Learning and Individual Differences 74 (2019) 101754

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Judging own and peer performance when using feedback in elementary school

Mariëtte van Loon[a,*], Janneke van de Pol[b]

[a] Department of Developmental Psychology, University of Bern, Fabrikstrasse 8, CH-3012 Bern, Switzerland
[b] Department of Educational Sciences, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Self-monitoring
Peer-monitoring
Feedback
Elementary school
Development

ABSTRACT

Children find it challenging to self-monitor the quality of their own test responses, and are typically overconfident. Inaccurate self-monitoring may not only be due to a metacognitive deficit, but also to self-protective biases. Therefore, monitoring peer performance and detecting others' errors may be easier than monitoring oneself. This study investigated 97 children's (52 fourth and 45 sixth grade) feedback use when scoring their own and their peers' concept learning. Children completed a concept-learning task, took a test, and then scored their own responses and the responses of one of their peers with use of feedback standards. Error detection was better for peer- than for self-score judgments. Further, monitoring was more accurate for older than younger children, and inaccurate prior knowledge led to less accurate peer and self-judgments. Findings imply that, when implementing co-scoring activities, it is important to be aware that its accuracy is affected by children's age and prior knowledge.

1. Introduction

In elementary school, integral to the process of developing knowledge in different school domains, children have to develop self-regulated learning skills. That is, they need to be able to adapt to increasing learning demands when studying in the classroom and when making homework. Effective self-regulated learning includes planning, prioritizing study tasks, allocating study time, and making use of appropriate study strategies (Bjork, Dunlosky, & Kornell, 2013). To be able to self-regulate learning, children need to self-monitor whether they are making progress (Roebers, Krebs, & Roderer, 2014). Self-monitoring predicts self-regulation; for instance, after identifying that certain materials are not yet well-learned, learning actions can be taken to achieve learning goals, such as re-studying information or seeking help (Zimmerman, 2000). However, self-monitoring of performance is challenging for children; they have difficulties distinguishing between well-learned and less well-learned items, and are typically overconfident (Baars, Van Gog, De Bruin, & Paas, 2014). If learners cannot accurately judge their current state of learning, adaptive self-regulation is impossible because they do not know whether learning materials should be selected or discarded from further study (Dunlosky & Rawson, 2012). Therefore, inaccurate monitoring is disadvantageous for learning (Destan & Roebers, 2015; Thiede, Anderson, & Therriault, 2003).

To improve self-monitoring and subsequent regulation and learning, providing feedback on task performance seems beneficial. Research with adults (Dunlosky & Rawson, 2012) and with fourth and sixth grade elementary school children (Van Loon & Roebers, 2017) asked learners to self-score their own responses when making use of feedback standards pointing out what correct performance should look like. This feedback helped to compare responses with the standards, and this improved monitoring (as measured with self-score judgments – SSJs) and subsequent regulation (as measured with restudy selections). However, some degree of overconfidence remained, indicating that even when using feedback standards, incorrect or only partially correct responses were not always recognized.

Children as well as adults are particularly overconfident when judging the quality of their own commission errors (i.e., entirely incorrect responses; Dunlosky, Rawson, & Middleton, 2005; Finn & Metcalfe, 2014). Also for partially correct responses, it is often believed that these are fully correct (Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013a). However, when inspecting work of others, it seems easier to detect errors and incomplete responses. Already at a young age, children more often recognize others' errors than their own mistakes (Destan, Spiess, De Bruin, Van Loon, & Roebers, 2017; Ruble, Eisenberg, & Higgins, 1994; Stipek & Hoffman, 1980). If children are better able to judge their peers' than their own performance, this may have educational implications, because peer assessment activities could

* Corresponding author.
E-mail address: mariette.vanloon@psy.unibe.ch (M. van Loon).

https://doi.org/10.1016/j.lindif.2019.101754
Received 22 October 2018; Received in revised form 25 June 2019; Accepted 1 July 2019

then possibly enhance self-monitoring and self-regulation of learning (Hwang, Hung, & Chen, 2014; Okita, 2014).

The present study aims to acquire insight into elementary school children's judgments of their own and their peers' performance when they are presented with feedback standards. The theoretical background for this study is provided by Vygotsky (1978), who argued that learning is a social process, and that collaborative interactions with peers are needed for this. Receiving peer feedback gives a student the opportunity to acquire insights into a peer's interpretation of their responses, which task items have been mastered, and which items may need improvement. Further, when giving feedback to a peer, children can learn from demonstrating and applying their knowledge, and from making comparisons with their own work (Carless & Boud, 2018; Nicol, Thomson, & Breslin, 2014). Through collaborative metacognitive learning activities, children can learn to apply the acquired insights into peer performance to themselves (Okita, 2014). However, such activities can only be successful when children are able to make accurate judgments about their peers' actual ability (Sebanz, Bekkering, & Knoblich, 2006). Thus, before designing collaborative monitoring tasks, it is first important to investigate how well children can judge peer performance.

Importantly, during the course of elementary school, children's ability to monitor learning undergoes strong development (Schneider & Löffler, 2016). To acquire insight into the developmental trajectory of children's ability to make self and peer judgments, we compared judgment accuracy between fourth and sixth grade elementary school children. Before clarifying the design of the present study, we first give an overview of the development of monitoring skills in elementary school and potential reasons for children's inaccurate monitoring. Then, we address effects of feedback on the accuracy of self- and peer monitoring.

1.1. Development of self-monitoring skills

Children's monitoring depends on the type of task they work on. When they are asked to differentiate between correct and incorrect answers for simple learning materials (e.g., when learning word pairs or when learning spelling), self-monitoring skills seem well developed when children are eight years of age (Roebers & Spiess, 2017; Schneider & Löffler, 2016). However, children's ability to monitor more complex learning and comprehension seems to undergo development over the elementary school years, and metacognitive skills keep developing at least until adolescence. When studying expository texts, children are only able to monitor their learning to a moderate extent when they are approximately 12 years old (De Bruin, Thiede, Camp, & Redford, 2011). Further, children's ability to effectively regulate their learning, and appropriately select materials that are not yet well-learned for further study develops from age ten onwards (Schneider & Löffler, 2016). Most improvement in monitoring accuracy is seen in elementary school children's error recognition skills. Before entering school, children are able to judge their correct responses as being right (Lyons & Ghetti, 2013). However, error recognition is more demanding than reporting how sure one is that performance is correct, and therefore, accurate monitoring of incorrect performance seems to undergo development during the elementary school years (Roebers, 2002, 2014).

Overconfidence remains an issue across the life span; most persons believe they perform superior to others, while actually overestimating their ability (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008), and adults as well as children tend to believe that they are better-than-average (Alicke & Govorun, 2005). However, throughout elementary school, children become more skilled to strategically use valid indicators (i.e., cues) of actual performance when making judgments (Koriat, Ackerman, Lockl, & Schneider, 2009; Van Loon, De Bruin, Leppink, & Roebers, 2017). Cues such as study time, experienced ease of studying, the ease with which information comes to mind when making a retrieval attempt, and one's ability to apply studied information in a novel context can all be valid when these give insights into one's level of understanding and performance (Koriat, 1997; Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013b). From the age of ten onwards, children strategically base judgments on cues that are related to their study and retrieval experiences, and this improves their monitoring accuracy (Van Loon et al., 2017). Younger children are less able to discriminate between correct and incorrect performance, and are more overconfident for errors. Instead of attending to actual indicators of performance, they seem to be more biased by their wishes to perform well (Stipek, 1984). Their tendency to give credit to most of their incorrect responses may reflect their wishes to be rewarded for their effort (i.e., they use their effort as a cue for their judgments). During the course of elementary school, children become less biased by wishful thinking (Schneider, 1998) and by the tendency to conflate effort with ability (Kurtz-Costes, McCall, Kinlaw, Wiesen, & Joyner, 2005), and this seems to explain developmental improvements in monitoring accuracy.

1.2. Effects of prior knowledge on monitoring accuracy

Learners' prior knowledge about a learned topic contributes to their feelings of processing and retrieval fluency, and these cues affect their monitoring judgments (Van Loon et al., 2013a). It is well known that accurate prior knowledge benefits achievement; prior knowledge makes it easier to process information during study, and it supports the building of a mental representation (Braithwaite & Goldstone, 2015). This information is, subsequently, also easily retrieved when taking a test, and students are likely to be highly confident that these responses will be correct. College students who had high prior knowledge showed better judgment accuracy than learners with low prior knowledge (Griffin, Jee, & Wiley, 2009; Taub, Azevedo, Bouchet, & Khosravifar, 2014). However, to our knowledge, there are no documented effects of accurate prior knowledge on elementary school children's self-monitoring.

Moreover, prior knowledge may also be inaccurate. Even though this inaccurate prior knowledge is fluently retrieved when taking a learning test, using this inaccurate prior knowledge for a test response will result in a commission error. Children's inaccurate prior knowledge may not only be disadvantageous for learning achievement, but also for self-monitoring accuracy (Van Loon et al., 2013a). Inaccurate prior knowledge is hard to correct; even when the correct and contradicting information is presented during study, misconceptions often remain (Van den Broek & Kendeou, 2008). A study with third to sixth grade elementary school children showed that inaccurate prior knowledge was often not corrected during study, and that children were overconfident and judged these pervasive errors as being correct (Van Loon et al., 2013a).

1.3. Effects of feedback on self-monitoring

One promising method to improve monitoring accuracy for children and adults alike is to ask them to complete a self-test, and to score their answers with use of feedback standards (Lipko et al., 2009; Rawson & Dunlosky, 2007). Feedback is most beneficial when it gives sufficient detail about the ideas that learners should have mentioned in their test responses (Lipko et al., 2009; Miller & Geraci, 2011). This way, learners can become aware of the discrepancies between their actual performance and the learning goals, and feedback can be a helpful tool to support recognition and correction of errors (Hattie & Timperley, 2007; Mory, 2004).

However, although inspection of feedback standards improved SSJs and restudy selections for fourth and sixth graders (Van Loon & Roebers, 2017), even after receiving feedback, children often expected that they would get more credit points for their responses than they would objectively receive. This persistent overconfidence seems at least partially due to inaccurate prior knowledge, which could hinder feedback processing (Butler & Winne, 1995). Even though learners are often

receptive to feedback and correct their misunderstandings, findings that students are not always detecting errors when using feedback indicate that in some cases feedback is ignored, rejected, or interpreted such that it fits inaccurate prior knowledge (Mory, 2004).

1.4. Children's peer-monitoring

Judgments that are made by peers may be more useful to give input for regulation of learning than self-monitoring judgments. Although self-monitoring seems to be based on experiential cues and self-serving biases (Destan et al., 2017), when monitoring performance of peers, more objective reasoning about performance quality may occur (Efklides, 2008; Koriat & Ackerman, 2010; Thomas & Jacoby, 2013). To successfully implement peer-scoring activities in elementary school, ideally, peer-score judgments (PSJs) should be accurate. However, even though learners seem to apply different judgment criteria for peers than for oneself, peer-score judgments do not always match objective performance. After receiving feedback, error recognition may remain challenging, not only when judging own performance, but also when scoring peers (Ruble et al., 1994). When second and fourth grade children inspected their own and a peer's writing task, they more often detected errors that were made by their peers (Cameron, Edmunds, Wigmore, Hunt, & Linton, 1997). Nevertheless, error detection was far from perfect, fourth graders only corrected 30% of their own errors and 50% of their peer's errors. Further, children may apply more conservative criteria when judging peers than when judging themselves, and this does not necessarily benefit judgment accuracy. For instance, when 7th grade middle school students gave themselves and their peers grades for different school tests, they underestimated peer performance more than own performance (Sadler & Good, 2006).

The accuracy of PSJs seems to be affected by developmental factors; in addition to children's self-monitoring ability, their monitoring accuracy for others' learning improves with age (Paulus, Tsalas, Proust, & Sodian, 2014). Developmental improvement in error recognition was observed by Cameron et al. (1997); fourth graders recognized more of their peers' and their own errors than second graders. Moreover, children's use of feedback for self- and peer-scoring improves with age (Destan et al., 2017).

One's prior knowledge may also affect the accuracy of PSJs. Cameron et al. (1997) showed that accurate prior knowledge did not only influence self-scoring but also peer-scoring; children who had better writing abilities were better at error monitoring both for themselves, but also for their peers. And even though processing and retrieval fluency cues can lead to overconfidence when prior knowledge is inaccurate, presumably, these misleading experiential cues should not bias PSJs as much as SSJs, because one has no insight into a peers' subjective study and test experiences (Koriat & Ackerman, 2010).

Interestingly, although peer judgments seem to be based on different factors than self-judgments, research suggests a tight relation between one's own monitoring processes and an understanding of other's learning and performance (Koriat & Ackerman, 2010). As shown by research with adult participants, estimates of performance of others are based on judgments for oneself (Thomas & Jacoby, 2013). That is, there may be consistency between judgments, such that judgments for oneself and a peer are similar, even when there are differences in performance. If this is the case, this may imply that children do not necessarily base their judgments on objective performance criteria, but instead, that for tasks for which they give themselves high scores they also give their peers high scores, and vice versa. If it is the case that judgments of peer performance are linked to estimates of own performance, then overconfidence for oneself may also be related to overconfidence for a peer. However, although research indicates similarity between judgments made for oneself and for others, it is unclear whether this is also the case for children. To acquire insight into this, we assess the intra-individual consistency of judgments made for oneself and for peers.

2. Present study

With this study, we aim to investigate the accuracy of self- and peer-monitoring for fourth and sixth grade students. These age groups are selected because in mid-elementary school, children need to become skilled to self-regulate their learning. From age 10 onwards, children start to base their study decisions on their monitoring (Dufresne & Kobasigawa, 1989), and hence, monitoring accuracy becomes an important predictor for learning success (Dunlosky & Rawson, 2012). However, children's ability to monitor learning still develops between the mid and late elementary school years (De Bruin et al., 2011; Krebs & Roebers, 2010). It is thus likely that developmental differences affect children's scoring of their own and their peers' responses; to acquire insights into this, the two age groups are compared. Moreover, the study investigates effects of children's accurate and inaccurate prior knowledge, and the consistency within individuals between their SSJs and PSJs.

After taking a pretest to assess prior knowledge, children studied difficult concepts, then they were tested for their understanding of these concepts. After taking the posttest, children received feedback standards, and judged the quality of their own answers and the quality of the answers that were given by a peer.

Indices of monitoring accuracy can be calculated with measures of relative and absolute accuracy (Dunlosky & Metcalfe, 2009). Relative accuracy is measured with correlations between a person's self-monitoring and performance, and indicates to what extent a person can discriminate between test responses with high and low quality. To further assess whether a learner was over- or underconfident, a measure of absolute accuracy is needed. For this measure, the magnitude of judgments is compared with the objective performance scores; judgments are accurate when deviations between judgment magnitudes and performance are low. Test responses can range in quality (i.e., incorrect responses, partially correct responses, and fully correct responses); learners would be overconfident when they give themselves more credit for responses than they would objectively receive, and underconfident when monitoring judgments are lower than objective performance (Pieschl, 2009). To acquire insights into relative and absolute accuracy, both measures are used in the present study.

Our first research question addresses whether there are differences in monitoring judgments and monitoring accuracy between SSJs and PSJs. This research question is addressed with measures of relative accuracy, absolute accuracy, and for the different test response types. We expect differences between SSJs and PSJs, such that children are more conservative when judging their peers' than their own responses. However, we do not have specific hypotheses about whether making more conservative judgments also improves relative and absolute judgment accuracy for peers in comparison to oneself; this is addressed exploratively. For analyses of the separate test response types, we expect to replicate findings by Van Loon and Roebers (2017) that recognizing one's own commission errors is most challenging, and that children are overconfident for these errors. Moreover, children seem better to detect their peers' than their own errors, we therefore hypothesize that the difference between SSJs and PSJs is most pronounced for commission errors, such that judgments are lower, and thus more accurate for errors made by peers.

The second research question concerns developmental differences between fourth and sixth graders in SSJs and PSJs. We expect better relative and absolute judgment accuracy for sixth than for fourth graders when comparing SSJ accuracy and PSJ accuracy between grades.

Moreover, our third question addresses whether prior knowledge affects SSJs and PSJs. We predict that prior knowledge affects SSJs, such that high accurate prior knowledge scores are related to more accurate relative and absolute self-monitoring, whereas inaccurate prior knowledge is expected to be related to less accurate relative and absolute self-monitoring. The effect of accurate and inaccurate prior knowledge on PSJs is addressed as an explorative question.

Finally, research question four concerns the consistency between SSJs and PSJs. Even though self- and peer-judgments are based on different judgment processes, at the same time, there may be consistency between SSJs and PSJs. We therefore exploratively address to what extent the magnitudes of the judgments, as well as judgment accuracy, are related between self- and peer-monitoring judgments within individuals.

3. Method

3.1. Participants

Participants were 97 children from the German speaking part of Switzerland, 52 fourth graders (30 males, 22 females; $M = 10.37$ years, $SD = 6.3$ months) and 45 sixth graders (19 males, 26 females; $M = 12.30$ years, $SD = 5.2$ months). The sample was recruited from public schools; family backgrounds were lower to upper middle class. The study was realized in accordance with the APA ethical principles and the declaration of Helsinki, and the Faculty Ethic Review Board of the University of [information withheld] approved the research project. With a letter, participants and their caretakers were informed about the study, including information about confidentiality of the data, that participants could decline to participate or withdraw after starting without any consequences, and that they could contact the first author if they had questions. Parental written consent was obtained, and when introducing the study in the classroom, it was again emphasized that participation was voluntary.

Participants were tested in their classroom, all children were used to follow class instructions in the German language. Participation was rewarded with a small gift.

3.2. Materials

Difficult concepts served as stimuli for the concept-learning task, an example of such a concept is '*Ivory: Ivory is a white material, which comes from the tusks of elephants and is often used to make small sculptures*'. Similar types of materials have successfully been used in previous studies with young learners (e.g., Lipko et al., 2009; Van Loon et al., 2013a; Van Loon & Roebers, 2017). The task consisted of 14 concepts for fourth graders and 16 concepts for sixth graders; selection of these concepts was based on teaching materials and the learning objectives outlined in the curriculum description for teachers. All concept definitions consisted of 3 idea units (cf. Van Loon & Roebers, 2017); these definitions were derived from a dictionary for elementary school children. Before selecting concepts for the task, a pilot study was conducted with participants in the same age range as the final sample (17 fourth graders, $M = 10.18$ years, $SD = 0.44$; 18 sixth graders, $M = 11.98$ years, $SD = 0.24$). For this pilot, children completed a pretest, study phase, and posttest for 24 concepts. Based on the pilot study, concepts were selected that were difficult, such that prior knowledge would be low. However, for all selected concepts, posttest scores were higher than pretest score, indicating that children in these age groups were able to learn at least part of the concept definitions.

The task consisted of six phases, as shown in Fig. 1: (1) Pretest. To-be-learned concepts were listed, with lines where children could write down the meaning of the concepts. This pretest quantified prior knowledge (i.e., for pretest items containing correct idea units, we scored the percentage of correct ideas as accurate prior knowledge; the pretest items which only contained incorrect idea units reflected the percentage of inaccurate prior knowledge) (2) Learning phase. The concepts and the concept definitions were presented to the children. Furthermore, for each concept, an example sentence clarified use of the concept. (3) Posttest. Similar as in the pretest, children could write down the meaning of each studied concept. (4) Restudy selections. Concepts were printed in a grid (a 2×7 grid in fourth grade and a 2×8 grid in sixth grade). With a checkmark, children could indicate which concepts they would like to restudy. Note that there was no actual restudy trial, and because the main focus of this research is on monitoring accuracy, results on restudy selections are not further addressed. (5) Feedback phase. Children received feedback standards, these listed the concepts and the definition as shown in the learning phase; each definition was parsed into three idea units (in line with Dunlosky, Hartwig, Rawson, & Lipko, 2010). (6) Judgment phase. With use of the feedback standards, children made judgments about the quality of their own and their peer's posttest performance (through exchange of response sheets), by indicating in a check box (ranging from 0 to 3 ideas) how many correct idea units they identified in the given response. Judgments were made after the test responses were provided and feedback was inspected, which is in line with research by Dunlosky et al. (2010), Lipko et al. (2009) and Van Loon and Roebers (2017).

All task materials were printed on paper and handed out in a ring binder. There were three different versions of the concept task for each age group, the order of the items was different in these versions. For each subtask, the order of the concepts was changed.

3.3. Procedure

Two experimenters visited the classroom to test the children. Before starting, they introduced the concept learning task, and they showed children examples of concepts. They told children that each to-be-learned concept consisted of three idea units, and that the aim was that children would try to remember all three ideas, so they could reproduce this when tested. Then, they showed the children the ring binders they would receive, and they told them that, when they saw a blank page with a large number on it, they should wait for further instructions. Then the concept-learning task folders were handed out with blue pens to complete the pre- and posttest. After each phase, there was a numbered page that indicated children to stop and wait for experimenter instructions before they went to the next phase. This way, the test time, the learning time and the time provided for inspecting feedback and making judgments was similar for all children.

When starting with the pre-test, children were told to write down the meaning of the concept if they thought they knew it, and they were told that they could leave it blank if they did not know the concept. The pre-test phase took 8 min. Then the learning task started, for 10 min, children studied concepts with definitions. After the learning phase, children could mark the concepts they would like to restudy. They then took the open-ended test, this phase lasted 12 min. Subsequently, they received feedback standards which contained the correct answers to the test questions, separated in three idea units. Children were told to compare the feedback standards with the test responses, and to use the feedback to make judgments about their own and their peer's responses. They were instructed that they had to assess the quality of these test responses, by indicating how many of the idea units were presented. To make sure that they would not change their responses after receiving feedback, they were told to make the judgments with a red pen; the blue pens were collected, and red pens were handed out. Children had 10 min to make the SSJs and 10 min to make PSJs. The order of making SSJs and PSJs was counterbalanced, so that 50% of the children first made SSJs, the others first made PSJs. That is, the children who had an S printed on their booklet were told to first make self-score judgments, and after instructions, exchange their test booklet with the peer who was sitting diagonally behind them (or diagonally in front of them, respectively). The children who had a P on their booklet first exchanged their booklet with their peer, and after instructions, they received their own booklet back to make SSJs. Children were instructed about the exchange and whether they either had to score their own answers or their peer's answers first. They did not receive instructions about whether they should compare their own with their peer's judgments. Ring binders were handed out such that the matched pairs had the same task version, to ensure that the similar feedback standard could be used

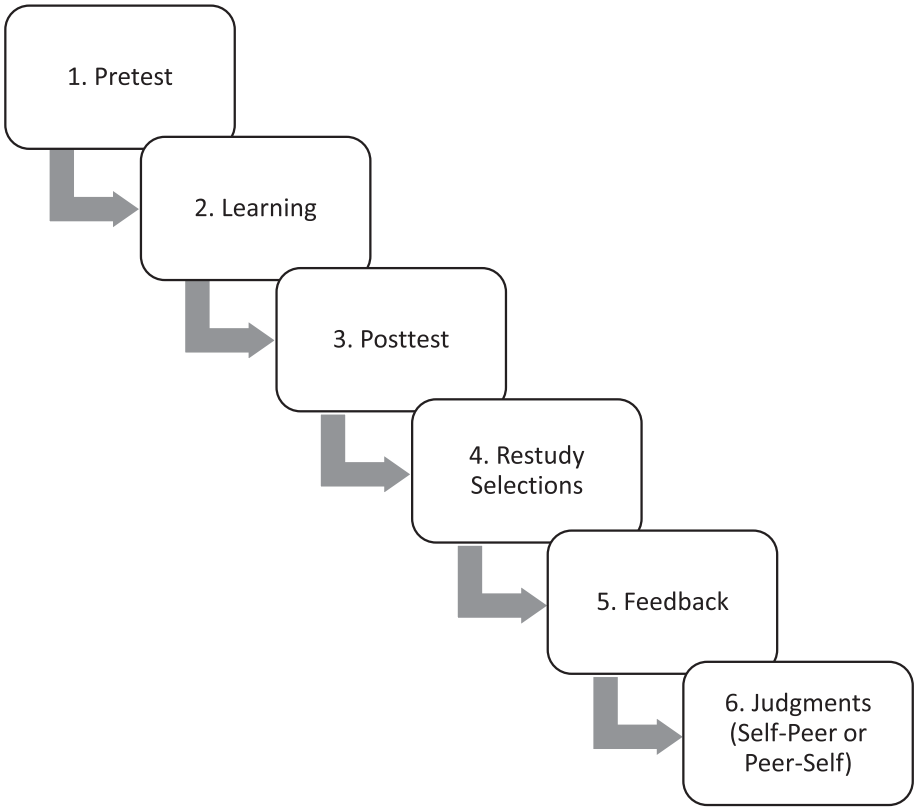

Fig. 1. Task phases and procedure.

when scoring oneself and when scoring the peer. The full procedure took approximately 60 min.

3.4. Scoring

Each concept definition consisted of three idea units. Similar as in Van Loon and Roebers (2017), the number of correct ideas (ranging from 0 to 3) were scored both in the pretest and the posttest responses. Idea units were scored when either provided literally or when containing the gist of the original idea unit. When no response was given at all, the answer was scored as an omission. When a completely incorrect response was given, this answer was scored as a commission error. After extensive discussion about scoring criteria, the experimenters coded the responses. To assess interrater reliability, 20% of the pre- and posttest responses were scored by both raters, agreement was high: ICC for the pretest = 0.87; ICC for the posttest = 0.86.

3.5. Analyses

For each participant, based on pretest data, the percentage of accurate prior knowledge and the percentage of inaccurate prior knowledge were calculated, and used as independent variables in the analyses to predict judgment accuracy. The correlations between the percentage of accurate prior knowledge responses and inaccurate prior knowledge responses were low to moderate (overall $r = -0.24$; $r = -0.31$ for the fourth grade and -0.04 for the sixth grade), indicating no issues with multicollinearity (Mansfield & Helms, 1982).

To assess performance, the number of correct idea units in test performance scores was calculated. Dependent variables in our analyses are magnitudes of SSJs and PSJs, and the accuracy of SSJs and PSJs. SSJs were related to objective performance to assess self-monitoring accuracy. Further, because tests were exchanged between children, the accuracy of children's PSJs could also be related to performance for this particular peer.

Both for fourth and sixth graders, there were no significant effects of task version on performance, mean SSJs, and mean PSJs (all $ps > 11$). Therefore, data from all versions are combined in the reported analyses. Moreover, the order of making SSJs and PSJs did not affect relative and absolute judgment accuracy (all $ps > 0.17$), therefore, the data from the group who first made SSJs and the group who first made PSJs are collapsed.

4. Results

In this section, we first present preliminary analyses on prior knowledge, test performance, and SSJs and PSJs. Then we address differences between self and peer monitoring, effects of accurate and inaccurate prior knowledge, and the consistency between SSJs and PSJs. Results are reported and compared for the two grade levels.

4.1. Pre- and posttest performance and judgment magnitudes

Table 1 shows mean prior knowledge, posttest performance, and SSJs and PSJs. As visible in this table, the sixth graders had more accurate prior knowledge than fourth graders, $t(95) = 2.71$, $p = .008$. Moreover, sixth graders had significantly less inaccurate prior knowledge than fourth graders, $t(95) = 2.24$, $p = .027$.

Further, Table 1 shows that posttest performance was significantly higher for sixth than for fourth graders, $t(95) = 8.44$, $p < .001$. As a measure of actual learning, corrected for pre-test performance, we computed a percentage difference score between post- and pre-test performance. In total, fourth graders produced 13.79% ($SD = 11.28$) more concept idea units at the posttest than at the pretest; the sixth graders showed 32.34% ($SD = 14.35$) improvement from pre- to posttest. Sixth graders showed more improvement than fourth graders, $t(95) = 7.12$, $p < .001$.

Monitoring judgments ranged from 0 (no correct ideas) to 3 (3 correct ideas). Table 1 shows that both for fourth and sixth graders,

Table 1
Prior knowledge, posttest performance, and judgment magnitudes.

	Prior knowledge: Correct idea units (%)	Inaccurate prior knowledge (%)	Posttest: Correct idea units (%)	Self-score judgments (range 0–3)	Peer score judgments (range 0–3)
Grade 4	9.65 (5.73)	18.54 (18.01)	23.44 (12.94)	1.09 (0.60)	0.88 (0.59)
Grade 6	13.15 (6.96)	11.68 (10.62)	45.48 (12.69)	1.56 (0.50)	1.45 (0.48)

Note. Standard deviations of the mean in parentheses.

PSJs were lower than SSJs. To investigate whether judgment magnitudes differ between SSJs and PSJs (RQ1), general linear models (GLM) for repeated measures were used to compare SSJs and PSJs. Magnitudes of SSJs and PSJs were included as a within-subject factor, grade was included as a between-subject factor to account for potential age differences (RQ2). The GLM for repeated measures confirms that magnitudes of PSJs were significantly lower than SSJs, $F(1, 93) = 17.76$, $p < .001$, $\eta_p^2 = 0.16$. Further, the main effect of Grade, $F(1, 93) = 23.52$, $p < .001$, $\eta_p^2 = 0.20$, shows that sixth graders made higher judgments than fourth graders.

4.2. Relative accuracy for self and peer judgments

To address RQ1 on whether there are differences in relative monitoring accuracy between SSJs and PSJs, intra-individual Goodman and Kruskal gamma correlations were calculated between SSJs and performance, and between PSJs and performance (cf. Nelson, 1984). This correlation indicates how strongly pairs of ordinal variables (judgments and performance per item) are associated for each participant, and can be interpreted as the proportion of ranked pairs in agreement. A gamma correlation of +1 would show that items for which judgments were high consistently received higher objective scores than items for which judgments were lower.

Mean gamma correlations between performance and SSJs were 0.77 ($SD = 0.41$) for fourth grade and 0.92 ($SD = 0.15$) for sixth grade. Gamma correlations between performance and PSJs were 0.69 ($SD = 0.42$) for fourth grade and 0.89 ($SD = 0.18$) for sixth grade. Fig. 2 shows the relative accuracy for the self- and peer-judgments. GLM

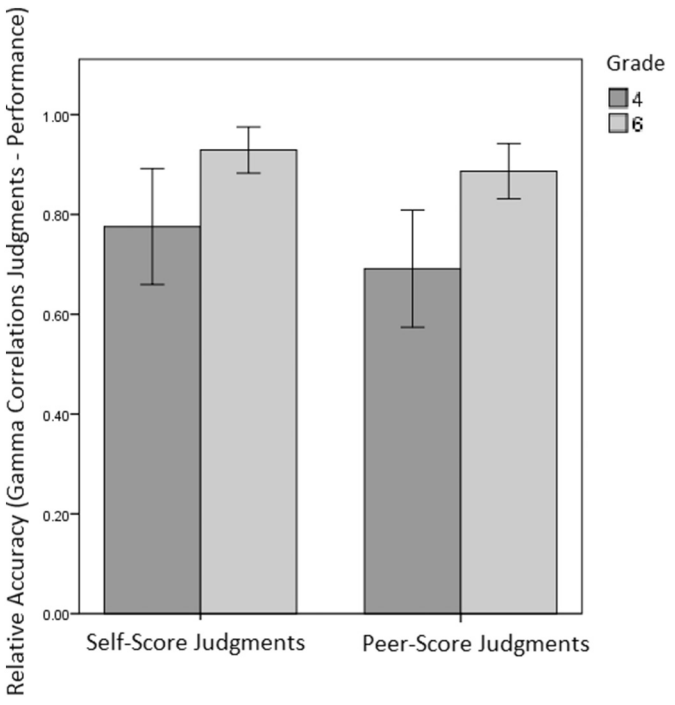

Fig. 2. Relative accuracy of self- and peer score judgments. (Error bars indicate the 95% confidence interval.)

analyses with relative accuracy for SSJs and PSJs as within-subjects factor (RQ1) and grade as between-subjects factor (RQ2) did not show differences between relative accuracy of SSJs and PSJs, $F(1, 93) = 3.07$, $p = .08$. However, there was a significant main effect of grade, $F(1, 93) = 8.97$, $p < .001$, $\eta_p^2 = 0.09$. As shown in Fig. 2, sixth graders showed higher relative accuracy for self and peer monitoring judgments than fourth graders.

Furthermore, follow-up regression analyses were used to investigate evidence for RQ3 concerning effects of accurate and inaccurate prior knowledge on relative judgment accuracy. Accurate and inaccurate prior knowledge are interval variables (since these are calculated as percentage values per individual); the percentage of accurate prior knowledge and the percentage of inaccurate prior knowledge were added to the regression model. Moreover, to account for grade, this variable was dummy coded and simultaneously entered. There were no effects of prior knowledge (accurate and inaccurate) on relative accuracy. That is, neither the percentage accurate prior knowledge ($b = -0.06$, $p = .56$), nor the percentage of inaccurate prior knowledge ($b = -0.13$, $p = .22$) predicted the relative accuracy of self-scoring. Further, accurate ($b = 0.01$, $p = .90$) and inaccurate prior knowledge ($b = 0.10$, $p = .37$) did not predict the relative accuracy of peer-scoring.

4.3. Absolute accuracy of self and peer score judgments

4.3.1. Overall absolute accuracy

To assess absolute monitoring accuracy (to address RQ1), for each participant, a measure of mean absolute accuracy was calculated by assessing the deviation between judgments and objective performance scores (cf. Schraw, 2009). Absolute accuracy can range from −3 to +3. A smaller absolute accuracy value (i.e., closer to zero) indicates better judgment accuracy; values above zero indicate that the participant was overconfident, whereas values below zero show underconfidence. Fig. 3 shows absolute accuracy for both age groups. Mean absolute accuracy of the SSJs was 0.25 ($SD = 0.35$) for fourth graders and −0.67 ($SD = 0.25$) for sixth graders. Mean accuracy of the PSJs was 0.09 for fourth graders ($SD = 0.33$) and −0.11 for sixth graders ($SD = 0.22$). The finding that absolute accuracy values for fourth graders are above zero, whereas values for sixth graders are below zero, indicates that fourth graders had a tendency for overconfidence when scoring self- and peers, whereas sixth graders tended to be underconfident. Independent sample *t*-tests show that absolute accuracy of fourth graders SSJs was indeed significantly higher than zero, that is, they were overconfident when monitoring their own responses, $t(51) = 4.52$, $p < .001$, 95% CI = 0.13–0.34. Although not significant, when judging peers, the trend seems to indicate that they were overconfident, $t(50) = 1.93$, $p = .06$, 95% CI = 0.00–0.18. In contrast to fourth graders, the non-significant trend for sixth graders seems to indicate that they were underconfident when scoring their own responses, $t(43) = -1.92$, $p = .06$, 95% CI = −0.15–0.00, and moreover, they were significantly underconfident for PSJs, $t(42) = -3.21$, $p = .003$, 95% CI = −0.18–−0.04 (Fig. 3).

A GLM was used to investigate whether there are differences between the absolute accuracy of SSJs and PSJs (RQ1); accuracy of SSJs and PSJs were included as within-subjects repeated measurements, and to address RQ2, grade was included as a between-subject factor.

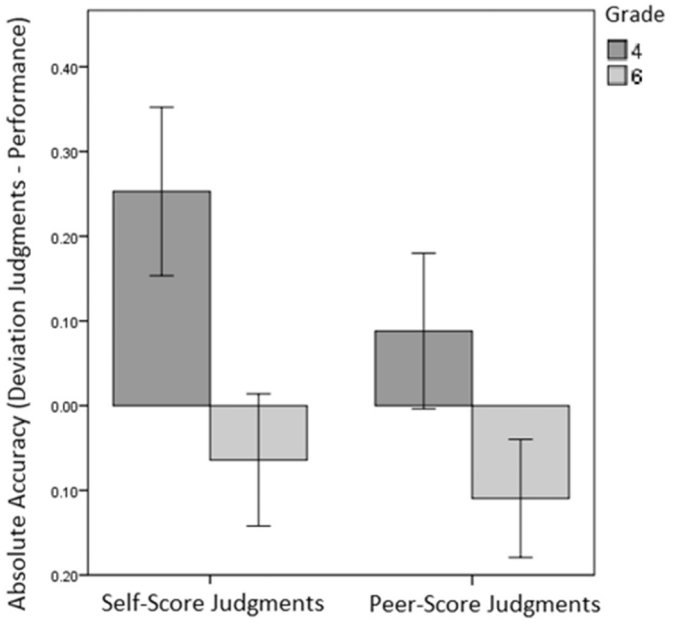

Fig. 3. Absolute accuracy of self and peer score judgments. (Error bars indicate the 95% confidence interval.)

Absolute accuracy values were higher for SSJs than PSJs, indicate higher deviation of judgments from performance, F (1, 91) = 7.67, p = .007, η_p^2 = 0.08. Moreover, the significant effect of grade, F(1, 91) = 27.29, p < .001, η_p^2 = 0.23, shows that values for absolute accuracy were higher for fourth than for sixth graders.

Moreover, regression analyses investigating effects of prior knowledge on absolute accuracy (RQ3) show that accurate prior knowledge was not related to absolute accuracy of SSJs (b = 0.07, t = 0.72, p = 47). However, inaccurate prior knowledge significantly predicted absolute accuracy (b = 0.20, t = 2.08, p = .04). That is, children with more inaccurate prior knowledge made less accurate judgments. For PSJs, effects of accurate prior knowledge were not-significant (b = −0.05, t = −0.509, p = .612), but inaccurate prior knowledge significantly predicted the accuracy of PSJs (b = 0.25, t = 2.513, p = .014), such that children with more inaccurate prior knowledge were less accurate when making judgments about peer performance.

4.3.2. Judgment magnitudes for the different response types

To address RQ1 for children's ability to monitor the quality of different test response types, SSJs and PSJs were investigated for omissions, commission errors, partially correct responses (containing either one or two correct idea units) and fully correct responses (containing all correct idea units) separately. Table 2 shows the SSJs and PSJs as a function of response type. As indicated by this table, even though omission errors were blank responses that did not contain any ideas at all, for some omissions, children judged that these contained correct ideas. A GLM analysis with SSJs and PSJs for omission errors as within-subjects factor and grade as between-subjects factor shows that there was no difference between SSJs and PSJ for omissions, F(1, 73) = 2.88, p = .09, η_p^2 = 0.04. However, the effect of grade was significant, F(1, 73) = 9.73, p = .003, η_p^2 = 0.12, such that judgments for omissions were lower for the sixth than for the fourth graders.

Children were overconfident when making SSJs for commission errors; both for fourth and sixth graders, judgments were significantly higher than zero; for fourth grade t(45) = 9.74, p < .001, for sixth grade t(16) = 5.49, p < .001. A GLM analysis showed a significant difference between SSJs and PSJs for commission errors, F(1, 60) = 7.33, p = .009, η_p^2 = 0.11, such that PSJs for commission errors were lower than SSJs. There was no significant effect of grade, p = .42.

For the partially correct responses consisting of one correct idea unit, a GLM shows that the difference between SSJs and PSJs was non-significant, F(1, 86) = 3.54, p = .06, η_p^2 = 0.04, although the trend seems to indicate that PSJs were lower than SSJs. The effect of grade was non-significant, p = .48. For the partially correct responses with two correct idea units, a GLM shows that PSJs were significantly lower than SSJs, F(1, 83) = 8.29, p = .005, η_p^2 = 0.04. Although the effect of grade was not significant, p = .51, there was a significant interaction between self-peer scoring and grade, F(1, 83) = 4.05, p = .048, η_p^2 = 0.05. Follow-up paired t-tests show that the SSJs were significantly higher than PSJs for fourth graders, t(41) = 2.82, p = .007, whereas there was no difference between these judgments for sixth graders, t(42) = 0.85, p = .40. For the fully correct responses, a GLM showed no differences between SSJs and PSJs, p = .59, and no effect of grade, p = .92.

4.4. Consistency

To investigate RQ4 about intra-individual consistency between judgments, we investigate relations between the magnitudes, as well as relations between judgment accuracy for self and peers. To investigate the relation between the magnitude of SSJs and PSJs, we calculated intra-individual gamma correlations between judgments made for oneself and one's peer per item. These correlations were 0.11 (SD = 0.64) for fourth graders and 0.10 (SD = 0.56) for sixth graders, and both gamma correlations were not higher than zero, ps > 0.23. This shows that there was no significant relation between judgment magnitudes for oneself and for peers. There was no difference between the grades, t(87) = 0.02, p = .99.

If relative accuracy for SSJs and PSJs would be consistent, this would indicate that children who have higher relative monitoring accuracy when scoring own responses also have higher monitoring accuracy when scoring peers, and vice versa. However, Pearson correlations between relative accuracy of SSJs and PSJs were low and not significant; overall r = −0.028, p = .791; r = −0.10 for fourth grade, and −0.11 for sixth grade. Further, we investigated whether measures of absolute accuracy were related between one's SSJs and PSJs. Pearson correlations between absolute accuracy values of SSJs and PSJs were significantly higher than zero, overall r = 0.365, p < .001; r = 0.15, p = .29 for fourth grade and r = 0.52, p < .001 for sixth grade. This implies consistency between absolute accuracy when scoring self and when scoring peers for sixth, but not for fourth graders.

Table 2
Self- and peer-score-judgments per test response type.

		Omission error	Commission error	1 idea (partially correct)	2 ideas (partially correct)	3 ideas (fully correct)
Grade 4	N	47	46	51	45	10
	SSJ	0.56 (0.67)	0.95 (0.66)	1.22 (0.47)	1.91 (0.53)	2.47 (0.83)
	PSJ	0.38 (0.49)	0.75 (0.52)	1.08 (0.55)	1.66 (0.68)	2.67 (0.70)
Grade 6	N	35	19	40	44	38
	SSJ	0.13 (0.40)	0.82 (0.61)	1.10 (0.36)	1.87 (0.31)	2.72 (0.34)
	PSJ	0.12 (0.30)	0.66 (0.60)	1.10 (0.38)	1.82 (0.34)	2.62 (0.45)

Note. Standard deviations of the mean in parentheses.

5. Discussion

With the present study, we aimed to investigate whether elementary school children are able to make accurate monitoring judgments when scoring test responses of themselves and their peers. To support children to make accurate judgments, they could inspect detailed feedback parsed into separate idea units. Lipko et al. (2009) showed that this idea-unit feedback can be highly beneficial for children when self-scoring; yet, it remains an open question to what extent it benefits peer judgment accuracy. To investigate potential developmental differences in judgment accuracy, fourth and sixth graders were compared. With use of pretest scores, we investigated effects of individual differences in prior knowledge on judgment accuracy. Finally, we addressed to what extent there was consistency between children's self and peer judgments.

5.1. Self and peer monitoring

RQ1 addresses whether there are differences in monitoring judgments and monitoring accuracy between SSJs and PSJs. To answer this question, children's judgment accuracy was investigated with measures of relative accuracy (a correlational measure assessing how well children can discriminate between responses of high quality and responses of lower quality, Nelson, 1984) and measures of absolute accuracy (quantifying whether judgments deviate from performance, these measures either indicate accurate monitoring, over- or under-confidence, Pieschl, 2009). Combining these two indices of judgment accuracy may give fine-grained insights into children's skills and limitations when assessing performance quality.

Our hypothesis that judgments are more accurate for peers than for oneself is only confirmed with measures of absolute accuracy, and not with measures of relative accuracy. Replicating Lipko et al. (2009) and Van Loon and Roebers (2017), children's relative accuracy was high when scoring their own and their peers' performance. This shows that when inspecting feedback, children were well able to distinguish between responses of high and low quality for themselves as well as for peers. However, even though children could use feedback standards to score the test responses, only few children were able to accurately score all responses (8.3% of the sample was fully accurate when scoring own responses and 7.3% when scoring peer responses). Measures of absolute accuracy may give more fine-grained insights into the match between children's judgment magnitudes and the objective scoring criteria. Findings show that children gave higher scores to their own answers than to the answers that were given by their peers, and absolute judgment accuracy was better for PSJs than for SSJs.

The fact that judgment magnitudes were lower and more conservative for peers than for oneself benefitted judgment accuracy for commission errors when scoring peers. The analyses for the different test response types show that children were particularly overconfident for their own commission errors. Even though these responses were entirely inaccurate, children gave themselves credit for these errors when making SSJs. Although the feedback standards showed the exact idea units a response should consist of, it seemed that children may have interpreted it such that it fitted their answers. Interestingly, when scoring the responses of their peers, children were better able to recognize that commission errors were incorrect, and that these should not receive credit. This may indicate that they were better able to apply the content of the feedback standards when scoring PSJs.

Judgments for oneself seem to be based on experiential cues related to subjective insights into learning and retrieval (Koriat & Ma'ayan, 2005). Possibly, when making commission errors, experiences that test responses are retrieved from memory and subsequently written down may hinder children to make accurate judgments (Benjamin, Bjork, & Schwartz, 1998). That is, when an answer is easily produced, persons often believe that this answer then must be correct. However, such experiential cues do not seem to play a similar role when children judge a peer (as they had no information about the learning and retrieval experiences of their peer). This may make it easier to recognize commission errors made by others. Furthermore, when scoring peer performance, children may have been less vulnerable to a wishful thinking bias (Destan et al., 2017). Even though it may have been challenging to make effective use of feedback when scoring own answers due to wishes to perform well, these wishes may not have hindered children to apply the feedback when judging peer performance. Further, better-than-average biases (Alicke & Govorun, 2005) and a tendency to conflate effort with ability (Kurtz-Costes et al., 2005) may only play a role when judging oneself, but not when judging others.

Also when scoring partially correct responses, children gave themselves more credit than they gave their peers. But although judgments for others' responses were lower than judgments for themselves, this did not necessarily reflect a better ability to judge performance for peers. That is, when judging the quality of partially correct responses, children became too strict and underconfident when scoring their peers' performance.

5.2. Developmental differences in judgment accuracy

We investigated whether there were developmental differences in judgment accuracy (RQ2) and hypothesized that the sixth grade children would make more accurate judgments for themselves and for their peers than fourth graders. This hypothesis was confirmed with measures of relative accuracy. Sixth graders were more accurate than fourth graders, such that they more systematically assigned low credit points to less learned concepts, and higher points to better learned concepts. However, although measures of relative accuracy show better discrimination skills for the older age group, recognition of commission errors did not improve for the older children, indicating that this remained challenging for both age groups. For complex metacognitive judgment tasks, developmental effects are often not seen until children are 12 years and older (Roebers, 2017). The finding that overconfidence for commission errors did not improve with age, may indicate that for children, commission errors are the hardest to notice and to correct. To address this assumption, future research could include older age groups (e.g., compare elementary school children with adolescents) to investigate development of error recognition skills.

Moreover, analyses of absolute accuracy show effects of grade for self and peer judgments. For SSJs, absolute accuracy was better for sixth than for fourth graders. Fourth graders were overconfident and gave themselves more credit when judging themselves than sixth graders. However, although children were more conservative when scoring their peers than when scoring themselves, for fourth graders, being more conservative for peers led to lower, and more accurate PSJs than SSJs, whereas sixth graders were underconfident when judging performance of their peers. That is, although older children gave lower judgments when judging performance than younger children, this did not necessarily benefit judgment accuracy. Instead, lowering judgments for peers led sixth graders to become too strict.

5.3. Effects of prior knowledge on judgment accuracy

RQ3 concerns effects of prior knowledge on SSJs and PSJs; to address this question, we investigated effects of accurate and inaccurate prior knowledge. The hypothesis that prior knowledge would affect judgments for oneself was only partially confirmed. Relative accuracy was not affected by children's prior knowledge. However, although absolute accuracy of SSJs was not affected by accurate prior knowledge, children's inaccurate prior knowledge affected absolute accuracy. That is, inaccurate prior knowledge led to more confident scoring and overestimation of the quality of answers when making judgments about one's own performance. Interestingly, and in line with Cameron et al. (1997), inaccurate prior knowledge did not only lead to more confidence when monitoring oneself, but also when judging peers. This

study is the first to show that children's inaccurate prior knowledge may not only be disadvantageous for self-monitoring (cf. Van Loon et al., 2013a), but that these disadvantages may also transfer when children judge performance of others.

5.4. Consistency between self- and peer-monitoring

Our fourth research question addresses consistency within students of self- and peer judgment magnitudes and judgment accuracy. The magnitudes of self- and peer judgments for the different test items were not consistent within persons. This shows that children clearly differentiated between their own and their peers' performance, and did not give peers similar scores as they gave themselves. Further, the finding that indices of relative accuracy were not related when scoring oneself and when scoring a peer indicate that children who were well able to discriminate between correct and incorrect responses for themselves could not necessarily do so for their peers, and vice versa.

Interestingly, measures of absolute judgment accuracy were consistent between SSJs and PSJs for sixth, but not for fourth graders. Although relations between absolute accuracy of self and peer judgments were low and nonexistent for fourth graders, for sixth graders, the correlation between absolute accuracy values for SSJs and PSJs was moderate to high. Possibly, consistency in absolute accuracy may develop between fourth and sixth grade. Van der Stel and Veenman (2008) showed that around the age of 12, metacognitive skills seem to develop into a general competence that is applicable across tasks and domains. Although previous research showed domain generality of metacognitive accuracy when scoring one's own performance on different kinds of tasks (Kleitman & Moscrop, 2010), the present research may imply that this general metacognitive competence may also carry on when scoring others. That is, our findings showed that around the age of 12, children who were more optimistic when scoring themselves were also more optimistic when scoring their peers. Future research should investigate whether these findings on potential development in consistency of absolute accuracy can be replicated.

5.5. Limitations and suggestions for future research

In the present study, we used actual educational content for both age groups. Therefore, task materials were different, and further, sixth graders studied more concepts than fourth graders. Although use of these tasks may make our findings educationally relevant, using different materials for both age groups also brings limitations when aiming to compare the groups. Although the findings on better self-monitoring accuracy for older than younger children replicate previous developmental research on monitoring accuracy (Schneider & Löffler, 2016), our findings on monitoring accuracy need to be interpreted with care. Although we had a large sample size and analyses of relative and absolute accuracy included the full sample, the analyses for the different response types only included a part of the sample. That is, few fourth graders made fully correct responses that contained all the three learned idea units, and several sixth grade students did not make commission errors. Further research is needed to acquire insight into replicability and generalizability of these findings, particularly findings for the different response types.

Moreover, it should be noted that metacognitive development cannot be considered in disconnection from cognitive development. That is, although sixth graders were less overconfident than fourth graders, they also learned more concepts as reflected by higher task performance. Further, although not measured, presumably sixth graders were more advanced than fourth graders when it comes to reading and writing skills, memory capacity, attention span, and processing speed (Bjorklund & Causey, 2017). It also seems likely that older children were better able to utilize the feedback in a more efficient way than the younger children. Fourth graders seemed to make more mistakes than sixth graders when scoring responses, as indicated by the finding that they sometimes awarded credit when no response was produced at all (i.e., omissions errors). This may indicate that for the younger children, it was more demanding to simultaneously inspect feedback and task responses and to mark the check box. Because cognitive and metacognitive development are likely to interact, assumptions about improvement of monitoring accuracy skills for the older age group need to be evaluated cautiously, and these need to be replicated with similar as well as other types of tasks.

Research shows that first judging the outcomes of a peer may support children to subsequently monitor their own learning progress and to improve learning achievement (Hwang et al., 2014). In this study, the order of making self- and peer judgments was randomized, that is, half of the children first made SSJs whereas the other half first made the PSJs. Comparison between these two groups did not show any differences in judgment accuracy. This seems to indicate that only asking children to engage in peer scoring before self-scoring is not sufficient to improve monitoring judgments. Findings also imply that children did not automatically adjust their own judgments after being presented with judgments that were made by a peer. In the present study, children were not specifically instructed to reflect on their judgment experiences, and to inspect and make use of the judgments made by their peer. Possibly, to benefit from collaborative scoring activities, children need specific instructions about how to use the judgments that are made by others (Hurme, Palonen, & Järvelä, 2006). Future research should investigate effects of instructions to explicitly compare judgments on scoring accuracy and collaborative learning. Moreover, future research should address whether children would also continue to inspect, judge and compare their own and their peers' learning when they are not explicitly prompted to do so.

Furthermore, in this study, we did not acquire insights into the social relations between peers who scored each other's work, and whether they considered each other as friends. This may influence to what extent children are more critical about their peers than about themselves; elementary school children may only show a self-serving bias when assessing a non-friend, but not when assessing a friend's performance (Posey & Smith, 2003). Future research could investigate how interpersonal relations between peers affect the judgment process and outcomes.

5.6. Implications and conclusions

When aiming to target better recognition of commission errors, designing collaborative learning activities during which children have to score each other's answers may be promising. Commission errors are hard to detect and improve (Rawson & Dunlosky, 2007), and our findings show that peer-judgments for these errors are clearly more accurate. It seems that children may find it more motivating to catch their peers' mistakes than their own (confirming findings by Okita, 2014). Therefore, peer-scoring activities may be valuable to support and improve error detection. The finding that children can accurately discriminate between correct and incorrect performance of their peers' and recognize others' errors indicates that performance does not always need to be judged by teachers, but that students themselves could collaborate on scoring. This could potentially free up time resources for teachers. Moreover, when using online tutoring systems, making use of peer feedback could be a valuable possibility to give children insight into how well they are progressing on tasks. Online tutoring systems can give detailed performance feedback when using multiple choice questions and when learning memory materials (such as when learning second language vocabulary). However, when more complex learning materials are used and students have to show that they understood the gist, such systems are less effective in giving feedback. Probably, peer feedback may be more useful with such kinds of tasks.

Further, learning from and with peers may also have long-term impacts on children's social learning skills, and their ability to consider different perspectives. Nevertheless, it is important to note that our

findings show that children are sometimes too strict with scoring when their peers' test responses have high quality. This indicates that, although such activities are likely to be very useful to improve collaborative learning, at the same time, teachers should instruct children how to use scoring criteria, and they should closely monitor the quality of self and peer scoring.

To conclude, children were better at error detection when inspecting others' work than when scoring their own performance. However, they were also more conservative for their peers, and underestimated peer performance more than their own performance. Although sixth graders were more accurate in discrimination between high and low quality responses than fourth graders, at the same time, they were too strict for their peers. Moreover, children's inaccurate prior knowledge led to less accurate and more overconfident judgments, both when judging themselves and their peers. Further, consistency of judgment accuracy may be developing; absolute accuracy of judgments was consistent between scores for oneself and a peer for sixth, but not for fourth graders. In sum, the present findings imply that, when inspecting feedback standards, children are able to judge both their own and their peers' responses. However, when implementing co-scoring activities, it is important to be aware that its accuracy is affected by individual differences in children's age and prior knowledge, particularly inaccurate prior knowledge.

Acknowledgement

The authors would like to thank Nike Tsalas for comments on a previous version of the manuscript, and Christian Baur and Alena Jeremias for support with task design and data collection.

References

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning, & J. I. Krueger (Vol. Eds.), *The self in social judgment. Vol. 1. The self in social judgment* (pp. 85–106). New York: Psychology Press.

Baars, M., Van Gog, T., De Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*(3), 382–391. https://doi.org/10.1002/acp.3008.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology. General, 127*(1), 55–68. https://doi.org/10.1037//0096-3445.127.1.55.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823.

Bjorklund, D. F., & Causey, K. B. (2017). *Children's thinking: Cognitive development and individual differences.* Sage Publications.

Braithwaite, D. W., & Goldstone, R. L. (2015). Effects of variation and prior knowledge on abstract concept learning. *Cognition and Instruction, 33*(3), 226–256. https://doi.org/10.1080/07370008.2015.1067215.

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

Cameron, C. A., Edmunds, G., Wigmore, B., Hunt, A. K., & Linton, M. J. (1997). Children's revision of textual flaws. *International Journal of Behavioral Development, 20*(4), 667–680. https://doi.org/10.1080/016502597385117.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354.

De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294–310. https://doi.org/10.1016/j.jecp.2011.02.005.

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning, 10*(3), 347–374. https://doi.org/10.1007/s11409-014-9133-z.

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, 12*(3), 315–336. https://doi.org/10.1007/s11409-017-9170-5.

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology, 47*(2), 274–296. https://doi.org/10.1016/0022-0965(89)90033-7.

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. https://doi.org/10.1080/17470218.2010.502239.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Thousand Oaks, CA: Sage Publications.

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

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*, 551–565. https://doi.org/10.1016/j.jml.2005.01.011.

Efklides, A. (2008). Metacognition defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277–287. https://doi.org/10.1027/1016-9040.13.4.277.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002.

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction, 32*, 1–9. https://doi.org/10.1016/j.learninstruc.2014.01.001.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*(7), 1001–1013. https://doi.org/10.3758/MC.37.7.1001.

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

Hurme, T.-R., Palonen, T., & Järvelä, S. (2006). Metacognition in joint discussions: An analysis of the patterns of interaction and the metacognitive content of the networked discussions in mathematics. *Metacognition and Learning, 1*(2), 181–200. https://doi.org/10.1007/s11409-006-9792-5.

Hwang, G.-J., Hung, C.-M., & Chen, N.-S. (2014). Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach. *Educational Technology Research and Development, 62*(2), 129–145. https://doi.org/10.1007/s11423-013-9320-7.

Kleitman, S., & Moscrop, T. (2010). Self-confidence and academic achievements in primary-school children: Their relationships and links to parental bonds, intelligence, age, and gender. In A. Efklides, & P. Misailidi (Eds.). *Trends and prospects in metacognition research* (pp. 293–326). Boston, MA: Springer.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology. General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349.

Koriat, A., & Ackerman, R. (2010b). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition, 19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010.

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The easily learned, easily remembered heuristic in children. *Cognitive Development, 24*(2), 169–182. https://doi.org/10.1016/j.cogdev.2009.01.001.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*(4), 478–492. https://doi.org/10.1016/j.jml.2005.01.001.

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. https://doi.org/10.1348/000709910X485719.

Kurtz-Costes, B., McCall, R. J., Kinlaw, C. R., Wiesen, C. A., & Joyner, M. H. (2005). What does it mean to be smart? The development of children's beliefs about intelligence in Germany and the United States. *Journal of Applied Developmental Psychology, 26*(2), 217–233. https://doi.org/10.1016/j.appdev.2004.12.005.

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. https://doi.org/10.1037/a0017599.

Lyons, K. E., & Ghetti, S. (2013). I don't want to pick! Introspection on uncertainty supports early strategic behavior. *Child Development, 84*(2), 726–736. https://doi.org/10.1111/cdev.12004.

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician, 36*, 158–160. https://doi.org/10.1080/00031305.1982.10482818.

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. https://doi.org/10.1007/s11409-011-9083-7.

Mory, E. H. (2004). Feedback research revisited. In M. J. Spector, D. M. Merrill, J. Van Merriënboer, & M. P. Driscoll (Vol. Eds.), *Handbook of research on educational communications and technology. Vol. 2. Handbook of research on educational communications and technology* (pp. 745–783). New York: Taylor & Francis Group.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102–122. https://doi.org/10.1080/02602938.2013.795518.

Okita, S. Y. (2014). Learning from the folly of others: Learning to self-correct by monitoring the reasoning of virtual characters in a computer-supported mathematics learning environment. *Computers & Education, 71*, 257–278. https://doi.org/10.1016/j.compedu.2013.09.018.

Paulus, M., Tsalas, N., Proust, J., & Sodian, B. (2014). Metacognitive monitoring of oneself and others: Developmental changes during childhood and adolescence. *Journal of Experimental Child Psychology, 122*, 153–165. https://doi.org/10.1016/j.jecp.2013.12.011.

Pieschl, S. (2009). Metacognitive calibration-an extended conceptualization and potential applications. *Metacognition and Learning, 4*(1), 3–31. https://doi.org/10.1007/

s11409-008-9030-4.
Posey, E., & Smith, R. A. (2003). The self-serving bias in children. *Psi Chi Journal of Undergraduate Research, 8*, 153–156.
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. https://doi.org/10.1080/09541440701326022.
Roebers, C. M. (2002). Confidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology, 38*(6), 1052–1067. https://doi.org/10.1037/0012-1649.38.6.1052.
Roebers, C. M. (2014). Children's deliberate memory development: The contribution of strategies and metacognitive processes. In P. J. Bauer, & R. Fivush (Vol. Eds.), *The Wiley handbook on the development of children's memory. Vol. Volume I/II. The Wiley handbook on the development of children's memory* (pp. 865–894). Chichester: John Wiley & Sons, Ltd.
Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review, 45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001.
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. https://doi.org/10.1016/j.lindif.2012.12.003.
Roebers, C. M., & Spiess, M. (2017). The development of metacognitive monitoring and control in second graders: A short-term longitudinal study. *Journal of Cognition and Development, 18*(1), 110–128. https://doi.org/10.1080/15248372.2016.1157079.
Ruble, D. N., Eisenberg, R., & Higgins, E. T. (1994). Developmental changes in achievement evaluation: Motivational implications of self-other differences. *Child Development, 65*(4), 1095–1110. https://doi.org/10.1111/j.1467-8624.1994.tb00805.x.
Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1.
Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking. *Developmental Science, 1*(2), 291–297. https://doi.org/10.1111/1467-7687.00044.
Schneider, W., & Löffler, E. (2016). The development of metacognitive knowledge in children and adolescents. In J. Dunlosky, & R. A. Bjork (Eds.). *The Oxford handbook of metamemory* (pp. 191–517). Oxford, UK: Oxford University Press.
Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3.
Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences, 10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009.
Stipek, D. J. (1984). Young children's performance expectations: Logical analysis or wishful thinking. *Advances in Motivation and Achievement, 3*(3).
Stipek, D. J., & Hoffman, J. M. (1980). Development of children's performance-related judgments. *Child Development, 51*(3), 912–914. https://doi.org/10.2307/1129485.
Taub, M., Azevedo, R., Bouchet, F., & Khosravifar, B. (2014). Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior knowledge in hypermedia-learning environments? *Computers in Human Behavior, 39*, 356–367. https://doi.org/10.1016/j.chb.2014.07.018.
Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66.
Thomas, R. C., & Jacoby, L. L. (2013). Diminishing adult egocentrism when estimating what others know. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 473–486. https://doi.org/10.1037/a0028883.
Van den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: The role of co-activation in confronting misconceptions. *Applied Cognitive Psychology, 22*(3), 335–351. https://doi.org/10.1002/acp.1418.
Van der Stel, M., & Veenman, M. V. (2008). Relation between intellectual ability and metacognitive skillfulness as predictors of learning performance of young students performing tasks in different domains. *Learning and Individual Differences, 18*(1), 128–134. https://doi.org/10.1016/j.lindif.2007.08.003.
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. https://doi.org/10.1016/j.jecp.2017.01.008.
Van Loon, M. H., De Bruin, A. B., Van Gog, T., & Van Merriënboer, J. J. (2013b). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning, 8*(2), 173–191. https://doi.org/10.1007/s11409-013-9100-0.
Van Loon, M. H., De Bruin, A. B. H., Van Gog, T., & Van Merriënboer, J. J. G. (2013a). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15–25. https://doi.org/10.1016/j.learninstruc.2012.08.005.
Van Loon, M. H., & Roebers, C. M. (2017). Effects of feedback on self-evaluations and self-regulation in elementary school. *Applied Cognitive Psychology, 31*(5), 508–519. https://doi.org/10.1002/acp.3347.
Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children, 23*(3), 34–41.
Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In W. J. McKeachie (Ed.). *Handbook of self-regulation* (pp. 13–39). Elsevier.