Gender Differences in Mental Simulation During Sentence and Word Processing

Stephanie I. Wassenburg, Björn B. de Koning, Meinou H. de Vries, A. Marije Boonstra and Menno van der Schoot

Department of Educational Neuroscience and LEARN! Research Institute for Learning and Education, Faculty of Psychology and Education, VU University Amsterdam, The Netherlands

Abstract

Text comprehension requires readers to mentally simulate the described situation by reactivating previously acquired sensory and motor information from (episodic) memory. Drawing upon research demonstrating gender differences, favouring girls, in tasks involving episodic memory retrieval, the present study explores whether gender differences exist in mental simulation in children (Grades 4 to 6). In Experiment 1, 99 children performed a sentence-picture verification task measuring mental simulation at sentence level. In Experiment 2, 97 children completed a lexical decision task in which imageability of words was manipulated to measure mental simulation at word level. Only for girls we found faster reaction times for matching versus mismatching sentence-picture pairs (Experiment 1) and high-imageability versus low-imageability words (Experiment 2). The results suggest that girls construct more coherent and vivid mental simulations than boys and rely more heavily on these representations. The results emphasize the importance of including gender into reading comprehension research.

Keywords: Mental simulation, gender, sentence-picture verification, lexical decision

Gender Differences in Mental Simulation During Sentence and Word Processing

According to embodied theories of cognition, readers derive meaning from text through the reactivation of previously acquired real-world perceptual, motor, and affective experiences which are stored in the brain regions that govern perception, action, and emotion (Barsalou, 1999). By mentally (re)enacting and integrating these perceptual, motor, and affective experiences, the reader constructs a so-called mental simulation of the events described in the text (Zwaan, 2003). In recent years, research has provided ample evidence that perceptual simulation is an integral part of language comprehension (Barsalou, 2008). Although a coherent mental representation of text typically consists of information from all sensory modalities (Sadoski & Paivio, 2013), the visual modality in particular has incited much empirical work (de Koning & van der Schoot, 2013). For example, the majority of research on the mental simulation of the objects, situations, and events described within a sentence using the sentence-picture verification task examines visual characteristics of the described situation (e.g., object orientation, Stanfield & Zwaan, 2001).

Especially in the last decade, the sentence-picture verification task has become a popular and widely used task in the mental simulation literature, probably due to its elegance and simplicity (Zwaan & Pecher, 2012). In this task, readers are required to read a sentence implying a perceptual attribute of an object (e.g., shape, orientation) and subsequently have to indicate whether or not the object shown in a subsequently presented picture was mentioned in the sentence. To be able to make an accurate decision, the reader is required to mentally simulate the (visual appearance of the) described objects and events and compare this to the appearance of the object shown in the picture (Zwaan & Pecher, 2012). When the visual characteristics of the

depicted object match that of the reader's mental simulation, readers are faster to verify matching than mismatching pictures. This is due to a relatively larger overlap between activated brain patterns resulting from (re)enacting perceptual information of the described situation and seeing the object picture. To date, such a match advantage has been demonstrated in numerous studies employing this paradigm in order to affirm the assumptions of the theory. Results suggest that readers automatically represent the implied orientation (Stanfield & Zwaan, 2001), shape (Zwaan, Stanfield, & Yaxley, 2002), visibility (Yaxley & Zwaan, 2007), colour (Zwaan & Pecher, 2012), number (Patson, George, & Warren, 2014), and distance (Vukovic & Williams, 2014) of described objects. Furthermore, it has been shown that mental simulation processes are robust, in that they, among other things, function independently of negation (Kaup, Yaxley, Madden, Zwaan, & Lüdtke, 2007), result in a detailed mental representation that is retained for at least 45 minutes (Pecher, van Dantzig, Zwaan, & Zeelenberg, 2009), are involved in multiple perceptual dimensions (Zwaan & Pecher, 2012), and are already observed in children of 7 years old (Engelen, Bouwmeester, de Bruin, & Zwaan, 2011).

Despite these seemingly straightforward and consistent findings, evidence is accumulating that the mental simulation process appears to be subject to the influence of several factors. For example, mental simulations have been shown to be constrained by certain linguistic aspects like verb aspect (Madden & Zwaan, 2003). Moreover, the mental simulation of implied object shape provides much stronger effects than that for object colour (Zwaan & Pecher, 2012). Furthermore, recent studies show that mental simulation processes may be more context dependent than previously assumed (e.g., van Dam, van Dijk, Bekkering, & Rueschemeyer, 2012). The context in which a word is presented emphasizes, for example, certain modality-specific properties and places constraints upon activation of associated mental simulations.

Recently, Rommers, Meyer, & Huettig (2013) have suggested that the construction of a mental simulation is mediated by task demands, like switching between sensory modalities (Pecher, Zeelenberg, & Barsalou, 2003) and use of word association strategies (Solomon & Barsalou, 2004). Consistent with reading comprehension research more generally (e.g., Clinton et al., 2012), there is now increasing awareness that the mental simulation processes should also be studied from an individual differences perspective (e.g., Hirschfeld, Feldker, & Zwitserlood, 2012).

To date, studies investigating individual differences using the sentence-picture verification task have mostly investigated the role of spatial or imagery skills. Behavioural studies have found no systematic relation between the magnitude of the match advantage and other cognitive skills (Pecher et al., 2009; Stanfield & Zwaan, 2001). Studies using neuroimaging and electrophysiological measures, however, have reported individual differences in mental simulation processes related to vividness, preference for visual imagery, and verbal versus visuo-spatial abilities (Cui, Jeter, Yang, Montague, & Eagleman, 2007; Hirschfeld et al., 2012; Kraemer, Rosenberg, & Thompson-Schill, 2009; Reichle, Carpenter, & Just, 2000). It has been suggested that neuroimaging or electrophysiological measures reveal differences at specific stages of individual processing that are lost in reaction time measures. Other individual processing factors, however, have received little or no attention from researchers investigating mental simulation processes. A factor that is worth exploring in this respect is gender as this factor has been linked to differences in reading comprehension in general and the construction of mental representations from text in particular (Clinton et al., 2012).

According to the simple view of reading proposed by Gough and Tunmer (1986), reading comprehension consists of two basic components: decoding and language comprehension.

Whereas the identification and recognition of printed words is required for reading, language comprehension skills are necessary to derive meaning from those words. Generally, research on reading comprehension consistently provides evidence for gender differences that favour females (e.g., Halpern et al., 2007; Logan & Johnston, 2009). This so-called gender gap already exists in fourth-graders and holds for different languages and countries, independently of educational system (Mullis, Martin, Kennedy, & Foy, 2007). Moreover, these early inequalities are likely sustained or even further exacerbated throughout secondary education (Warrington, Younger, & Bearne, 2006), and may form a fundamental problem for boys in later academic achievement and employment (Wood, 2003).

These findings are consistent with those obtained in research on gender differences in more specific cognitive tasks of which the outcome can be influenced by both decoding and comprehension processes. For example, studies using meaningful language show that girls outperform boys on verbal fluency, perceptual speed, and spelling tasks (Burns & Nettelbeck, 2005; Harshman, Hampson, & Berenbaum, 1983). Studies of pure decoding skills (e.g., nonsense words), however, found no gender differences (Below, Skinner, Fearrington, & Sorrell, 2010; Logan & Johnston, 2010). This suggests that gender differences observed in reading comprehension and other language-based processing tasks are most likely the result of individual differences in comprehension processes and abilities, rather than individual differences in decoding skills (Clinton et al., 2012).

It has been shown that comprehension processes like constructing, (re)activating, and updating involved in building a rich and coherent mental representation or simulation are memory-based processes (Gerrig & O'Brien, 2005). Episodic (as well as semantic) memory plays a crucial role in the retrieval of world knowledge and information about personal

experiences (van den Broek, Young, Tzeng, & Linderholm, 1999). Research has shown that there are gender-based differences in tasks that require retrieval of information from episodic memory (e.g., Clinton et al., 2012; Herlitz & Rehnman, 2008). Clinton et al. (2012) showed, for example, that girls outperform boys when it comes to making reinstatement inferences (i.e., connecting text information with previously read text) based on easier access to information from episodic memory, which contributes to differential task performance on reading comprehension. This suggests that girls were able to construct more coherent mental representations by drawing upon their prior experiences. The notion that mental representations generated by males and females qualitatively differ when it comes to characteristics such as shape, size, and colour (for an overview, see Richardson, 1991) may be considered a more specific interpretation of this latter suggestion. Furthermore, females tend to generate higher ratings than males when asked to evaluate vividness or controllability of imagery (Richardson, 1991). Self-reported vividness of imagery, in turn, has been shown to be related to differences in activation of perceptual processing areas in the brain and stronger shape-match-advantages for participants with vivid imagery (Hirschfeld et al., 2012). Therefore, it would be expected for girls to outperform boys on a sentence-picture verification task.

In sum, girls seem to outperform boys on tasks involving a strong linguistic component, requiring the comprehension of language, and thus the construction of mental representations. The recruitment of prior experiences appears to play an important role in explaining these differences. This nicely fits with current thinking about the mental simulation processes involved in sentence comprehension, which is assumed to rely on reactivation of prior experiences. Hence, it seems timely to investigate whether there are gender differences in the mental simulation process when processing language. Such gender-based differences in mental simulation might

indirectly impact on reading comprehension and therefore would have important implications for education, particularly if they exist already early in life.

The present study

The aim of the present study was to investigate gender differences in mental simulation during language comprehension. More specifically, we used the sentence-picture verification task to investigate whether gender differences could be observed when mentally simulating the information described in a sentence (Experiment 1). In doing so, we extend previous research by exploring gender differences in underlying cognitive processes of reading comprehension, advance the research on mental simulation with the sentence-picture verification paradigm in a new direction (i.e., the role of gender differences), and do so in an age group (i.e., primary school children) that—in research using this task—has received little attention (for an exception, see Engelen et al., 2011). Based on research showing that males and females differ in visual imagery and mental simulation processes (e.g., Clinton et al., 2012; Richardson, 1991), we expect girls to construct more coherent and vivid mental representations than boys. In the sentence-picture verification task, this would be evidenced by a stronger match-advantage for girls than for boys. In a follow-up experiment (Experiment 2), we addressed the same question for single word processing using a lexical decision task. This experiment was designed to exclude alternative explanations for the results obtained in Experiment 1. Sentence comprehension involves processing of both syntax and semantics. Understanding syntax, however, depends on different cognitive processes than understanding semantics. Moreover, the sentence-picture verification task requires the reader to compare visual objects to language which encourages the reader to use visual imagery. In order to conclude that potential gender differences are due to fast and automatic mental simulation processes specifically, rather than task demands, strategic imagery

use or other involved cognitive processes (e.g., syntactic understanding), it was deemed necessary to include an additional experiment. Repeating the results of Experiment 1 with a lexical decision task, a non-imagery task which taps earlier into the mental simulation process and does not require syntax processing, would provide converging evidence for the hypothesis. More specifically, in Experiment 2 it was expected that girls would show a stronger effect of imageability of words on their lexical decision speed than boys, even in the absence of task demands requiring imagery. In both experiments, measures for, presumably, gender-biased cognitive processes were considered as control variables (e.g. decoding, reading comprehension, mental rotation, and visual working memory).

Experiment 1

Method

Participants. The participants were 99 children (M_{age} = 11.0 years, SD = .9, range 8.7–13.1) from Grades 4 through 6 in five regular primary schools in different areas in the Netherlands in order to acquire a representative sample. Overall (52 boys, M_{age} = 11.0, SD = 1.0, and 47 girls, M_{age} = 11.0, SD = 1.0) and within each grade, boys and girls were matched on age. There were 29 children (13 boys) from Grade 4 (age range 8.7–11.8, M_{boys} = 10.0, SD = .8, M_{girls} = 10.1, SD = .7), 39 children (23 boys) from Grade 5 (age range 10.2–13.1, M_{boys} = 10.9 years, SD = .6, M_{girls} = 10.9, SD = .5), and 31 children (16 boys) from Grade 6 (age range 11.2–13.1, M_{boys} = 12.0, SD = .6, M_{girls} = 11.9, SD = .5). All children came from schools with relatively high concentrations of native Dutch students and were fluent Dutch speakers. Exclusion criteria were behavioural problems and developmental or intellectual disabilities based on school reports, because of their possible influence on reading comprehension and general cognitive processing.

Participation was voluntary and children received a small present after the experiment. All children's parents provided informed consent for participation.

Materials.

Control measures. Decoding skill was assessed by the grade-appropriate Reading Speed Test ('Leestempo'; for a scientific justification of the test, see Krom & Kamphuis, 2001). The Reading Speed Test is commonly used by primary schools in the Netherlands as a part of their pupil monitoring system and has high reliability (range .85–.91). It is a standardized test which is purposefully designed to acquire information about a student's decoding skills when reading silently under time pressure. It is not a measure of pure decoding speed, but rather provides an indication of how many words were read within the given time, while making sure that what was read silently, was read accurately and comprehensively. The test shows relatively high correlations (range r = .69–.73) with other measures of decoding skill (e.g., reading out loud a list of words). During the Reading Speed Test, children read a text and were required to fill in blanks by circling one out of three answer options at a time. The alternatives could be incorrect on syntactic level, on semantic level, or on both levels. They had 5-7 minutes to fill in as many blanks as possible, depending on grade. The total score was the total number of correctly filled in blanks. These scores were converted into a reading-age equivalent (RAE; van der Leij, Struiksma, & Vieijra, 1995). The RAE reflects the child's actual reading level, expressed in the number of months of reading instruction, with ten months of reading instruction being equal to one academic year. RAE scores on the Reading Speed Test were retrieved from the participating school administrations.

Children's reading comprehension level was assessed by the grade-appropriate standardized Test for Reading Comprehension of the Dutch National Institute for Educational

Measurement (CITO), which is designed to provide scores for general reading comprehension level in primary school children (‘Toets Begrijpend Lezen’, for a scientific justification of the test for Grade 4, see Feenstra, Kamphuis, Kleintjes, & Krom, 2010, and for Grade 5 and 6, see Weekers, Groenen, Kleintjes, & Feenstra, 2011). Each version of the test contained two modules, each consisting of a narrative or expository text and 25 multiple-choice questions. The questions were designed to tap both the text-base and situation model representation which can be constructed from the text (e.g., Kintsch, 1988) and pertained to either the word, sentence or text level. Children were instructed to read texts and answer questions about the texts by choosing the correct option. There was no time limit. Children’s total raw test score, which was obtained by summing the correct answers on the 50 items, were transformed into normed proficiency scores. These scores enabled comparisons between children from different grades. Both measures described above are part of the standard pupil monitoring system of the Dutch Institute for Educational Measurement (CITO) and are evaluated by the Commission on Testing Matters (COTAN) of the Dutch Institute of Psychologists (NIP). These tests are judged to be “good” on all applicable criteria indicating that the reliability coefficients were > .80, the construct validity is good, and the quality of the principles on which the tests were constructed as well as the quality of the test material and the manual are high (see their respective scientific justifications).

Participants’ spatial abilities were measured with an adapted version of Quaiser-Pohl's (2003) Picture Rotation Test which was specifically designed for primary school children. That is, we only used the coloured pictures of animals as stimuli and the task consisted of 30 items. In each item, a target picture was presented together with three comparison pictures. Two of these comparison pictures were mirror images of the target picture, whereas one was a picture that, compared to the target picture, was rotated in the plane. Participants were required to choose

which of these three pictures the rotated version of the original picture was. The dependent variable was the number of correct answers.

Participants' short-term memory span was assessed by the tapping backwards subtask of the visual memory span (VMS) task of the Wechsler Memory Scale–Revised (WMS-R) (Wechsler, 1987). Participants were required to repeat a tapping sequence given by the experimenter in reversed order by tapping on a card with nine green squares. Tapping sequences were of increasing length, ranging from 2 to 8 taps. There were two sequences of every length. A participant's score was the number of correctly backwards repeated sequences.

Sentence-picture verification task. A set of 24 experimental sentences and pictures adapted from Zwaan, Stanfield, and Yaxley (2002) were created for this experiment (see Appendix A for examples of sentence-picture pairs). Of all 24 experimental sentences, there were two versions only differing in the prepositional phrase, which suggested a distinct shape of the critical noun (e.g., 'the eagle [in the nest]' versus 'the eagle [in the sky]'). Participants read one of the two versions of each sentence from the computer screen, followed by either a matching or mismatching picture with regard to the implied shape (e.g. an eagle with folded wings versus an eagle with outstretched wings). All pictures were coloured line drawings, created by a professional artist for this experiment, and were scaled to occupy a square of approximately 15x15 cm on the computer screen.

By crossing the two versions of experimental sentences and the two versions of pictured objects, four experimental lists were created. Each list contained 24 experimental sentence-picture pairs, of which shape matched for half of the pairs and mismatched for the other half. In other words, each list contained 12 items involving a match between the shape implied in the sentence and the shape depicted in the picture and 12 items involving a mismatch between the

implied and pictured shape of an object. Across the four versions, all item combinations were used equally often. Because all 24 experimental trials required a yes-response, 24 filler trials requiring a no-response were added to balance responses. In filler trials a sentence, similarly structured as experimental sentences, was followed by a picture of an object that was not mentioned in the preceding sentence (e.g., the sentence "The girl put the ring on her finger" followed by a picture of a garbage bin).

Procedure. All tests were administered at the participants' own schools. The Reading Speed Test and reading comprehension test were administered by the school, as part of the standard pupil monitoring system of CITO. Both were paper-pencil tasks administered in the classroom using a whole-class test taking approach. These measures were retrieved from the participating school administrations.

For administration of all other tasks, participants were tested in two sessions, once using the whole-class test taking approach in the classroom and once individually in a quiet room. The whole-class test session only consisted of the picture rotation test. For this paper-pencil task, participants were instructed to rotate the pictures in their mind by showing them an enlarged example item which could be rotated like the hands of a clock. The instructor stressed that this had to be done in their minds, because they were not allowed to rotate the paper task on their desk. By underlining a picture, the participants indicated which of the three comparison pictures matched the target picture. They had 1.5 minutes to answer as many items as possible. In total, this session took about 10 minutes.

The individual session always consisted of the VMS task and the sentence-picture verification task. For the VMS task, the instructor provided instructions before showing the card with the nine green squares. The task started with two practice sequences of two taps. For every

sequence, the instructor started tapping a number of squares in sequence (approximately one square per second) and after two seconds the participant was required to repeat this sequence backwards by tapping the corresponding squares in reversed order. When the participant failed to correctly repeat both sequences within a certain length, the task was aborted by the instructor.

The sentence-picture verification task was completed on a 15 inch laptop. Following prior research, children were instructed to read aloud each sentence at their own pace (cf. Engelen et al., 2011, Experiment 2). This enabled the experimenter to check that no words were skipped or pronounced incorrectly and that there were no misunderstandings about the sentences. Moreover, it reduced the chance of certain strategic behaviour like rereading or focusing on specific parts of the sentence by slowing down. So, by letting children reading out loud the sentence we ensured that they had read the whole sentence and moved on immediately by pressing the space bar. Importantly, prior research using a sentence-picture verification task has shown that reading sentences out loud yields comparable results as when reading the sentences silently (Engelen et al., 2011). The participants had to indicate as fast and accurately as possible whether or not the depicted object had been mentioned in the sentence. Each trial started with a horizontally and vertically centred sentence on the screen, displayed in a black 24-point Courier New Bold font against a white background. When participants had understood the sentence, they pressed the space bar after which a 500 ms fixation cross appeared, followed by a picture. By pressing the keys on the laptop keyboard marked by a green and red sticker for yes- and no-responses respectively, participants indicated whether the depicted object was mentioned in the preceding sentence or not. To warm up and familiarize participants with the task, the experiment began with two practice trials consisting of one related and one unrelated sentence-picture pair. Trials

were presented in random order. The sentence-picture verification task took about 15 minutes. In total, the individual testing session took about 25 to 30 minutes.

Results

Preliminary analyses. For the sentence-picture verification task, data of two experimental items were removed prior to the statistical analyses because of low accuracy scores due to the pictures being difficult to recognize. Furthermore, data of two participants were excluded from the dataset. One participant only gave no-responses on all trials, whereas another participant gave no-responses on all corresponding pictures in the mismatch condition reflecting inaccurate understanding of the task. The average proportion of correct answers for all remaining trials, including fillers, was high (M = .96, SD = .04). This indicated that participants had understood the procedure and were not biased towards either affirmative or negative responses. In further reaction time analyses, filler items were not included and incorrect responses were eliminated. Reaction times (RTs) <300 ms and >3000 ms were considered as outliers and were removed. The remaining responses were additionally trimmed by removing RTs which were more than 2.5 standard deviations from the overall condition's mean. This is the mean for either all matching or all mismatching trials across all subjects. In total, this constituted removal of less than 5% of the data. Table 1 shows the mean picture-verification latencies and accuracy scores for boys and girls in the match and mismatch conditions. As can be seen in this table, the accuracy scores and RTs for match trials and mismatch trials together did not show evidence for a speed-accuracy trade-off. That is, the faster response times for matching versus mismatching items did not coincide with reduced accuracy on those items. Therefore, comparison and interpretation of these RTs is warranted.

Table 1
Picture verification latencies and accuracy

	Condition			
	Match		Mismatch	
Measure	Boys	Girls	Boys	Girls
Reaction time (ms)	1182 (255)	1118 (263)	1169 (271)	1227 (350)
Percentage correct	97 (5)	97 (5)	95 (7)	93 (13)

Note. Standard deviations are given in parentheses.

Descriptive statistics and *t* tests for all control variables by gender are shown in Table 2. Independent samples *t* tests showed that boys performed significantly better than girls on mental rotation (d = .50). Effect sizes (Cohen's d) of .2, .6, and .8 are considered to be small, medium, and large, respectively (Cohen, 1988). With respect to decoding RAE, there was a marginally significant effect of decoding skill indicating that girls scored better than boys (d = .38). The crossed random effects analyses presented later, however, showed that these differences did not influence the RTs. Furthermore, there were no significant gender differences on reading comprehension scores and visual memory span.

Table 2
Means, standard deviations and t tests for control measures by gender

	Boys			Girls				
Control Measure	*n*	*M*	*SD*	*n*	*M*	*SD*	*t*	*p*
Decoding RAE	52	36.04	13.75	46	41.26	13.69	-1.88	.063
Reading Comprehension	51	45.96	17.17	46	48.93	21.09	-.77	.446
Mental Rotation	52	14.52	7.08	47	11.11	6.60	2.47	.015
Visual Memory Span	50	8.54	1.94	47	8.15	1.67	1.06	.291

Note. RAE = reading-age equivalent

Analyses of reaction times. Reaction times data were analysed by using mixed-effects modelling, where random effects can be included to model dependencies between both participant's and item's replications. More specifically, a one-level crossed random effects

model (CREM) was used with the Full Maximum Likelihood estimation procedure (Carson & Beeson, 2013). Advantages of this method over the more conventional analyses of variance are an increased statistical power, a better account of heterogeneity of variance, and a better use of available data. In the present analysis, condition, gender, grade, and the control variables (i.e., decoding RAE, reading comprehension, mental rotation, and VMS) were treated as fixed effects. All control variables were grand mean centred. To find the best fitting model, predictors were added using a stepwise procedure. In Table 3 only the best fitting model is displayed.

First of all, to test whether participants and items should be considered random effects, we tested a totally unconditional model. This model only included participants and items as random effects. The results showed that both random effects were significant, indicating that, as expected, RTs differed across participants; $Wald(1) = 6.16$, $p < .001$, and items; $Wald(1) = 2.75$, $p = .006$. This justifies the inclusion of both random effects in the model.

In the next step, the control measures were added to the model, together with the main predictors (the results are displayed in Table 3, Model 1). The predictors of theoretical interest were grade, gender, condition, and the interaction term of gender with condition. Controlling for decoding RAE, reading comprehension, mental rotation, and VMS, there were no main effects of grade and gender. Although the coefficient estimates of condition were significant, $t(1728.98) = -2.48$, $p = .013$, indicating faster RTs for matching than for mismatching trials, the fixed effect of condition did not reach significance, $F(1, 1724.25) = 2.46$, $p = .117$. Importantly, the interaction term of gender with condition was significant, $F(1, 1734.88) = 4.13$, $p = .042$. Pairwise comparisons with Bonferroni correction revealed that there was a significant match advantage for girls; $F(1, 1728,98) = 6.16$, $p = .013$ (with faster reaction times for matching than for

mismatching trials, see Table 1), but not for boys; $F(1, 1731) = .12$, $p = .726$. None of the control variables were significant (all ps > .05) and they were therefore removed from the model.

In the final step, two interaction terms were added to the model (i.e., a two-way interaction of grade with condition, and a three-way interaction of grade, gender, and condition). The purpose of this addition was to explore the influence of grade on the magnitude of the match advantage and its dependence of gender. None of these interaction terms reached significance (Fs > 1) and were therefore removed from the model. Hence, Model 1 (Table 3) remained the best fitting model.

Table 3

Regression Coefficient Estimates and Variance-Covariance Estimates for CREM Predicting Observed Reaction Time

	Model 1	
Parameters	Estimate	SE
Fixed parameters		
Intercept	1156.31***	40.20
Grade 4	101.80	6.26
Grade 5	109.13	56.59
Grade 6 (reference)		
Gender: boy	-28.92	50.08
Gender: girl (reference)		
Condition: Match	-52.04*	20.96
Condition: Mismatch (reference)		
Interaction terms		
Match*boy	58.96*	29.03
Random parameters		
Residual	91861.17***	3139.98
Subjects	45964.35***	7487.35
Items	6793.56**	2481.48
Deviance (-2 log likelihood)	26399.36	

Note. Model 1 was hierarchically built and represents the best fitting model.
* $p < .05$ ** $p < .01$ *** $p < .001$.

Discussion

The results showed that only some children benefited from a picture that matched the target object's shape as implied by the preceding sentence context. Interestingly, this benefit only appeared to be present for girls given that they responded faster to matching trials than to mismatching trials, whereas boys did not show such a match advantage. A likely interpretation of these findings is that girls (but not boys) mentally simulated the situation described in the sentence. Girls' processing of the sentences, by relying on the construction of mental simulations, seems to be based more on—(re)activation of—sensory referents of the described situation that were acquired during previous experiences and that are now stored in the brain. Because of this, matching sentence-picture trials supposedly induced a larger overlap in (re)activated brain patterns in girls than in boys, which speeds up the recognition process and thereby influences the magnitude of the match advantage. Importantly, the advantage for girls over boys was not due to a general reading ability advantage as no gender differences were found in the reading comprehension scores and girls' slightly better decoding skills did not confound the RTs on the sentence-picture verification task. It may be possible, however, that not finding a relation between decoding skill and RTs was due to using words and syntactic constructions that were too easy for individual differences in decoding skills to emerge. Moreover, boys' advantage on spatial ability did not seem to help them construct a more vivid perceptual simulation during reading. Mental rotation tests have even appeared to be negatively correlated with other measures of visual imagery (Greenberg & Knowlton, 2014). This could be due to their focus on manipulation of spatial imagery, instead of activation or construction of a mental simulation. Moreover, it has been argued that visual and spatial imagery are dissociative systems (Farah, Hammond, Levine, & Calvanio, 1988). Apparently, the skill to mentally manipulate objects does not help to activate or access visual images. Finally, in accordance with results from Engelen et

al. (2011) grade neither had an impact on reaction times nor on the magnitude of the match advantage. Children as young as fourth-graders seemed to be able to mentally simulate the situation described in a sentence.

Whereas these results seem straightforward, the particular task we used (i.e., sentence-picture verification task) leaves some space for alternative explanations. First, the sentence-picture verification task taps into the comprehension process relatively late, which makes the mental representation susceptible to post-access influences at the time of response (Simpson, 1994). Second, due to the repeating structure of the sentences (subject, verb, critical noun [direct object], and prepositional phrase), it is possible that children strategically directed their attention to the critical noun. Third, only highly concrete words were used as critical nouns in this task. Concrete words recruit visual imagery processes more easily than abstract words (Barber, Otten, Kousta, & Vigliocco, 2013). Therefore, concrete words usually elicit faster response times than abstract words. Moreover, it has been suggested that females' ratings of concrete words tend to be more based on affective experiential traces than males' ratings, whereas no gender differences were found in ratings of abstract and emotion words (Bauer & Altarriba, 2008). These alternative explanations were addressed in Experiment 2.

Experiment 2

In Experiment 2, we employed a lexical decision task to study gender differences in mental simulation for single word processing. This enabled us to study the effects of gender independently of specific task demands and exclude alternative explanations provided by the sentence-picture verification task. That is, whereas the sentence-picture verification task involves imagery and further higher-processing over and above activating a single word's meaning, like syntax understanding and integrating information, lexical decision does not require such

processing demands (Bottini et al., 1994). Both tasks, however, involve activating meaning by means of mental simulation (Chumbley & Balota, 1984; Willems, Hagoort, & Casasanto, 2010). Importantly, lexical decision taps into the comprehension process earlier than sentence-picture verification, because no further processing is required after successfully activating a word's meaning. Moreover, participants react immediately to the linguistic stimuli as no pictures are used in this task. Therefore, RTs reflect the mental simulation process more directly. Conducting a second experiment using the lexical decision task thus enabled us to investigate whether the gender differences in mental simulation processes are due to fast and automatic mental simulation processes (i.e., lexical decision) or are more likely the result of imagery task demands and other processes involved in sentence processing (Gullick, Mitra, & Coch, 2013; Willems et al., 2010).

To study mental simulation at the word level, which logically excludes the possibility of strategic processing of the stimuli, imageability of words was manipulated. Imageability refers to the ease with which a word or concept can be mentally simulated and represented, and seems to be influenced by familiarity (Barber et al., 2013). It is important to note that imageability should not be confused with concreteness, even though concrete words are usually rated higher on imageability than abstract words because they have more direct sensory referents and it is easier to access the corresponding mental image (Gullick et al., 2013; Schwanenflugel & Stowe, 1989). Because the present study focuses on mental simulation skills, imageability of words, but not concreteness, was manipulated. Research shows that processing of high- and low-imageability words is influenced by familiarity and frequency (Colombo, Pasini, & Balota, 2006; Nation & Snowling, 1998). Therefore, in the present experiment, high- and low-imageability words were matched on word frequency and age of acquisition. Based on research showing that for high-

imageability words it is easier to recruit visual imagery processes (Barber et al., 2013; Schwanenflugel & Stowe, 1989), we expected high-imageability words to be verified faster than low-imageability words (Morrison & Ellis, 2000). Furthermore, if the gender differences found in Experiment 1 were specifically due to mental simulation processes, rather than to other processes specific to sentence processing and/or the task, we would expect girls to show a stronger advantage of imageability than boys.

Method

Participants. The participants were 97 children (M_{age} = 11.0 years, *SD* = 1.0, range 8.7–13.1) from Grades 4 through 6 in five regular primary schools in different areas in the Netherlands of which 77 children also participated in Experiment 1. The present sample was acquired at the same schools as in Experiment 1. Again, all children that were present in class at the time of administration were tested. As in Experiment 1, boys and girls were matched on age overall (53 boys, M_{age} = 11.0, *SD* = 1.0, and 44 girls, M_{age} = 10.9, *SD* = 0.9) and within each grade. There were 30 children (14 boys) from Grade 4 (age range 8.7–11.8, M_{boys} = 10.0, *SD* = .8, M_{girls} = 10.1, *SD* = .7), 39 children (23 boys) from Grade 5 (age range 10.2–13.1, M_{boys} = 10.9, *SD* = .6, M_{girls} = 10.9, *SD* = .5), and 28 children (16 boys) from Grade 6 (age range 11.5–13.1, M_{boys} = 12.2, *SD* = .4, M_{girls} = 11.9, *SD* = .4). As in the first experiment, exclusion criteria were behavioural problems or developmental or intellectual disabilities based on school reports. Participation was voluntary and children received a small present after the experiment. The children's parents provided informed consent for participation.

Materials.

Control measures. The same measures for decoding skill, reading comprehension, spatial rotation, and visual memory span as in Experiment 1 were used.

Lexical decision task. For the lexical decision task, 60 words with a transparent orthography were taken from a list generated by Van Loon-Vervoorn's (1985) containing, among others, 4600 Dutch nouns which are rated (on a scale from 1 to 7) on their imageability. Two lists of 30 words were created, each consisting of 15 high-imageability and 15 low-imageability nouns, $t(39.89) = 31.60$, $p < .001$, $d = 8.16$. See Table 4 for the descriptives. High- and low-imageability words were matched on word frequency (Schrooten & Vermeer, 1994), age of acquisition (Van Loon-Vervoorn, 1985), number of syllables, and word length (all *p*-values > .60).

Table 4

Descriptives for high- and low-imageability words

	High		Low	
	M	*SD*	*M*	*SD*
Imageability ***	6.44	.23	3.10	.53
Age of Acquisition (months)	81.30	13.50	81.67	12.61
Word Frequency	32.40	42.99	33.03	58.38
Syllables	1.87	.63	1.80	.61
Word length	5.80	1.54	5.97	1.54

*** $p < .001$

To create two experimental lists of 60 items, each list of 30 words was complemented with 30 pseudo words which served to balance yes- and no-responses. Pseudo words were obtained by changing one or two letters from the words in the other list. All pseudo words were orthographically and phonologically legal. Participants always saw one of the two experimental lists. To exclude potential influence of decoding speed on reaction times, words were recorded and presented auditorily by using 'Praat' software (Boersma & Weenink, 2011).

Procedure. As in Experiment 1, scores for the Reading Speed Test and the reading comprehension test were retrieved from school administrations. All other measurements were assessed in two separate sessions. The classroom session only consisted of the Picture Rotation

task. The individual session consisted of the VMS and lexical decision task. Procedures of all control tests were the same as in Experiment 1. However, children that had already participated in Experiment 1, were not tested again. For these children, their earlier acquired scores on control tasks were used, completed with their performance on the lexical decision task.

For the lexical decision task, participants were tested individually on a 15 inch laptop while wearing headphones, placed in a quiet room within their own school. Every trial started with a 3000 ms, horizontally and vertically centred, black fixation cross against a white background. After the fixation cross disappeared (i.e., the white computer screen was presented), participants heard a spoken word, after which immediately an answer screen appeared with a red square on the left side, showing the word 'NO', and a green square on the right side of the screen, showing the word 'YES'. Participants had to indicate as fast and accurate as possible whether the word was an existing word or not, by pressing the corresponding side of the mouse keys on the laptop keyboard; the right mouse key for yes-responses (marked by a green sticker) and the left mouse key for no-responses (marked by a red sticker). RTs were recorded automatically from the onset of the answer screen until the participants pressed the answer key. The experiment began with two practice trials, consisting of one real and one pseudo word. Words were presented in random order. In total, the lexical decision task took about 10 minutes.

Results

Preliminary analyses. For the lexical decision task, data of two participants were excluded from the dataset, because of low total proportion of correct responses (< .75). The average proportion of correct responses to all remaining trials, including pseudo words, was high (M = .95, SD = .08). This indicated that participants had understood the procedure and were not biased towards either affirmative or negative responses. In further RT analyses, pseudo words

were not included and incorrect responses were eliminated. Of the remaining data, responses were trimmed by removing reaction times more than 2.5 standard deviations from the overall condition's mean (i.e., the mean for either low imageability or high imageability words across all subjects). This constituted of less than 5% of the data. Table 5 shows the mean lexical decision latencies and accuracy scores for boys and girls in each condition. The accuracy scores and reaction times for high and low imageability words together showed no evidence of a speed-accuracy trade-off, warranting further comparisons of RTs.

Table 5
Lexical decision latencies and accuracy

	Imageability			
	High		Low	
Measure	Boys	Girls	Boys	Girls
Reaction time (ms)	752 (297)	685 (297)	776 (322)	814 (360)
Percentage correct	98 (5)	98 (5)	96 (6)	94 (7)

Note. Standard deviations are given in parentheses.

Descriptive statistics and *t* tests for all control variables by gender are shown in Table 6. The *t* tests for independent samples indicated that boys performed significantly better than girls on mental rotation (d = .56). There were, however, no gender differences on decoding RAE, reading comprehension, and VMS.

Table 6
Means, standard deviations and t tests for control measures by gender

	Boys			Girls				
Control Measure	*n*	*M*	*SD*	*n*	*M*	*SD*	*t*	*p*
Decoding RAE	53	36.75	12.85	43	38.84	12.99	-.79	.434
Reading Comprehension	53	47.55	18.00	44	47.91	21.13	-.09	.928
Mental Rotation	53	14.83	7.36	44	10.84	6.90	2.73	.007
Visual Memory Span	51	8.59	1.76	44	8.32	1.79	.74	.461

Note. RAE = reading-age equivalent

Analyses of reaction times. As in Experiment 1, the data was analysed by using a one-level CREM, with both participants and items as random effects. Imageability, gender, grade, and the control variables were treated as fixed effects. To test whether participants and items should be considered random effects, the totally unconditional model with random effects alone was tested first. The results showed that both random effects were significant, indicating that, consistently with Experiment 1, RTs differed across participants; *Wald*(1) = 6.11, $p < .001$, and items; *Wald*(1) = 3.42, $p = .001$. Hence, the inclusion of both random effects in the model was justified.

In the next step (see Model 1 in Table 7), the four control measures (grand mean centred) were added to the model as fixed factors, together with grade, gender, imageability, and the interaction of gender with imageability. Controlling for decoding RAE, reading comprehension, mental rotation, and VMS, there were no main effects of grade and gender. Although the coefficient estimates of imageability were significant, $t(112.18) = -2.49$, $p = .014$, indicating faster RTs for high-imageability than for low-imageability words, the fixed effect of imageability did not reach significance, $F(1, 54.99) = 2.40$, $p = .127$. Importantly, in line with the results of Experiment 1, the interaction term of gender with imageability was significant, $F(1, 2340.40) = 5.68$, $p = .017$. Again, pairwise comparisons with Bonferroni correction revealed that there was a significant imageability advantage for girls; $F(1, 112.81) = 6.19$, $p = .014$ (with faster RTs for high-imageability than for low-imageability words, see Table 5), but not for boys; $F(1, 89.89) = .01$, $p = .923$. With respect to the control variables, only VMS appeared to be associated with RTs. Children with a larger visual memory span were faster in making lexical decision. Decoding RAE, reading comprehension, and mental rotation were removed from the model.

In the final step, the interaction terms of grade with imageability, and grade with gender and imageability were added to the model in order to explore grade-related effects. The results can be found in Table 7 (Model 2). The interaction term of gender with imageability was still significant ($p = .034$). Looking at the interaction terms with grade, only the two-way interaction effect of grade with imageability was significant, $F(2, 2305.93) = 3.10$, $p = .045$. Post hoc analyses with Bonferroni correction revealed that only in Grade 4 an overall effect of imageability was found, $F(1, 172.11) = 4.78$, $p = .030$, with faster RTs for high-imageability than for low-imageability words. The three-way interaction term was not significant and therefore removed from the model. A χ^2 test revealed that the fit of Model 2 was not a significant improvement compared to the fit of Model 1; $\chi^2(2) = 5.62$, $p = .060$. Because a more parsimonious model is preferred, Model 1 remained the best fitting model.

Table 7
Regression Coefficient Estimates and Variance-Covariance Estimates for CREMs Predicting Observed Reaction Time

	Model 1		Model 2	
Parameters	Estimate	SE	Estimate	SE
Fixed parameters				
Intercept	693.46***	41.51	647.13***	42.38
VMS	-21.91*	10.82	-21.97*	10.83
Grade 4	-37.95	48.94	-6.60	50.87
Grade 5	-31.17	45.00	-8.03	46.82
Grade 6 (reference)				
Boy	-12.43	38.17	-11.85	38.21
Girl (reference)				
High-imageability	-51.75*	20.81	-13.96	26.48
Low-imageability (reference)				
Interaction terms				
High-imageability*boy	49.88*	20.81	48.90*	20.98
High-imageability*grade 4			-61.18*	26.81
Random parameters				
Residual	63765.39***	1886.73	63592.46***	1881.64
Subjects	26855.28***	4416.96	26894.45***	4421.77
Items	2839.48**	842.54	2882.43**	850.45
Deviance (-2 log likelihood)	34161.94		34156.32	

Note. Both models were hierarchically built. Model 1 represents the best fitting model.
* $p < .05$ ** $p < .01$ *** $p < .001$.

Discussion

As in Experiment 1, the results clearly showed an advantage for girls over boys regarding their task performance, even after correcting for visual memory span. Whereas for boys imageability did not influence the speed with which they made lexical decisions, girls responded significantly faster to high-imageability words than to low-imageability words. These results show that the ease with which a representation of a word can be accessed or activated influences the speed of girls' lexical decisions. This is consistent with previous research showing that high-imageability words have more direct sensory referents than words which are harder to imagine, and hence corresponding mental images are easier to access and verification latencies are shorter

(Schwanenflugel & Stowe, 1989). We therefore suggest that girls' advantage of high-imageability words is an indication that their lexical decisions highly depend on rapidly constructed mental simulations, whereas this may not be true for boys. As in Experiment 1, decoding skills did not influence RTs. Furthermore, the lack of relation between spatial rotation skill and lexical decisions (including imageability) provided additional evidence for the conclusion that spatial ability and mental simulation are two dissociative systems as suggested in Experiment1. Boys' advantage on spatial ability did not help them construct more vivid perceptual simulations. Overall, the results of Experiment 2 nicely replicated the results of Experiment 1.

General discussion

The aim of the present study was to investigate potential gender differences in mental simulation that could account for frequently reported general gender differences in reading comprehension. In Experiment 1, we focused on mental simulation during sentence processing, using the sentence-picture verification task, whereas in Experiment 2 we focused on faster mental simulation processes during single word processing, using a lexical decision task. The results from both experiments were strikingly similar, which suggests that both effects of matching and imageability presumably represent the mental simulation process during language processing. As most research on mental simulation processes has focused on adult readers, the present results provide an important contribution to the thus far small research base showing evidence that children, like adults, retrieve sensory and imaginal information for processing sentences and making lexical decisions (Engelen et al., 2011; Schwanenflugel & Stowe, 1989). These findings are in accordance with embodied theories of cognition which assume that in order to comprehend language, the reader is required to reactivate previously acquired experiences in

the form of, for example, a visual mental representation. Importantly, both experiments are supportive of the idea that this assumption not only applies to adults but also to children at a young age.

A major finding of the present study is that children's gender appears to have an important influence on the mental simulation process. Girls seemed to benefit more from their mental simulations regarding the speed with which they made decisions on both tasks than boys did. Surprisingly, boys did not show any match advantage or imageability effect during sentence and word processing respectively. Their performance, however, was comparable to that of girls with regard to overall speed and accuracy. This might indicate that boys rely less than girls on detailed perceptual simulations when making fast decisions. In line with prior research, girls' advantage in performance might be explained by their stronger ascription of sensory and emotional experiences to words (Bauer & Altarriba, 2008). Moreover, based on ERPs, it has been suggested that women perform a deeper semantic analysis of words, even when the task does not require this and in the absence of active response paradigms (Wirth et al., 2007).

Apparently, imaginal and sensory information is more important for the construction of a mental simulation and subsequent task performance than general visuo-spatial skills as indexed by a mental rotation task. This was evidenced by an advantage of boys in these skills, which had no influence on their task performance. Moreover, the difference in performance between girls and boys was not due to gender differences in decoding or reading comprehension. Therefore, it is likely that the observed effects validly reflect girls excelling more than boys with regard to the construction of a mental simulation and its influence on further decisions about processed words or sentences. Despite the gender difference in mental simulation, no such difference was found for general reading comprehension. This may not be so surprising when considering that mental

simulation is a component process of language comprehension. Many other processes (e.g., orthographic knowledge, syntax processing, reading strategies, but also task demands) can influence the final outcome on a general reading comprehension measure, which may diminish or compensate for a gender effect in mental simulation. It may therefore be interesting to study gender differences on other underlying component processes and their relative influence on general reading comprehension in future research. In addition, although the present results show that the found gender differences on semantic processing at sentence level were not due to specific task demands, future investigations should also aim at distinguishing between the role of task demands and semantic processing at word level.

A potential source for the present gender differences may reside in episodic memory, which is assumed to play a crucial role in the construction of a mental simulation already at an early stage of language processing (Elman, 2009). In line with this, Clinton et al. (2012) recently showed an advantage for girls in tasks using episodic memory. According to them, girls access their episodic memory and activate the associated perceptual and sensory referents more easily than boys do. This causes girls to construct more coherent mental simulations and improves their performance on a cognitive task compared to boys. For the sentence-picture verification task, this would mean that drawing upon episodic memory to construct a more coherent mental simulation results in a relatively larger overlap with the presented picture in the match condition, which logically facilitates the comparison process and speeds up the verification times. For the lexical decision task, being able to rapidly access episodic memory and reactivate a word's sensory referents, helps the reader make lexical decisions faster. High-imageability words, which elicit activation of more sensory referents than low-imageability words, facilitate this process

even further. Further research is required to investigate to what extent mental simulation, episodic memory, and language processing are (causally) related.

In line with the growing awareness of the importance of translating empirical cognitive findings on mental simulation and visualization to education (de Koning & van der Schoot, 2013), we provide some educational implications that arise from the present findings. Generally, our findings suggest that the development of reading comprehension methods in primary education may be facilitated by having knowledge of individual differences and the influence of gender on an early stage of language processing. Male students in particular may benefit from early intervention, as cognitive abilities are not an inherent aspect of gender and can be improved and trained in time (Law, Pellegrino, & Hunt, 1993). Furthermore, in our study we found gender differences in primary school children between 8 and 13 years old. It could well be that these differences change over time, because they are still being developed. Further research is needed to explore whether these differences still exist in later years and their potential influence on or relation to other cognitive abilities.

In sum, the reported studies show the importance of taking participant characteristics into account when studying mental simulation processes during language processing. Most research has been conducted with the assumption that concepts are represented similarly across gender, whereas the present study clearly indicates differences. The present study provided evidence for girls, but not boys, depending highly on the construction of mental simulations during language-related tasks. Presumably, girls construct more coherent and vivid mental simulations than boys do by accessing their episodic memory and activating experiential traces more easily during reading. The findings of the present study emphasize the importance of including gender into future research which is meant to expand our knowledge on the models of language

comprehension. Adopting a one-size-fits-all approach when investigating (or theorizing about) mental simulation does not seem to accurately represent existing individual differences and future research should, therefore, focus on potentially important individual differences.

References

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53. http://doi.org/10.1016/j.bandl.2013.01.005

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, *22*(4), 637–660. http://doi.org/10.1017/S0140525X99532147

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–45. http://doi.org/10.1146/annurev.psych.59.103006.093639

Bauer, L. M., & Altarriba, J. (2008). An investigation of sex differences in word ratings across concrete, abstract, and emotion words. *The Psychological Record*, *58*, 465–474.

Below, J. L., Skinner, C. H., Fearrington, J. Y., & Sorrell, C. A. (2010). Gender differences in early literacy: Analysis of kindergarten through fifth-grade dynamic indicators of basic early literacy skills probes. *School Psychology Review*, *39*(2), 240–257.

Boersma, P., & Weenink, D. (2011). PRAAT: Doing phonetics by computer.

Bottini, G., Corcoran, R., Sterzi, R., Paulesu, E., Schenone, P., Scarpa, P., … Frith, D. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language A positron emission tomography activation study. *Brain*, *117*(6), 1241–1253. http://doi.org/10.1093/brain/117.6.1241

Burns, N. R., & Nettelbeck, T. (2005). Inspection time and speed of processing: Sex differences on perceptual speed but not IT. *Personality and Individual Differences*, *39*(2), 439–446. http://doi.org/10.1016/j.paid.2005.01.022

Carson, R. J., & Beeson, C. M. L. (2013). Crossing Language Barriers : Using Crossed Random Effects Modelling in Psycholinguistics Research. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 25–41.

Chumbley, J. I., & Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Memory & Cognition*, *12*(6), 590–606. http://doi.org/10.3758/BF03213348

Clinton, V., Seipel, B., van den Broek, P., McMaster, K. L., Kendeou, P., Carlson, S. E., & Rapp, D. N. (2012). Gender differences in inference generation by fourth-grade students. *Journal of Research in Reading*. http://doi.org/10.1111/j.1467-9817.2012.01531.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Colombo, L., Pasini, M., & Balota, D. A. (2006). Dissociating the influence of familiarity and

meaningfulness from word frequency in naming and lexical decision performance. *Memory & Cognition*, *34*(6), 1312–1324. http://doi.org/10.3758/BF03193274

Cui, X., Jeter, C. B., Yang, D., Montague, P. R., & Eagleman, D. M. (2007). Vividness of mental imagery: individual variability can be measured objectively. *Vision Research*, *47*(4), 474–8. http://doi.org/10.1016/j.visres.2006.11.013

de Koning, B. B., & van der Schoot, M. (2013). Becoming part of the story! Refueling the interest in visualization strategies for reading comprehension. *Educational Psychology Review*, *25*(2), 261–287. http://doi.org/10.1007/s10648-013-9222-6

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*(4), 547–582. http://doi.org/10.1111/j.1551-6709.2009.01023.x

Engelen, J. A. A., Bouwmeester, S., de Bruin, A. B. H., & Zwaan, R. A. (2011). Perceptual simulation in developing language comprehension. *Journal of Experimental Child Psychology*, *110*(4), 659–75. http://doi.org/10.1016/j.jecp.2011.06.009

Farah, M. J., Hammond, K. M., Levine, D. N., & Calvanio, R. (1988). Visual and spatial mental imagery: Dissociable systems of representation. *Cognitive Psychology*, *20*(4), 439–462. http://doi.org/10.1016/0010-0285(88)90012-6

Feenstra, H., Kamphuis, F., Kleintjes, F., & Krom, R. (2010). *Wetenschappelijke verantwoording. Begrijpend lezen voor groep 3 tot en met 6 [Scientific justification of reading comprehension tests for Grades 5 to 8]*. Arnhem, The Netherlands: Cito.

Gerrig, R. J., & O'Brien, E. J. (2005). The scope of memory-based processing. *Discourse Processes*, *39*(2-3), 225–242. http://doi.org/10.1080/0163853X.2005.9651681

Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. *Remedial and Special Education*, *7*(1), 6–10. http://doi.org/10.1177/074193258600700104

Greenberg, D. L., & Knowlton, B. J. (2014). The role of visual imagery in autobiographical memory. *Memory & Cognition*. http://doi.org/10.3758/s13421-014-0402-5

Gullick, M. M., Mitra, P., & Coch, D. (2013). Imagining the truth and the moon: an electrophysiological study of abstract and concrete word processing. *Psychophysiology*, *50*(5), 431–440. http://doi.org/10.1111/psyp.12033

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51. http://doi.org/10.1111/j.1529-1006.2007.00032.x

Harshman, R. A., Hampson, E., & Berenbaum, S. A. (1983). Individual differences in cognitive

abilities and brain organization, part I: Sex and handedness differences in ability. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *37*(1), 144–192. http://doi.org/10.1037/h0080690

Herlitz, A., & Rehnman, J. (2008). Sex differences in episodic memory. *Current Directions in Psychological Science*, *17*(1), 52–56. http://doi.org/10.1111/j.1467-8721.2008.00547.x

Hirschfeld, G., Feldker, K., & Zwitserlood, P. (2012). Listening to "flying ducks": individual differences in sentence-picture verification investigated with ERPs. *Psychophysiology*, *49*(3), 312–321. http://doi.org/10.1111/j.1469-8986.2011.01315.x

Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A., & Lüdtke, J. (2007). Experiential simulations of negated text information. *Quarterly Journal of Experimental Psychology (2006)*, *60*(7), 976–990. http://doi.org/10.1080/17470210600823512

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. http://doi.org/10.1037/0033-295X.95.2.163

Kraemer, D. J. M., Rosenberg, L. M., & Thompson-Schill, S. L. (2009). The neural correlates of visual and verbal cognitive styles. *The Journal of Neuroscience*, *29*(12), 3792–3798. http://doi.org/10.1523/JNEUROSCI.4635-08.2009

Krom, R. S. H., & Kamphuis, F. H. (2001). Wetenschappelijke verantwoording van de toetsserie Leestechniek & Leestempo [Scientific justification of the Reading Speed Test]. Arnhem: Cito.

Law, D. J., Pellegrino, J. W., & Hunt, E. B. (1993). Comparing the tortoise and the hare: gender differences and experience in dynamic spatial reasoning tasks. *Psychological Science*, *4*(1), 35–40. http://doi.org/10.1111/j.1467-9280.1993.tb00553.x

Logan, S., & Johnston, R. (2009). Gender differences in reading ability and attitudes: examining where these differences lie. *Journal of Research in Reading*, *32*(2), 199–214. http://doi.org/10.1111/j.1467-9817.2008.01389.x

Logan, S., & Johnston, R. (2010). Investigating gender differences in reading. *Educational Review*, *62*(2), 175–187. http://doi.org/10.1080/00131911003637006

Madden, C. J., & Zwaan, R. A. (2003). How does verb aspect constrain event representations? *Memory & Cognition*, *31*(5), 663–672. http://doi.org/10.3758/BF03196106

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, *91*(2), 167–180. http://doi.org/10.1348/000712600161763

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). PIRLS 2006 international report. *TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.*

Nation, K., & Snowling, M. J. (1998). Semantic processing and the development of word-recognition skills: evidence from children with reading comprehension difficulties. *Journal of Memory and Language*, *39*(1), 85–101. http://doi.org/10.1006/jmla.1998.2564

Patson, N. D., George, G., & Warren, T. (2014). The conceptual representation of number. *The Quarterly Journal of Experimental Psychology*, *67*(7), 1349–1365. http://doi.org/10.1080/17470218.2013.863372

Pecher, D., van Dantzig, S., Zwaan, R. A., & Zeelenberg, R. (2009). Language comprehenders retain implied shape and orientation of objects. *The Quarterly Journal of Experimental Psychology*, *62*(6), 1108–1114. http://doi.org/10.1080/17470210802633255

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. *Psychological Science*, *14*(2), 119–124. http://doi.org/10.1111/1467-9280.t01-1-01429

Quaiser-Pohl, C. (2003). The Mental Cutting Test "Schnitte" and the Picture Rotation Test-two new measures to assess spatial ability. *International Journal of Testing*, *3*(3), 219–231. http://doi.org/10.1207/S15327574IJT0303_2

Reichle, E. D., Carpenter, P. A., & Just, M. A. (2000). The neural bases of strategy and skill in sentence-picture verification. *Cognitive Psychology*, *40*(4), 261–295. http://doi.org/10.1006/cogp.2000.0733

Richardson, J. T. E. (1991). Gender differences in imagery, cognition, and memory. In R. H. Logie & M. Denis (Eds.), *Mental images in human cognition* (pp. 271–303). New York, NY: Elsevier.

Rommers, J., Meyer, A. S., & Huettig, F. (2013). Object Shape and Orientation Do Not Routinely Influence Performance During Language Processing. *Psychological Science*, *24*(11), 2218–2225. http://doi.org/10.1177/0956797613490746

Sadoski, M., & Paivio, A. (2013). *Imagery and text: a dual coding theory of reading and writing* (2nd ed.). New York, NY: Routledge.

Schrooten, W., & Vermeer, A. (1994). *Woorden in het basisonderwijs. 15.000 woorden aangeboden aan leerlingen [Words in primary education. 15.000 words offered to pupils]*. Tilburg: Tilburg University Press.

Schwanenflugel, P. J., & Stowe, R. W. (1989). Context availability and the processing of abstract

and concrete words in sentences. *Reading Research Quarterly*, *24*(1), 114–126. http://doi.org/10.2307/748013

Simpson, G. B. (1994). Context and the processing of ambiguous words. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 359–374). San Diego, CA: Academic Press.

Solomon, K. O., & Barsalou, L. W. (2004). Perceptual simulation in property verification. *Memory & Cognition*, *32*(2), 244–259. http://doi.org/10.3758/BF03196856

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*(2), 153–156. http://doi.org/10.1111/1467-9280.00326

van Dam, W. O., van Dijk, M., Bekkering, H., & Rueschemeyer, S.-A. (2012). Flexibility in embodied lexical-semantic representations. *Human Brain Mapping*, *33*(10), 2322–2333. http://doi.org/10.1002/hbm.21365

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The landscape model of reading: inferences and the online construction of a memory representation. In H. van Oosterndorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 62–87). Mahwah, NJ: Lawrence Erlbaum.

van der Leij, A., Struiksma, A. J. C., & Vieijra, J. P. (1995). *Diagnostiek van technisch lezen en aanvankelijk spellen [Diagnostics of decoding and early spelling ability]*. Amsterdam: VU Uitgeverij.

Van Loon-Vervoorn, W. A. (1985). *Voorstelbaarheidswaarden van Nederlandse woorden: 4600 substantieven, 1000 verba en 500 adjectieven [Imageability values of Dutch words: 4600 nouns, 1000 verbs, and 500 adjectives]*. Swets en Zeitlinger.

Vukovic, N., & Williams, J. N. (2014). Automatic perceptual simulation of first language meanings during second language sentence processing in bilinguals. *Acta Psychologica*, *145*, 98–103. http://doi.org/10.1016/j.actpsy.2013.11.002

Warrington, M., Younger, M., & Bearne, E. (2006). *Raising boys' achievement in primary schools: towards a holistic approach*. Maidenhead: Open University Press.

Wechsler, D. (1987). *Wechsler memory scale - revised*. New York, NY.

Weekers, A., Groenen, I., Kleintjes, F., & Feenstra, H. (2011). Wetenschappelijke verantwoording papieren toetsen. Begrijpend lezen voor groep 7 en 8 [Scientific justification of reading comprehension tests for Grades 5 and 6]. Arnhem: Cito.

Willems, R. M., Hagoort, P., & Casasanto, D. (2010). Body-specific representations of action verbs: neural evidence from right- and left-handers. *Psychological Science*, *21*(1), 67–74.

http://doi.org/10.1177/0956797609354072

Wirth, M., Horn, H., Koenig, T., Stein, M., Federspiel, A., Meier, B., … Strik, W. (2007). Sex differences in semantic processing: event-related brain potentials distinguish between lower and higher order semantic analysis during word reading. *Cerebral Cortex*, *17*(9), 1987–97. http://doi.org/10.1093/cercor/bhl121

Wood, E. (2003). The power of pupil perspectives in evidence-based practice: the case of gender and underachievement. *Research Papers in Education*, *18*(4), 365–383. http://doi.org/10.1080/0267152032000176864

Yaxley, R. H., & Zwaan, R. A. (2007). Simulating visibility during language comprehension. *Cognition*, *105*(1), 229–236. http://doi.org/10.1016/j.cognition.2006.09.003

Zwaan, R. A. (2003). The Immersed Experiencer: Toward An Embodied Theory Of Language Comprehension. In B. H. Ross (Ed.), *Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 44, pp. 35–62). New York, NY: Academic Press. http://doi.org/10.1016/S0079-7421(03)44002-4

Zwaan, R. A., & Pecher, D. (2012). Revisiting Mental Simulation in Language Comprehension: Six Replication Attempts. *PLoS ONE*, *7*(12), e51382. http://doi.org/10.1371/journal.pone.0051382

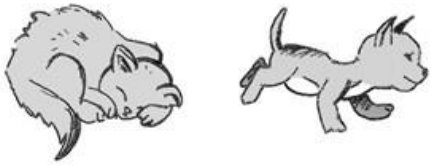
Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language Comprehenders Mentally Represent the Shapes of Objects. *Psychological Science*, *13*(2), 168–171. http://doi.org/10.1111/1467-9280.00430

Appendix A

Samples of sentence-picture pairs translated from Dutch

The boy put the banana in his [shopping cart/mouth].

The girl saw the cat [in the basket/on the street].



The woman put the book on the [shelf/copy machine].

The man kept/held the umbrella [in the closet/over his head].*

*In the Dutch version of the sentences, only the last part of every sentence (i.e., the prepositional phrase) differed between conditions.