two independent raters scored 37.5% of the pre- and posttest responses and showed good inter-rater reliability, Cohen's Kappa = .73 for the pretest and .75 for the posttest responses. Disagreements were solved through discussion; the scores of the first rater were used for the analyses and the remaining responses were scored by a single rater.

Analyses

In our analyses, we first present the overall calibration of the self-evaluations, by comparing subjective evaluations with actual performance scores (cf. Schraw, 2009). Learners are accurately calibrated when the deviation between subjective and objective performance scores is low; values above 0 indicate that learners were overconfident, values below 0 show underconfidence. Then, to further investigate children's ability to discriminate between different test response types (omissions, commission errors, and one, two, or three ideas correct), we report children's raw SSJs on item-level. Against the background of the accessibility theory outlined above, we had specific hypotheses about overconfidence for commission errors and partially correct responses; therefore, we analyzed SSJs and restudy selections for incorrect responses (commission errors) and responses for which only one idea was correct (partially correct responses) with a General Linear Model (GLM) for repeated

measures. This 2x2x2x2 GLM for repeated measures contained two within-subject factors as repeated measures, namely: 1) Timing of SSJs and Restudy Selections (Pre- and Post-Feedback) and 2) the test response type (commission errors and one-idea responses). Further, the GLM contained two between-subject factors: 1) Grade Level (Grade 4 vs. Grade 6) and 2) Feedback Type (Full Definition Feedback vs. Idea-Unit Feedback). Interaction effects between within-subject factors were followed-up with repeated measure GLMs for each response type separately; interaction effects between within- and between subject factors were followed up with MANOVAs.

Further, to investigate the relation between self-evaluations and self-regulation, gamma correlations were calculated between SSJs and restudy selections (cf. Thiede & Dunlosky, 1999). A negative gamma correlation of -1 shows that items for which SSJs were low were consistently more often selected than items for which SSJs were high; a gamma of 0 shows no relation between SSJs and restudy selections. For all significant effects, η_p^2 is reported to give an indication of the effect size.

Results

In this section, we present analyses to investigate the hypotheses about effects of feedback and grade on children's SSJs, GJs, and restudy selections. Firstly, we present descriptive statistics on pretest and posttest performance. Children's prior knowledge was low, 4th graders knew 5.71% of the ideas of the concepts (*SD* = 4.65), 6th graders knew 5.67% of the concepts, (*SD* = 4.4), and there was no difference between the age groups, $t(98) = .036$, $p = .97$. Test performance was 37.79% of idea units correct for the 4th graders (*SD* = 16.47) and 41.23% (*SD* = 17.16) for 6th graders; there was no significant difference between groups, $t(98) = 1.02$, $p = .31$. The finding that age groups did not differ in percentage of prior knowledge and test performance indicates that the difficulty level of the tasks was held comparable between the two age groups, as intended. Table 1 shows the test performance for

the different test response types (omission, commission error, one idea correct, two ideas correct, and three ideas correct) for the two age groups under investigation.

Item-Specific Self-Evaluations (SSJs)

Table 1 shows the raw SSJs for the different test response types. As visible in the Table, the SSJs became more accurate after receiving feedback for all response types. An analysis of the aggregated difference between raw SSJs and mean performance scores confirms this observation: Overall calibration accuracy, indicated by the mean deviation between SSJs and objective performance accuracy, was affected by SSJ Timing, $F(1, 99) = 15.18$, $p < .001$, $\eta_p^2 = .14$. That is, calibration for Mean SSJs was more accurate after feedback (2.7% overall overconfidence) than before feedback (6.9 % overconfidence). Note that even before receiving feedback, overall calibration of SSJs was accurate and only slightly overconfident, however, feedback even improved overall calibration of SSJs.

As expected and visible in Table 1, children were mainly overconfident for their commission errors and their responses for which they had only one idea unit correct. This overconfidence is indicated by the finding that their SSJs were higher than the objective number of idea units that were present in their test responses. Figure 1 shows the SSJs for commission errors and one-idea-unit responses before and after receiving feedback for the two grade levels. A 2x2x2x2 GLM for repeated measures for SSJ Timing (pre-and post-feedback) and Test Response Type (commission errors and one idea correct) as within-subject factors and Grade Level and Feedback Type as between-subject factors shows that SSJs for commission errors and one-idea-unit responses were lower after than before receiving feedback, $F(1, 77) = 39.41$, $p < .001$, $\eta_p^2 = .34$. The significant effect of Response Type $F(1, 77) = 15.43$, $p < .001$, $\eta_p^2 = .44$, shows that raw SSJs were lower for commission errors than for one-idea-unit responses, indicating that children discriminated in their judgments between these test responses. Notably, there was a significant interaction effect between SSJ Timing

and Response Type, $F(1, 77) = 12.26$, $p = .001$, $\eta_p^2 = .14$. A follow-up repeated measures GLM for both test response types separately shows a significant decrease in SSJs for commission errors after feedback, $F(1, 79) = 44.69$, $p < .001$, $\eta_p^2 = .36$. As well, SSJs for one-idea-unit responses significantly decreased, $F(1, 94) = 22.05$, $p < .001$, $\eta_p^2 = .19$. However, the interaction effect is stronger for the commission errors than for one-idea-unit responses, indicating that the feedback had stronger effects on lowering of SSJs for commission errors compared to partially correct responses.

There was no significant main effect of Grade ($p = .22$) and Feedback Type ($p = .95$). However, there was a significant interaction effect between Response Type and Grade, $F(1, 77) = 11.56$, $p = .001$, $\eta_p^2 = .13$. A follow-up MANOVA shows that there was no difference between 4th and 6th graders in SSJs for commission errors before and after receiving feedback (both ps > .65). Interestingly, as visible in Figure 1, 6th graders gave higher SSJs for one-idea-unit responses than 4th graders, both before, $F(1, 79) = 1.99$, $p = .022$, $\eta_p^2 = .06$, and after receiving feedback, $F(1, 79) = 6.17$, $p = .015$, $\eta_p^2 = .07$.

Global Judgments

Figure 2 shows the mean GJs for the 4th and 6th graders; mean GJs that were made before receiving feedback were 64.26% (*SD* = 24.1) for 4th graders and 61.17% (*SD* = 21.5) for 6th graders; mean GJs after receiving feedback were 54.39 (*SD* = 25.3) for 4th graders and 57.43% (*SD* = 18.3) for 6th graders. After receiving feedback, children's global judgments were significantly lower and more accurate than judgments before receiving feedback, $F(1, 96) = 17.97$, $p < .001$, $\eta_p^2 = .15$. There was no significant effect of Grade ($p = .95$) and Feedback Type ($p = .54$).

Consistency of Self-Evaluative Judgments

Since in the present approach, a number of different judgments were gathered, an interesting question concerns the consistency of participants' judgments across the course of

the experiment. The mean SSJs before receiving feedback were strongly related to the mean SSJs made after receiving feedback, Pearson's $r = .81$, $p < .001$. Further, the Pearson correlation between the GJs made before feedback and GJs made after receiving feedback was $r = .74$, $p < .001$, indicating a strong intra-individual consistency across self-evaluations made at different time points. Moreover, the GJs before receiving feedback were strongly related to the mean SSJs that were made before children received feedback, $r = .64$, $p < .001$, and as well, the mean SSJs and GJs that were made after feedback were significantly related, $r = .55$, $p < .001$.

Restudy Selections

Table 1 shows the percentage of restudy selections for the different test response types. First of all, there was no effect of Timing on the overall percentage of restudy selections ($p = .69$); before receiving feedback children decided to restudy 34.97% (SD = 16.56) of the concepts, after receiving feedback children decided to restudy 34.56% (SD = 17.56). Feedback type did not have any effect on the overall restudy percentage ($p = .653$), however, the effect of Grade approached significance, $F(1, 96) = 3.88$, $p = .052$, $\eta_p^2 = .04$. A MANOVA shows that before receiving feedback, there was a near-significant difference between grades, such that 4th graders decided to restudy more concepts ($M = 38.26\%$, SD = 17.52) than 6th graders ($M = 31.80\%$, $SE = 15.07$), $F(1, 98) = 3.91$, $p = .051$, $\eta_p^2 = .04$. After receiving feedback, there was no significant difference between grades in percentage of restudy selections (4th graders $M = 37.42\%$, $SE = 19.66$; 6th graders $M = 31.81\%$, $SE = 14.95$), $p = .11$.

Intra-individual gamma correlations between SSJs and Restudy Selections show that these were strongly related to each other, such that items for which children gave higher SSJs were less often selected for restudy. Gamma between pre-feedback SSJs and restudy selections was -.67 (SD = .44) for 4th graders and -.87 (SD = .16) for 6th graders; Gamma

between post-feedback SSJs and restudy selections was -.81 (*SD* = .29) for 4th graders and -.74 (*SD* = .35) for 6th graders. A repeated measures GLM with Timing (pre- and post-feedback correlations) as repeated measurement, and Grade and Feedback as between-subject factors did not show differences in correlations as an effect of Timing ($p = .79$), and there was no main effect of Grade ($p = .20$) and Feedback Type ($p = .87$). However, the interaction between Timing and Grade was significant, $F(1, 96) = 13.59$, $p < .001$, $\eta_p^2 = .12$. A MANOVA shows a significant effect of Grade on pre-feedback Gamma correlations between SSJs and restudy; 6th graders were substantially better able to relate restudy to their SSJs than 4th graders, $F(1, 98) = 9.78$, $p = .002$. After receiving feedback this age difference had vanished, there was no significant difference in gamma correlations between 4th and 6th graders anymore, $p = .29$.

GLM analyses for repeated measures were conducted to investigate restudy selections for commission errors and one-idea-unit responses. There was a main effect of Response Type, $F(1, 77) = 26.80$, $p < .001$, $\eta_p^2 = .26$, showing that children more often decided to select commission errors for restudy (49%) than one-idea-unit responses (27.5%). There was no significant effect of Timing ($p = .07$), although a trend shows that numerically, more commission errors and one-idea-unit responses were selected after feedback (40.9%) than before feedback (35.6%). There were no main effects of Feedback Type ($p = .52$) and Grade ($p = .47$), however, there was a significant interaction effect between Response Type and Grade, $F(1, 77) = 7.02$, $p = .010$, $\eta_p^2 = .08$. This interaction effect is shown in Figure 3. A follow-up MANOVA shows that, before receiving feedback, fourth graders less often selected their commission errors for further study than 6th graders, $F(1, 79) = 6.07$, $p = .016$, $\eta_p^2 = .07$. However, there was no difference between grades in restudy selections for commission errors after receiving feedback, $p = .38$. Moreover, there was no difference between 4th and 6th

graders in selection of one-idea-unit responses for restudy, neither before (p = .29) nor after (p = .41) receiving feedback.

Discussion

With this study, we investigated whether and to what extent feedback is beneficial for 4th and 6th grade school children to improve their self-evaluations and subsequent self-regulation. Children learned definitions of difficult concepts and then took a test. After completing the test, they were asked to self-evaluate their performance and select concepts for restudy. Then, they received feedback through a comparison standard. Thereby, two types of feedback were used. That is, half of the children received full definition feedback, while the other half received idea-unit based feedback for which the different idea units of the full definition were separated. After receiving feedback, children again self-evaluated their performance and made restudy selections.

Our first question addressed how well children can evaluate the quality of their test performance, and whether feedback improved self-evaluations. Children made item-specific self-evaluations (SSJs) for each response, and as well, they judged with global self-evaluations (GJs) how many responses they expected to be correct. As hypothesized, children were overconfident when self-scoring their performance; their SSJs and their GJs were too optimistic in comparison to their objective performance. Analyses of the item-specific self-evaluations showed that children were especially overconfident for their commission errors and the partially correct responses.

Importantly, we found that feedback reduced overconfidence. That is, global as well as item-specific self-evaluations became more accurate and less overconfident when children could compare their test responses with a feedback standard. Although the children were still overconfident for their commission errors, the degree of overconfidence was substantially reduced after receiving feedback. These findings indicate that both age groups were equally

overconfident for the incorrect test responses, and similarly benefitted from the feedback. With tasks asking children to memorize and retrieve information, research suggests that self-evaluation accuracy does not necessarily further improve in the later elementary school years (Schneider & Löffer, 2016). Our concept learning task is a memory-based task, and children are required to retrieve the idea-units of the definition when taking the test. Our lack of evidence for further improvement of self-evaluation ability when monitoring commission errors between 10 and 12 years of age confirms previous research.

Also for partially correct responses, overconfidence was reduced and self-evaluations became more accurate after children received feedback. A surprising finding is that 4th graders were less overconfident than 6th graders when self-evaluating partially correct responses, both before and after receiving feedback. Past research does not clearly show why evaluations for partially correct responses were better for 4th than for 6th graders. Processing fluency may be one reason for this finding. On average, 6th graders have higher cognitive capacity and faster processing speed than 4th graders, and these differences may have contributed – at least in part – to stronger subjective experiences of fluency. Possibly, for 6th graders it was less difficult to retrieve the studied idea units when giving partially correct responses. The ease of retrieving only partially correct ideas may have lead them to misinterpret how much of their response was truly correct (Finn & Tauber, 2015). Besides experiencing more fluent processing, older children may also be more sensitive to the experiences and feelings they perceive when completing tasks than younger children are. Research by Van Loon, De Bruin, Leppink, and Roebers (2017) shows that older elementary school children strongly rely on fluency experiences when making self-evaluative judgments, whereas younger elementary school children may be less sensitive to these experiences. This may be a reason why the older children were even somewhat more overconfident for partially known information. Future research should further investigate whether the differential use of processing fluency cues

may explain developmental differences in overconfidence. However, even though children were somewhat overconfident, it has to be noted that, in comparison to middle school children and adults, children were still very accurate when monitoring their partially correct performance. We will return to this finding below.

There was a strong relation between self-evaluations made before and after feedback and between global and item-specific judgments. This seems to indicate that judgments were consistent and that effects of feedback were similar for global and item-specific self-evaluations. Interestingly, although we expected idea-unit feedback standards to be more beneficial for self-evaluation accuracy than full-definition feedback standards, there were no differences between these two feedback types concerning reduction in overconfidence. We did not find an additional benefit for idea-unit feedback compared to full-definition feedback. In fact, both SSJs and GJs became more accurate, regardless of the type of feedback children received. This contrasts research with middle school and adult learners, showing that self-evaluations were more accurate after comparing one`s own performance with idea-unit standards than with full-definition standards (Dunlosky et al., 2010; Lipko, Dunlosky, Hartwig, et al., 2009). One possible interpretation is that the overall effect of feedback was strong and that over and above that general effect, the two different kinds of feedback did not yield to differential effects.

Another important but yet surprising observation is that for both types of feedback, children were very well able to implement the received information and thus extensively improve their self-evaluations. Although children stayed somewhat overconfident after receiving feedback, especially for commission errors, their SSJs were more accurate than SSJs in research with adults and adolescents. When translating the SSJs to a percentage of overconfidence, and thus make them comparable to previous research, the degree of overconfidence for commission errors after receiving feedback was only 22% for 6th graders,

and 23% for 4th graders. Research with middle school students (Lipko, Dunlosky, Hartwig, et al., 2009) shows that adolescents are more overconfident for commission errors, both after receiving full-definition feedback and idea-unit feedback. The same is true for adults who seem more overconfident for commission errors than the children in our study, particularly when receiving full-definition feedback (Dunlosky et al., 2010). A similar pattern is found for the scoring of partially correct responses, where children seem more accurate than adolescents and adults in previous research. As for the present study, when receiving feedback, the degree of overconfidence for partially correct responses was only 5% for 4th graders and 13% for 6th graders. Adults and adolescents were much more overconfident when receiving feedback on partially correct responses (Dunlosky et al., 2010; Lipko, Dunlosky, Hartwig, et al., 2009).

Although we do not have a definite explanation why children in this study were better calibrated in comparison to previous research, we have some speculations. We used a self-evaluation scale as to let children exactly indicate the number of idea units that were in their test responses. In contrast, previous research on self-scoring concept learning used less fine-grained scoring scales. Further, participants in previous studies were not asked to exactly indicate the number of correct idea units, but were only asked to self-score whether responses were incorrect, partially correct, or fully correct. Using the midpoint of such a scale (indicating that a response is partially correct), may actually indicate that a learner is unsure about the given answer he or she has recalled, instead of indicating that only a part of the studied concept has been correctly recalled (Dunlosky, Serra, Matvey, & Rawson, 2005; Zamary, Rawson, & Dunlosky, 2016). In that case, use of the partially correct response option in previous research may have reflected participants' uncertainty in the quality of their response, and not that they actually believed that their response was partially correct. Our SSJ scale was designed to make the SSJ correspond to the number of idea units and this may have

facilitated accurate self-evaluations (Dunlosky, Mueller, & Thiede, 2016; Metcalfe & Finn, 2013).

Furthermore, children made the self-evaluative judgments under conditions that have been demonstrated to be beneficial for judgment accuracy. Firstly, the judgments were delayed until after studying, because delayed judgments are typically much more accurate than judgments made during study (Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011). Secondly, the children made postdictions after completing the test; they could thus base their judgments on their test-experience, a factor that seems beneficial for self-evaluation accuracy (Hacker, Bol, Horgan, & Rakow, 2000). Thirdly, the children had access to their responses when making the judgments. Because there was no need to retrieve the given test response, making judgments was not overly demanding in terms of working memory load. This in turn may have benefitted judgment accuracy (Baddeley, 2003).

It has been proposed that learners need practice with feedback and self-scoring, in order to effectively self-evaluate and self-regulate learning (Winne, 1997). Possibly, children's previous educational experiences with self-scoring may have supported them to accurately self-evaluate performance in this concept learning task. However, it has to be noted that results on effects of practice are not always indicating beneficial effects of experience on self-evaluations and self-regulation. For instance, Cao and Nietfeld (2007) conducted a study during an entire semester, for which they required college students to make confidence judgments. Students did not improve their study strategies, even after they had gathered experience with making judgments and when they received feedback about their performance. Future research should further address potential benefits of fine-grained scoring scales and effects of experience on self-scoring.

With our second question, we addressed to what extent the improvements in self-evaluations through feedback are reflected in children's regulation of learning. After self-

scoring, children indicated which concepts they would like to restudy. We hypothesized that feedback would improve restudy selections, such that the selections would focus more on correction of erroneous and partially correct responses after receiving feedback than before. Interestingly, although feedback had no overall effect on the number of items that were selected for restudy, restudy selections became more strategic after receiving feedback. That is, children more often decided to restudy their commission errors and their partially correct responses, rather than responses that were correct already.

Although research on learning only rarely reports developmental differences in self-evaluation accuracy in the late elementary school years, age differences are usually observed when investigating regulation of learning. That is, regulation is more effective for older children (Krebs & Roebers, 2012; Metcalfe & Finn, 2013). In line with literature on developmental differences in effectiveness of regulation, 6th graders made more adaptive restudy selections than 4th graders did before receiving feedback. More so than 4th graders, 6th graders strongly used their self-evaluations as input for their restudy selections. Especially when selecting commission errors for restudy, 6th graders were more effective before receiving feedback, and more often attempted to correct their commission errors through further study. Before receiving feedback, the 6th graders selected more than half of their commission errors for further study, whereas 4th graders only selected one-third of those errors. Importantly, feedback proved to be a powerful tool for the 4th graders to improve restudy selections. The relation between SSJs and restudy selections became stronger, indicating that regulation became more strategic because 4th graders more often selected the items for which they gave low SSJs for further study. Interestingly, although 6th graders showed better regulation of learning before feedback, there were no more differences between the two age groups after having received feedback. Even though 4th graders may not have the fully developed skills to effectively regulate learning without feedback, the feedback helped

them to become equally effective as the older 6th graders. These promising findings may indicate that detailed, item-level feedback on performance can facilitate metacognitive development. However, it has to be noted that these positive conclusions are based on investigations of the correlation between SSJs and restudy selections; findings on restudy selections look troubling when considering the actual number of selected items. Most of the concepts were dropped from further study; hence, children decided that they would rather not spend more study time on learning these items. Not only the half of the commission errors was not further selected, but also two-thirds of the partially known items were dropped from further study. Although children seemed to be able to identify incorrect and incomplete test responses with use of feedback, mostly they would not be selected, thus making it unlikely that these would be corrected through future study. This finding is problematic. It indicates that, in a self-regulated learning environment, most children's achievement would be suboptimal, because of ineffective regulation. Possibly, children were not motivated to further study these items. For effective self-regulation, accurate self-evaluations are necessary but not sufficient; in fact, the motivation to learn is just as important (Chatzistamatiou, Dermitzaki, Efklides, & Leondari, 2015). Future research should investigate how children can be supported to more effectively regulate learning for commission errors and items that are only partially learned, and investigate interventions that may not only improve self-evaluation accuracy, but also motivation.

In sum, detailed feedback improved children's self-evaluations of their performance. Restudy selections were closely linked to these self-evaluations, indicating that feedback enables children in the late elementary school years to self-regulate learning, and to use their self-evaluations as a basis for study decisions. Although children remained overconfident for commission errors and items that are only learned partially, both full-definition feedback and idea-unit feedback helped them to recognize and select them for further study. Feedback was

especially beneficial to improve 4th-graders' study selections. However, a problematic finding is that learners in both age groups did often not select incorrect and partially learned items for restudy; not even when the children were able to identify them. Research should investigate how learners can be supported and motivated to further study items, for which performance is incorrect or incomplete.

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.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature reviews neuroscience*, *4*(10), 829-839. doi:10.1038/nrn1201

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The complexities of complex span: explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology: General*, *132*(1), 71. doi: 10.1037/0096-3445.132.1.71

Bjorklund, D. F., & Green, B. L. (1992). The adaptive nature of cognitive immaturity. *American Psychologist*, *47*(1), 46. doi: 10.1037/0003-066X.47.1.46

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*(2), 133-151. doi: 10.1080/00220970109600653

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269-290. doi: 10.3200/jexe.73.4.269-290

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

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2015). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*(2), 215-235. doi: 10.1007/s11409-015-9142-6

Cao, L., & Nietfeld, J. L. (2007). College Students' Metacognitive Awareness of Difficulties in Learning the Class Content Does Not Automatically Lead to Adjustment of Study Strategies. *Australian Journal of Educational & Developmental Psychology*, *7*, 31-46.

Case, R., & Griffin, S. (1990). Child cognitive development: The role of central conceptual structures in the development of scientific and social thought. *Advances in Psychology*, *64*, 193-230. doi: 10.1016/S0166-4115(08)60099-0

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*(3), 386-404. doi: 10.1016/0022-0965(82)90054-6

Chatzistamatiou, M., Dermitzaki, I., Efklides, A., & Leondari, A. (2015). Motivational and affective determinants of self-regulatory strategy use in elementary school mathematics. *Educational Psychology*, *35*(7), 835-850.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450-466. doi: 10.1016/S0022-5371(80)90312-6

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5-to 7-year-old children. *Journal of Experimental Child Psychology*, *126*, 213-228. doi: 10.1016/j.jecp.2014.04.001

Destan, N., Spiess, M. A., de Bruin, A., van Loon, M., & Roebers, C. M. (2017). 6-and 8-year-olds' performance evaluations: Do they differ between self and unknown others?. *Metacognition and Learning*, 1-22. doi: 10.1007/s11409-017-9170-5

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*(3), 467-484. doi: 10.1080/17470218.2010.502239

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271-280. doi: 10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *The Journal of General Psychology, 132*(4), 335-346. doi: 10.3200/genp.132.4.335-346

Dunlosky, J., Mueller, M. L., & Thiede, K. W. Methodology for Investigating Human Metamemory. In J. Dunlosky & S.K. Tauber (Eds), *The Oxford Handbook of Metamemory* (pp. 23-38). Oxford, UK: Oxford University Press.

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*(4), 567-586. doi: 10.1007/s10648-015-9313-7

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive psychology*, *1*(4), 324-340. doi: 10.1016/0010-0285(70)90019-8

Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology*, *54*(1), 1-34. doi: 10.1016/S0301-0511(00)00051-X

Hacker, D. J. (1997). Comprehension monitoring of written discourse across early-to-middle adolescence. *Reading and Writing*, *9*(3), 207-240. doi: 10.1023/A:1007989901667

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160. doi: 10.1037/0022-0663.92.1.160

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi: 10.3102/003465430298487

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189. doi: 10.1037/0096-3445.133.2.189

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*(2), 161-173. doi: 10.1016/j.lindif.2007.03.004

Kitsantas, A., & Zimmerman, B. J. (2006). Enhancing self-regulation of practice: The influence of graphing and self-evaluative standards. *Metacognition and Learning, 1*(3), 201-212. doi: 10.1007/s11409-006-9000-7

Koriat, A. (1993). How do we know that we know - the accesibility model of the feeling of knowing. *Psychological Review, 100*(4), 609-639. doi: 10.1037/0033-295x.100.4.609

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology*, *80*(3), 325-340. doi: 10.1348/000709910X485719

Krebs, S. S., & Roebers, C. M. (2012). The impact of retrieval processes, age, general achievement level, and test scoring scheme for children's metacognitive monitoring and controlling. *Metacognition and Learning*, *7*(2), 75-90. doi: 10.1007/s11409-011-9079-3

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*(2), 173-194. doi: 10.1007/s11409-010-9056-2

Lewandowsky, S. (2011). Working memory capacity and categorization: individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720. doi: 10.1037/a0022639

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*(4), 307-318. doi: 10.1037/a0017599

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology, 103*, 152-166. doi: 10.1016/j.jecp.2008.10.002

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science?. *Journal of Applied Research in Memory and Cognition*, *3*(3), 171-176. doi: 10.1016/j.jarmac.2014.04.002

Lipowski, S. L., Merriman, W. E., & Dunlosky, J. (2013). Preschoolers can make highly accurate judgments of learning. *Developmental Psychology, 49*(8), 1505-1516. doi: 10.1037/a0030614

Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgements. *International Journal of Behavioral Development*, *26*(4), 327-333.

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds), *Metacognition in Educational Theory and Practice* (pp. 117-144). New York: Routeledge

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174-179. doi: 10.3758/pbr.15.1.174

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and learning*, *8*(1), 19-46. doi: 10.1007/s11409-013-9094-7

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*(3), 303-314. doi: 10.1007/s11409-011-9083-7

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*(4), 267-270. doi: 10.1111/J.1467-9280.1991.Tb00147.X

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *Journal of Experimental Education, 74*(1), 7-28.

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*(2), 159. doi: 10.1007/s10409-006-9595-6

Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning*, *4*(1), 3-31. 10.1007/s11409-008-9030-4

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4-5), 559-579. doi: 10.1080/09541440701326022

Renner, C. H., & Renner, M. J. (2001). But i thought i knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology, 15*(1), 23-32. doi: 10.1002/1099-0720(200101/02)15:1<23::aid-acp681>3.0.co;2-j

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological Bulletin,* 137(1), 131-148. doi: 10.1037/a0021705

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *PLoS ONE, 9*(7), e98663. doi: 10.1371/journal.pone.0098663

Roebers, C. M. (2014). Children's deliberate memory development: The contribution of strategies and metacognitive processes. In P. J. Bauer & R. Fivush (Eds.), *The Wiley Handbook on the Development of Children's Memory, Volume I/II* (pp. 865-894). Chichester: John Wiley & Sons, Ltd.

Roebers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. *Learning and Individual Differences*, *29*, 141-149. doi: 10.1016/j.lindif.2012.12.003

Roebers, C. M., Moga, N., & Schneider, W. (2001). The role of accuracy motivation on children's and adults' event recall. *Journal of Experimental Child Psychology*, *78*(4), 313-329. doi: 10.1006/jecp.2000.2577

Salles, A., Ais, J., Semelman, M., Sigman, M., & Calero, C. I. (2016). The metacognitive abilities of children and adults. *Cognitive Development, 40*, 101-110. doi: 10.1016/j.cogdev.2016.08.009

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*(1), 33-45. doi: 10.1007/s11409-008-9031-3

Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking. *Developmental Science, 1*(2), 291-297. doi: 10.1111/1467-7687.00044

Schneider, W., & Löffler, E. (2016). The development of metacognitive knowledge in children and adolescents. In J. Dunlosky & S.K. Tauber (Eds), *The Oxford Handbook of Metamemory* (pp. 491-518). Oxford, UK: Oxford University Press.

Sedikides, C., & Green, J. D. (2009). Memory as a self-protective mechanism. *Social and Personality Psychology Compass*, *3*(6), 1055-1068. doi: 10.1111/j.1751-9004.2009.00220.x

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. doi: 10.1037/0022-0663.95.1.66

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*(4), 475-511. doi: 10.3102/0034654314564881

Van Loon, M. H., de Bruin, A. B., van Gog, T., & van Merriënboer, J. J. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction*, *24*, 15-25. doi: 10.1016/j.learninstruc.2012.08.005

Van Loon, M., de Bruin, A., Leppink, J., & Roebers, C. (2017). Why are children overconfident? Developmental differences in the implementation of accessibility cues when judging concept learning. *Journal of Experimental Child Psychology*, *158*, 77-94. doi: 10.1016/j.jecp.2017.01.008

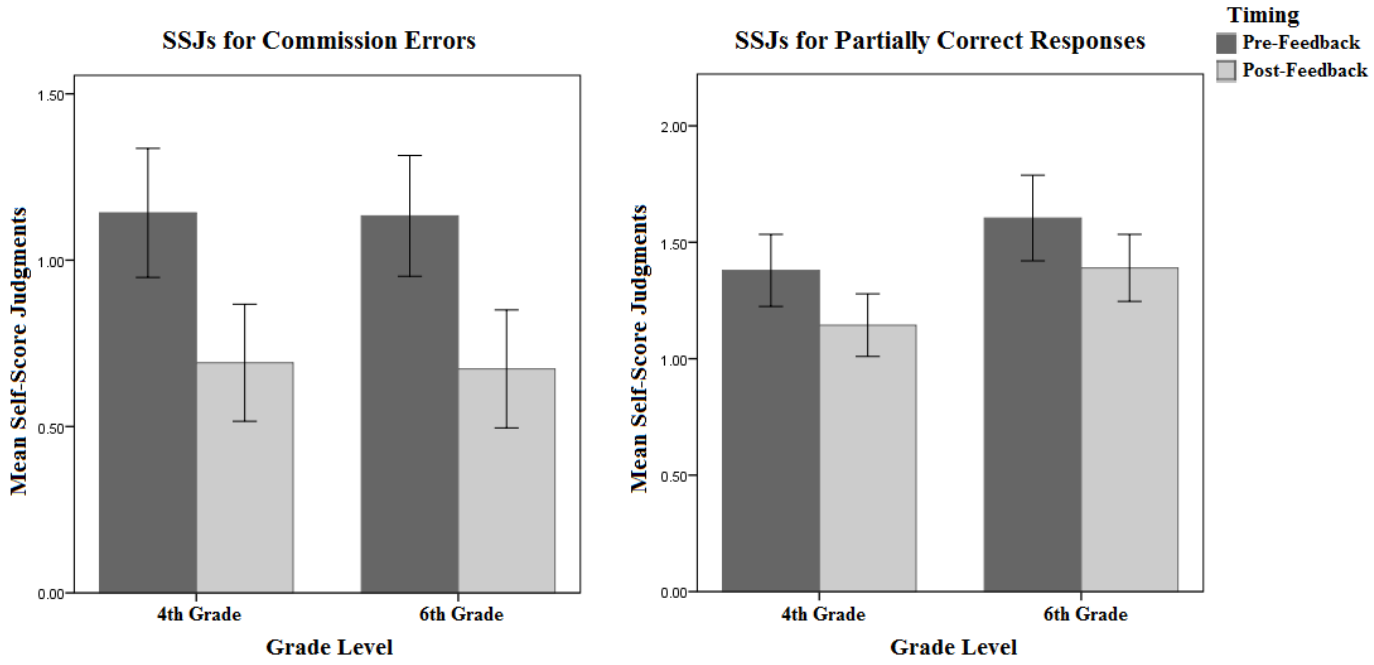
Winne, P. H. (1997). Experimenting to bootstrap self-regulated learning. *Journal of Educational Psychology, 89*(3), 397-410. doi: 10.1037/0022-0663.89.3.397

Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction, 46*, 12-20. doi: 10.1016/j.learninstruc.2016.08.002

Table 1. The percentage of omissions, commission errors, and test responses containing one, two, and three correct idea units for the two age groups, and mean SSJs and restudy selections for these test response types (SDs of the mean in parentheses).

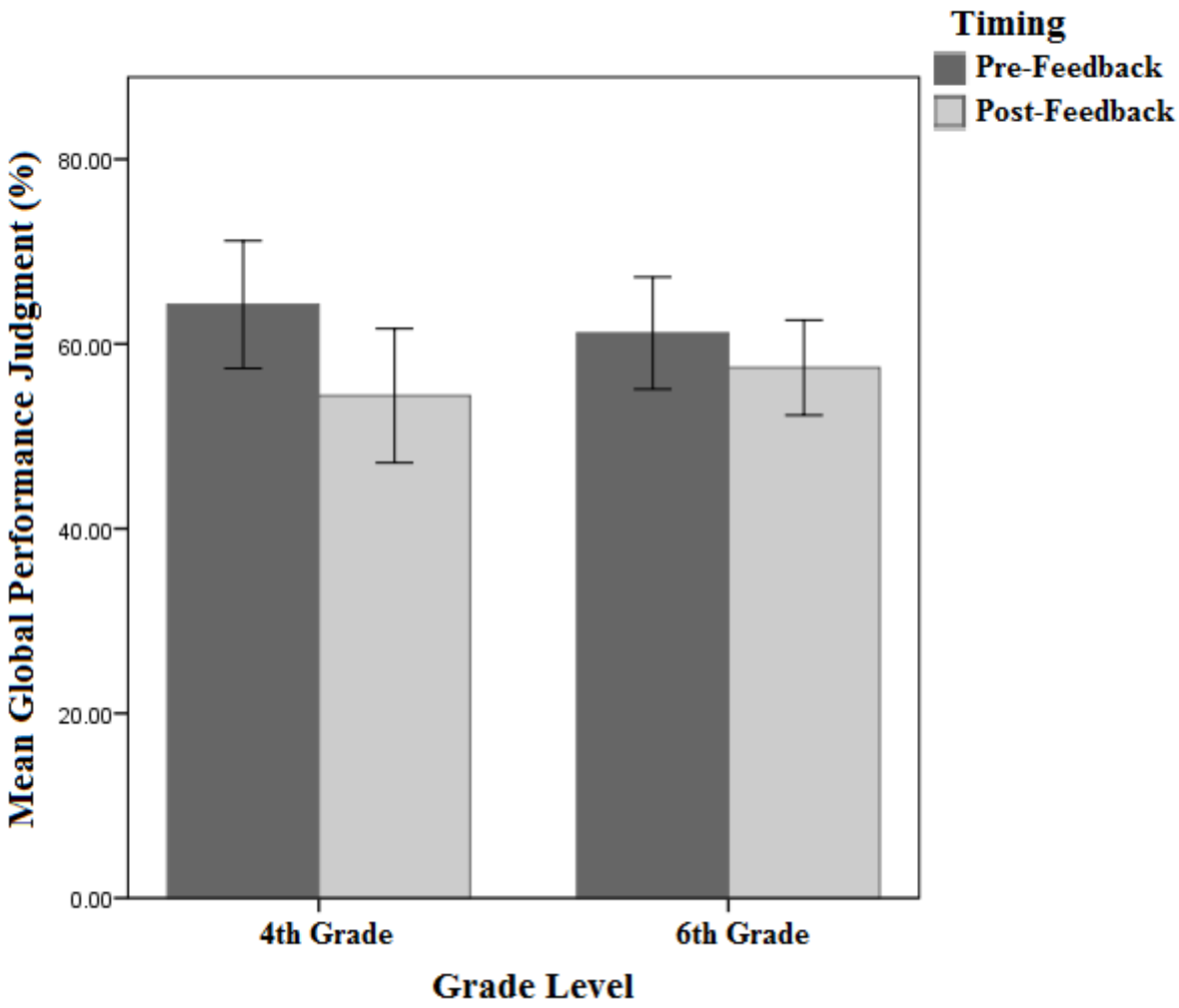
	Omissions	Commission Errors	One Idea Unit	Two Idea Units	Three Idea Units
			Percentage of Test Responses		
Grade 4	17.6%	16.1%	24.8%	26.2%	15.4%
Grade 6	23.6%	12.0%	25.8%	28.4%	10.2%
			SSJs pre-feedback		
Grade 4	.07 (.40)	1.14 (.58)	1.38 (.54)	2.02 (.75)	2.43 (.78)
Grade 6	.01 (.07)	1.13 (.61)	1.60 (.65)	1.97 (.72)	2.48 (.68)
			SSJs post-feedback		
Grade 4	.03 (.24)	.69 (.53)	1.14 (.47)	2.07 (.59)	2.75 (.48)
Grade 6	.01 (.07)	.67 (.60)	1.39 (.51)	1.75 (.64)	2.44 (.65)
			Restudy Selections pre-feedback		
Grade 4	87.03 (21.0)	34.05 (40.3)	26.67 (28.8)	20.75 (27.1)	11.79 (26.2)
Grade 6	90.88 (15.7)	54.78 (36.7)	22.56 (26.9)	5.12 (5.6	5.75 (11.2)
			Restudy Selections post-feedback		
Grade 4	80.20 (27.6)	50.68 (39.8)	31.38 (34.3)	17.71 (2)	8.76 (21.3)
Grade 6	74.93 (27.4)	56.41 (37.5)	30.52 (32.2)	9.45 (14.8)	9.20 (21.9)

Figure 1. Self-Score Judgments for Commission Errors and Partially Correct One-Idea-Unit Responses



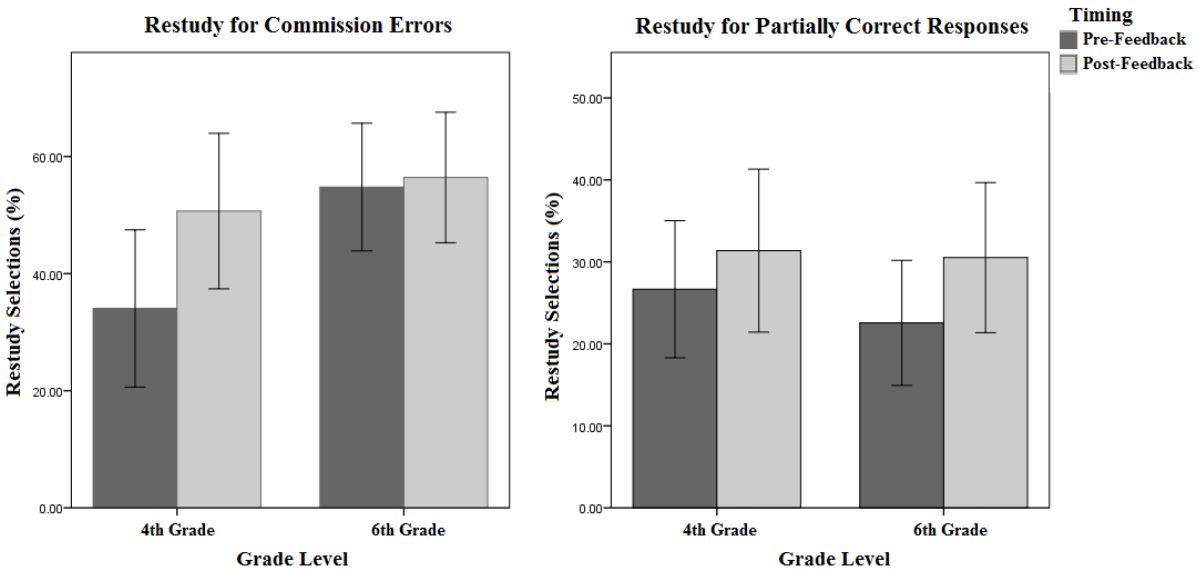
Note. Error bars indicate the 95% confidence interval.

Figure 2. Global Judgments made Before and After Receiving Feedback.



Note. Error bars indicate the 95% confidence interval.

Figure 3. Restudy Selections for Commission Errors and Partially Correct One-Idea-Unit Responses



Note. Error bars indicate the 95% Confidence Interval