

# The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study

Citation for published version (APA):

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merriënboer, J. J. G. (2013). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning, 8*(2), 173-191. <https://doi.org/10.1007/s11409-013-9100-0>

**Document status and date:**

Published: 01/08/2013

**DOI:**

[10.1007/s11409-013-9100-0](https://doi.org/10.1007/s11409-013-9100-0)

**Document Version:**

Publisher's PDF, also known as Version of record

**Document license:**

Taverne

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.umlib.nl/taverne-license](http://www.umlib.nl/taverne-license)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[repository@maastrichtuniversity.nl](mailto:repository@maastrichtuniversity.nl)

providing details and we will investigate your claim.

## The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study

Mariëtte H. van Loon · Anique B. H. de Bruin ·  
Tamara van Gog · Jeroen J. G. van Merriënboer

Received: 17 February 2012 / Accepted: 20 May 2013 /  
Published online: 7 June 2013  
© Springer Science+Business Media New York 2013

**Abstract** When studying verbal materials, both adults and children are often poor at accurately monitoring their level of learning and regulating their subsequent restudy of materials, which leads to suboptimal test performance. The present experiment investigated how monitoring accuracy and regulation of study could be improved when learning idiomatic phrases. Elementary school children (fourth and sixth-graders) were instructed to predict their test performance by providing judgments of learning (JOLs). They provided JOLs immediately after studying each idiom, after studying all idioms, or after studying all idioms followed by generating sentences in which the idioms were used. Correlations between JOLs and test performance showed that delayed-JOLs and delayed-JOLs with sentence generation were more accurate than immediate JOLs. JOLs after sentence generation also improved regulation of study compared to delayed-JOLs only. Analyses of JOL reaction times suggest that delayed-JOLs led children to retrieve the literal meaning of the idiom, whereas JOLs after sentence generation led children to focus on connections between studied information, contextual information, and prior knowledge. Sentence generation presents a promising method to improve regulation of study, and thus idiom learning.

**Keywords** Monitoring · Regulation · Idioms · Judgment of learning · Sentence generation · Reaction times

Children frequently study verbal materials, both in the classroom and when doing homework. Previous research has found that monitoring and regulation play an important role in learning (e.g., Dunlosky and Metcalfe 2009; Nelson and Narens 1990; Thiede et al. 2003). During self-

---

M. H. van Loon (✉) · A. B. H. de Bruin · J. J. G. van Merriënboer  
Department of Educational Development & Research and Graduate School of Health Professions  
Education, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands  
e-mail: m.vanloon@maastrichtuniversity.nl

T. van Gog  
Department of Psychology, Erasmus University Rotterdam, P.O. Box 1783, 3000 DR Rotterdam,  
The Netherlands

regulated study, learners attempt to reduce the discrepancy between their learning goals and their perceived current state of learning (e.g., Nelson and Narens 1990; Thiede and Dunlosky 1999). Monitoring refers to assessing this current state of learning (Dunlosky and Metcalfe 2009), which is—ideally—used to regulate further study activities. That is, learners should monitor how well they have learned materials, and based on that information, make a decision on whether they will study these materials further or discard them (Thiede and Dunlosky 1999). More accurate monitoring has been shown to lead to more effective restudy decisions when regulating further study (Metcalfe and Finn 2008a). In turn, more effective regulation leads to higher test performance and better academic achievement (Ertmer and Newby 1996; Thiede et al. 2003).

Unfortunately, when studying textual materials monitoring accuracy is often low, implying that learners often cannot distinguish between what they do and do not understand. Correlations between predicted and actual test performance usually do not exceed .25 (see the review by Dunlosky and Lipko 2007; Maki 1998). As a result, regulation of study is likely to be ineffective, and therefore test performance might be suboptimal (Thiede et al. 2003). Recently, there has been increased interest in improving metacognitive skills, both in adults and young learners (e.g., Dunlosky and Lipko 2007; Huff and Nietfeld 2009; Lipko et al. 2009; Von der Linden et al. 2011).

### **Improving monitoring accuracy through long-term memory retrieval**

To improve monitoring and regulation of study, it is essential to understand how poor monitoring skills can be explained. According to Koriat's (1997) cue-utilization framework, monitoring is based on a variety of cues, and accuracy is often low because judgments are based on superficial cues, such as ease of processing (Koriat et al. 2009), familiarity with the topic of study (Griffin et al. 2009), or inaccurate prior knowledge (Van Loon et al. 2013). Since superficial cues are not valid predictors of test performance, monitoring might not be accurate when students use these as a basis for judgments.

Instead, when learners are required to activate valid cues about their learning, their predictions of test performance improve (Nelson and Dunlosky 1991; Thiede and Anderson 2003). One of the most influential studies on the improvement of monitoring accuracy is the delayed Judgment of Learning (JOL) experiment by Nelson and Dunlosky (1991). Participants were asked to provide JOLs after studying unrelated word pairs: They had to predict the likelihood of recall of each target word on the future test. Learners made JOLs either immediately after studying each item or after a delay, when a short interval was situated between study and JOLs. The delayed-JOLs improved monitoring accuracy to a great extent: The gamma correlation (Nelson 1984) between predicted and actual test performance was .90, compared with only .38 for immediate JOLs. Presumably, for immediate JOLs retrieval of targets that remained active in working memory is dominant. Delayed-JOLs are more accurate, because when predicting performance after studying, people attempt to retrieve the target from long-term memory. Since retrieval from long-term memory is also needed on the test, the delayed-JOL instruction provides learners with valid cues, which in turn leads to more accurate judgments (Nelson and Dunlosky 1991; see Rhodes and Tauber 2011, for a meta-analysis).

### **Improving monitoring accuracy through delayed generation**

Even though delayed-JOLs can improve monitoring for associative learning tasks, monitoring usually remains inaccurate when delayed-JOLs are made after study of more complex

and educationally relevant materials (Dunlosky and Lipko 2007; Maki 1998). When taking a test about complex materials, such as expository texts or figurative language, learners have to show that they understand information instead of simply retrieving the studied target (Thiede et al. 2003). Thus, retrieval of literal information from long-term memory does not necessarily provide learners with valid cues about performance.

Models explaining the understanding of content-rich materials agree that complex information can be processed literally, but importantly, can also be processed on a higher level with a focus on *gist* (Graesser et al. 1994; Kintsch and Van Dijk 1978; Van Den Broek and Kendeou 2008). To accurately understand complex, content-rich materials, the reader should not only focus on literal processing of information but also on gist, in order to integrate the novel information with contextual information and with prior knowledge (Kintsch 1998; Levorato et al. 2004; Perfetti 2007). For example, when learners study about blood circulation, they may have prior knowledge about the heart, lungs, and arteries, and have an understanding of the dynamics of oxygen transportation. The novel information about blood circulation needs to be linked to this prior knowledge in order to reach understanding. Because integration of studied information with prior knowledge is important for learning of content-rich materials, to improve monitoring learners should not only focus on literal retrieval, but also on cues that give them indications about gist, that is, understanding of the target (Anderson and Thiede 2008; Thiede et al. 2003).

Research has shown that generation instructions are an effective way to help learners focus on the gist (Anderson and Thiede 2008; De Bruin et al. 2011; Thiede and Anderson 2003; Thiede et al. 2005). For instance, instructing learners to generate summaries (Anderson and Thiede 2008; Thiede and Anderson 2003), or keywords about a text (De Bruin et al. 2011; Thiede et al. 2005) can focus their attention on connections between studied information, contextual information, and prior knowledge. These cues, indicating their understanding of the gist, are still available when learners provide JOLs and therefore lead to more accurate judgments (Thiede and Anderson 2003). It is important to note, though, that these generation tasks *only* have a beneficial effect on monitoring accuracy when these are delayed after study (Thiede et al. 2003, 2005). Memory for textbase representations decays more rapidly than memory for gist (Fletcher and Chrysler 1990; Kintsch et al. 1990; Anderson and Thiede 2008). When generating keywords or summaries is done immediately after study, details from the text are more activated within working memory (Cowan 1999). Therefore, when performing a generation task immediately after study, learners can rely more on the accessibility of details from the text. When learners generate keywords or summaries after a delay, such details have decayed and they need to base their judgments more on the gist of texts stored in long-term memory. The latter provide better cues, due to higher similarity with the test context (Thiede et al. 2005).

### Monitoring and regulation of study when learning idioms

A type of complex and educationally relevant verbal learning material that has –to the best of our knowledge– not yet been studied in research on monitoring and regulation, is material containing figurative language, such as idioms. Idioms are metaphorical phrases, such as ‘to throw in the towel’, whose meaning cannot be derived from the meaning of individual words that make it up (Swinney and Cutler 1979). Idioms are an important topic of study, because of their high frequency of occurrence in spoken and written language (Brinton et al. 1985; Nippold and Taylor 2002; Tabossi et al. 2009). Theories on idiom learning demonstrated that this metaphorical language can also, as other complex text materials, be processed on two

levels. First of all, idioms can be processed on a literal level. An idiom-learning task has characteristics of other associative learning tasks, because learners can literally retrieve the meaning of the idiom when presented with the figurative expression. Traditional theories of idiom learning suggest that idioms can be processed in much the same manner as individual words (Bobrow and Bell 1973; Swinney and Cutler 1979). As single words, each idiom has its own concept node, and one concept can activate multiple connected concept nodes (Cutting and Bock 1997).

More recent research suggests that idioms are not only stored as simple word-word associations, and processed with a focus on the literal meaning, but that the mental representation for idioms is much more complex than that for simple words (Langlotz 2006). Idiom comprehension is to a certain extent comparable to comprehension of expository texts (Langlotz 2006). When idioms are correctly understood, a learner also needs to focus on processing of the gist of this idiom. An idiomatic phrase is assumed to activate the same lexical items in a mental representation that are involved in the comprehension of non-metaphorical discourse (Sprenger et al. 2006). When idioms are stored and activated in memory, the idiom is activated with connections to contextual information, and the learner's prior knowledge (Cain and Towse 2008; Langlotz 2006; Levorato and Cacciari 1992, 1999). When the idiom's concept node is activated, this activation spreads to the lexical-syntactic level (which are the words of which the idiom consists), and also to the conceptual level (prior knowledge and contextual information). The knowledge about the gist of idioms is stored through these connection patterns (Cutting and Bock 1997; Langlotz 2006; Sprenger et al. 2006). Thus, to understand and use idiomatic expressions, learners should not only focus on retrieval of the literal meaning, but they also have to focus on connections which indicate their understanding of the gist.

Learning idioms is an interesting activity for metacognition research, because it has characteristics of associative learning tasks like study of related word pairs, and also characteristics of content-rich educational materials. If idioms can be processed with or without attention to gist, presumably the nature of idiom learning has implications for monitoring and regulation. As yet, it is unknown how well children can monitor and regulate study when learning idioms. Because idiom learning has characteristics of associative learning tasks, learners' monitoring accuracy and subsequent regulation are expected to benefit from a delayed-JOL instruction when compared with an immediate JOL instruction. A delayed-JOL instruction is assumed to lead children base JOLs on long-term memory retrieval (Nelson and Dunlosky 1991), which provides them with a more accurate indication of future test performance than when making immediate judgments when the information is still present in working memory. Because idioms can be processed with a focus on literal retrieval of the meaning, when judging learning of idioms, an attempt to retrieve the meaning can provide learners with a valid cue regarding future performance. Therefore, delayed-JOLs should be more accurate than immediate JOLs, and subsequent regulation should be more accurate following delayed-JOLs than following immediate JOLs.

Moreover, to fully understand the gist of idioms, they also need to be connected to prior knowledge and contextual information (Levorato and Cacciari 1999; Sprenger et al. 2006). If, for example, a student understands the idiom 'throw in the towel', then the mental representation of the idiom should not only capture the literal meaning, but also the gist of the idiom, because of connections with prior knowledge and possible contexts in which this idiom can be used.

Previous research has shown that sentence generation facilitates memory performance (Graf 1982). The present study investigates whether sentence generation could also improve metacognitive skills. When a learner has good idiom understanding, he should be able to create a novel sentence in which he correctly uses the studied idiom. Because sentence

generation is assumed to focus learners' attention on valid cues regarding gist information, when making JOLs this instruction is expected to have a beneficial effect on monitoring accuracy when compared to immediate JOLs and delayed-JOLs without this generation instruction (Anderson and Thiede 2008; Thiede et al. 2003). Moreover, we hypothesize that regulation would be more accurate following sentence generation than following immediate JOLs and mere delayed-JOLs.

### Reaction times when providing JOLs

As indicated above, immediate JOLs, delayed JOLs, and JOLs after sentence generation are expected to be based on different cues. Measures of reaction times (RTs) when children make JOLs (hereafter referred to as JOL RTs) provide an opportunity to further investigate assumptions about the use of cues.

First of all, retrieval of accessible information from working memory is a faster process than long-term memory retrieval (Conway and Engle 1994; Sternberg 1969). Thus, if children base JOLs on targets that remained active in working memory when monitoring idiom understanding immediately after learning, these immediate JOL RTs should be shorter than JOL RTs for the judgments which are provided after a delay.

When investigating possible differences in cue use between monitoring following the delayed-JOLs with or without sentence generation, it is especially interesting to look at the JOLs at the low end of the scale. When children provide low JOLs, they indicate that they have low confidence to produce the correct meaning of the idiom at the test. Research on JOL RTs shows that low JOLs are usually accompanied by slow JOL RTs, because information which learners attempted to retrieve did not come to mind easily (e.g., Benjamin et al. 1998; Koriat and Ma'ayan 2005; Metcalfe and Finn 2008b; Son and Metcalfe 2005). Findings by Metcalfe and Finn (2008b) showed that when learners made self-paced JOLs about an associative learning task, RTs for the low JOLs were longer than RTs for higher JOLs.

However, JOL RT research provided evidence that not all delayed-JOLs are necessarily based on a retrieval attempt (Metcalfe and Finn 2008b; Son and Metcalfe 2005). When investigating delayed-JOLs for an associative name learning task, Son and Metcalfe (2005) found that learners did not spontaneously attempt retrieval, and the low JOLs were provided quickly. Moreover, Metcalfe and Finn (2008b) found that when participants had to provide JOLs under time constraints, low JOLs were provided fast. This finding led to the assumption that two different processes can underlie delayed-JOLs: Retrieval of studied information, and a quick assessment stage in which participants do not attempt to literally retrieve information (Metcalfe and Finn 2008b; Son and Metcalfe 2005).

In the present study, learners provide self-paced JOLs about the studied idioms. Differences in JOL RTs between the sentence generation and uninstructed delayed-JOL conditions are expected to occur at the low end of the JOL scale. When learners provide JOLs following sentence generation, low JOLs are presumed to result from cues that the idiom could not be connected to prior knowledge and contextual information, instead of from an unsuccessful attempt to retrieve the literal meaning. Therefore we expect low JOLs after sentence generation to be fast. The processes involved in making uninstructed delayed-JOLs are expected to differ from those involved when the person first generates sentences. In line with the assumption that delayed-JOLs are based on a retrieval attempt (Rhodes and Tauber 2011), we expect to replicate findings by Metcalfe and Finn (2008b) that low delayed-JOLs are accompanied by slow RTs.

## Development of metacognitive skills

For associative learning tasks, the delayed-JOL effect has also been demonstrated in studies with young learners (Koriat and Shitzer-Reichert 2002; Schneider et al. 2000). It has been shown that monitoring skills for associative learning tasks can develop at a young age; children 6 years of age could more accurately predict their recall after a delay rather than immediately after study (Schneider et al. 2000). Research on metacognitive skills when learning paired associates showed that age differences are more apparent when investigating children's regulation of study. Findings indicate that sixth-grade learners show better regulation abilities than fourth-grade learners (e.g. Dufresne and Kobasigawa 1989; Krebs and Roebbers 2010). Thus, even though monitoring for associative learning tasks can be accurate for young children, the ability to use JOLs to guide subsequent study does not seem to be developed until around 10 years of age (Dufresne and Kobasigawa 1989; Zimmerman and Martinez-Pons 1990).

Research on development of idiom learning indicates that children aged seven and over start to develop the ability to understand the meaning of idioms, and idiom comprehension skills continue to develop into adolescence (Brinton et al. 1985; Cain et al. 2009). Until present, it is unclear whether developmental differences affect monitoring and regulation of idiom study. To date, most research on metacognition when studying *complex* verbal materials has focused on adult learning. While there are some studies on paired associate learning in young children (e.g. Schneider et al. 2000), to the best of our knowledge, only a study by De Bruin et al. (2011) investigated how monitoring and subsequent regulation of study could be improved when children had to study expository text. They investigated whether generating keywords improved fourth-, sixth-, and seventh-graders' JOL accuracy. De Bruin et al. (2011) showed that generating keywords after a delay improved monitoring accuracy only for sixth- and seventh-graders. For the fourth-graders, there was no effect of the delayed generation task. These findings imply that age differences in monitoring may occur for more complex learning tasks (Brown et al. 1983; De Bruin et al. 2011; Krebs and Roebbers 2010). De Bruin et al. (2011) asked children, in addition to providing JOLs, to select texts that they would like to reread. The results showed that participants did use their monitoring to decide which texts to reread, even the fourth-graders. Thus, even though fourth-graders could not accurately monitor their learning, they did use the results of their monitoring to regulate their study.

These findings indicate that, when learning more complex, educationally relevant verbal materials, it is not yet clear whether and how monitoring and regulation can be improved for fourth-graders. Because idiom learning can be considered a complex task, we explored whether sixth-graders would show more accurate monitoring and more accurate regulation than fourth-graders, in line with the findings by De Bruin et al. (2011).

## Present study

The aim of the present study was to investigate whether a delayed-JOL and a sentence generation instruction improve monitoring accuracy and subsequent regulation of idiom study compared to an immediate JOL instruction. To explore whether developmental differences affect monitoring and regulation of study with an idiom learning tasks, we tested both fourth- and sixth-graders. JOL RTs were measured in order to acquire information about the effects of instructions on the speed of providing JOLs, which might provide insight into cues used when judging idiom learning.



We tested the following hypotheses: Regarding monitoring accuracy, we expected that delayed-JOLs would be more accurate than immediate JOLs (Hypothesis 1a), that JOLs after sentence generation would be more accurate than delayed-JOLs (Hypothesis 1b), and that monitoring would be more accurate for sixth-grade than fourth-grade learners (Hypothesis 1c). Because monitoring accuracy has been shown to affect regulation, we expected that restudy selections would be more accurate following delayed-JOLs than immediate JOLs (Hypothesis 2a), and moreover that restudy selections would be more accurate following sentence generation than following delayed-JOLs (Hypothesis 2b). Furthermore we expected restudy selections to be better for sixth-grade than fourth-grade learners (Hypothesis 2c). With regard to JOL RTs we expected immediate JOLs to be faster than delayed-JOLs and JOLs after sentence generation (Hypothesis 3a), and we expected that the low JOLs would be provided faster following sentence generation than for uninstructed delayed-JOLs (Hypothesis 3b).

## Method

### Participants and design

Participants were 90 primary school pupils, 49 fourth-grade students (18 boys and 32 girls,  $M=9.9$  years,  $SD=0.6$  years) and 41 sixth-grade students (15 boys and 26 girls,  $M=12.0$  years,  $SD=0.6$  years). They were recruited from five elementary schools in the south of the Netherlands. For all participants, parental consent was obtained. The study was based on a 2 (grade level)  $\times$  3 (instruction) between-subjects design. For the idiom study task, participants were assigned randomly to one of three groups, which were immediate JOL ( $n=30$ , 17 fourth-grade, 13 sixth-grade); uninstructed delayed-JOL ( $n=28$ , 15 fourth-grade, 13 sixth-grade); and delayed-JOL after sentence generation ( $n=32$ , 17 fourth-grade, 15 sixth-grade).

### Materials

*Idioms* Twenty Dutch idioms with their meanings were selected for the idiom learning task, Table 1 shows examples of the idioms. A pilot study involving 24 elementary school children who did not participate in the final experiment confirmed that these idioms were unfamiliar to students of this age. In this pilot study, learners (10 4th grade and 14 6th grade pupils) were asked if they knew the meaning of 35 idioms. These idioms were presented to them without their meaning. They were told to write down the meaning of the idiom if they thought they knew this, and leave a question-mark if they did not know the meaning of the idiom. From the 35 idioms in this pilot study, 20 idioms were selected for the present study. For the selected idioms, none of the 4th grade learners wrote down the correct meaning (of their responses 55.0 %,  $SD=23.1$ , was a question mark; 45.0 %,  $SD=23.1$ , was an incorrect

**Table 1** Examples of Dutch idioms

Dutch idiom	English translation	Meaning
Korte armen hebben	To have short arms	To have a lack of money
Een bok schieten	To shoot a goat	To do something stupid
De hoogte hebben	To have height	To be drunk
Met spek schieten	To shoot with bacon	To brag about something



response). Only 3.6 % ( $SD=3.4$ ) of the responses of the 6th grade learners was correct (whereas 72.5 %,  $SD=16.6$ , was a question mark; and 25.0 %,  $SD=15.5$ , an incorrect response).

In the study phase, the idiom with the meaning and an example sentence in which the idiom was used were presented on a laptop computer screen. An example of an English equivalent of the learning task is ‘Step on it (*idiom*) – Hurry up (*meaning*) – Step on it, the taxi will be here any time and you’re not even dressed (*example sentence*)’.

*JOLs* When requested to provide JOLs, participants were presented with the idiom on the screen, and the question: “How confident are you that you will be able to provide the meaning of this idiom at the test?” JOLs were given with a mouse click on a thermometer with nine squares, comparable to the thermometer used by Koriat and Shitzer-Reichert (2002). The scale was depicted as a color drawing of a thermometer ranging from deep blue (the first square, no chance to recall the meaning) to deep red (the ninth square, completely certain to recall the meaning).

*Sentence generation instruction* Participants were instructed to generate 20 novel sentences in which the 20 previously studied idioms occurred. The idiom was provided on the screen as a cue, and the children were encouraged to generate a meaningful novel sentence in which they used the idiom. The children orally produced the novel sentences, the experimenter typed the sentences.


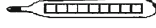

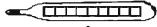
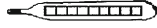

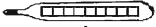

*Selection for restudy* Participants were shown a sheet of paper with a  $2 \times 10$  array in which each cell was filled by a previously studied idiom. The format of an array was chosen because this places less demands on working memory capacity than an item-by-item format (Thiede and Dunlosky 1999). The order of items was randomized, because Ariel et al. (2011) demonstrated that decisions for re-study are influenced by presentation order. Participants were required to mark with a pen, which phrases they would like to re-study, and 0 to 20 items could be selected.

*Test* The test consisted of self-paced recall of the 20 idioms, in which participants were asked to type the correct meaning when provided with the idiom.

*Apparatus* The idioms, sentence generation task, JOL task, and recall test were presented on a PC compatible laptop and E-prime software (Schneider et al. 2002) was used for stimulus presentation as well as response and reaction time recording. For each participant, the order of items in the subtasks was randomized anew by the software from study to test.

## Procedure

Figure 1 depicts an overview of the procedure of the study. All children were tested individually in quiet rooms of their schools. Testing time was about 40 min per child. The children were instructed that idioms with their meaning would appear on the computer screen together with an example sentence in which the idiom was used. They were told they had to study the idioms so they would be able to show that they comprehended the meaning when cued with the idiom. They were presented with an example of a study trial and an example of a test trial before the idiom learning task started. The study phase was self-paced, that is, the idiom, its meaning, and the example sentence remained on the screen until the participant pressed the space bar. No feedback was provided during idiom learning.

Instructions Idiom Learning Task		
Immediate JOL Group	Delayed JOL Group	Sentence Generation Group
Study Idiom 1 JOL Idiom 1 	Study Idiom 1 Study Idiom 2 ⋮ Study Idiom 20	Study Idiom 1 Study Idiom 2 ⋮ Study Idiom 20
Study Idiom 2 JOL Idiom 2 ⋮ Study Idiom 20 JOL Idiom 20 	<b>Filler Task</b>	Sentence Idiom 1 Sentence Idiom 2 ⋮ Sentence Idiom 20
<b>Filler Task</b>	JOL Idiom 1  JOL Idiom 2  ⋮ JOL Idiom 20 	JOL Idiom 1  JOL Idiom 2  ⋮ JOL Idiom 20 
<b>Restudy Selections</b>		
To crack the whip		To kick the bucket
Down in the dumps		To jump the gun
<b>Test</b>		
What is the meaning of 'To kick the bucket'?		
What is the meaning of 'To crack the whip'?		

**Fig. 1** Representation of the procedure of the study

Before providing JOLs, children received an explanation of the thermometer and practiced with an example. Depending on assigned condition, JOLs were provided either immediately after studying each idiom (immediate JOL instruction), after studying all idioms and performing a filler task (delayed-JOL instruction), or after studying and generating sentences for all idioms (sentence generation instruction). Note that following the sentence generation instruction, JOLs were also provided after a delay. Participants in the sentence generation condition first practiced the sentence generation task with an example idiom. After generating an example sentence they received feedback once on the quality of this sentence. Then the actual generation task started. When the children indicated that they could not generate a sentence for an idiom, the experimenter encouraged them a maximum of three times to try to generate a sentence. No feedback was provided on the quality of the generated sentences. Participants in the other conditions performed a filler task (a mental rotation task), which was done to create the same delay for all participants between the self-paced study phase and the recall test.

After studying and providing JOLs, interspersed with a sentence generation or filler task depending on assigned condition, participants had the opportunity to select which items they would like to restudy. They were reminded that they had to take a test at a later point, on which they had to correctly recall the meaning of the 20 idioms. Note that they did not actually get to restudy those items because restudy would have intervened between JOLs and recall on the test (Thiede and Dunlosky 1999).

The test was self-paced, after pressing the F1 key, the next idiom appeared. After completing the test, all the children were told that they performed well and were thanked for their participation.

## Analyses

*Test performance scoring* For 52 participants (57 %), two raters, who were blind to the experimental condition, coded the responses on the test items as incorrect (0), partially correct (0.5), or completely correct (1). Each meaning of an idiom consisted of 1–3 idea units and credit was given for responses that expressed these ideas. Comprehension was emphasized, so if the participant responded with a paraphrase of the original meaning, for example ‘make haste’ instead of ‘hurry up’ when cued with ‘step on it’, this was scored as correct. Partial credit was given if a response expressed one or two but not all of the idea units contained in a definition. The raters agreed on coding 94 % of the time. Because inter-rater agreement was high, one rater scored the test performance of the remaining participants, and these scores were used in the analysis.

*Monitoring and regulation accuracy* The gamma correlation was used to measure monitoring accuracy and regulation of study, because this non-parametric statistic is considered the most appropriate measure of relative accuracy (Nelson 1984). To assess monitoring accuracy, intra-individual gammas were calculated between participants’ JOLs and their actual score for each of the 20 idioms. For six participants, monitoring gammas could not be computed, because of invariance in test performance (i.e., they had no items correct on the test). For regulation of study, the intra-individual gamma correlation between JOLs and whether an idiom was selected for restudy (yes = 1, no = 0) was calculated, as operationalized by Thiede et al. (2003). More accurate regulation was indicated by a stronger negative correlation, a gamma of  $-1$  indicates that previous JOLs are adequately translated into the selection of idioms for restudy. For three participants gammas could not be computed, because of invariance in restudy selections.

For monitoring accuracy gammas and regulation of study gammas, the ANOVA assumption of equality of variance was violated. Therefore, gammas were transformed using the Fisher-Z transformation (Fisher 1915). Note that the Fisher-Z transformation did not change the pattern of results in any substantive manner. For the analyses described below, all reported post-hoc tests are Bonferroni-corrected *t*-tests, and partial eta squared is reported as a measure of effect size.

*RT data* To investigate the effects of JOL instructions on JOL RTs, JOLs were coded into three levels: Low (JOL rating 1–3); Intermediate (JOL rating 4–6), and High (JOL rating 7–9). Because of the hierarchical structure of the data (each participant provided 20 JOLs, and participants were nested in grade levels), the effect of instructions on JOL RTs was investigated with a mixed model multilevel regression analysis. For this analysis, a random intercept was included for every participant. JOL instruction, Grade level (grade 4 and grade 6) and JOL level (low, intermediate, high) were included as fixed factors, and in addition the interaction JOL Instruction \* JOL level was added to the model. JOL RTs were measured in ms, for significant effects the regression coefficients, standard errors, and 95 % confidence intervals are reported.

## Results

Table 2 presents Mean study time, Mean JOLs and Mean of individual Median JOLs for each experimental condition and grade, and JOL skewness is reported with *z*-scores to

**Table 2** Mean study time per idiom (in seconds), Mean JOLs, mean of individual Median JOLs, JOL distribution, selection for restudy, and test performance (standard deviations of the mean in parentheses)

JOLs	Fourth grade			Sixth grade		
	Immediate	Delayed	Sentence generation	Immediate	Delayed	Sentence generation
<i>n</i>	17	15	17	13	13	15
Mean Study Time per Idiom	7.8 (2.7)	8.9 (4.4)	8.2 (2.3)	8.7 (2.9)	8.8 (3.5)	7.8 (3.0)
Mean JOLs	5.6 (2.2)	5.3 (2.8)	5.2 (3.0)	5.7 (2)	5.8 (2.6)	4.8 (3.2)
Median JOLs	5.5 (1.4)	5.2 (2.1)	5.2 (2.1)	5.6 (1.4)	5.7 (1.7)	4.8 (2.7)
<i>z</i> -score skewness JOLs	−2	−0.71	−0.67	−2.27	−1.85	1.03
Selection for restudy (%)	55.6 (20.7)	47 (17)	51.8 (20.1)	35.8 (22.1)	43.8 (10.2)	55.3 (23.5)
Correct test performance (%)	26.8 (23.5)	23.5 (19.4)	34.3 (23.7)	36 (20.5)	33 (19.4)	36.8 (23.5)

Mean study time per idiom, Mean JOLs, Mean of individual Median JOLs, *z*-scores indicating skewness of JOLs, percentage of items selected for restudy, and percentage correct performance for the JOL instructions and grade levels

provide insight into the JOL distributions. In addition, Table 2 presents the percentage of idioms selected for restudy, and the percentage of correct test performance. In the analyses described below, interaction effects are only reported when significant.

ANOVAs showed no significant difference between the Mean study time per idiom for the different conditions,  $F(2, 89)=.56, p=.57$ . Because the JOL distribution was skewed, Median JOLs were used to investigate whether there was a relation between JOL instructions and JOL magnitudes. ANOVAs exhibited no significant relation between JOL instructions and Median JOLs,  $F(2, 87)=.70, p=.50$ . In addition, there was no significant relation between JOL instructions and test score,  $F(2, 87)=.89, p=.41$ .

Monitoring accuracy and regulation of study

Table 3 shows the regular and the Fisher-Z transformed gamma correlations for monitoring accuracy. A 3 (JOL Instruction) × 2 (Grade) ANOVA revealed a significant main effect of JOL

**Table 3** Monitoring accuracy and regulation of study (standard deviations of the mean in parentheses)

JOL instructions	Fourth grade			Sixth grade		
	Immediate JOL	Delayed-JOL	Sentence generation	Immediate JOL	Delayed-JOL	Sentence generation
Monitoring Accuracy						
Regular Gammas	.25 (.38)	.75 (.26)	.78 (.24)	.22 (.43)	.70 (.29)	.89 (.09)
Fisher-Z Gammas	.31 (.49)	1.35 (.87)	1.35 (.76)	.26 (.50)	1.2 (.82)	1.7 (.67)
Regulation of Study						
Regular Gammas	−.14 (.37)	−.73 (.24)	−.79 (.42)	−.41 (.3)	−.78 (.21)	−.93 (.07)
Fisher-Z Gammas	−.18 (.46)	−1.17 (.73)	−1.7 (.97)	−.48 (.36)	−1.38 (.81)	−2.01 (.68)

Regular gamma correlations and Fisher-Z transformed gamma correlations indicating monitoring accuracy and regulation of study for the JOL instructions and the grade levels

instruction on monitoring accuracy,  $F(2, 81)=24.61$ ,  $p<.001$ ,  $\eta_p^2=.39$ . Post hoc tests showed higher monitoring accuracy for participants who received the delayed-JOL instruction ( $p<.001$ ) and the sentence generation instruction ( $p<.001$ ) than for participants who received the immediate JOL instruction. This finding that delayed-JOLs led to more accurate monitoring than immediate JOLs confirms Hypothesis 1a. Results did not show a significant difference between the sentence generation and the delayed-JOL instruction ( $p=.259$ ; Hypothesis 1b). Moreover, there was no main effect of Grade level on monitoring,  $F(1, 78)=.09$ ,  $p=.77$  (Hypothesis 1c).

Table 3 shows the regular and the Fisher-Z transformed gamma correlations for regulation of study. The 3 (JOL instruction) \* 2 (Grade level) ANOVA on regulation of study showed a significant effect of JOL instruction,  $F(5, 81)=34.87$ ,  $p<.001$ ,  $\eta_p^2=.39$ . Post hoc tests showed that the delayed-JOL instruction improved regulation of study when compared to the immediate JOL instruction ( $p<.001$ ), confirming Hypothesis 2a. The sentence generation instruction improved regulation of study when compared to both the immediate JOL ( $p<.001$ ) and the delayed-JOL ( $p=.01$ ) instructions, confirming Hypothesis 2b, which predicted that regulation would be most accurate following sentence generation.

The effect of Grade is not significant,  $F(5, 81)=1.67$ ,  $p=.07$ ,  $\eta_p^2=.04$ , although there is a tendency indicating that regulation of study may be more accurate for sixth-graders than for fourth-graders. This finding, although only marginally significant, seems to be in line with Hypothesis 2c, which predicted better regulation for the older age group.

There was no significant effect of JOL instruction on percentage of idioms selected for restudy,  $F(2, 87)=1.34$ ,  $p=.27$ ; nor of Grade on percentage of items selected for restudy,  $F(1, 88)=2.09$ ,  $p=.15$ . This shows that the above mentioned effects were not due to differences in the number of items selected.

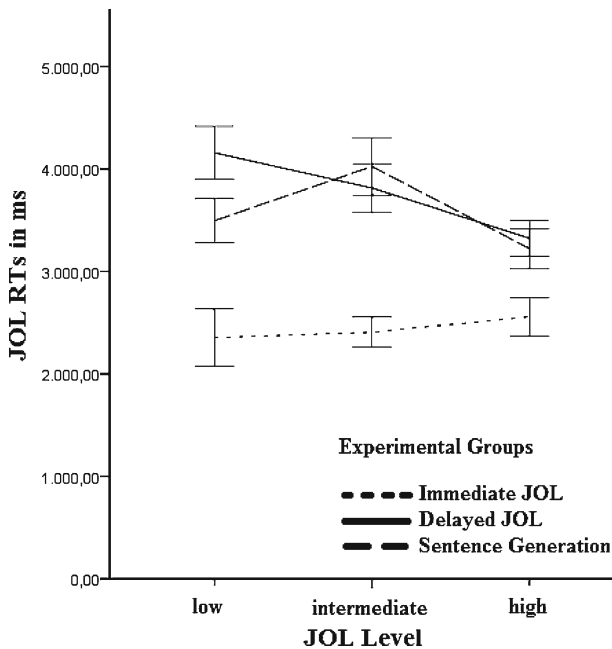
### JOL reaction times

Figure 2, in which JOL RTs are visible as a function of JOL instructions and JOL levels, provides a visual representation of the results of the analyses described below.

The multilevel regression analysis revealed significant main effects of JOL instruction, JOL level, and Grade level, and moreover a significant JOL Instruction \* JOL level interaction effect. The significant effect of JOL instruction on JOL RTs,  $F(2, 1773)=120.06$ ,  $p<.001$ , shows, in line with Hypothesis 3a, that immediate JOLs were provided faster than delayed-JOLs ( $b=-1287.89$ ,  $SE=87.86$ ,  $CI=-1460.21$  to  $-1115.57$ ), and JOLs after sentence generation ( $b=-632.22$ ,  $SE=139.32$ ,  $CI=-905.48$  to  $-358.97$ ). There was no significant difference between mean JOL RTs following the sentence generation and the delayed-JOL instruction,  $p=.192$ .

Even though we did not have specific hypotheses regarding the effect of grade level on overall JOL RTs, we will report the analyses in order to provide a complete picture of the data. The significant effect of Grade level,  $F(1, 1773)=46.41$ ,  $p<.001$ , shows JOLs were provided faster by the sixth-graders than the fourth-graders ( $b=479.50$ ,  $SE=.70.30$ ,  $CI=341.44$  to  $617.55$ ).

Importantly, as visible in Fig. 2, there was an interaction effect between JOL instructions and JOL levels, and this interaction effect was significant,  $F(4, 1773)=8.91$ ,  $p<.001$ . Further investigation of this interaction effect shows, in line with Hypothesis 3b, that the JOL RTs differ at the low end of the JOL scale (JOLs 1 – 3). As expected, for low JOLs there was a significant difference between the delayed-JOL condition and the sentence generation condition,  $F(2, 479)=36.87$ ,  $p<.001$ . Low JOLs were provided faster following the sentence generation instruction than following the delayed-JOL instruction ( $b=662.97$ ,  $SE=164.77$ ,  $CI=339.20$  to  $986.74$ ).



**Fig. 2** JOL RTs (in ms) as a function of JOL instruction and JOL level. Error bars indicate the 95 % Confidence Interval. JOL RTs are fastest for Immediate JOLs. For the low JOLs, JOL RTs are faster following sentence generation than following uninstructed delayed-JOLs

## Discussion

This study investigated children's metacognitive skills when learning idioms. We aimed to address the question of whether these skills would improve by focusing attention on literal retrieval cues and gist cues prior to monitoring. Therefore, we investigated the effects of immediate JOL, delayed-JOL, and sentence generation instructions on monitoring and regulation of fourth- and sixth-grade learners.

First of all, we hypothesized that delayed-JOLs would be more accurate than immediate JOLs (Hypothesis 1a). Results on monitoring accuracy show that, in line with our expectations, providing JOLs after a delay improved monitoring. Also when providing JOLs following sentence generation, monitoring accuracy significantly improved when compared to immediate JOLs. Contrary to our expectations (Hypothesis 1b), monitoring accuracy did not differ between the delayed-JOL and the sentence generation group. Monitoring of idiom study was highly accurate following both instructions: Mean gamma correlations were .73 following delayed-JOLs and .84 following sentence generation. Note that both following delayed-JOL and sentence generation instructions, monitoring judgments were provided after a delay. Thus, the benefits of the sentence generation instruction on monitoring may be attributable to the delayed-JOL procedure. In addition, we expected that JOLs would be more accurate for sixth-graders than for fourth-graders (Hypothesis 1c). This expectation was not confirmed, monitoring accuracy did not differ between the two age groups.

In line with our expectations (Hypothesis 2b), the sentence generation instruction improved regulation of study, compared to both delayed-JOL and immediate JOL instructions. In addition, as predicted by Hypothesis 2a, the delayed-JOL instruction led to better regulation

than the immediate JOL instruction. Moreover, the nearly significant effect of grade level on regulation of study implies that regulation was more accurate for sixth-graders than for fourth-graders, in line with our expectations (Hypothesis 2c). Older children tended to regulate study more accurately than the younger ones, by more often selecting idioms for restudy that they had judged as more difficult, and vice versa.

The measurements of JOL reaction times provided additional insights into effects of the three instructions on cue utilization when monitoring idiom learning. Our findings suggest that indeed, the JOL instructions led children to rely on different cues. In line with our expectations, JOLs were made fastest following the immediate JOL instruction (Hypothesis 3a), presumably because immediate judgments were based on monitoring of information that remained available in working memory (Nelson and Dunlosky 1991). Moreover, the speed of providing monitoring judgments differed for the low JOLs, with which participants indicated that they had low confidence that they would provide the correct meaning of the idiom at the test. Low JOLs following sentence generation were provided faster than low JOLs following uninstructed delayed-JOLs (Hypothesis 3b). The JOLs following the sentence generation task show the same pattern as the judgments in the speeded JOL condition by Metcalfe and Finn (2008b). These findings suggest that, when monitoring learning following sentence generation, participants did not attempt to retrieve the meaning of the idiom. On the other hand, following the delayed-JOL instruction, the RTs for the low JOLs were longer. The slope for JOL RTs following the delayed-JOL instruction shows the same pattern as findings on RTs for delayed-JOLs by Benjamin et al. (1998), and the uninstructed delayed-JOL condition of the research by Metcalfe and Finn (2008b). This suggests that, when children are asked to monitor idiom learning after a delay, they spontaneously attempt to retrieve the meaning. On the other hand, monitoring idiom learning following sentence generation seems to lead learners not to focus on retrieval fluency, but on a different cue. Based on studies by Thiede et al. (2003) and De Bruin et al. (2011), we assume that the sentence generation task leads learners to focus on valid cues related to the ability to connect the idiom with prior knowledge, instead of retrieval fluency cues. The reaction time findings thus seem to provide additional support for the argument that instructions can affect young learners cue utilization.

Our results expand findings on the effect of instructions on metacognitive skills (e.g., De Bruin et al. 2011; Nelson and Dunlosky 1991; Thiede et al. 2003) to children's idiom learning. It can be concluded that the delayed-JOL effect that was found with associative learning tasks (Nelson and Dunlosky 1991) also applies to more complex and educationally relevant tasks such as studying idioms. The memory retrieval account, which states that JOLs are accurate because they rely on information from long-term memory (Nelson and Dunlosky 1991; Rhodes and Tauber 2011) provides an explanation for more accurate monitoring when judging idiom learning after a delay. This implies that metacognition for idiom learning has characteristics of metacognition for associative learning tasks. This might also explain why there was no difference in monitoring accuracy between fourth- and sixth-graders. Note that the reaction times of the JOLs were longer for the fourth-graders than for the sixth-graders, which might indicate that judging idiom learning was more difficult for the younger learners. However, monitoring accuracy following the delayed-JOL instructions and the sentence generation task was accurate for the fourth-graders as well as for the sixth-graders. This finding is in line with findings by Schneider et al. (2000), who showed that even young elementary school students can monitor learning of paired associates very well when judging their memory after a delay.

Importantly, our findings that sentence generation led to more accurate regulation of study than the delayed-JOL instruction, indicate that sentence generation does not only improve



language learning (Graf 1982), but can also have a beneficial effect on metacognition. Our findings on the beneficial effect of sentence generation can be explained by the nature of the idiom learning task. In addition to having characteristics of associative learning tasks in which target information can be retrieved literally, idiom learning also has characteristics of more complex and educationally relevant text comprehension tasks. Because idiom processing can be gist-based, with a focus on connections with contextual information and prior knowledge (Cain and Towse 2008; Langlotz 2006; Levorato and Cacciari 1999; Sprenger et al. 2006), the sentence generation instruction had a beneficial effect on regulation of study. Moreover, the faster JOL reaction times for the low JOLs following sentence generation imply that retrieval of the meaning of idioms is not needed for accurate JOLs and restudy selections.

Previous research by De Bruin et al. (2011) showed that tasks which helped learners focus on the gist of novel information can improve monitoring and regulation for sixth-graders. However, monitoring of fourth-graders was inaccurate. Thus, even though De Bruin et al. (2011) found that a generation task affected regulation of fourth-graders, this regulation was based on inaccurate monitoring and therefore this regulation was not expected to have a beneficial effect on achievement. The present study is the first to show that after accurate monitoring, fourth-grade learners' regulation of study can be improved with the use of a sentence generation task. This is an important finding because more effective regulation following accurate monitoring leads to better academic achievement (Ertmer and Newby 1996; Thiede et al. 2003).

Even though the marginally significant effect of grade level on regulation has to be interpreted with some caution, the findings on metacognition when learning idioms are in line with previous research findings that developmental differences do not affect monitoring accuracy, but do affect regulation of study (Dufresne and Kobasigawa 1989; Krebs and Roebers 2010). Results by Krebs and Roebers (2012) suggested that effects of cognitive capacity are likely to be found with tasks requiring higher metacognitive demands. A possible explanation for the present findings is that older children have more cognitive capacity and are thus better at remembering how they judged their idiom learning. Therefore, for these learners it might be less difficult to decide which idioms need extra study before taking the test. Because sentence generation affected learners' restudy decisions, but did not substantially affect the accuracy of monitoring, the assumption arises that tasks which direct learners' attention to gist cues have more influence on regulation than on monitoring. Future research should further investigate the effect of developmental differences on metacognitive skills when learning idioms.

Based on established theories (e.g., Thiede et al. 2003), we presume that learners focus on gist when judging their learning following generation tasks. We did not directly measure the cues used when learners made metacognitive judgments following Sentence Generation. Another possible explanation for the effect of the sentence generation task could be that it leads learners to attempt visualizing their semantic knowledge (Bransford 1979). Maybe this could have affected their metacognitive judgments and JOL reaction times. Research by Begg et al. (1989) indicated that metacognitive judgments are based on the ease of processing. Possibly, the ease of sentence processing, that is, the ease of generating a correct and complete sentence, informed learners about the quality of their gist representation. Because we did not directly test cue utilization, future research should further investigate what causes the exclusive influence of sentence generation on regulation of study.

Since this is the first study to examine the effect of delayed-JOLs and sentence generation on monitoring and regulation of idiom learning, we should be careful with interpretations. First of all, it is important to note that the measures of JOL RTs are indirect measures of cue use. It could be that the faster RTs for the low JOLs following sentence generation do not result from a focus on cues related to gist, but instead result from a retrieval attempt that learners made when generating the sentences prior to providing JOLs. In that case, the low

reaction times may have been due to the retrieval attempt which was already made at the time of the sentence generation task, instead of the cues that the learners derived about their understanding of the gist of the idiom. However, if learners provide faster JOLs because they focus on the earlier retrieval attempt, we would also expect faster JOL RTs for intermediate and high JOLs following the sentence generation task. We did not find an overall difference in JOL RTs between the Delayed-JOL and the Sentence Generation conditions. This makes the plausibility of the explanation that the fast low JOL RTs are a result of a previous retrieval attempt during sentence generation unlikely.

Possibly, the beneficial effect of the sentence generation condition could also be due to the extra time the participants spent with the materials. Even though this explanation seems unlikely, because in that case sentence generation would also be expected to improve monitoring, it cannot be completely ruled out. Additionally, the less effective regulation in the delayed-JOL group could possibly be explained by the distraction of the filler task, rather than by the benefit of the sentence generation task. However, this explanation also seems improbable, because if the learners were distracted prior to providing metacognitive judgments we would also expect higher overall RTs for the delayed JOL group, whereas we found that only the RTs for the low judgments were higher. However, future research could benefit from a design in which learners who do not participate in a generation task are still presented with the target materials before providing JOLs, instead of with a completely different filler task.

We did not include a condition in which JOLs were provided immediately after the sentence generation task, because Thiede et al. (2005) showed that the delay between study and generation is the only critical lag when investigating monitoring accuracy, and that the lag between generation tasks and judgments is not that important. However, it might be interesting to include an immediate sentence generation condition in future research.

Moreover, in the present study, participants selected items for restudy relatively short after learning the idioms, whereas in educational contexts there is often more time between study sessions. Therefore, in future research it would be interesting to investigate whether the lag between study and restudy decisions influences the effect of the sentence generation instruction.

Importantly, the current work contributes to insights into the validity of cues on which metacognitive judgments are based. Moreover, the findings add to the knowledge about potentially effective techniques to improve regulation when young learners study educationally relevant materials. As to implications for education, our findings suggest that sentence generation is preferable over delaying JOLs, as sentence generation improved regulation of study more, and we know that regulation of study influences learning and performance (Ertmer and Newby 1996). Thus, sentence generation instructions seem promising to improve self-regulated learning of educationally relevant materials such as idioms.

**Acknowledgments** This research was supported by the Netherlands Organization for Scientific Research (NWO), grant 411.07.151. We thank Eva van Eekelen for assistance with data collection, Arno Muijtjens for help with data analysis, and Martine Baars, Luciana Falbo, Peter Stiers, and Bob Wilkinson for comments on previous drafts of this manuscript.

## References

- Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, *128*, 110–118.
- Ariel, R., Al-Harthy, I. S., Was, C. A., & Dunlosky, J. (2011). Habitual reading biases in the allocation of study time. *Psychonomic Bulletin & Review*, *18*, 1015–1021.

- Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68.
- Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory & Language*, *28*, 610–632.
- Bobrow, A., & Bell, M. (1973). On catching on to idiomatic expressions. *Memory & Cognition*, *1*, 343–346.
- Bransford, J. D. (1979). *Human cognition: Learning, understanding, and remembering*. Belmont: Wadsworth.
- Brinton, B., Fujiki, M., & Mackey, T. A. (1985). Elementary school age children's comprehension of specific idiomatic expressions. *Journal of Communication Disorders*, *18*, 245–257.
- Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markham (Eds.), *Handbook of child psychology: Cognitive development* (Vol. 3, pp. 77–166). New York: Wiley.
- Cain, K., & Towse, A. S. (2008). To get hold of the wrong end of the stick: reasons for poor idiom understanding in children with reading comprehension difficulties. *Journal of Speech, Language, and Hearing Research*, *51*, 1538–1549.
- Cain, K., Towse, A. S., & Knight, R. S. (2009). The development of idiom comprehension: an investigation of semantic and contextual processing skills. *Journal of Experimental Child Psychology*, *102*, 280–298.
- Conway, A. R. A., & Engle, R. W. (1994). Working memory and retrieval: a resource-dependent inhibition model. *Journal of Experimental Psychology: General*, *123*, 354–373.
- Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge: Cambridge University Press.
- Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition*, *25*, 57–71.
- 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*, 294–310.
- 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., & Lipko, A. R. (2007). Metacomprehension: a brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*, 228–232.
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks: Sage Publications.
- Ertmer, P. A., & Newby, T. J. (1996). The expert learner: strategic, self-regulated, and reflective. *Instructional Science*, *24*, 1–24.
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, *10*, 507–521.
- Fletcher, C. R., & Chrysler, S. T. (1990). Surface forms, textbases, and situation models: recognition memory for three types of textual information. *Discourse Processes*, *13*, 175–190.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395.
- Graf, P. (1982). The memorial consequences of generation and transformation. *Journal of Verbal Learning and Verbal Behavior*, *21*, 539–548.
- Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition*, *37*, 1001–1013.
- Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition & Learning*, *4*, 161–176.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: University Press.
- Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394.
- Kintsch, W., Welsch, D. M., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: a theoretical analysis. *Journal of Memory and Language*, *29*, 133–159.
- Koriat, A. (1997). Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.
- Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The easily learned, easily remembered heuristic in children. *Cognitive Development*, *24*, 169–182.
- Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*, 478–492.
- Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy. In P. Chambres, M. Iztaute, & P. J. Marescaux (Eds.), *Metacognition: Process, function and use* (pp. 1–17). Boston: Kluwer Academic Publishers.

- Krebs, S. S., & Roebbers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology*, *80*, 325–340.
- Krebs, S. S., & Roebbers, C. M. (2012). The impact of retrieval processes, age, general achievement level, and test scoring scheme for children's metacognitive monitoring and controlling. *Metacognition and Learning*, *7*, 75–90.
- Langlotz, A. (2006). *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom variation*. Amsterdam/Philadelphia: John Benjamins.
- Levorato, M. C., & Cacciari, C. (1992). Children's comprehension and production of idioms: the role of context and familiarity. *Journal of Child Language*, *19*, 415–433.
- Levorato, M. C., & Cacciari, C. (1999). Idiom comprehension in children: are the effects of semantic analysability and context separable? *European Journal of Cognitive Psychology*, *11*, 51–66.
- Levorato, M. C., Nesi, B., & Cacciari, C. (2004). Reading comprehension and understanding idiomatic expressions: a developmental study. *Brain and Language*, *91*, 303–314.
- 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.
- Maki, R. H. (1998). Predicting performance on text: delayed versus immediate predictions and tests. *Memory & Cognition*, *26*, 959–964.
- Metcalfe, J., & Finn, B. (2008a). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*, 174–179.
- Metcalfe, J., & Finn, B. (2008b). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *34*, 1084–1097.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133.
- Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: the "Delayed-JOL effect". *Psychological Science*, *2*, 267–270.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–171). New York: Academic Press.
- Nippold, M. A., & Taylor, C. L. (2002). Judgments of idiom familiarity and transparency: a comparison of children and adolescents. *Journal of Speech, Language, and Hearing Research*, *45*, 384–391.
- Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*, 357–383.
- Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological Bulletin*, *137*, 131–148.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh: Psychology Software Tools Inc.
- 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.
- Son, L. K., & Metcalfe, J. (2005). Judgments of learning: evidence for a two-stage process. *Memory & Cognition*, *33*, 1116–1129.
- Sprenger, S. A., Levelt, W. J. M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, *54*, 161–184.
- Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction-time experiments. *American Scientist*, *57*, 421–457.
- Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, *18*, 523–534.
- Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory & Cognition*, *37*, 529–540.
- Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*, 129–160.
- Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of text. *Journal of Educational Psychology*, *95*, 66–73.
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 1024–1037.
- Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1267–1280.
- Van den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: The role of co-activation in confronting misconceptions. *Journal of Applied Cognitive Psychology*, *22*, 335–351.

- Van Loon, M. H., De Bruin, A. B. H., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15–25.
- Von der Linden, N., Schneider, W., & Roebers, C. (2011). The effects of summary production and encoding condition on children's metacognitive monitoring. *Metacognition & Learning, 6*, 3–23.
- Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology, 82*, 51–59.