The impact of retrieval processes, age, general achievement level, and test scoring scheme for children's metacognitive monitoring and controlling

Saskia Susanne Krebs · Claudia Maria Roebers

Abstract This multi-phase study examined the influence of retrieval processes on children's metacognitive processes in relation to and in interaction with achievement level and age. First, $N=150$ 9/10- and 11/12-year old high and low achievers watched an educational film and predicted their test performance. Children then solved a cloze test regarding the film content including answerable and unanswerable items and gave confidence judgments to every answer. Finally, children withdrew answers that they believed to be incorrect. All children showed adequate metacognitive processes before and during test taking with 11/12- year-olds outperforming 9/10-year-olds when considering characteristics of on-going retrieval processes. As to the influence of achievement level, high compared to low achievers proved to be more accurate in their metacognitive monitoring and controlling. Results suggest that both cognitive resources (operationalized through achievement level) and mnemonic experience (assessed through age) fuel metacognitive development. Nevertheless, when facing higher demands regarding retrieval processes, experience seems to play the more important role.

Keywords Metacognitive development · Achievement level · Strategic regulation · Test performance · Monitoring · Control processes

Theories of self-regulated learning emphasize the importance of monitoring, control processes, and individual differences, for improving learning outcomes and for describing the learner's activities during learning and test taking (e.g., Boekaerts 1999; Hong et al. 2009; Pintrich 2004). Monitoring relates to an on-line attentiveness towards learning progress and current level of mastery (Nelson and Narens 1994; Schneider and Pressley 1997). It is considered to be one of the most important aspects of metamemory, as it can stimulate the execution of self-regulated control processes (Goldsmith and Koriat 2008; Nelson and Narens 1994). There is ample evidence that monitoring and control processes (metacognitive processes) influence adults' and children's test performance (e.g., Hacker et

S. S. Krebs (✉) · C. M. Roebers
Department of Psychology, University of Bern, Bern, Switzerland
e-mail: saskia.krebs@psy.unibe.ch

al. 2008; Hacker et al. 2000; Thiede et al. 2003; Veenman and Spaans 2005; Veenman et al. 2004). Yet, only a few studies investigated children's metacognitive processes and its impact on academic outcomes in naturalistic, school-related contexts, such as memory reporting during test taking. Further, studies have shown that metacognitive processes undergo substantial improvements during the elementary school years (for a review, see Schneider and Lockl 2008; Veenman et al. 2006). However, the exact nature of this development is not yet fully understood. The question arises whether it is the increasing mental capacity or the growing experience leading to superior metamnemonic awareness that drives the development of metacognitive skills. Contextual factors, such as teachers' mnemonic orientation in their instructions, have also been shown to play a role in metacognitive development (e.g., Coffman et al. 2008). The purpose of this study was to examine the development of metacognitive processes and its potentially influencing factors among children during test taking from an individual differences approach.

Self-regulation is more than metacognition; it is a complex interactive process involving metacognitive, motivational, and behavioural components which are affected by contextual variables such as task features and setting conditions (Zimmerman 1995). One theoretical framework addressing metacognitive processes in memory reporting including influences of contextual demands was proposed by Koriat and Goldsmith (1996). It postulates that memory accuracy depends on the situational demands, such as accuracy incentives and report option, and on monitoring effectiveness. The product of retrieval and monitoring results in a "best candidate" answer. Whether this answer will be given or withheld (controlling) depends on the subjective judgment of correctness and on the situational demands. Accordingly, metacognitive processes have an impact on test behaviour and on academic outcomes and are sensitive to situation and task specifics (Krebs and Roebers 2010; Nietfeld et al. 2006a; Roderer and Roebers 2009). Both in children (e.g., Schneider et al. 2000) and in adults (e.g., Hacker et al. 2000), metacognitive processes during learning have been investigated, however, only few studies applied these aspects to test taking. The following paragraphs will briefly review the literature on possible factors impacting metacognitive processes and their development with a focus on the concepts included in the abovementioned framework.

Koriat and Goldsmith's framework (1996) suggests accuracy of monitoring processes to be influenced by retrieval. Similarly, theories of self-regulated learning claim that memory or test performance are a function of underlying knowledge being activated through test questions. Based on this theoretical background, it is assumed that higher memory strength results in superior retrievability leading to better monitoring. Consequently, monitoring can be expected to be better in easy compared to difficult and in answerable compared to unanswerable items. Studies confirmed this assumption by showing that monitoring processes substantially depend on the quality of the information available for monitoring, such that monitoring will be poorer if participants only retrieved undifferentiated or vague information (e.g., Ackerman and Goldsmith 2008). Differences in retrieval processes, operationalized through various degrees of item familiarity (Metcalfe and Finn 2008), item difficulty (Nietfeld et al. 2005), or memory quantity (Rhodes and Kelley 2005), have shown to impact adults' monitoring. For children, it was documented (Lockl and Schneider 2004) that students aged 7 years and older can metacognitively differentiate (monitoring resolution) in terms of judgments of learning between easy- and hard item pairs (high versus low associative word pairs). Moreover, children's calibration—referring to the precision with which a person can postdict or predict (test) performance by means of confidence judgments or judgments of learning, respectively—tends to be more accurate for easy than for difficult items (Howie and Roebers 2007).

Similarly, regarding metacognitive controlling, learners have been shown to use retrieval processes such as assessments of item or task difficulty (Goldsmith and Koriat 2008; Koriat and Goldsmith 1996; Maki et al. 2005; Nelson and Narens 1990; Nietfeld et al. 2005) in deciding whether to allocate further cognitive resources (e.g., allocation of study time) and in predicting their performance. The decision of providing or withholding retrieved information, as well as the detection and correction of errors (e.g., Koriat and Goldsmith 1996) mirror possible control processes during test taking. In such test situations, adults substantially increase their test accuracy through efficient controlling (Goldsmith and Koriat 2008; Koriat and Goldsmith 1996). Given that controlling typically develops later than corresponding monitoring in ontogeny (for a review, see Schneider and Lockl 2008), it is not surprising that younger school children (8-year-olds) were found to effectively control their test accuracy only under optimal conditions (Roderer and Roebers 2009). However, 11/12-year-olds showed more effective and goal-orientated controlling by selectively maintaining and withdrawing answers to the overall benefit of test accuracy (Krebs and Roebers 2010; Roebers et al. 2009) and by adjusting this behaviour relative to situational demands (e.g., scoring scheme; Roebers et al. 2009) and retrieval processes (Krebs and Roebers 2010).

In sum, retrieval processes have a direct and strong impact on monitoring and controlling. Yet, analogues to the development of metacognition, much remains to be learned about the nature of the impact of item difficulty on metacognitive processes (Dunlosky and Bjork 2008; Rhodes and Kelley 2005). Moreover, this issue has not yet been addressed within the research context of educational psychology. The present study tries to shed further light on the influence of retrieval processes on children's developing metacognitive skills in an educational context by contrasting monitoring and controlling for easy, difficult, and unanswerable items in relation to achievement level and age. Including these two important sources for individual differences allows investigating the interactions between retrieval and age and/or achievement level for accurate memory reporting.

As to the question of which factors fuel developmental progression in metacognition, there are findings suggesting that schooling and the resulting increasing metacognitive experience exert a powerful influence on metacognitive skills. One way of investigating this influence is to realize a training in which children are exposed to massed experience. In this context, Melot (1998) provided 6- and 9-year-olds who participated in a training session with metacognitive knowledge about the efficiency of a strategy on their recall performance. When this knowledge was being elaborated during training it turned out to substantial improve strategy use and performance. Also, socio-communicative interactions between preschoolers and adults have been shown to be significantly related to the development of strategic skills in memory (e.g. Ornstein et al. 2004). Further, repeated testing (e.g. Koriat 1997) and repeated learning (e.g. Lovelace 1984) can positively influence adults' monitoring accuracy. In a similar vein, fifth and sixth graders showed better test performance after having been exposed to a monitoring training compared to no training or to training with no monitoring aspect. Moreover, monitoring exercises and feedback on performance substantially showed to affect class performance among undergraduate students (Nietfeld et al. 2006b).

Another source fuelling the development of metacognitive skills that is being debated in the literature are children`s growing cognitive resources. One approach of looking at the influence of cognitive resources on metacognitive skills is by comparing the performance of lower and higher achieving students. In this regard, studies with adults consistently documented that calibration is associated with individuals' achievement level. Performance

in terms of general academic ability predicts calibration (e.g., Hacker et al. 2000; Maki et al. 2005; Nietfeld et al. 2005). Typically, lower achieving students show higher overestimation (in their predictions and postdictions of performance), are thus less accurate than high achievers (e.g., Maki et al. 2005). When trying to improve calibration through incentives, Hacker et al. (2008) found that low achieving student's miscalibration could to some extent be decreased, implying that monitoring accuracy and achievement level are partly independent. However, only a few studies including children investigated the influence of achievement level on metacognitive skills. Schwenk et al. (2009) examined 4- to 8- year-olds' metamnemonic awareness regarding strategy use. Results suggested systematic individual differences in recall separating low from high performers. Importantly however, children's general achievement level is likely to interact with item or task difficulty. Thus, without additionally including item difficulty, differences between high and low achievers in metacognition are likely to be confounded with item difficulty. Further, since both age and achievement level have shown to influence metacognitive processes, the interplay between achievement level, age, and item difficulty is also of interest; depending on the degree of item difficulty, the influence of achievement level, age, and the interaction between the two on metacognition may vary. In sum, by comparing high and low achievers of two different age groups, the current study aims to shed light on the impact of experience (measured through age) and of cognitive resources (operationalized through achievement level) on the metacognitive development in children in relation to retrieval processes.

The research questions of the current research are: Among 8/9- and 11/12-year-olds, what impact have achievement level and age on the ability to monitor memory strength and to control retrieved answers? Therefore, 9/10- and 11/12- year old high and low achievers were involved in a multi-phase experiment. After having watched an educational film, children were informed about an upcoming test on the film's topic and asked to predict their test performance. One week later, children first watched the film again before filling in every blank of a cloze test (forced report) that consisted of easy, difficult, and unanswerable items regarding the film. On a 7-point Likert scale, children gave confidence judgments to every answer indicating how certain they were that the answers were correct. Finally, children had the option to cross out answers they believed to be incorrect (free report). By comparing test accuracy in the forced and free report phase and by contrasting the monitoring judgments across items of different degrees of difficulty (easy, difficult, and unanswerable items), it is not only possible to investigate the relation between monitoring and controlling and to quantify gains in memory performance, but also to shed light on the impact of retrieval (operationalized through item difficulty) on metacognitive processes. In order to increase children's controlling, a scoring scheme for maintaining or crossing out answers was implemented (e.g., Roderer and Roebers 2009; Roebers et al. 2009).

Based on abovementioned findings, monitoring was hypothesized to be relatively adequate in 9/10- and 11/12-year-olds, while controlling undergoes important further improvements in this age range. In more details, prediction of performance, calibration, and monitoring resolution were expected to be better in easy compared to difficult items. As to the impact of general achievement level, high achievers were expected to predict their performance more precisely, and to metacognitively distinguish (monitoring resolution) more adequately between easy, difficult, and unanswerable items as well as between correct and incorrect responses than low achievers. High achievers were also expected to better control their test performance by maintaining and withdrawing answers more selectively compared to low achievers. Moreover, when comparing the two age groups, differences between high and low achievers' performance in relation to retrieval processes were of

major interest since they may uncover factors fueling metacognitive development. While main effects of achievement level independent of age would suggest a growing cognitive capacity to be a possible major developmental drive, a main effect of age independent of achievement level would indicate that experience might be one of the main developmental sources. Interactions between achievement level and age would suggest that both aspects are relevant regarding developing metacognitive skills and interactions between item difficulty and achievement level and/or age would specify the developmental advantage of age and achievement level respectively.

Method

Sample

Out of a larger sample including 283 children, $N=150$ (49% female, 51% male) children from public schools of Bern and Luzern, Switzerland, completed the study. Seventy-seven children were 9/10- years (mean age=9 years; 7.7 months; $SD=6.9$ months) and $N=73$ 11/12- years old (mean age=11 years; 6.4 months; $SD=6.3$ months). Based on teacher ratings in various domains, children were divided into low, medium, and high achievers (lower, medium, and upper 33% of the achievement level distribution), accordingly, analyses revealed that recall performance of high compared to low achievers was better, $F (1, 146)=17.55$, $p<.001$, $\eta^2=.11$. In the study, only low ($N=67$) and high achievers ($N=83$) were included. In the run-up to the study, parents gave written consent and children provided oral approval of participation. Not all participants were native Swiss German speakers; however, all children had a sufficient level of German language ability to follow school instruction. Children were randomly assigned to the control ($N=77$) and scoring ($N=73$) conditions. In terms of experimental condition and achievement level, the groups did not differ in their chronological age, sex, and native language. This resulted in cell sizes of $N=15$ to 22 participants.

Material

Prior to the study, for every class to be tested, on scales from 1 to 5 (1 = sufficient; 2 = slightly under average; 3 = average, 4 = slightly above average; 5 = very good), teachers rated every student in linguistic, scientific, and mathematical based subjects (Cronbach's $\alpha=.90$).

A major aim of this study was to establish an experimental paradigm within the field of educational psychology that allows the investigation of metacognition. Therefore, based on the data of previous studies (Krebs and Roebers 2010; Roderer and Roebers 2009; Roebers et al. 2009), a written school-like cloze test with 24 items was used. This test was based on a German educational film (7 min) on sugar production from the children's educational TV program 'Die Sendung mit der Maus' which the children saw prior to test taking. Shortly before solving the test, by means of a questionnaire and without seeing the actual items, children predicted their performance for easy and difficult questions respectively through estimating the number of items that they would solve correctly (e.g., Afterwards, we will ask you eight easy questions regarding the movie you just saw. How many of these eight questions do you think will you answer correctly?) . The cloze test included 17 answerable items (e.g., The first step in harvesting sugar beets is to cut off _____.) subdivided into $N=8$ easy, $N=6$ difficult, and $N=3$ items of medium difficulty. Analyses of item

difficulty revealed 0.83, 0.63, and 0.29 mean item difficulties for easy, medium, and difficult items, respectively. Additionally, $N=7$ unanswerable items regarding details not provided in the film (eg. ____ pieces of sugar beets are being cut per hour), were integrated. On 7-point Likert scales with Smiley illustrations (Roebers 2002) from 1 to 7 (1 = absolutely unconfident; 4 = indecisive; 7 = very confident) displayed next to every item, children gave confidence judgments of correctness to every answer.

Design and procedure

A 2 (age: 9/10- and 11/12-year-olds) × 2 (experimental condition: control and scoring conditions) × 3 (item format: easy vs. difficult vs. unanswerable items) × 2 (achievement level: low vs. high) factorial design was employed. By means of the G* Power program (Faul et al. 2007) a power analysis was conducted. Based on previous findings (Krebs and Roebers 2010; Roderer and Roebers 2009; Roebers et al. 2009) we anticipated medium (to large) effects for the between- subjects factor experimental condition and for the within-subject factor item difficulty. Regarding achievement level, we also expected medium to large effects (Hacker et al. 2008). Results of the power analyses revealed that for detecting medium to large effects with a likelihood of 80%, a sample size of $N=70$–156 is needed.

Two experimenters visited the class and presented the educational film. Given that earlier findings (Michel et al. 2007) with the same paradigm revealed superior knowledge acquisition for later recall when showing the educational movie twice compared to once, the movie was shown again 7 days later in smaller groups of 5 to 8 children. Children then predicted their memory performance regarding easy and difficult items on the film topic in terms of estimated numbers of correct answers. During test taking, for each phase pencils of different colours were used in order to avoid measurement confusions. In the first and forced report phase, children filled in every blank of a cloze test with a blue pencil before unfolding a part of each page to uncover the confidence scales. After the experimenters had explained how to use the 7-point Likert scales, participants gave confidence judgments for every question with a green pencil. Finally, the green pencils were replaced by red ones in order to allow crossing out answers that were thought to be incorrect (free report). At that time, the experimental manipulation was implemented and either a scoring (scoring condition) or no scoring scheme (control condition) was introduced.

While children in the control condition were informed that they could cross out answers they believed to be incorrect, a scoring scheme was introduced in the scoring condition. To ensure that the scoring scheme was incentive enough to improve controlling, and based on recent findings (Krebs and Roebers 2010; Roebers et al. 2009), following scoring scheme was chosen: get 1 credit for each correct answer, lose 3 credits for each incorrect answer (+1: −3). Children were told they could earn credits, which would be exchanged with a present at the end of the experiment. Presents of different levels of attractiveness were shown. The experimenters explained that participants with the most credits may choose a present first, while those with fewer credits would have to choose among the remaining ones. Children were told that for each correct answer one credit would be given, for each incorrect answer 3 credits would be subtracted, and for each crossed-out answer no credits would be received. The experimenter emphasized to thoroughly think about how to earn as many credits as possible but no direct strategy was given. Altogether, children understood the instructions straightaway. At last, all participants received positive feedback and picked a gift of their choice.

Dependent measures

Test accuracy in the forced report phase was assessed through the number of correct answers divided by the total number of items in the cloze test. For test accuracy in the free report phase, the number of correct answers maintained was divided by the overall number of maintained answers (correct and incorrect answers). Based on previous studies (Koriat 2008; Stankov and Lee 2008), calibration (prediction accuracy) was calculated through the absolute value of the difference between performance prediction and actual recall performance in the forced report phase with positive values indicating overestimation and negative values underestimation. Therefore, predictions and the amount of correct answers were converted into percentages. To examine monitoring resolution, the difference in mean confidence judgments between correct and incorrect answers was computed. Distinction of item format was computed through subtracting mean confidence judgments for answers to unanswerable items from mean confidence judgments for answers to answerable item. As to controlling for answerable items, the number of answers correctly maintained or withdrawn was calculated in percentages. To allow comparison between controlling for unanswerable and controlling for answerable items, the mean percentage of incorrect answers crossed-out in the free report phase was calculated.

Results

Preliminary analyses indicated no systematic gender differences; therefore data were collapsed across gender. To allow direct effect comparisons across dependent variables, estimators of effect sizes (partial eta²- values) will be reported. Since previous studies have shown substantial quality increase in control processes when an incentive is available (Krebs and Roebers 2010; Roebers et al. 2009), the focus of this study lies on the scoring condition. When it comes to controlling, readers are reminded that the experimental manipulation only took place during free report; hence, whereas monitoring was investigated for the entire sample, controlling was examined for the scoring condition only in order to present a focused view on the targeted issues (this was also done to facilitate presentation of results). In general, in the forced report phase, between 16.7% and 90.5% of the items were answered correctly providing a satisfactory database for examining the involved metacognitive processes. Additionally, children's overall test accuracy benefitted from the withdrawing option: test accuracy from forced to free report increased significantly by .07, $F (1, 149)=81.77$, $p<.001$, $\eta^2=.35$. To shed light on the possible causes of this accuracy boost, following analyses will investigate the involved metacognitive processes in relation to age, achievement level, and item difficulty in more details.

In Fig. 1, monitoring prior to test taking, namely calibration, is displayed. It can be seen that all children, except older high achievers within easy items, showed overestimation. Results revealed better calibration for easy than for difficult items, $F (1, 145)=66.91$, $p<.001$, $\eta^2=.32$, and high achievers were overall better calibrated than low achievers, $F (1, 145)= 12.61$, $p<.01$, $\eta^2=.08$. Moreover, a main effect of age, $F (1, 145)=124.23$, $p<.001$, $\eta^2=.12$, was found which was specified through a two-way interaction between item difficulty and age, $F (1, 145)=5.07$, $p<.05$, $\eta^2=.04$, indicating that 11/12-year-olds' overall superior calibration was due to better performance in difficult compared to easy items.

Readers are reminded that the experimental manipulation took place in the free report only, therefore, regarding monitoring during test taking, all children are included in the analyses regardless of the experimental condition. In order to examine children's
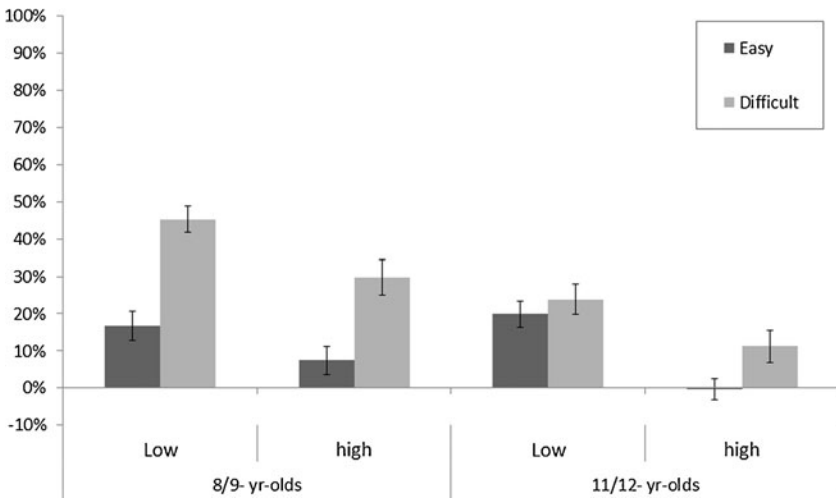
Fig. 1 Calibration in percentage as a function of age, achievment level, and item difficulty. All scores differed significantly from zero at the 0.01 level. Error bars represent standard errors of the mean

monitoring of answerable items during test taking, monitoring resolution indexed by the difference score between mean confidence judgments for correct and incorrect answers was assessed (see Fig. 2). Results revealed a three-way interaction between item difficulty, age, and achievement level, $F (1, 78)=5.04$, $p<.05$, $\eta^2=.06$, showing that overall monitoring resolution seemed to be more pronounced in high compared to low achievers and in easy compared to difficult items. However, the latter was not the case among 11/12-year-olds high achievers where the opposite pattern was found; monitoring resolution was higher in easy than in difficult items.
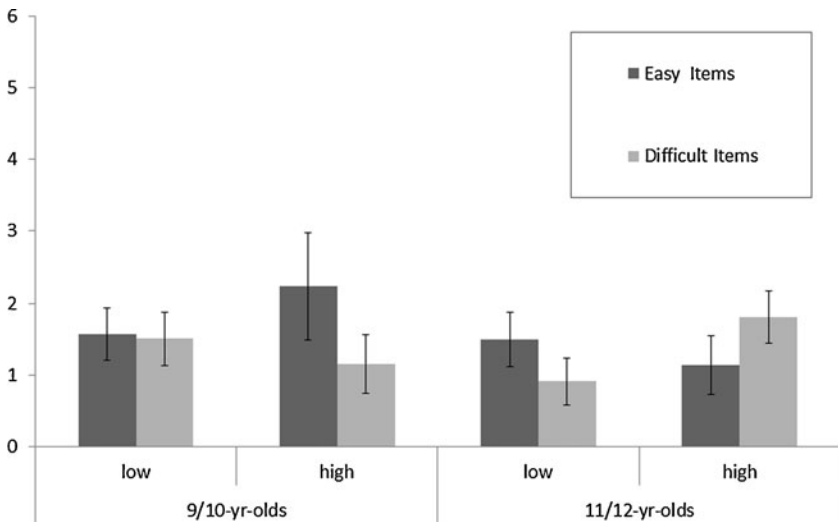

Fig. 2 Monitoring resolution in difference scores between mean confidence judmgent for correct and incorrect answers. All scores differed significantly from zero at the 0.01 level. Error bars represent standard errors of the mean

Another way to investigate monitoring during test taking was to explore children's ability to metacognitively differentiate between answerable and unanswerable items. The latter, which we shall call "distinction" in item format in the following, was measured through the difference scores and analysed as a function of age and achievement level. ANOVA revealed a stronger distinction between answerable and unanswerable items in 11/12- compared to 9/10-year-olds, F (1, 149)=17.51, $p<.001$, $\eta^2=.11$, and in high as opposed to low achievers, F (1, 149)=9.96, $p<.01$, $\eta^2=.07$. No interactions were found.

As to controlling of answerable items during test taking, adequate controlling (% answers correctly crossed-out or maintained) in memory performance was examined for the scoring condition as a function of age, achievement level, and item difficulty. As can be seen in Fig. 3, 11/12-year-olds showed better controlling than 9/10-year-olds, F (1, 69)= 10.91, $p<.001$, $p<.001$, $\eta^2=.14$, and high achievers outperformed low achievers, F (1, 69)= 14.14, $p<.01$, $p<.001$, $\eta^2=.17$. Further, controlling was found to be better in easy than in difficult items independent of age and achievement level, F (1, 69)=121.22, $p<.001$, $\eta^2=.64$. Moreover, a two-way interaction was found between item difficulty and age, F (1, 69)=8.11, $p<.01$, $\eta^2=.11$, indicating that the superior controlling of 11/12- compared to 9/10-year-olds was mainly due to better performance in difficult as opposed to easy items.

To examine whether participants' control processes in free report defer across answerable and unanswerable items, the number of incorrect answers crossed-out for answerable and for unanswerable items were compared among children assigned to the scoring condition. ANOVA was conducted revealing, as can be seen in Fig. 4, main effects of question format, F (1, 69)=139.77, $p<.001$ $\eta^2=.70$, achievement level, F (1, 69)=4.64, $p<.05$ $\eta^2=.06$, and age, F (1, 69)=5.66, $p<.05$, $\eta^2=.08$. These main effects were qualified through two-way interactions between item format and age, F (1, 69)=9.40, $p<.01$, $\eta^2=.12$, and item format and achievement level, F (1, 69)=5.46, $p<.05$, $\eta^2=.07$. These effects mirror better controlling in answerable than in unanswerable items which was more pronounced in 11/12- as opposed to 9/10-year-olds and in high compared with low achievers.

In order to investigate the relation between the respective monitoring and control processes, correlations were computed for 9/10- and for 11/12-year-olds separately. As
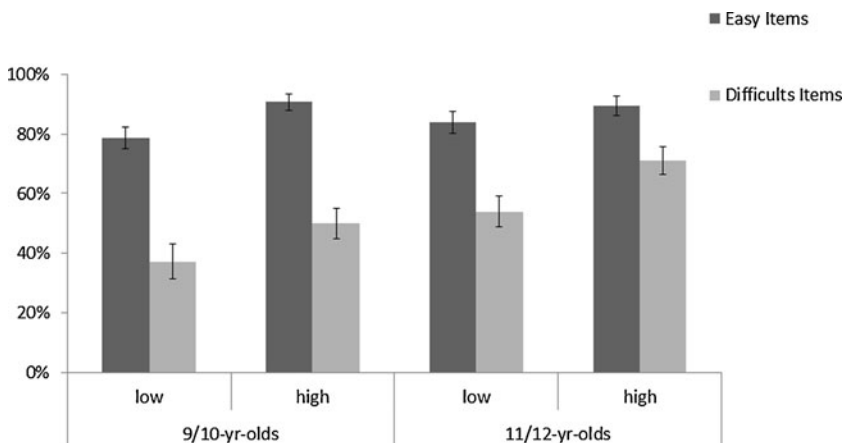

Fig. 3 Adequate controlling in the scoring group computed through the amount of correct answers maintained and incorrect answers withdrawn. All scores differed significantly from zero at the .01 level. Error bars represent standard errors of the mean
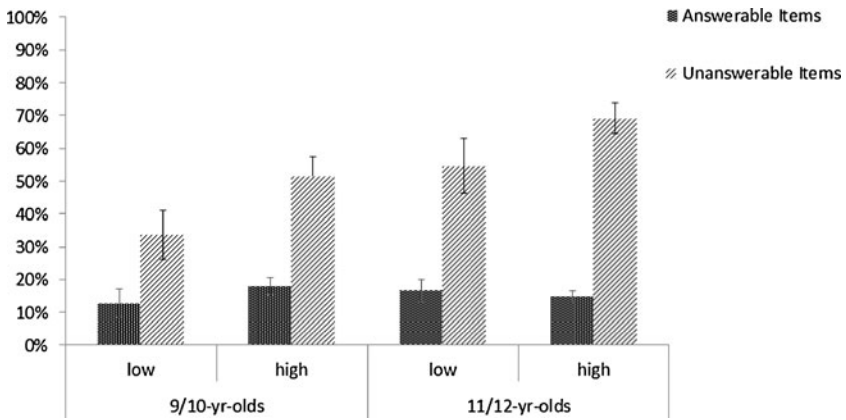
Fig. 4 Mean percentage of withdrawn incorrect answers for all items in the scoring group as a function of age and achievement level. All scores differed significantly from zero at the 0.01 level. Error bars represent error of them mean

displayed in Table 1, among 9/10-year-olds, significant correlations were found between monitoring resolution and adequate controlling, between calibration and distinction of item format, and between adequate controlling in answerable items and controlling in unanswerable items (% answers crossed-out). Concerning 11/12-year-olds, significant correlations were found between monitoring resolution and distinction of item format, and between calibration and adequate controlling.

Discussion

The current study examined the impact of achievement level and age on self-regulation and the involved metacognitive micro-processes in an educational context. This object of investigation appears an important issue for future work in order to improve classroom activities, teachers' instructions, and training programs regarding metacognitive skills and

Table 1 Correlations between monitoring and controlling as a function of age in the scoring group

	Monitoring resolution	Distinction of item format	Calibration for answerable items	Adequate controlling for answerable items	Controlling for unanswerable items
Monitoring resolution	–	.35* (77)	.05 (77)	.79** (47)	.30** (77)
Distinction of item format	.73** (73)	–	−.33** (77)	.12 (47)	.28* (77)
Calibration for answerable items	−.06 (72)	.06 (72)	–	.27 (47)	.06 (77)
Adequate controlling for answerable items	.70** (35)	.17 (35)	−.20 (35)	–	.14 (47)
Controlling for unanswerable items	.17 (73)	.25* (73)	−.14 (72)	.01 (35)	–

9/10-year-olds are displayed above and 11/12-year-olds below the diagonal line. * $p<.05$; ** $p<.01$; *** $p<.001$. Sample sizes for each correlation are presented in parentheses

performance in achievement tests. More precisely, the present work examined 9/10- and 11/12-year-olds metacognitive monitoring and controlling in relation to retrieval processes in a test situation with a focus on the possible factors that may facilitate metacognitive development and performance such as mnemonic experience (assessed through age) and cognitive resources (operationalized through achievement level). As documented for adults (Goldsmith and Koriat 2008), it was found in this study that already 9- to 12-year-olds' memory accuracy appears to benefit from a withdrawing option during test taking reflecting accurate controlling. Based on our multi-phase approach where the order of monitoring and control processes was determined, the latter suggests that controlling might be based on monitoring. The following section, will describe the involved metacognitive processes in more detail by focusing on developmental differences in relation to retrieval processes (operationalized through item difficulty).

Regarding answerable items, as to monitoring prior to test taking, children were somewhat overconfident regarding their own performance (see Fig. 1). Expectedly, children's calibration was more accurate for easy compared to difficult items and high achievers outperformed low achievers. Although no age effects were expected for this age range, results showed that 11/12-year-olds were more accurate in their predictions than 9/10-year-olds showing rather strong overestimation. This age-related advantage turned out to be mostly due to a more realistic view of their performance when facing higher demands in difficult items. Concerning monitoring during test taking, well developed monitoring resolution was found among all children (see Fig. 2) reflected by higher confidence judgments for correct compared to incorrect answers. Moreover, the degree to which children were able to differentiate between correct and incorrect responses was overall more pronounced for easy than for difficult items and among high compared to low achievers, except among older high achievers where monitoring resolution was more pronounced in difficult as opposed to easy items. Taken together, between the age of 9 and 12 years, mnemonic experience seems to facilitate the process of taking retrieval into account when monitoring answerable items. While cognitive resources possibly drive the overall development of monitoring processes, experience may further fuel the "item difficulty"-aspect of this development. Increasing experience with declarative metacognitive knowledge and practice regarding the relation between item difficulty, learning behavior, and memory performance may be an explanation of these age-dependent findings (Schneider and Lockl 2008).

Within unanswerable items, 11/12-year-olds and high achievers showed a more pronounced distinction between answerable und unanswerable items than 9/10-year-olds and low achievers, respectively, indicating somewhat more sophisticated monitoring at the lower end of the certainty-uncertainty continuum. Since for unanswerable items, no memory representations are being created or older representations of general knowledge are activated, one could also say 9/10-year-olds and low achievers exhibited deficits in monitoring memory absence or in differentiating older from newer memory traces. Possibly, younger children and low achievers might have difficulties in recognizing unanswerable items as such, or in interpreting the lack of memory strength due to a lack of mnemonic experience. Further, detection of unanswerable items or memory absence may demand more cognitive resources than answerable items leading to poorer monitoring among lower achieving children. Seemingly, when it comes to retrieval processes, growing mental capacity as well as increasing mnemonic experience may facilitate the progression in metacognitive monitoring. When comparing the effect sizes, however, experience appears to play a stronger developmental role than mental capacity which is in line with the results found in monitoring for answerable items.

Thus, results suggest that during elementary school years, growing cognitive resources and increasing mnemonic experience fuel overall accuracy of monitoring processes prior to and whilst test taking. As was found previously (Roebers et al. 2007), uncertainty monitoring seems to be primarily age-related suggesting that further advancements in metacognitive monitoring may at least in part be due to an increasing mnemonic experience leading to emerging ability to consider retrieval processes as a valuable information source. On the one side, uncertainty monitoring constitutes an important prerequisite for avoiding or correcting errors in educational and other settings (Roderer and Roebers 2009). On the other side, overestimation of performance before test taking could be functional and protective for younger and/or low achievers in terms of motivation, self-esteem, and development (e.g., Shin et al. 2007). Thus, we leave the interpretation open as to whether improvements in uncertainty monitoring in the course of development are progress in its literal sense.

While previous studies have mostly studied monitoring and controlling separately, a strength of the present study is that monitoring and controlling in relation to retrieval processes are examined together. Given the impact of retrieval processes on monitoring that was found in the study, it is not surprising that retrieval was also found to substantially influence young participants' self-initiated controlling. Concerning the involved control processes, we looked at adequate maintaining of correct and adequate withdrawing of incorrect answers to answerable items (see Fig. 3). As expected, controlling was better for easy than for difficult items. Also, high achievers outperformed low achievers. Regarding the influence of mnemonic experience on controlling, 11/12-year-olds showed better performance when it comes to higher demanding retrieval processes. Thus corresponding to the results regarding monitoring, older children's controlling of answers with a higher degree of uncertainty was superior. As to controlling for answerable compared to unanswerable items (see Fig. 4), all children showed better controlling for answerable than for unanswerable items (more incorrect answers to unanswerable than to answerable items were crossed-out) which was stronger pronounced in older than in younger children and in high compared to low achievers. Hence, older children and high achievers showed better controlling of answers with a higher degree of uncertainty or with weaker memory traces compared to younger children and low achievers, respectively. With this respect, comparisons of the effect sizes, however, suggests that experience play a more important role than mental capacity.

Summing up, as for controlling during test taking, both mnemonic experience and cognitive resources seem to be important developmental factors. However, when comparing the effect sizes, experience appear to play a more important role than mental capacity. Analogous to monitoring processes, although experience and mental capacity seem to explain variance in the development of metacognitive processes, experience seems to be the primary developmental factor when it comes to the ability to taking memory strength into account for efficient controlling. The similar pattern of findings for monitoring and controlling appears to further suggest that monitoring may fuel and inform control processes. This was underlined by the mainly significant correlations between the respective monitoring and control processes (see Table 1). To our knowledge, the present study is the first documenting these effects which may also be influential in educational psychology and practice. In elementary school children with poor metacognitive skills, while the accuracy of general metacognitive skills possibly can be improved through cognitive trainings, metacognitive processes regarding the relevance of memory strength in retrieval, may effectively be enhanced by involving children in explicit mnemonic experience. Whether such experience ought to be specific and metacognitive in nature,

such as drawing individual's attention to differences in retrieval processes and by helping to consider this source of information during monitoring and controlling in a test situation, or general like overall mnemonic experience, remains to be investigated in future research.

Whereas earlier studies (e.g., Roebers et al. 2009) suggested that by the age of 12, a flexible and adaptive use of control strategies is observable, the present findings show a more differentiated picture of metacognitive development: Between the age of 9 and 12, metacognitive skills seem to become more accurate and more fine-tuned in relation to available mental capacity and mnemonic experience. Thereby, the influence of experience and cognitive resources on metacognitive development varies. In this respect, the answers to our research questions are as follows: Overall, both experience and cognitive resources seem to affect the development of the increasing quality of monitoring and control processes in a test situation. However, concerning metacognitive processes relying on different strengths of memory traces, experience appear to be the more crucial developmental factor. The similar findings for monitoring and controlling and the significant correlations between them further suggest the assumption that controlling relies on monitoring, at least in a learning setting similar to the paradigm used in the current study. The present work's findings suggest that metacognitive processes are not strictly dependent on growing mental capacity and, hence, are not resistant against intervention. These findings are in line with results found in a recent study (Ceci et al. 2010) suggesting that at least from preschool age the nature of memory representation, rather than neural maturation of the prefrontal cortex, is the primary driving force in memory and metamemory development.

Certainly, there are two limitations of the present study that related to the measurement and operationalization of achievement level. It is possible that reliability of the achievement level measurement was suboptimal and that there is a confound between achievement level and experience. But teachers' ratings regarding student's ability have repeatedly been found to be accurate (e.g., Feindberg and Shapiro 2003)—and the consistent pattern of results presented here corresponds well with these findings. Of course, complete achievement testing, would have been a more reliable and less confounding source for classification. Further, the level of perceived item difficulty is likely to vary systematically depending on individuals' achievement level and intelligence. Children and adults with higher cognitive abilities encode more information, store information in more elaborated and organized networks of knowledge, and are better able to use retrieval cues for their search in long-term memory than individuals with lower cognitive abilities (Schneider and Pressley 1997). Thus, our classification of item difficulty based on previous data might not exactly correspond to individual's perceived difficulty. In this study, children were asked to give overall estimations of their degree of learning and confidence judgments for every given answer. Moreover, confidence judgments focused on the correctness of an answer and not on perceived item difficulty. Hence, perceived difficulty and confidence judgments might share significant amounts of variance. In future studies, subjective judgments of item difficulty and confidence judgments for correctness of one's answer should be included and compared with each other. Possibly, this approach would increase our understanding about the influence of item difficulty and intellectual ability on metacognition. Further, although, when it comes to the ability to use retrieval processes as a cue for memory strength, results support the assumption that experience plays the more important role in metacognitive development than cognitive resources, one should not neglect that the age ranges of 9/10 and 11/12 years are rather narrow and may be too small to detect to detect strong effects of growing mental capacity.

In the present study, controlling has been experimentally manipulated only. It would also be interesting to target—with the aim of improving—monitoring through manipulations of the situational test demands. Although the present study and recent work on metacognition and self-regulation indicate that monitoring often guides controlling, possibly the former also may be based on feedback of control processes and feedback loops from retrieval (Hoffmann-Biencourt et al. 2010; Koriat et al. 2009).

Other sources for individual differences besides achievement level, such as dispositional factors, may impact metacognitive processes as well, for example personality traits (Stankov and Lee 2008), school-related behavior (Normandeau and Guay 1998), test experience (e.g., Nietfeld et al. 2006b), subjective task values (Wigfield et al. 2008), perceived psychological cost (Wigfield et al. 2008), goal orientation (Theodosiou and Papaioannou 2006), and task interest (Schraw et al. 2001). Most of these studies have been realized with adults only and a different pattern of results may emerge including children. Further research in this domain is necessary.

References

Ackerman, A., & Goldsmith, M. (2008). Control over grain size in memory reporting—with and without satisfying knowledge. *Journal of Experimental Psychology, 34*, 1224–1245.

Boekaerts, M. (1999). Metacognitive experiences and motivational state as aspects of self-awareness: Review and discussion. *European Journal of Psychology of Education, 14*, 571–584.

Ceci, S. J., Fitneva, S. A., & Williams, W. M. (2010). Representational constraints on the development of memory and metamemory: A developmental-representational theory. *Psychological Review, 117*, 464–495.

Coffman, J. L., Ornstein, P. A., McCall, L. E., & Curran, P. J. (2008). Linking teacher's memory-relevant language and the development of children's memory skills. *Developmental Psychology, 44*, 1640–1654.

Dunlosky, J., & Bjork, R. A. (2008). *Handbook of metamemory and memory.* New York: Taylor and Francis Group, LLC.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Feindberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18*, 52–65.

Goldsmith, M., & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. S. Benjamin & B. H. Ross (Eds.), *Skill and strategy in memory use* (pp. 1–53). Amsterdam: Academic.

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*, 160–170.

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*, 101–121.

Hoffmann-Biencourt, A., Lockl, K., Schneider, W., Ackerman, A., & Koriat, A. (2010). Self-paced study time as a cue for recall predictions across school age. *British Journal of Development Psychology, 28*, 767–784.

Hong, E., Peng, Y., & Rowell, L. L. (2009). Homework self-regulation: Gender, and achievement-level differences. *Learning and Individual Differences, 19*, 269–276.

Howie, P., & Roebers, C. M. (2007). Developmental progression in the confidence-accuracy relationship in event recall: Insights provided by a calibration perspective. *Applied Cognitive Psychology, 21*, 871–893.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology, 126*, 349–370.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*, 945–959.

Koriat, K., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation or memory accuracy. *Psychological Review, 103*, 490–517.

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The easily learned, easily remembered heursitic in children. *Cognitive Development, 24*, 169–182.

Krebs, S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology, 80*, 325–340.

Lockl, K., & Schneider, W. (2004). The effects of incentives and instructions on children's allocation of study time. *The European Journal of Developmental Psychology, 1*, 153–169.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology, 10*, 756–766.

Maki, R. M., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*, 723–731.

Melot, A.-M. (1998). The relationship between metacognitive knowledge and metacognitive experiences: Acquisition and re-elaboration. *European Journal of Psychology of Education, 13*, 75–89.

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*, 1084–1097.

Michel, E., Roebers, C. M., & Schneider, W. (2007). Educational films in the classroom: Increasing the benefit. *Learning and Instruction, 17*, 172–183.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125–173). San Diego: Academic.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge: MIT Press.

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

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006a). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*, 258–271 [Journal; Peer Reviewed Journal].

Nietfeld, J. L., Cao, L., & Osboren, J. W. (2006b). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition Learning, 1*, 159–179.

Normandeau, S., & Guay, F. (1998). Preschool behavior and first-grade school achievement: The mediational role of cognitive self-control. *Journal of Educational Psychology, 90*, 111–121.

Ornstein, P. A., Haden, C. A., & Hedrick, A. M. (2004). Learning to remember: Social-communicative exchanges and the development of children's memory skills. *Developmental Review, 24*, 374–395.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*, 385–407.

Rhodes, M., & Kelley, C. (2005). Executive processes, memory accuracy, and memory monitoring: An aging and individual difference analysis. *Journal of Memory and Language, 52*, 578–594.

Roderer, T., & Roebers, C. M. (2009). Children's strategic regulation of memory accuracy. In M. R. Kelley (Ed.), *Applied memory*. Hauppauge: Nova.

Roebers, C. M. (2002). Confidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology, 38*, 1052–1067.

Roebers, C. M., Von der Linden, N., Schneider, W., & Howie, P. (2007). Children's metamemorial judgments in an event recall task. *Journal of Experimental Child Psychology, 97*, 117–137.

Roebers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *British Journal of Development Psychology, 79*, 749–767.

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: Evidence for developmental trends. In J. B. R. A. Dunlosky (Ed.), *Handbook of metamemory and memory* (pp. 391–409). New York: Psychology Press.

Schneider, W., & Pressley, M. (Eds.). (1997). *Memory development between two and twenty* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring evidence from a judgment-of-learning task. *Cognitive Development, 15*, 115–134.

Schraw, G., Flowerday, T., & Lehman, S. (2001). Increasing situational interest in the classroom. *Educational Psychology Review, 13*, 211–224.

Schwenk, C., Bjorklund, D. F., & Schneider, W. (2009). Developmental and individual differences in young children's use and maintenance of a selective strategy. *Developmental Psychology, 45*, 1034–1050.

Shin, H. E., Bjorklund, B., & Beck, F. E. (2007). The adaptive nature of children's overestimation in a stretegic memory task. *Cognitive Development, 22*, 197–212.

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology, 4*, 961–976.

Theodosiou, A., & Papaioannou, A. (2006). Motivational climate, achievement goals and metacognitive activity in physical education and exercise involvement in out-of-school settings. *Psychology of sports and exercises, 7*, 361–379.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73.

Veenman, M. V. J., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences, 18*, 159–176.

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14*, 89–109.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14.

Wigfield, A., Hoa, L. W., & Klauda, S. L. (2008). The role of achievement values in the regulation of achievement behaviors. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 169–195). Mahwah: Lawrence Erlbaum Associates Publishers.

Zimmerman, B. J. (1995). Self-regulation involves more than metacognition: A social cognitive perspective. *Educational Psychologist, 30*, 217–221.