Developmental Progression in Performance Evaluations: Effects of Children's Cue-Utilization and Self-Protection

Van Loon, M., Destan, N., Spiess, M., de Bruin, A., & Roebers, C. M.

(2017).

Learning and Instruction, 51, 47-60.

https://doi.org/10.1016/j.learninstruc.2016.11.011

Abstract

To effectively self-regulate learning, children need to self-evaluate whether they meet learning goals. Unfortunately, self-evaluations are often inaccurate, typically, children are overconfident. We investigated two explanations for developmental progression in self-evaluations related to children's (48 5/6-year-olds and 53 7/8-year-olds) interpretations of performance: Improved reliance on item difficulty, and reduced sensitivity to self-protection biases. Self-evaluations were more accurate for 7/8-year-olds than for 5/6-year-olds. There was no developmental increase in reliance on item difficulty; even 5/6-year-olds made adaptive use of this cue. Both age groups were overconfident for incorrect responses, but were able to use performance feedback to improve confidence judgments. However, when self-rewarding, 5/6-year-olds were less likely to take negative performance feedback into account than 7/8-year-olds. The 5/6-year-olds were able to base confidence judgments on performance feedback, but did not use feedback to the same extent when self-rewarding. This may indicate that self-protective biases are an important cause of overconfidence in children.

Keywords: Confidence Judgments; Self-Reward; Children; Development; Overconfidence

Developmental Progression in Performance Evaluations: Effects of Children's Cue-Utilization and Self-Protection

One of the most important developing skills in childhood is the ability to self-regulate learning (Lyons & Ghetti, 2013; McClelland & Cameron, 2012; Roebers, 2014). Adaptive self-regulation is goal-directed, that is, a person regulates learning in order to reduce the discrepancy between the learning goals and the current state of learning (Nelson & Narens, 1990). In order to adaptively self-regulate learning, a person should self-evaluate which tasks have and have not yet been mastered, plan and prioritize study tasks, differentially allocate attention and study time, use appropriate study strategies, and if needed, seek for help to complete tasks (Boekaerts, 1997; Vohs & Baumeister, 2011; Winne & Hadwin, 2008; Zimmerman & Schunk, 2001). To self-regulate learning in school, it is necessary that students can accurately self-evaluate whether they meet learning goals, whether their performance is correct or incorrect, and how many credit points they would receive for their performance from their teacher (Schneider & Pressley, 1997). Importantly, the relation between such self-evaluations, self-regulation, and resulting performance is robust (Dunlosky & Rawson, 2012; Krebs & Roebers, 2010).

In school, children have to engage in a variety of self-regulatory actions in order to improve learning and performance. Even though the ability to adaptively self-evaluate and self-regulate learning develops until late adolescence (Steinbeis & Crone, 2016), preschool children show emerging skills to engage in self-regulated learning such as self-evaluating which of their answers were and were not correct (Destan & Roebers, 2015; Roebers & Fernandez, 2002), and seeking help to achieve task goals (Coughlin, Hembacher, Lyons, & Ghetti, 2015). However, making accurate self-evaluations proves to be difficult for children

and therefore self-regulation is often not adaptive, this has detrimental effects on academic performance (Artelt & Schneider, 2015; Hacker, Bol, & Bahbahani, 2008).

Failures with self-evaluations most often occur in the form of overconfidence when confidence is higher than justified by performance (Finn & Metcalfe, 2014). Especially young children are overconfident (Destan, Hembacher, Ghetti, & Roebers, 2014; Lipko, Dunlosky, Lipowski, & Merriman, 2012; Lipowski, Merriman, & Dunlosky, 2013). Although in general, a slight degree of overconfidence can improve motivation and task persistence (Shin, Bjorklund, & Beck, 2007), in education, extensive overconfidence has negative effects on the learning process. When students of various ages are overconfident, they prematurely cease studying, do not improve test items for which performance was incorrect, and choose inefficient study strategies (De Bruin, Thiede, Camp, & Redford, 2011; Destan & Roebers, 2015; Krebs & Roebers, 2010; Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013a). Together and consistently, overconfidence leads to less efficient self-regulation and ultimately to underachievement (Dunlosky & Rawson 2012; Roderer & Roebers, 2010; Serra & Metcalfe, 2009).

Most research on self-evaluation accuracy focuses on self-evaluations in adults (e.g., Dunlosky & Rawson, 2012; Griffin, Wiley, & Thiede, 2008). Unfortunately, reasons for children's inaccurate self-evaluations, and developmental factors underlying age-related differences, are yet poorly understood. With this study we aimed to shed light on the bases of children's self-evaluations, by investigating their ability to discriminate between correct and incorrect performance.

Self-Evaluating Performance in Childhood

In the preschool years, children start to acquire insights into their own learning, and they begin to differentiate between items for which they are able and items for which they are unable to provide a response, suggesting an early but dichotomous concept of evaluating

performance (Lyons & Ghetti, 2011; 2013). During the kindergarten years, the magnitude of overconfidence declines steadily (Lipko, Dunlosky, & Merriman, 2009; Roebers, 2014), because children increasingly learn to be *"less-than-absolutely certain*" concerning incorrect performance. Age differences are most pronounced when self-evaluating uncertainty, suggesting that children learn to increasingly differentiate on a certain-uncertain continuum (Flavell, 2000). In the present study, we included kindergartners (5/6-year-olds) and 2nd graders (7/8-year-olds) to address the effects of development on self-evaluations. Children studied associated images, took a test, and after test-taking they self-evaluated their test performance by making confidence judgments (CJs), with which they indicated how sure they were that their response was correct. In line with research on children's self-evaluations (e.g., Destan et al., 2014), we expected that for both age groups, self-evaluations about the correctness of their answers would discriminate between correct and incorrect performance (Hypothesis 1a). Furthermore, we know that within the tested age range, strong developmental changes in accuracy of self-evaluations are observed (Roebers, 2014, Schneider, Vise, Lockl, & Nelson, 2000). Therefore, older children were expected to discriminate more accurately than younger children when making self-evaluations (Hypothesis 1b). Moreover, after taking the test and giving CJs for each test response, children received objective feedback about the accuracy of their performance, and they gave Confidence Judgments when facing Feedback (hereafter CJ-FBs). Kindergartners are already able to take performance feedback into account; standards that show actual performance accuracy help them to improve self-evaluations and learning (Muis, Ranellucci, Trevors, & Duffy, 2015). Therefore, in the present study, feedback about performance accuracy was expected to improve self-evaluations for both age groups (Hypothesis 1c).

Importantly, besides investigating effects of development on self-evaluations, we aimed to address underlying reasons for this developmental progression. Self-evaluations are

to a large extent determined by learners' interpretations of the difficulty of the task and invested effort (Koriat & Nussinson, 2009). Decreasing overconfidence might be due – at least to some extent – to children's understanding that performance depends on the difficulty of the task. Further, it is important to note that young children may not yet understand the difference between effort and ability (Folmer et al., 2008; Nichols, 1978); children younger than 8 years of age seem to be self-protective and base their self-evaluations more on the desire to be rewarded for their effort than on their evaluations of true ability and actual performance, and this may cause overconfidence (Folmer et al., 2008; Kurtz-Costes, McCall, Kinlaw, Wiesen, & Joyner, 2005; Miele, Son, & Metcalfe, 2013, Nichols, 1978; Ruble, Eisenberg, & Higgins, 1994). Older children may have an increasing understanding that effort is not a sufficient prerequisite for correct performance. In the present study, we investigate to what extent developmental differences in self-evaluations are due to age differences in interpretations of item difficulty and effects of self-protective biases, by addressing these two distinct, but not mutually exclusive explanations.

Item difficulty Cue-Utilization in Children

With the cue-utilization framework, Koriat (1997) has put forward the idea that accurate self-evaluating relies on the use of valid cues or heuristics, as for example fluency of memory retrieval ("*The answer came quickly to my mind, so I am sure about the correct answer*") or the perceived ease of learning ("*I have easily learned this material, so I will easily remember it, too*"). In adults, valid cues have consistently been found to yield a strong influence on the accuracy of self-evaluations; at the same time, the use of invalid cues (such as using font size as a basis for judgments, Mueller, Dunlosky, Tauber, & Rhodes, 2014) plays a major role for explaining adults' failures in self-evaluations (Griffin et al., 2008).

According to the cue-utilization explanation, young children's self-evaluations are inaccurate because they do not consider the most valid task cues. Research on children`s cue

use is still in its early stages; the few existing studies suggest that the degree to which they rely on valid cues increases over the elementary school years. For example, there is an inverted relationship between retrieval fluency and confidence in children (the longer it takes children to come up with an answer, the less confident they are; Koriat & Ackerman, 2010), but this association is stronger in older compared to younger children. Similarly, the "easily-learned-easily-remembered" heuristic is present in children, but tends to be more influential for self-evaluations in older compared to younger children (Hoffmann-Biencourt, Lockl, Schneider, Ackerman, & Koriat, 2010). Thus, younger children may not yet be able to take valid cues into account, at least not to the same extent as older children and adults, and this may constitute a major reason for younger children`s inaccurate self-evaluations.

For adults, awareness of the feeling of difficulty, an indicator of study effort due to the intrinsic cognitive load imposed by the task, is an important cue that gives an indication about whether performance might or might not be correct (DeLeeuw & Mayer, 2008). Older elementary school children, like adults, clearly base their self-evaluations on item difficulty. Koriat, Ackerman, Lockl and Schneider (2009a; 2009b) showed with word-learning and general knowledge learning tasks that reliance on study time, a cue indicating encoding effort, affected self-evaluations for children older than 8 years. The easier it felt to process new information, the more confident children were. In contrast, younger elementary school children (6- and 7-year olds) did not yet seem to use perceived item difficulty as a cue, suggesting that children's reliance on item difficulty cues develops around the age of 8 to 9 years. However, evidence on younger children`s item difficulty cue use is inconsistent; some research shows that kindergartners do not yet discriminate between easy and difficult task items (Dufresne & Kobasigawa, 1989; Koriat et al., 2009a, 2009b; Paulus, Tsalas, Proust, & Sodian, 2014). When image learning tasks are used, instead of reading tasks, information may be easier to process for children and item difficulty cues may be more salient. When learning

associated images, kindergartners (5- and 6-year-olds) are already able to adaptively rely on item difficulty cues when making self-evaluations (Destan et al., 2014; Roebers, von der Linden, Schneider, & Howie, 2007). In sum, it is unclear at what age children become able to take item difficulty cues into account, and to what extent this cue use explains age differences and accuracy of self-evaluations. In the present study, drawing from the adult literature and applying the cue-utilization framework (Koriat, 1997), we address the effect of item difficulty as an index of the study effort. Item difficulty is a valid cue for improving accuracy of self-evaluations, independent of an individual`s age, but this cue may be used inefficiently (if at all) by young children. When evaluating the accuracy of their test responses with CJs, we expect that 7/8-year-olds take item difficulty to a stronger degree into account in their self-evaluations than 5/6-year-olds (Hypothesis 2a). Because children need to infer performance when making initial CJs, we assume that item difficulty mainly affects these self-evaluations. When feedback about performance accuracy is provided, adults flexibly incorporate this information to come to more valid cues, and this improves accuracy of self-evaluations (Castel, 2008). In a similar vein, when children are provided with feedback on performance, this gives them an objective indication of performance. They then no longer need to base their evaluations on inferential cues, such as item difficulty. Therefore, we hypothesize that, when presented with performance feedback and making CJ-FBs, children no longer base their self-evaluations on item difficulty cues (Hypothesis 2b).

The Self-Protection Explanation

In contrast to the task difficulty cue explanation, which entails that young children are hindered to accurately evaluate performance because they lack insight into valid cues, the self-protection explanation proposes that children *can* accurately evaluate their performance. That is, children have the ability to discriminate between incorrect and correct performance (Lipowski, Merriman, & Dunlosky, 2013; Schneider, 1998). However, they may not always

base their self-evaluations on actual performance accuracy (Lipko et al., 2012; Schneider, 1998). Children may not take negative performance into account because it poses a threat to their sense of confidence and self-esteem (Shin et al., 2007). Overestimating performance then helps children to stay motivated and to persist, even in new and difficult tasks (Bjorklund & Green, 1992). Young children intend to do well, and when making self-evaluations they may conflate their wishes for correct performance with actual performance (if performance is suboptimal). That is, they base their evaluations more on the desire to have good performance: The so-called wishful thinking bias (Schneider, 1998; Shin et al., 2007; Stipek, 1984). According to this self-protection explanation, young children make limited use of their insights into objective performance accuracy when self-evaluating performance, as they want to feel good about themselves and be rewarded for effort. This tendency is likely to affect self-evaluative judgments.

In the present study we address to what extent children's overconfidence can - at least in part - be explained by their tendency to use self-protective biases. Even though we presume that both age groups are able to take feedback into account (conform Hypothesis 1c), when facing that their response is incorrect, the younger age group is hypothesized to be more overconfident than older children. That is, 7/8-year-olds are presumed to give lower CJ-FBs for incorrect responses than 5/6-year-olds (Hypothesis 3a). Furthermore, we included a self-rewarding judgment (RJ) in which children were allowed to give themselves credit points for their answers while being presented with performance feedback. Comparing children's CJ-FBs with RJs can give us insight into whether children distinguish between confidence and reward. If children are able to take negative performance feedback into account in their CJ-FBs, but have the tendency to disregard insights into performance when making RJs, this seems to indicate that children can accurately evaluate performance with feedback, but do not do so when self-rewarding due to their desire to be rewarded for effort instead of actual

performance accuracy. When overconfidence is due to self-protection and the desire to be rewarded, younger children's RJs might discriminate less accurately between correct and incorrect performance than their CJ-FBs (Hypothesis 3b).

Present Study

In sum, two explanations have been given for children's inaccurate self-evaluations: a) young children lack the ability to base self-evaluations on valid cues; and b) young children are self-protective and therefore do not accurately evaluate performance. In this study, we aimed to investigate whether age differences in accuracy of self-evaluations result from cue-utilization of item difficulty, self-protective biases, or both. Children (5/6-year-olds and 7/8-year-olds) studied associated Asian ideograms (Japanese Kanji) and their meanings; these paired associates varied in difficulty. After study they took a test, and then self-evaluated their test performance with three self-evaluation measures: CJs, CJ-FBs, and RJs. To our knowledge, this is the first study that simultaneously explores both the cue-utilization and the self-protection explanations. Most previous studies included only one single self-evaluation measure (e.g. Schneider et al., 2000), leaving the question of the relative importance of such influences unanswerable. Insight into effects of self-protective biases and use of item difficulty cues on self-evaluations are important for instructional practice. When developmental differences are observed, this would indicate that interventions to improve self-evaluations need to be tailored for different age groups.

The following hypotheses are tested: Firstly, we expected that for both age groups, self-evaluations would discriminate between correct and incorrect performance (Hypothesis 1a); 7/8-year-olds were expected to discriminate more accurately than 5/6-year-olds (Hypothesis 1b). Further, for both age groups, feedback about response accuracy was expected to improve self-evaluations (Hypothesis 1c). Moreover, we expect that 7/8-year-olds take item difficulty to a stronger degree into account in their self-evaluations than 5/6-year-

olds (Hypothesis 2a), and we hypothesize that after receiving feedback, children no longer base their self-evaluations on item difficulty cues (Hypothesis 2b). Finally, we hypothesize that, even with feedback, the younger age group will be more overconfident for incorrect performance than the older age group (Hypothesis 3a), and we expect that younger children's RJs discriminate less accurately between correct and incorrect performance than their CJ-FBs (Hypothesis 3b).

Method

Participants

The sample consisted of 101 children; 48 5/6-year-olds who were kindergartners (*M* 6.0 years, *SD* = 4.8 months; 26 girls) and 53 7/8-year-olds who were 2nd graders (*M* 8.1 years, *SD* = 4.2 months, 26 girls). The children were tested individually in a quiet room of their kindergarten or school. Participants were tested in the German-speaking part of Switzerland; participants predominantly came from middle-class families and all participants were sufficiently fluent in the German language to understand task instructions. None of the participants was familiar with reading Asian ideograms. In order to ensure enough data points for each child (variance in difficulty of Kanji, a sufficient number of correct and incorrect test responses, and variance in self-evaluative judgments) children were tested twice; the two testing sessions were separated by one week. Parental informed consent was acquired prior to testing.

Materials

Task phases were: 1) study phase; 2) recognition test; 3) giving Confidence Judgments without feedback (CJs); 4) giving Confidence Judgments with Feedback (CJ-FBs); and 5) giving Self-Rewarding Judgments with feedback (RJs). Figure 1 depicts an overview of the task phases and the materials. Stimuli were Japanese symbols (Kanji); this Kanji-image learning task was used because it proves a suitable study task for which outcomes do not

depend on prior knowledge and reading ability. Kanji have been successfully used in previous research on children's self-evaluations (Destan et al., 2014; Destan & Roebers, 2015, and Roderer & Roebers, 2010; 2014); the Appendix lists the used Kanji and gives an index of the difficulty of each Kanji.

Kanji were presented randomly through all phases of the task, and in every task phase, all Kanji were shown before moving on to the next phase. Children were tested in two sessions, with different stimuli in both sessions. All tasks were completed on a touch screen tablet (Acer Iconia W700, 11.6''), equipped with E-Prime.

In the study phase, children studied the meaning of Kanji; 5/6-year-olds studied 10 Kanji, 7/8-year-olds studied 12 Kanji per session (so across both sessions, 5/6-year-olds studied 20 Kanji, 7/8-year-olds studied 24 Kanji). For study, each Kanji was shown with a picture representing the meaning. Kanji were presented in two consecutive fixed-length study phases with fixed study times (3 and 2 s, respectively).

When taking the recognition test, the previously studied Kanji were presented with four pictures. One of these pictures represented the correct meaning, the remaining three pictures represented incorrect meanings of other Kanji studied in the session; children had to choose the correct meaning.

Note that all self-evaluations were made after the recognition test was completed. When making CJs, children rated their confidence in performance accuracy. They were successively presented with each Kanji of the previous test and the four test response options. Their own test response was indicated with a black frame around the previously chosen picture. To indicate confidence in correctness, children pressed one of seven buttons on a thermometer scale (cf. the scale used by Van Loon et al., 2013b). The buttons on the thermometer scale formed a continuum of colors from blue to red; colors on the thermometer

buttons ranged from dark blue (symbolizing cold – very unsure) to dark red (symbolizing hot – very sure).

When giving CJ-FBs, children again rated their confidence in performance when facing both their own response and the objectively correct response. Children were presented with each Kanji and the four response options; their own test response was surrounded by a black frame and the objectively correct response was surrounded by a green frame. If the chosen test response was the correct meaning of the Kanji, the picture was surrounded by a black and a green frame. Children pressed one of seven buttons on the thermometer scale similar to the one used for initial CJs.

When making RJs, children rewarded their own test responses with credits points on a 7-point scale, representing dice points ranging from 0 – 6. Similar as in the CJ-FB phase, children were shown each Kanji and the four response options; their own test response was surrounded by a black frame and the objectively correct response surrounded by a green frame. The dice-scale was presented below the Kanji and the response options.

Procedure

Children were familiarized with the tablet prior to testing, and were familiarized with the Kanji and practiced the task procedure (study phase, recognition test, CJs; CJ-FBs; RJs) with three examples. The thermometer scale that was used for the CJs and the CJ-FBs was introduced using the cold/warm story that was well-known to the participating children; when seeking a hidden object, children are told "cold" as they move away from, and "warm" as they get closer to it. The further away from the object, the less certain one is about its correct location. In this analogy, the color blue represents cold/uncertainty, and red represents warm/certainty. Accordingly, children practiced pressing the reddish buttons to indicate certainty, and the bluish buttons to indicate uncertainty; the darker the color (or the closer to the two poles), the higher the degree of un/certainty. The experimenter assured participants'

understanding of the scale using an example for the two poles and the middle button by asking three questions: An easy one (*What's the color of grass?),* a difficult one (*In numbers, how much hair do I have on my head?)*, and one of medium difficulty (*How old am I?).* After responding, children had to indicate how sure they were to have responded correctly by pressing on the thermometer scale. All children learned the use of the thermometer scale on the tablet with ease.

When giving CJs about their test responses, the experimenter explained that they would see their own test response surrounded by a black frame. The child was asked to indicate how sure s/he was that the response was correct, by pressing the according button on the thermometer. When giving CJ-FBs, they were explained that they would now see their own response (surrounded by the black frame), but that also something new was added, and that they would see the correct response in a green rectangle. They were then asked to give a next confidence judgment on the similar thermometer indicating how sure they now were about performance accuracy given the feedback.

When making RJs, children were instructed to consider themselves in the role of a teacher, to give themselves credit points on the dice scale for each of their test responses. They were explained that the dice scale depicted reward points, ranging from no reward (0 points) to high reward (six points). They saw their own answer (surrounded by the black frame) and the correct answer (surrounded by the green frame). They were told that they could decide for each response how many points they would like to give.

The second session was one week apart from Session 1; during Session 2, children practiced with one Kanji to familiarize themselves again with the task phases. After completing the second testing session, children received a small gift.

Analyses

Based on a separate sample of 92 participants in the same age range as the sample in this study, difficulty indexes for each Kanji were calculated according to the method by Lienert and Raatz (1998). The difficulty of a Kanji is equal to the percentage of correct answers for this particular item across the total sample, correcting for the probability of guessing. Thus, difficult Kanji have a low difficulty index, whereas easy Kanji have a high difficulty index. The resulting item difficulty indices for the Kanji varied between 10.87 and 66.67. Kanji and difficulty indices are listed in the Appendix. Similar to Destan et al. (2014), based on a median split (Median = 40.58), Kanji were categorized as easy or difficult.

All self-evaluations (CJs, CJ-FBs, and RJs) were given on a 7-point scale ranging from 0 – 6. To investigate whether participants' self-evaluations discriminate between correct and incorrect performance, judgments across correct and incorrect performance were contrasted when item difficulty was additionally taken into account. Further, mean intra-individual Gamma correlations between performance and self-evaluations were calculated to indicate discrimination accuracy. Gamma is a non-parametric measure of correlation, and is considered an appropriate measure of the relation between performance and scale-based self-evaluations (Nelson, 1984; Thiede, Anderson, & Therriault, 2003). Gamma can range between -1 and +1, more accurate self-evaluations are indicated by stronger positive Gamma correlations. Mixed ANOVAs were used to investigate the impact of the nature of the self-evaluation, item difficulty, and age differences; for significant effects, partial eta squared gives an indication of the effect sizes. Data across the two sessions was collapsed, to get a sufficient data base for confidence and self-reward judgements for correct and incorrect responses. Even though we do not have specific hypotheses about differences between session 1 and 2, for explorative reasons we additionally report effects of session on CJs, CJ-FBs, and RJs.

Results

In this section, we report discrimination between correct and incorrect responses when making CJs, CJ-FBs, and RJs, and evaluate evidence for age differences in item difficulty cue-utilization and self-protection. Table 1 shows mean performance and mean self-evaluative judgments. The 7/8-year-olds had better recognition performance than 5/6-year-olds, F (1, 99) = 33.57, $p < .001$, $\eta_p^2 = .25$.

Discrimination between Correct and Incorrect Performance

Confidence Judgments. Table 1 shows children's CJs for correct and incorrect test responses. A mixed ANOVA for CJs with Performance (correct, incorrect) as within-subjects variable and Age (kindergarten, 2nd grade) as between-subjects variable, shows that mean CJs were significantly higher for correct than for incorrect responses, F (1, 98) = 39.20, $p < .001$, $\eta_p^2 = .35$. There was no significant effect of Age, $p = .21$, but the relation between Performance and Age was qualified by a two-way interaction effect, F (1, 98) = 5.48, $p = .009$, $\eta_p^2 = .07$, such that older children had more confidence in correct performance than younger children. Gamma correlations indicate the strength of the relation between self-evaluations and performance; Gammas for both age groups are shown in Figure 2. Both for 5/6-year-olds and 7/8-year-olds, Gamma correlations for CJs were significantly higher than zero; for 5/6-year-olds $G = .18$, $SD = .53$, $t(44) = 2.33$, $p = .025$, for 7/8-year-olds $G = .59$, $SD = .37$, t (50) = 11.30, $p < .001$. As visible in Figure 2, the Gamma correlation for 7/8-year-olds was significantly higher than the Gamma for 5/6-year-olds, F (1, 94) = 3.93, $p < .001$, $\eta_p^2 = .17$, indicating that older children better discriminated between correct and incorrect performance. A mixed ANOVA to investigate potential effects of Session on Gamma correlations for CJs did not show differences between Session 1 and Session 2, $p = .51$.

Confidence Judgments with Feedback. From Table 1 it appears that both age groups accurately discriminated between incorrect and correct test responses in their CJ-FBs. A mixed ANOVA confirmed that mean CJ-FBs were higher for correct responses than for

incorrect responses, F (1, 98) = 256.33, $p < .001$, $\eta_p^2 = .72$. The main effect of Age, F (1, 98) = 5.60, $p = .020$, $\eta_p^2 = .05$, indicates that overall CJ-FBs were higher for 7/8-year-olds than for 5/6-year-olds. The lack of a significant interaction between Performance and Age, $p = .76$, shows that both age groups equally discriminated between correct and incorrect performance. Figure 2 shows that Gamma correlations between CJ-FBs and Performance were high for both age groups. These Gammas significantly differed from zero; for 5/6-year-olds, $G = .81$, $SD = .37$, $t(44) = 14.47$, $p < .001$; for 7/8-year-olds $G = .99$, $SD = .03$, t (50) = 251.81, $p < .001$. However, even though for both age groups the high Gammas indicate good discrimination, the Gamma correlation for 7/8-year-olds was significantly more accurate than Gamma for 5/6-year-olds, F (1, 94) = 12.37, $p < .001$, $\eta_p^2 = .12$. There was no effect of Session on the Gamma correlations between CJ-FBs and Performance, $p = .33$.

To investigate whether performance feedback led to improved accuracy of self-evaluations, with a mixed ANOVA we compared intra-individual Gamma correlations for CJs with Gamma correlations for CJ-FBs (within-subjects repeated measurement) as an effect of age. The main effect of Gamma shows that discrimination accuracy of CJ-FBs was higher than discrimination accuracy of CJs, F (1, 92) = 82.21, $p < .001$, $\eta_p^2 = .47$. The main effect of Age, F (1, 92) = 34.93, $p < .001$, $\eta_p^2 = .28$, shows that 7/8-year-olds more accurately discriminated between correct and incorrect performance than 5/6-year-olds. The lack of a significant interaction effect between Gamma correlations and Age, $p = .061$, shows that both age groups were similar in the extent they improved confidence judgments with use of performance feedback. That is, both age groups appropriately adjusted their confidence after receiving feedback.

Reward Judgments. Table 1 shows that, when giving self-rewards for performance, children discriminated between correct and incorrect responses, F (1, 98) = 158.86, $p < .001$, $\eta_p^2 = .62$. Even though there was no main effect of Age, $p = .079$, the significant interaction

between Performance and Age, F (1, 98) = 24.92, $p < .001$, $\eta_p^2 = .15$, shows that 7/8-year-olds gave themselves more reward for correct responses and less reward for incorrect responses than 5/6-year-olds. For both age groups, Gamma correlations differed from zero; for 5/6-year-olds $G = .42$, $SD = .63$, $t(47) = 4.62$, $p < .001$; for 7/8-year-olds G = .94, $SD = .12$, $t(50) = 56.43$, $p < .001$. Figure 2 shows that Gammas for RJs were more accurate for 7/8-year-olds than for 5/6-year-olds; a main effect of Age, F (1, 97) = 33.15, $p < .001$, $\eta_p^2 = .26$, confirms that 5/6-year-olds discriminated less between correct and incorrect responses than 7/8-year-olds in their RJs. There was no difference between Session 1 and Session 2 in Gammas for RJs, $p = .37$.

In sum, both age groups were able to discriminate between correct and incorrect performance (confirming Hypothesis 1a), the 7/8-year-olds discriminated more accurately between correct and incorrect performance than 5/6-year-olds (confirming Hypothesis 1b), and feedback improved self-evaluations (confirming Hypothesis 1c).

Utilization of Item Difficulty Cues

Gamma correlations between difficulty level (1 = easy, 2 = difficult) and performance were -.29 ($SD = .40$) for 5/6-year-olds and -.34 ($SD = .45$) for 7/8-year-olds. The negative Gamma correlations show that correct performance was less likely for Kanji categorized as difficult. Gammas differed from zero; for 5/6-year-olds, $t(47) = 5.14$, $p < .001$; for 7/8-year-olds $t(52) = 5.48$, $p < .001$, showing that item difficulty was a valid cue that was diagnostic of children's performance. There was no difference in cue validity of item difficulty cues for the two age groups, $p = .59$.

To investigate children's ability to take item difficulty cues into account in their CJs, we conducted a mixed ANOVA with CJs for easy and difficult items for the correct and incorrect responses (within-subject factors) and age (between-subjects factor). Figure 3 shows CJs for easy and difficult Kanji, for correct and incorrect test performance separately. As

visible in this Figure and confirmed by the analysis, CJs were lower for difficult than for easy items (for easy items $M = 4.58$, $SD = 1.09$, for difficult items $M = 4.03$, $SD = 1.17$; for correct easy items $M = 4.96$, $SD = 1.31$, for incorrect easy items $M = 3.89$, $SD = 1.69$; for correct difficult items $M = 4.29$, $SD = 1.37$, for incorrect difficult items $M = 3.6$, $SD = 1.46$); F (1, 91) = 23.20, $p < .001$, $\eta_p^2 = .20$. There was no main effect of Age, $p = .23$, and the lack of a significant interaction between Age and Item Difficulty, $p = .43$, shows that age did not predict reliance on item difficulty when self-evaluating performance. The lack of a developmental increase in reliance on item difficulty shows that even 5/6-year-olds were able to make adaptive use of item difficulty for their CJs (contrasting Hypothesis 2a). There was no difference in effects of Item Difficulty on CJs between Session 1 and Session 2, $p = .10$. Further, in line with our expectations (Hypothesis 2b), item difficulty did not affect self-evaluations after children were provided with performance feedback; Item Difficulty did not affect CJ-FBs (for easy items $M = 4.63$, $SD = 1.13$, for difficult items $M = 4.02$, $SD = 1.23$; for correct easy items $M = 5.73$, $SD = .66$, for incorrect easy items $M = 2.56$, $SD = 1.92$; for correct difficult items $M = 5.64$, $SD = .88$, for incorrect difficult items $M = 2.48$, $SD = 1.79$), $p = .26$, and further, there was no effect of Item Difficulty on RJs (for easy items $M = 4.51$, $SD = .98$, for difficult items $M = 4.04$, $SD = 1.04$; for correct easy items $M = 5.19$, $SD = 1.28$, for incorrect easy items $M = 2.98$, $SD = 1.52$; for correct difficult items $M = 5.08$, $SD = 1.24$, for incorrect difficult items $M = 2.93$, $SD = 1.32$), $p = .22$.

Evidence for the Self-Protection-Hypothesis

To investigate overconfidence, we compared CJs, CJ-FBs, and RJs for incorrect responses, Table 1 shows these self-evaluations for the two age groups. A mixed ANOVA with CJs, CJ-FBs, and RJs for incorrect responses as within-subjects factors, and Age as between-subjects factor investigated children's differentiation between self-evaluations. Firstly, the main effect of Self-Evaluations, F (2, 196) = 21.66, $p < .001$, $\eta_p^2 = .18$, shows that

self-evaluations for incorrect responses were highest for CJs, followed by RJs, and lowest for CJ-FBs. Thus, when making CJ-FBs and RJs, children were less overconfident than when making CJs, and RJs for incorrect responses were higher than CJ-FBs. There was no significant main effect of Age, $p = .964$, however, there was a significant Self-Evaluation by Age interaction, F (2, 196) = 3.241, $p = .041$, $\eta_p^2 = .03$. For 5/6-year-olds' incorrect responses, CJ-FBs were lower than CJs ($p < .001$), but RJs were only marginally lower than CJs ($p = .059$), and most interestingly, 5/6-year-olds' RJs were significantly higher than their CJ-FBs ($p = .002$). This finding shows that younger children are especially optimistic when self-rewarding incorrect performance. For 7/8-year-olds, CJs for incorrect responses were higher than CJ-FBs ($p < .001$) and RJs ($p < .001$); they did not differentiate between CJ-FBs and RJs ($p = .840$). These findings support the self-protection explanation: Even though 5/6-year-olds are equally able as 7/8-year-olds to take performance accuracy into account when making CJ-FBs, they tend to reward themselves even for incorrect performance when making RJs (confirming Hypothesis 3a). There was no significant main effect of Session, $p = .84$, however, there was a significant interaction effect between Session and Self-Evaluation. F (2, 91) = 5.655, $p = .005$, $\eta_p^2 = .11$. Follow-up mixed ANOVAs for each of the Self-Evaluations separately show that Session did not have an effect on CJs ($p = .11$) and CJ-FBs ($p = .29$) for incorrect performance, however, the effect of Session on RJs was significant, F (1, 93) = 8.33, $p = .005$, $\eta_p^2 = .08$. Children gave themselves more reward for incorrect performance in Session 2 ($M = 3.09$, $SD = 1.54$) than in Session 1 ($M = 2.69$, $SD = 1.46$).

Moreover, to investigate whether discrimination between correct and incorrect responses is similar when children evaluate self-rewards and confidence, we compare Gamma correlations between CJ-FBs and RJs. The main effect of Gamma shows that CJ-FBs were more accurate than RJs, F (1, 93) = 27.71, $p < .001$, $\eta_p^2 = .23$. The main effect of Age, F (1, 93) = 27.00, $p < .001$, $\eta_p^2 = .23$, shows that overall, Gammas were higher for 7/8-year-olds

than for 5/6-year-olds, indicating better discrimination. Interestingly, the significant interaction between Gamma and Age, F (1, 93) = 17.11, $p < .001$, $\eta_p^2 = .16$, shows that 5/6-year-olds were differentiating between their CJ-FBs and their RJs; their RJs were less related to performance than their CJ-FBs (confirming hypothesis 3b). In contrast, for 7/8-year-olds, the Gamma for RJs was similar to the Gamma for CJ-FBs. Session did not affect differentiation between CJ-FBs and RJs, $p = .25$.

Discussion

This study shows that 7/8-year-olds make more accurate self-evaluations than 5/6-year-olds when judging their confidence in performance (with CJs), when evaluating performance while facing feedback (with CJ-FBs), and when rewarding themselves for their performance (with RJs). Children of both age groups were able to discriminate between correct and incorrect performance (confirming Hypothesis 1a), and feedback improved discrimination accuracy (confirming Hypothesis 1c). The 7/8-year-olds were somewhat better able to use feedback for confidence judgment than 5/6-year-olds, for the older children the relation between confidence and performance was very strong, showing that they were consistently less confident for incorrect than for correct responses. For all three self-evaluations, the relation between performance and evaluations was stronger for 7/8-year-olds; 5/6-year-olds' self-evaluations were less accurate and more overconfident. This shows a developmental increase in the ability to self-evaluate performance (confirming Hypothesis 1b). This evidence of developmental progression in accuracy of self-evaluations corroborates previous research (Lipko et al., 2009; Lipowski et al., 2013; Roebers, 2014; Schneider et al., 2000; Schneider & Lockl, 2008).

The novelty of this research lies in our aim to explain these age differences in the accuracy of self-evaluations. In the literature, two explanations have been put forward for children's inaccurate self-evaluations when evaluating incorrect performance: a) young

children lack the ability to base self-evaluations on valid cues; and b) young children are self-protective and therefore do not accurately evaluate performance. We evaluated both explanations; to our knowledge, this is the first study investigating effects these distinct, but not mutually exclusive cue utilization and self-protection explanations on children's self-evaluations.

Our assumption that accuracy of self-evaluations could be due to older children's increased ability to implement valid item difficulty cues was not confirmed: Both age groups made adaptive use of item difficulty cues (contrasting Hypothesis 2a). The finding that even 5/6-year-olds take item difficulty into account does not support the assertion that a lack of adaptive utilization of item difficulty is the main reason for developmental progression in self-evaluation accuracy. Instead, findings imply that young children already base CJs on their learning experiences. These findings are in line with indications that 5/6-year-olds are able to use item difficulty cues for self-evaluations (Destan et al., 2014); findings show that development of use of item difficulty cues already occurs in the pre-school years. However, the findings do not confirm research showing less adaptive item difficulty cue use for young children (Koriat et al., 2009a, 2009b; Koriat & Ackermann, 2010). From previous research, it has been inferred that children younger than 8 years of age cannot yet base self-evaluations on cues derived from study experiences (Koriat et al., 2009a, 2009b). In the present study, we used an image learning task, whereas research showing developmental progression in cue utilization used textual learning tasks, such as word-pair (Koriat et al., 2009b), and general knowledge learning (Koriat & Ackerman, 2010).

It has to be noted that our operationalization of item difficulty (items were categorized as easy or difficult based on performance of a separate sample), and the timing of the self-evaluations may have implications for the comparability between this research and previous studies on children's reliance on item difficulty. For instance, in research by Koriat et al.,

(2009a; 2009b), children made prospective judgments predicting performance (i.e., judgments of learning – JOLs), whereas in the present study children made retrospective CJs. Further, in our study, the study time was fixed, whereas study was self-paced in previous research (Koriat et al., 2009a, 2009b; Koriat & Ackerman, 2010). In this study, item difficulty seems to be due to the visual complexity of the Kanji and the relatedness between the Kanji and the meaning. Studying images with a fixed study time may give learners clear indications of item difficulty, because they do not need to simultaneously use multiple task cues. Possibly, self-paced study of text materials provides less salient cues related to study experiences and cognitive load, because children simultaneously have to implement multiple cues such as task difficulty, study time, and processing fluency (Mueller, Tauber, & Dunlosky, 2013). In children, self-evaluating tends to be easier and more accurate for tasks that have low cognitive demands compared to more demanding tasks (Howie & Roebers, 2006; Roebers et al., 2007). When children get older, they can increasingly cope with higher cognitive task demands, and self-evaluations become more accurate for difficult tasks (Flavell, 2000). Possibly, the relatively low task demands in the present study may explain why 5/6-year-olds were already able to implement item difficulty cues.

From this study, as well as from previous research (e.g. Destan et al., 2014; Kobasigawa & Metcalf-Haggert, 1993), it can be derived that item difficulty seems a powerful cue which, when the task is easy and the cue is salient, may be incorporated into young children`s self-evaluations. The effect size shows that item difficulty had strong effects on CJs, however, the difference between CJs for easy and difficult items is not that big. CJs were made after taking the recognition test, and encoding effort due to item difficulty may be less important after test-taking than before test taking, when learners do not yet have the test experiences and need to rely on encoding experiences. Future research should investigate

whether the effect of item difficulty is stronger when learners make prospective, rather than retrospective judgments.

Interestingly, when self-evaluating performance while presented with feedback, item difficulty no longer played a role (confirming Hypothesis 2b). This indicates that these cues were only important for "true metacognitive" judgments; children thus had to infer accuracy based on their subjective learning experiences, that is, they had to take a meta-perspective (Nelson & Narens, 1990). When the feedback was presented, they received objective information about their performance, and did not need to make inferences about accuracy based on their study experiences. This indicates that children flexibly switched to more valid cues indicating the actual cognitive accuracy. Even 5/6-year-olds were clearly aware that objective performance is a crucial factor when judging confidence. Also when rewarding, item difficulty no longer played a role, indicating that performance feedback was more important than study experiences. However, we presume that, without feedback, item difficulty will have a strong effect on RJs, because these are supposed to be based on interpretations of effort (Folmer et al., 2008; Nichols, 1978; Ruble et al., 1994).

To evaluate whether children are self-protective when evaluating performance, confidence was compared with self-reward. For incorrect responses, self-evaluations would be accurate only if children went down on the evaluation scale and give a judgment of zero. When receiving feedback, both age groups were less overconfident in comparison to making CJs without feedback. However, mean CJ-FBs were still inappropriately high; children were not using the lowest point on the scale, indicating self-protection for both age groups. To further evaluate the self-protection hypothesis, children's confidence was compared with self-rewards for performance. Gamma correlations show that 7/8-year-olds discriminated very well between correct and incorrect performance when self-rewarding; their RJs discriminated equally well as their CJ-FBs. However, 5/6-year-olds' RJs for incorrect responses were higher

than CJ-FBs (confirming Hypothesis 3a). Even though they accurately discriminated between correct and incorrect performance when making CJ-FBs, for RJs, discrimination was less accurate. This indicates that, when self-rewarding, 5/6-year-olds did not take performance accuracy to the same extent into account as did 7/8-year-olds (confirming Hypothesis 3b). The finding that 5/6-year-olds are more inclined to reward themselves for incorrect responses supports the assertion that developmental differences in accuracy of self-evaluations are partly due to self-protection. A further interesting finding is that 5/6-year-olds were giving themselves lower rewards for correct responses than 7/8-year-olds. This finding may give a further indication that children rewarded themselves more for effort than for actual performance accuracy. When a response was correct they may have had the opinion that they put less effort into learning and retrieving this item, and therefore, they may have given themselves less reward (Folmer et al., 2008; Kurtz-Costes et al., 2005; Nichols 1978). In a future study, it may be interesting to let children estimate the invested effort into learning each specific item to test this interpretation.

In play-oriented learning environments such as kindergarten, children are often rewarded for effort instead of accuracy (Stipek & Tannatt, 1984). Therefore, 5/6-year-olds might lack understanding of the school principle that reward is given for correct performance. Possibly our 5/6-year-olds – still in kindergarten - did not yet fully understand reward principles, and therefore give themselves high rewards for incorrect performance. A potential limitation of the present study is that we did not assess children's understanding of giving reward points. However, a lack of understanding of reward principles does not seem to fully explain our findings. Firstly, the findings show that even 5/6-year-olds understood the principle that correct performance should be rewarded with more points than incorrect performance. Further, conform our findings, research in which children had to reward themselves as well as their peers shows that even preschoolers understand that reward is given

for correct performance (Schneider, 1998). Even 3-year-olds were able to use negative performance feedback when scoring peers; however, they gave their peers lower rewards for incorrect performance than they gave to their own incorrect performance. This suggests that 5/6-year-olds understand reward principles, but do not apply it when rewarding themselves. Of additional interest is our finding that both 5/6-year-olds and 7/8-year-olds gave themselves more rewards for incorrect responses in the second than in the first testing session, providing a further indication that more experience with the task, which may be an indication of perceived effort, led to more self-protection. Future research should investigate effects of task experience and time-on-task on self-rewarding and self-protection.

We only investigated self-evaluations, and based on our findings, we cannot draw conclusive statements about effects of item difficulty cues and self-protection on self-regulation. The relation between confidence judgments and regulation is robust in children (Destan et al., 2014; Destan & Roebers, 2015; Koriat & Ackerman, 2010; Krebs & Roebers, 2010). What remains unknown is to what extent item difficulty cues that affect self-evaluations, would subsequently affect self-regulation of study. Research shows that even kindergartners base study decisions on item difficulty and spend more time studying difficult than easy materials (Destan et al., 2014; Kobasigawa & Metcalf-Haggert, 1993). Therefore, we presume that when children base CJs on item difficulty, they also base self-regulatory decisions on this cue; this assumption should be addressed in future research.

Children were tested individually and the learning task was designed for research purposes, it was not actually a part of the children's school learning curriculum. Possibly in school learning situations, children may have different ways of self-evaluating and interpreting feedback. Furthermore, especially in a self-regulated learning context, learners' prior knowledge and their motivation have extensive effects on learning (Pekrun, Hall, Goetz

& Perry, 2014). Future research should address self-evaluation with actual classroom learning tasks, and address effects of learning motivation on self-evaluations and performance.

This study seems to indicate that children implement item difficulty cues, but that they often do not accurately evaluate performance because of self-protection. Self-protection thus seems to be a major cause of overconfidence in young children. There are clear examples in which overconfidence is harmful, for instance, when giving witness reports in court rooms, high confidence in incorrect information can have detrimental effects (Roebers et al., 2014). However, when overconfidence gives children the motivation to persist with the task it can be beneficial, and some self-protection could even be encouraged (Shin et al., 2007). For kindergartners, overconfidence might not be harmful because they do not need to self-regulate; most of their learning activities are co-regulated by their teacher (Hadwin, Järvelä, & Miller, 2011). However, when children have to self-regulate, children typically discard most of the materials from study for which they are confident, even when these are not yet well learned (Van Loon et al., 2013a). In education, overconfidence does not seem to improve motivation to learn; instead, children usually allocate their study to materials for which they are not confident that they are well learned. Thus, when children have to self-regulate learning, inaccurate self-evaluations likely have negative effects on learning. Therefore, teachers should teach the importance of accurate self-evaluations and support children to implement valid cues such as item difficulty. Future research should further address at what age inaccurate self-evaluations are beneficial for motivation to learn, and at what age overconfidence starts to have harmful effects on self-regulation and learning.

References

Artelt, C., & Schneider, W. (2015). Cross-Country Generalizability of the Role of Metacognitive Knowledge in Students' Strategy Use and Reading Competence. *Teachers College Record*, *117*(1)

Bjorklund, D. F., & Green, B. L. (1992). The adaptive nature of cognitive immaturity. *American Psychologist, 47*(1), 46-54. doi: 10.1037/0003-066x.47.1.46

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction, 7*(2), 161-186. doi: http://dx.doi.org/10.1016/S0959-4752(96)00015-1

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429-437. doi: 10.3758/MC.36.2.429

Coughlin, C., Hembacher, E., Lyons, K. E., & Ghetti, S. (2015). Introspection on uncertainty and judicious help-seeking during the preschool years. *Developmental Science, 18*(6), 957-971. doi: 10.1111/desc.12271

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. doi: 10.1016/j.jecp.2011.02.005

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223-234. doi: 10.1037/0022-0663.100.1.223

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5-to 7-year-old children. *Journal of experimental child psychology, 126*, 213-228.

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, 1-28.

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology, 47*, 274-296.

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

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction, 32*, 1-9.

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development, 24*(1), 15-23. doi: 10.1080/016502500383421

Folmer, A. S., Cole, D. A., Sigal, A. B., Benbow, L. D., Satterwhite, L. F., Swygert, K. E., & Ciesla, J. A. (2008). Age-related changes in children's understanding of effort and ability: Implications for attribution theory and motivation. *Journal of Experimental Child Psychology*, *99*(2), 114-134. doi:10.1016/j.jecp.2007.09.003

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*(1), 93-103. doi: 10.1758/MC.36.1.93

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*(2), 101-121. doi: 10.1007/s11409-008-9021-5

Hadwin, A. F., Järvelä, S., & Miller, M. (2011). Self-regulated, co-regulated, and socially shared regulation of learning. *Handbook of self-regulation of learning and performance*, *30*, 65-84.

Hoffmann-Biencourt, A., Lockl, K., Schneider, W., Ackerman, R., & Koriat, A. (2010). Self-paced study time as a cue for recall predictions across school age. *British Journal of Developmental Psychology, 28*(4), 767-784.

Howie, P., & Roebers, C. M. (2006). Developmental progression in the confidence-accuracy relationship in event recall: Insights provided by a calibration perspective. *Applied Cognitive Psychology, 21*, 871-893. doi: 10.1002/acp.1302

Kobasigawa, A., & Metcalf-Haggert, A. (1993). Spontaneous allocation of study time by first- and third-grade children in a simple memory task. *The Journal of Genetic Psychology, 154*(2), 223-235. doi: 10.1080/00221325.1993.9914736

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. doi: 10.1037/0096-3445.126.4.349

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

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

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009b). The memorizing effort heuristic in judgments of learning: A developmental perspective. *Journal of Experimental Child Psychology, 102*(3), 265-279. doi: 10.1016/j.jecp.2008.10.005

Koriat, A., & Nussinson, R. (2009). Attributing study effort to data-driven and goal-driven effects: Implications for metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1338-1343. doi: 10.1037/a0016374

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology, 80*(3), 325-340. doi: 10.1348/000709910x485719

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. doi: http://dx.doi.org/10.1016/j.appdev.2004.12.005

Lienert, G., & Raatz, U. (1998). Testaufbau und Testkonstruktion: Weinheim: Beltz.

Lipko, A. R., Dunlosky, J., Lipowski, S. L., & Merriman, W. E. (2012). Young children are not underconfident with practice: The benefit of ignoring a fallible memory heuristic. *Journal of Cognition and Development, 13*(2), 174-188. doi: 10.1080/15248372.2011.577760

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

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

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child development, 82*(6), 1778-1787.

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. doi: 10.1111/cdev.12004

McClelland, M. M., & Cameron, C. E. (2012). Self-regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives, 6*(2), 136-142. doi: 10.1111/j.1750-8606.2011.00191.x

Miele, D. B., Son, L. K., & Metcalfe, J. (2013). Children's naive theories of intelligence influence their metacognitive judgments. *Child Development, 84*(6), 1879-1886. doi: 10.1111/cdev.12101

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1-12. doi: 10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*(2), 378-384. doi: 10.3758/s13423-012-0343-6

Muis, K. R., Ranellucci, J., Trevors, G., & Duffy, M. C. (2015). The effects of technology-mediated immediate feedback on kindergarten students' attitudes, emotions, engagement and learning outcomes during literacy skills development. *Learning and Instruction*, *38*, 1-13. doi: 10.1016/j.learninstruc.2015.02.001

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

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125-141.

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child Development, 49*(3), 800-814. doi: 10.2307/1128250

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. doi: 10.1016/j.jecp.2013.12.011

Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. P. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology*, *106*(3), 696. doi: 10.1037/a0036006

Roderer, T., & Roebers, C. M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: An eye-tracking approach. *Metacognition and Learning*, *5*(3), 229-250. doi: 10.1007/s11409-010-9059-z

Roderer, T., & Roebers, C. M. (2014). Can you see me thinking (about my answers)? Using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance. *Metacognition and learning*, *9*(1), 1-23. doi:10.1007/s11409-013-9109-4

Roebers, C. M. (2014). Children's deliberate memory development: The contribution of strategies and metacognitive processes *The Wiley Handbook on the Development of Children's Memory, Volume I/II* (pp. 865-894).

Roebers, C. M., & Fernandez, O. (2002). The effects of accuracy motivation on children's and adults' event recall, suggestibility, and their answers to unanswerable questions. *Journal of Cognition and Development, 3*(4), 415-443. doi: 10.1080/15248372.2002.9669676

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. doi: 10.1016/j.jecp.2006.12.006

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. doi: 10.1111/j.1467-8624.1994.tb00805.x

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

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: Evidence for developmental trends. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (Vol. 14, pp. 391-409). Mahwah, NJ: Lawrence Erlbaum.

Schneider, W., & Pressley, M. (1997). *Memory Development Between Two and Twenty*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schneider, W., Vise, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring - Evidence from a judgment-of-learning task. *Cognitive Development, 15*(2), 115-134. doi: 10.1016/s0885-2014(00)00024-1

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. *Handbook of Metacognition in Education*, 278 - 298

Shin, H., Bjorklund, D. F., & Beck, E. F. (2007). The adaptive nature of children's overestimation in a strategic memory task. *Cognitive Development, 22*(2), 197-212. doi: 10.1016/j.cogdev.2006.10.001

Steinbeis, N., & Crone, E. A. (2016). The link between cognitive control and decision-making across child and adolescent development. *Current Opinion in Behavioral Sciences*, *10*, 28-32. doi: 10.1016/j.cobeha.2016.04.009 .

Stipek, D. J. (1984). Young children's performance expectations: Logical analysis or wishful thinking. *Advances in Motivation and Achievement, 3*(3).

Stipek, D. J., & Tannatt, L. M. (1984). Children's judgments of their own and their peers' academic competence. *Journal of Educational Psychology, 76*(1), 75-84. doi: 10.1037/0022-0663.76.1.75

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

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

Van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merrienboer, J. J. G. (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. doi: 10.1007/s11409-013-9100-0

Vohs, K. D., & Baumeister, R. F. (2011). *Handbook of Self-Regulation: Research, Theory, and Applications*: Guilford Press.

Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and Self-Regulated Learning: Theory, Research, and Applications* (pp. 297-314). New York: Lawrence Erlbaum Associates.

Zimmerman, B. J., & Schunk, D. H. (2001). Reflections on theories of self-regulated learning and academic achievement. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives* (Vol. 2, pp. 289-307): Lawrence Erlbaum Associates.

Table 1. Mean Performance and Mean Self-Evaluative Judgments for Correct and Incorrect Responses

	5/6-year-olds	**7/8-year-olds**
Performance (%)	44.69% (17.7)	64% (15.8)
Confidence Judgments (CJs)		
Correct Responses	4.26 (1.4)	4.89 (.8)
Incorrect Responses	3.70 (1.5)	3.65 (1.4)
Confidence Judgments with Feedback (CJ-FBs)		
Correct Responses	5.47 (.9)	5.84 (.5)
Incorrect Responses	2.28 (1.8)	2.77 (1.7)
Reward Judgments (RJs)		
Correct Responses	4.60 (1.4)	5.62 (.7)
Incorrect Responses	3.14 (1.3)	2.73 (1.3)

Note. Performance and self-evaluations (confidence judgments; confidence judgments facing feedback; reward judgments) for correct and incorrect recognition responses for the two age groups. Self-evaluations were given on a 7-point scale and range from 0 – 6. Standard deviations of the mean in parentheses.

Figure 1. Procedure

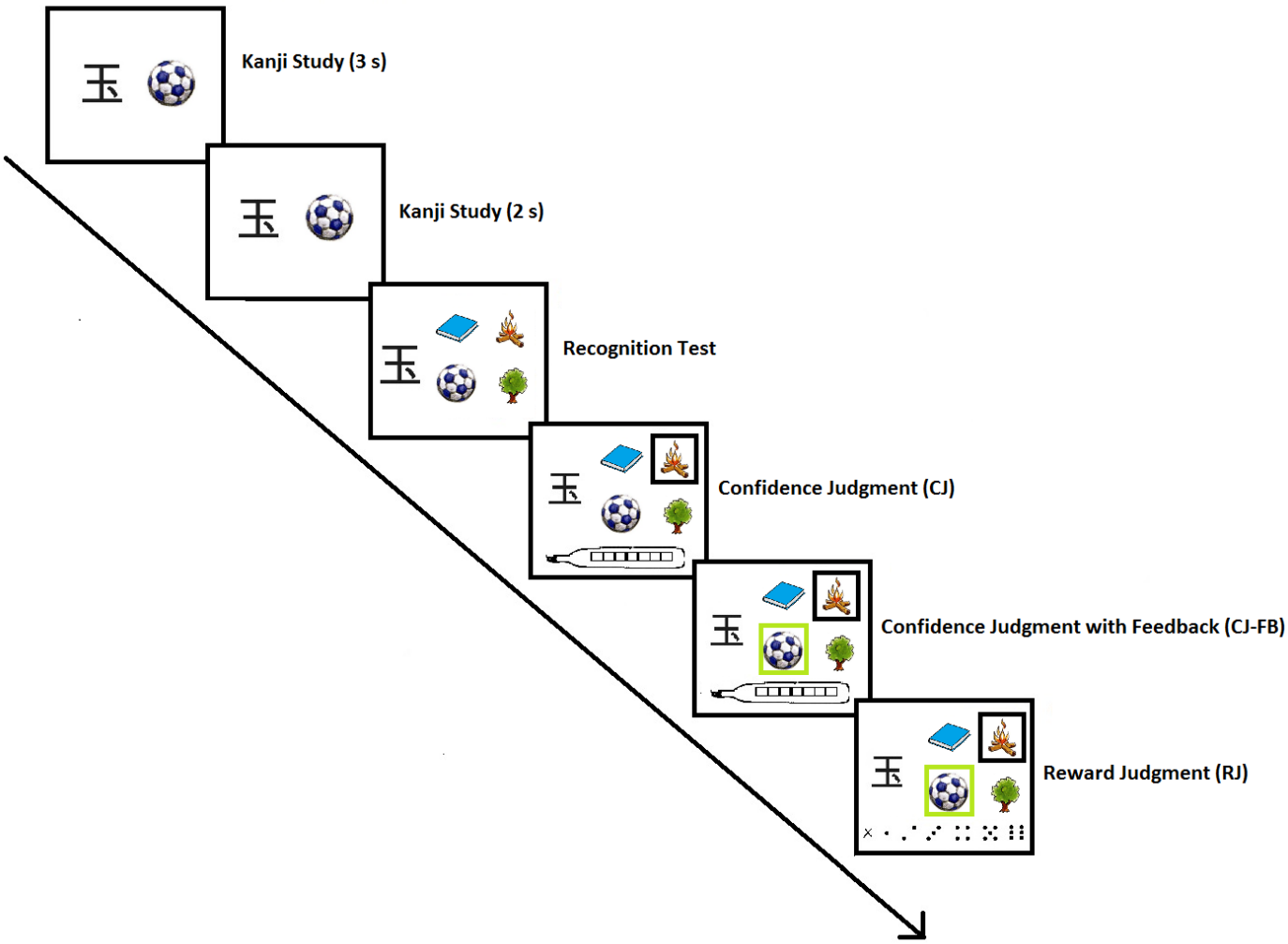
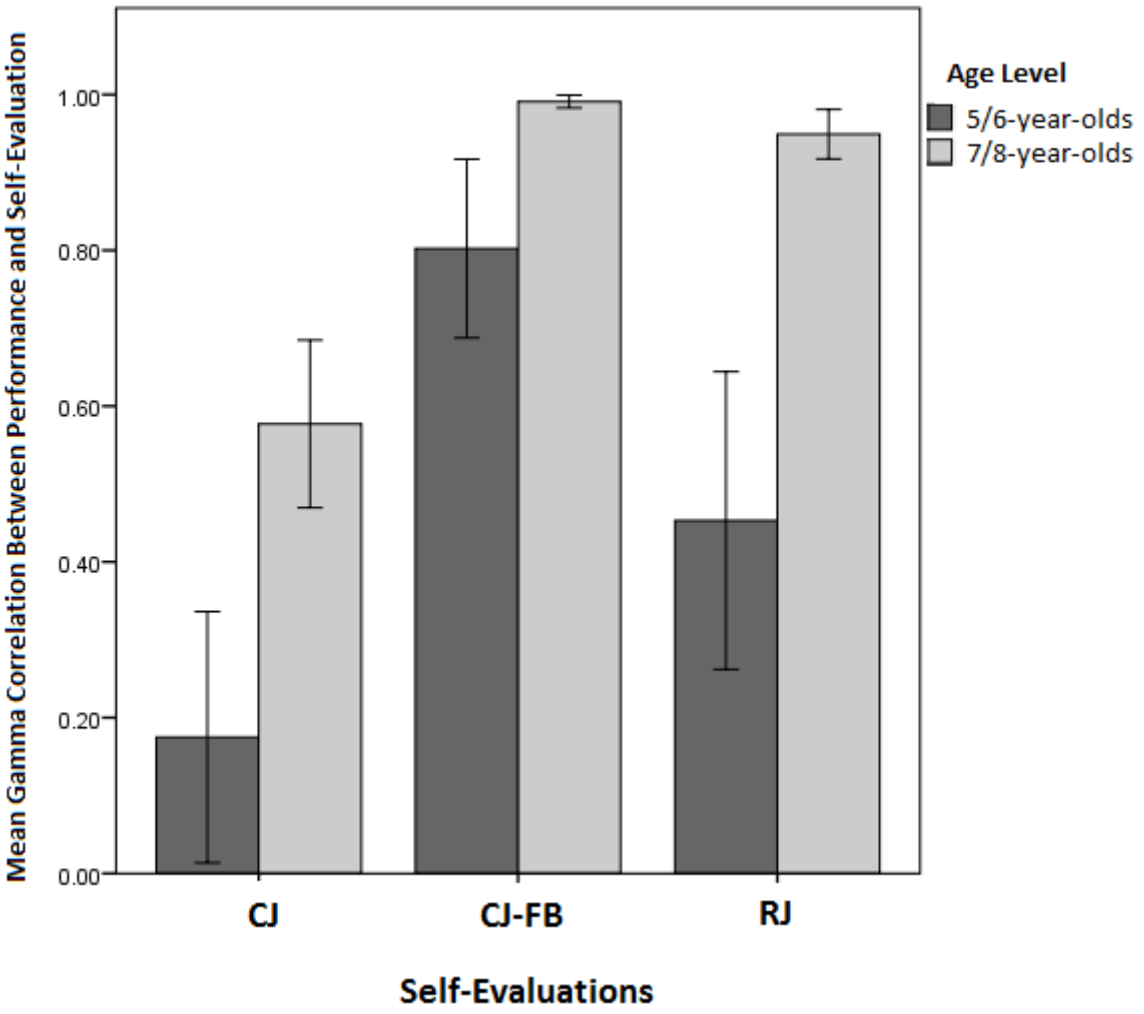


Figure 2.

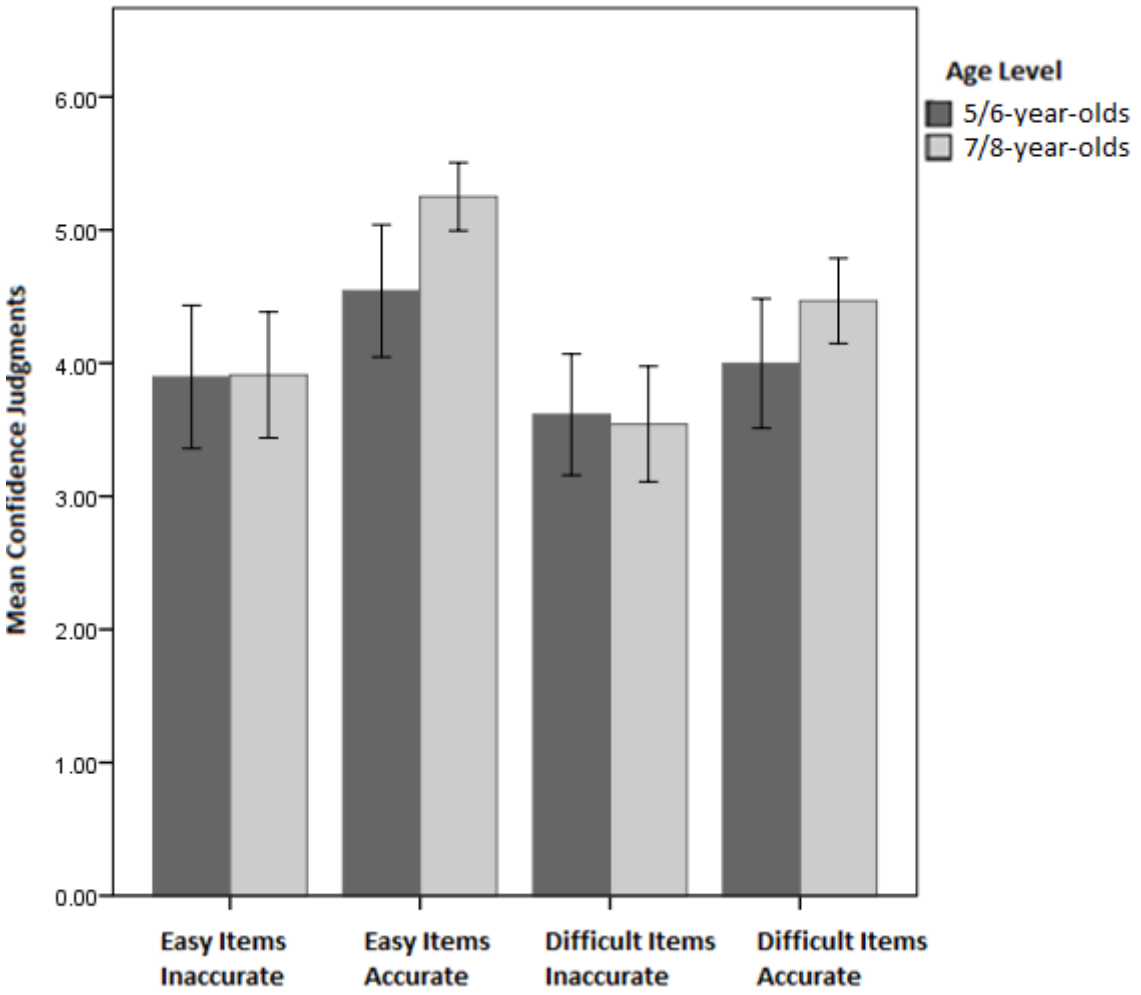
Gamma Correlations between Self-Evaluative Judgments and Performance



Gamma Correlations between Performance and Self-Evaluations: Confidence Judgments (CJs); Confidence Judgments facing Feedback (CJ-FBs) and Reward Judgments (RJs), for 5/6-year-olds and 7/8-year-olds. Error bars indicate the 95% confidence interval.

Figure 3.

Confidence Judgments as a Function of Item Difficulty



Confidence judgments for easy and difficult items for incorrect and correct responses for the two age groups. Error bars indicate the 95% confidence interval.