PRESCHOOL CHILDREN'S JUDGMENTS OF LEARNING: THE EFFECTS OF DELAY AND PRACTICE

A dissertation submitted to Kent State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

by

Stacy L. Lipowski

August, 2011

Dissertation written by

Stacy L. Lipowski

B.A., Mount Union College, 2005

M.A., Kent State University, 2008

Ph.D., Kent State University, 2011

Approved by

William E. Merriman John Dunlosky Katherine Rawson Beth Wildman Christopher Was Janice Kroeger Chair, Doctoral Dissertation Committee Member, Doctoral Dissertation Committee

Accepted by

Maria S. Zaragoza John R. D. Stavley Chair, Department of Psychology Dean, College of Arts and Sciences

TABLE OF CONTENTS

LIST OF FIGURES		
ACKNOWLEDGMENTS		
CHAPTER 1	1	
NTRODUCTION		
1.1 Preschool children's judgments of learning	1	
1.2 Accuracy of delayed JOLs in children	6	
1.3 Overview of current investigation	7	
CHAPTER 2	9	
EXPERIMENT 1		
2.1 Introduction	9	
2.2 Method	13	
2.2.1 Participants	13	
2.2.2 Materials and procedure	13	
2.3 Results and discussion	15	
2.3.1 Judgment Accuracy	15	
2.3.2 Judgment magnitude and final recall as a function of Pre-		
JOL recall	18	
2.3.3 Judgment magnitude and final recall as a function of Item		
Relatedness	22	
CHAPTER 3	26	
EXPERIMENT 2	-	

3.1	Introduction	
3.2	Method	27
	3.2.1 Participants	27
	3.2.2 Materials and procedure	28
3.3	Results and discussion	29
	3.3.1 Judgment Accuracy	29
	3.3.2 Judgment magnitude and final recall as a function of Pre-	
	JOL recall	30
	3.3.2 Judgment magnitude and final recall as a function of Item	
	Relatedness	31
	3.3.4 Forced choice questions	32
CHAPTER	4	34
EXPERIM	ENT 3	
4.1	Introduction	34
4.2	Method	37
	4.2.1 Participants	37
	4.2.2 Materials and procedure	38
4.3	Results and discussion	40
	4.3.1 Judgment Accuracy	40
	4.3.2 Judgment magnitude and final recall as a function of Pre-	
	JOL recall	42

4.3.2 Judgment magnitude and final recall as a function of Item					
Relatedness	43				
4.3.4 Forced choice questions	44				
CHAPTER 5	46				
GENERAL DISCUSSION					
REFERENCES	55				
APPENDIX	58				

LIST OF FIGURES

Figure 1	18
Mean final recall as a function of pre-JOL recall in Experiment 1.	
Figure 2	20
Mean JOL magnitude as a function of pre-JOL recall in Experiment 1.	
Figure 3	23
Mean final recall as a function of item relatedness in Experiment 1.	
Figure 4	24
Mean JOL magnitude as a function of item relatedness in Experiment 1.	

ACKNOWLEDGMENTS

This dissertation is dedicated to Matthew and Evan Lipowski, whose unconditional love and support made this journey possible. I would like to thank my advisors, Dr. William E. Merriman and Dr. John Dunlosky for their continuous support and guidance. I would also like to thank the members of the Merriman lab and RADlab, as well as all of my committee members – Dr. Katherine Rawson, Dr. Christopher Was, Dr. Beth Wildman and Dr. Janice Kroeger – for their support during this process. Finally, I would like to thank my family and friends for their constant love and support.

CHAPTER 1

INTRODUCTION

1.1 PRESCHOOL CHILDREN'S JUDGMENTS OF LEARNING

If students are able to accurately monitor their memory, they could be more efficient learners. For example, when studying for an exam, a student may try to make judgments about how well he has learned the relevant pieces of information. Accurate judgments will allow the student to make good decisions about how to allocate study time. Ideally, he will spend less time studying well-learned material and more time studying material that is less well-learned. Accurate metacognitive judgments and the efficient use of study time can lead to better performance on an exam (Dunlosky, Hertzog, Kennedy, & Thiede, 2005).

Throughout the school years, students have many opportunities to refine their metacognitive judgments; however, not much is known about these abilities in young children. The primary goal of the current studies is to examine metacognitive monitoring in preschool children. Before introducing the current experiments, I will discuss a type of judgment that has been used to measure the accuracy of metacognitive monitoring. Then, I will discuss the accuracy of these judgments in adults and one factor (i.e., timing of the judgment) that has been shown to influence their accuracy. To motivate the current experiments, I will discuss how timing influences children's metacognitive judgments. Finally, I will discuss the primary goal and the general method used in the current studies.

Judgments of learning (JOLs), which are judgments about the likelihood of recalling a studied item on a subsequent memory test, provide one way for researchers to examine the accuracy of metacognitive monitoring. In the typical paradigm, adults are instructed to learn a set of paired associates. Sometime after studying each item, JOLs are made. The researchers may ask, for example, "Using a scale from 0 to 100, what is the likelihood that you will recall this item on a later memory test?" The JOLs are followed by a final recall test. The relative accuracy of JOLs is evaluated with the Goodman-Kruskal gamma correlation (G), which measures how accurately an individual has predicted the likelihood of recalling one item relative to another. Many of the original studies concerning the relative accuracy of JOLs found that adults' judgments were only moderately accurate (e.g., Vesonder & Voss, 1985; Begg, Duft, Lalonde, Melnick, & Sanvito, 1989). These findings suggested that adults' monitoring accuracy had room for improvement.

Nelson and Dunlosky (1991) proposed that the timing of the judgments may be one factor that influences their accuracy. In previous studies, JOLs were typically made immediately after studying target items. Nelson and Dunlosky (1991) predicted that JOLs would be more accurate with a delay between the study phase and when JOLs were made. To test this hypothesis, they had undergraduates study unrelated noun pairs. Individuals made immediate JOLs for half of the items and delayed JOLs for the remaining items. Immediate JOLs were made immediately after each word pair was studied and delayed JOLs were made after all items in a block had been studied. JOLs were more accurate in the delayed condition than in the immediate condition. The average gamma correlation between delayed JOLs and recall was +.90, but the correlation between immediate JOLs and recall was only +.38 (see Rhodes & Tauber, 2011 for a review).

To explain this delayed-JOL effect, Nelson and Dunlosky (1991) proposed the Monitoring-Dual-Memories (MDM) principle. This principle is based on Atkinson and Shiffrin's (1968) modal model of memory, and also relies on the assumption that individuals monitor what they can retrieve from memory when making JOLs. The MDM principle hypothesizes that individuals monitor information in both short-term memory (STM) and long-term memory (LTM). If a JOL is made immediately after studying a target item, the judgment may be based on information that is present in STM and LTM. However, only information that has been transferred to LTM will be available during the final recall test. Therefore, when immediate judgments are made, the information that is available in STM may add noise to the individual's evaluation of what information is actually available in LTM. Hence, delayed JOLs are more accurate than immediate JOLs because individuals can only rely on information about the target item that is available in LTM when making delayed judgments.

Dunlosky and Nelson (1992) found support for the Monitoring-Dual-Memories principle by manipulating the cue individuals received when making judgments of learning. Undergraduates studied and made JOLs for unrelated, noun-noun pairs (e.g., ocean-tree). All students made immediate JOLs for half of the items and delayed JOLs for the remaining items. Immediate JOLs were made immediately after the target item was studied. Delayed JOLs were made after all word pairs were studied. Also, a minimum of 10 items between study and JOL ensured a sufficient delay for items in the delayed JOL condition. The type of cue given when making JOLs was a betweensubjects variable. Half of the students were presented only with the stimulus when making JOLs (e.g., ocean- ?) and the other half were presented with the complete stimulus-response pair (e.g., ocean-tree). A 10 minute filler task followed the JOLs. Finally, individuals were given a final recall test. They were presented with the first word of each pair and instructed to provide the second word.

Dunlosky and Nelson (1992) replicated the delayed-JOL effect, but only in the stimulus alone condition. When the students were cued with only the stimulus, delayed JOLs were significantly more accurate than immediate JOLs (median G = +.93 and +.45, respectively). When the students were cued with the complete stimulus-response pair, the accuracy of delayed JOLs was not significantly different from the accuracy of immediate JOLs (median G = +.60 and +.55, respectively). These results show that attempting to retrieve the target response is a key component of the delayed-JOL effect. When cued with the stimulus-response pair, the individual is making the JOL with the response available in short-term memory. This adds noise to the monitoring of information in long-term memory, which is what happens when individuals make immediate JOLs. Although these findings provide support for the MDM principle, other theories have been proposed to explain the high accuracy of delayed JOLs.

In response to the MDM principle, Spellman and Bjork (1992) proposed the selffulfilling-prophecy (SFP) hypothesis. Consistent with the MDM principle, Spellman and Bjork (1992) assume that participants attempt to retrieve the target item when making JOLs. Previous research has shown that retrieval practice can improve long-term retention of an item, even when the individual is given no feedback (see Roediger & Karpicke, 2006 for a recent review). Based on this finding, successful retrieval of an item when making delayed JOLs should lead to better memory for that item at final recall. If a target item is not successfully retrieved when a delayed JOL is made for that item, successful retrieval at final recall is highly unlikely. Given that both delayed-JOLs and final recall are assumed to be based on retrieval success, Spellman and Bjork (1992) argue that the high correlation typically found between these two measures is simply a self-fulfilling-prophecy.

Although the SFP hypothesis and the MDM principle offer different explanations for the delayed-JOL effect, both theories provide useful insight into how individuals make delayed JOLs. Most important for the current experiments, both theories emphasize the central role that retrieval plays when adults make delayed JOLs. Nelson, Narens, and Dunlosky (2004) found evidence suggesting that adults attempt to retrieve the target information prior to making JOLs. They had undergraduates study unrelated word pairs. The participants made both immediate and delayed JOLs. Immediately before making their JOLs, half of the participants were asked to recall the target item when cued with the stimulus. The remaining individuals were not asked to recall the target item before making their JOLs. Nelson et al. (2004) found that explicitly asking individuals to retrieve the target item before making their JOLs did not influence their responses. Both groups gave similar JOLs and had similar levels of final recall. Also, the accuracy of their immediate and delayed JOLs was similar and both groups showed the delayed JOL effect. These findings suggest that adults attempt retrieval of the target information when making JOLs, even when they are not explicitly instructed to do so.

1.2 ACCURACY OF DELAYED JOLS IN CHILDREN

Although the delayed-JOL effect is well established in the adult literature, not much research has been conducted concerning this phenomenon in children. Schneider, Visé, Lockl, and Nelson (2000) examined whether the delayed-JOL effect could be found in a group of elementary school children. Kindergarteners, second graders, and fourth graders were shown pairs of unrelated, concrete objects. To insure equal levels of recall across the age groups, each group was given a different set size to learn. Based on a pilot study, kindergarteners, second graders, and fourth graders studied 8, 10, and 12 pairs, respectively. The children were told that they would have to recall the second word of a pair when shown a picture of the first word. Half of the children in each age group were in the immediate JOL group and the remaining children were in the delayed JOL group. Immediate JOLs were made immediately after studying the target item and delayed JOLs were made after all pairs had been studied. Schneider et al. (2000) found the delayed-JOL effect in each age group. Overall, the accuracy of delayed JOLs (mean G = +.74) was much higher than the accuracy of immediate JOLs (mean G = +.18). Even kindergarteners could make accurate delayed JOLs.

Koriat and Shitzer-Reichert (2002) also found evidence showing that grade school children can make accurate delayed JOLs. Second and fourth graders studied and made JOLs for 24 word pairs. The timing of the judgments (immediate vs. delayed) was a within-subject variable and the type of cue used for the judgments (stimulus alone vs. stimulus-response pair) was a between-subjects variable. Consistent with Dunlosky and Nelson (1992), Koriat and Shitzer-Reichert (2002) found a delayed-JOL effect in the stimulus only condition, but not in the stimulus-response condition. This outcome provides further evidence that children can accurately monitor their learning, but only when the judgments are made after a delay and are cued with the stimulus alone. Consistent with Dunlosky and Nelson's (1992) research with adults, this study emphasizes the importance of attempting to retrieve the target item from memory when making JOLs.

1.3 OVERVIEW OF CURRENT INVESTIGATION

A major goal of the current investigation was to explore whether preschoolers can make accurate delayed JOLs. In three experiments, preschoolers were taught names for several different animals. Half of the names were related to the animal in some way and half were not. For instance, "This pig's name is Brian." or "Frogs hop. So this frog's name is Hoppy." For their JOLs, the children predicted whether they would remember the name for each of the animals. Before each JOL was made, the children were asked to recall the animal's name. Although it is assumed that individuals automatically attempt to retrieve the target information before making a JOL, pre-JOL recall made this

7

information more salient to the preschoolers. It also made it possible to explore whether children based their JOLs on their ability to recall the items. If children attempt to retrieve the animal's name immediately before making a JOL, they should be more likely to use this recall attempt as a cue. Lastly, the JOLs were followed by a final recall test.

In Experiment 1, the accuracy of delayed JOLs was compared to the accuracy of immediate JOLs. Along with the preschoolers, a group of 3rd graders was tested to serve as a comparison group. The performance of the 3rd graders in the current experiment should be similar to that of the 2nd and 4th graders in the experiments discussed above. In Experiments 2 and 3, only preschoolers were tested and the children only made delayed JOLs. The procedures were similar to those used in Experiment 1; however, an additional task was added. After all of the JOLs were made, the children were shown two of the animals and asked to judge which animal's name they would be more likely to remember later. This forced choice paradigm was used to assess whether the children's JOLs were influenced by a response bias, specifically, the tendency to report that all of the animals' names would be recalled. In Experiment 2, the preschoolers also received an additional study trial, including a retrieval attempt and feedback. In Experiment 3, to explore the influence of practice on preschoolers' JOLs, half of the children received the additional study trial and half did not.

CHAPTER 2

EXPERIMENT 1

2.1 INTRODUCTION

Previous research has shown that children can make accurate delayed JOLs under some conditions (Schneider, et al., 2000; Koriat & Shitzer-Reichert, 2002). However, no studies have examined the accuracy of immediate or delayed JOLs in preschool-age children. Based on Schneider et al.'s (2000) finding that kindergartners can make accurate delayed JOLs, it is possible that preschoolers will be able to make accurate JOLs in a delayed condition. However, previous research has shown that preschoolers can be extremely overconfident when making predictions about their own memory performance (e.g, Flavell, Friedrichs, & Hoyt, 1970; Yussen & Levy, 1975; Lipko, Dunlosky, & Merriman, 2009). Such overconfidence could cause preschoolers' predictions to be inaccurate in both the immediate and delayed conditions. That is, they may not distinguish between well-learned items and poorly learned items in their judgments because they are confident that all of the items will be recalled later.

In a recent study concerning overconfidence in preschoolers, Lipko, Dunlosky, and Merriman (2009) had 4- and 5-year-olds study 10 pictures of familiar objects. The children named each picture as it was presented and then were given a 10 second study period. Next, the experimenter asked the children how many pictures they would be able to recall once they were covered up. Finally, children completed a final recall test in which they were asked to recall as many pictures as they could remember. After the experimenter told the child how many pictures he or she had recalled, the procedure was repeated two more times with different sets of pictures. The children were overconfident across all three trials. Even on the final trial, after two experiences of recalling fewer items than predicted, children's predicted recall (M = 7.05) was significantly higher than their actual recall (M = 3.52).

In another experiment, Lipko et al. (2009) found that preschoolers' predictions about their memory remained overconfident across *five* trials. As in the previous experiment, 4- and 5-year-olds studied 10 pictures of familiar objects for 10 seconds, predicted their future recall, and completed a final recall test. This procedure was repeated four times with different pictures. Along with overconfidence, Lipko et al. (2009) were also interested in whether children remembered their previous recall performance and whether they used this information when making predictions about future recall performance. To determine whether the preschoolers' accurately remembered their previous recall performance, half of the children were asked how many pictures they remembered on the last trial immediately before making their predictions.

The preschoolers remained overconfident throughout the experiment. Even after four trials of practice, children's predicted recall on the final trial (M = 7.20) was significantly higher than their actual recall (M = 3.13). Even though the preschoolers were able to accurately remember how many items they had recalled on the previous trial, they did not use this information when making predictions about how many items they would recall on future trials. The predictions of children who were asked to recall how many items they had just recalled on the previous trial were not different from the predictions of children who were not asked to make these postdictions. These findings provide evidence that preschoolers do not use past performance on a memory task as a cue when making predictions regarding how they will perform on that task in the future. Consistent with these findings, preschoolers in the current studies may not use performance on pre-JOL recall as a cue when making predictions about their performance at final recall. Ignoring this cue would cause preschoolers' judgments to be overconfident.

Even though Lipko et al. (2009) examined aggregate judgments (i.e., judgments of the total number of items to be recalled), their findings still suggest that children's item-by-item judgments may be overconfident in the current studies. In fact, Schneider, Visé, Lockl, and Nelson (2000) found that children were more overconfident when making item-by-item judgments than aggregate judgments. As described above, kindergartners, 2nd graders, and 4th graders studied and made JOLs for pairs of unrelated objects. All children made both item-by-item and aggregate judgments. For the aggregate judgments, children were asked how many of the items they would be able to remember overall. For the item-by-item judgments, children were asked if they would remember the target item of each pair individually. Children's item-by-item judgments were higher than their aggregate judgments. Based on these findings, preschoolers are expected to be overconfident in both the immediate and delayed conditions. Third

graders are expected to make accurate JOLs in the delayed condition, but not in the immediate condition.

Along with timing of the JOLs, item relatedness was manipulated in the current studies. Half of the names the children were taught were associated with the animal in some way (e.g., Pigs "oink." So this pig's name is Oinky.) In contrast, half of the names were unrelated to the animal (e.g., "The cat's name is Molly.) The primary reason for manipulating item relatedness was to ensure that recall was not at floor or ceiling. However, this manipulation also provided a way to examine the effects of item relatedness on preschoolers' JOLs. Previous research has shown that both adults and children give higher JOLs to easy items than difficult items (e.g., Koriat, 1997; Dunlosky & Matvey, 2001, Koriat & Shitzer-Reichert, 2002).

As described above, Koriat and Shitzer-Reichert (2002) had 2^{nd} and 4^{th} graders study a list of word pairs then make JOLs on a scale from 1 (i.e., "no chance to recall the response word" to 5 (i.e., "completely certain to recall the response word"). Along with examining the effects of timing, they looked at the influence of item difficulty on recall and JOLs. Half of the word pairs were judged to be easy to remember and the other half were considered difficult to remember. Item difficulty influenced both recall and JOLs. Recall was significantly higher for easy items (M = .73) than difficult items (M = .12). Children gave higher JOLs to easy pairs than difficult pairs; however, there was an Age x Difficulty interaction. The 2^{nd} and 4^{th} graders gave equally high JOLs to easy pairs (M =4.06 and 3.92, respectively), but the 2^{nd} graders were more overconfident than the 4^{th} graders when making JOLs for difficult items (M = 3.51 and 2.99, respectively). Consistent with Koriat and Shitzer-Reichert's (2002) findings, it is possible that the 3rd graders will distinguish between related and unrelated items when making their JOLs. They are expected to give higher JOLs to related items than unrelated items in the current experiment. However, considering the overconfidence of the 2nd graders, it is unlikely that the preschoolers in the current experiments will use item relatedness as a cue when making JOLs. The preschoolers' JOLs are expected to be highly overconfident. If this hypothesis is supported, preschoolers will not distinguish between related and unrelated items when making JOLs because they will give "yes" judgments to most or all of the items.

2.2 METHOD

2.2.1 Participants

Children were recruited from middle class, suburban schools in Northeast Ohio. The participants were 29 preschoolers (M age = 4-9; range = 4-2 to 5-4; 18 girls and 11 boys) and 39 3rd graders (M age = 8-10; range = 8-1 to 9-9; 18 girls and 21 boys). Each child received a small prize for participating.

2.2.2 Materials and procedure

Children were tested individually at their schools. The children were taught names for various stuffed animals. The preschoolers learned names for 12 animals and the 3rd graders learned names for 18 animals. The degree of relatedness between the animals and their names and the timing of the JOLs were within-subject variables. Half of the names were associated with the animal in some way and half of the names were not associated with the animal. For example, the pig's name was "Oinky" for half of the participants and "Brian" for the remaining participants (see Appendix A for a list of animals and names). Immediately before making each JOL, the experimenter asked the children to recall the animal's names. For half of the trials, children made JOLs immediately after being told an animal's name. For the remaining trials, judgments were made after a two minute delay. Immediate and delayed judgments were grouped into two separate blocks. The order in which the blocks were presented was counterbalanced.

For the immediate trials, the animals were presented to the child one at a time. Each time an animal was presented, the experimenter would ask what type of animal it was and state the animal's name. If the name was related to the animal, the experimenter would explicitly point out the association (e.g., "Elephants like to eat peanuts. So this elephant's name is Peanut.") The child was asked to repeat the animals name to insure it was heard correctly. Next, the child was asked to recall the animal's name and then make an immediate JOL. The experimenter asked the child, "Do you think you will remember the name I taught you for this animal if I ask you later?" The child had to answer with either a "yes" or "no" response. The child made a JOL for each animal. After a thirty second delay, the experimenter presented each animal to the child one last time for a final recall test. The animals were presented in the same order for all parts of the task, including study, JOLs, and final recall. For the delayed trials, the remaining animals were presented to the child one at a time. The procedure was the same as in the immediate trials, except for the timing of the JOLs. The child was asked about the type of animal, told the name and any associations, and then was asked to repeat the name. A two minute delay followed this study phase. The delay was filled with an age appropriate distractor task (i.e., coloring). Next, the experimenter presented each animal for a second time and asked the child to recall the name before making a delayed JOL. After a thirty second delay, the child completed a final recall test.

2.3. RESULTS AND DISCUSSION

2.3.1 Judgment Accuracy

As expected, preschoolers' immediate JOLs were extremely overconfident. Twenty-four of the 29 children said they would remember every animal's name, but none recalled every name. Mean final recall in the immediate condition was only .41, compared to mean predicted recall of .87. There was no evidence that the preschoolers could make accurate immediate JOLs.

Preschoolers' delayed JOLs were also overconfident, but less so than their immediate JOLs. Fourteen of 29 said they would recall every animal's name; only one did so. Mean predicted recall in the delayed condition was .74, whereas mean recall itself was only .33, t(28) = 7.65, p < .001. Predicted recall was not correlated with actual recall in this condition, r(27) = .11, p > .10.

Each child who made uniformly positive delayed JOLs (i.e., said they would remember every animal's name) also made uniformly positive immediate JOLs. Ten other children predicted recalling every name only when making immediate judgments and none showed this pattern when making delayed judgments, p = .001. Thus, preschoolers were more likely to show awareness of failure to learn when making a delayed judgment than when making an immediate judgment. However, a minority of the preschoolers were responsible for this effect.

Were the delayed JOLs of those preschoolers who made differential judgments (i.e., predicted recalling some, but not other names) accurate? That is, were they more likely to recall a name at the end of the session if they had predicted they would recall it than if they had predicted they would not recall it? Only 13 children made differential delayed JOLs (two made uniformly negative ones.) A Goodman-Kruskal gamma correlation (*G*) was computed between each of these children's delayed JOLs and final recall (Nelson, 1984). *G* ranges from -1.0 to +1.0 with values closer to +1.0 meaning more accurate judgments. It is based on the difference between the proportions of concordant and discordant pairs while ignoring ties. If an individual gives a higher JOL to an item that they recall than to an item that they fail to recall, it is considered a concordant pair. If the individual gives a higher JOL to an item that they fail to recall than to an item that they correctly recall, it is a discordant pair.

G could not be computed for three of the 13 children because their data included only tied pairs. Of the remaining children, six had gammas of 1.0, but four had gammas of -1.0. This pattern is not significantly different from chance. There was no evidence that the minority of preschoolers who did make differential judgments had made them accurately.

In contrast to the preschoolers, the 3^{rd} graders showed the delayed JOL effect that has been found in previous research with older children and adults. The accuracy of their delayed JOLs (mean G = .93) was significantly higher than the accuracy of their immediate JOLs (mean G = .31), t(16) = 2.77, p < 05. This result is very similar to what Nelson and Dunlosky (1991) found in undergraduates (where mean Gs were .90 and .38 for delayed and immediate JOLs, respectively). These results are also consistent with the delayed JOL effect reported by Schneider et al. (2000) for 2^{nd} and 4^{th} graders.

Also, in the current study, the 3rd graders were less likely to show response biases than preschoolers. Twenty-two of the 39 3rd graders said they would recall every animal's name at final recall in the immediate condition, compared to 24 of 29 preschoolers, $\chi^2 = 5.28$, p < .05. Three 3rd graders said they would recall every animal's name in the delayed condition, compared to 14 of 29 preschoolers, $\chi^2 = 14.61$, p < .001. Like the preschoolers, some 3rd graders showed this response bias only in immediate judgments (N =18) and none showed it only in delayed judgments, p < .001. Thus, in both age groups, children were more likely to show awareness of failure to learn when making a delayed judgment than when making an immediate judgment. However, delayed JOLs were more accurate in the older than in the younger group.



Figure 1. Mean final recall as a function of pre-JOL recall in Experiment 1.

2.3.2 Judgment Magnitude and Final Recall as a Function of Pre-JOL Recall

It is assumed that adults attempt to recall an item in order to make a JOL about it (Nelson, Narens, & Dunlosky, 2004). In the current experiment, the children were asked to explicitly recall each animal's name immediately before making a JOL about it. One possible explanation for the low accuracy of delayed JOLs in preschoolers is that their judgments were not influenced by whether this attempt to recall the name succeeded or failed. A second possible explanation is that they were influenced by pre-JOL recall, but it was just not a diagnostic cue for final recall in this group. These proposals are only relevant in the delayed condition because children always correctly recalled the animals' names during pre-JOL recall in the immediate condition.

For delayed JOLs, gamma between pre-JOL recall and final recall could be computed for 21 of 29 preschoolers. It could not be computed for six who did not prerecall any names correctly or for two others who had all tied pairs. Of the 21 children, pre-JOL recall was perfectly correlated with final recall (mean G = 1.00), which provides evidence that pre-JOL recall would have been a highly diagnostic cue even for the preschoolers who predicted recalling every name (N = 12 of the 21) or who made differentiated, but inaccurate JOLs (N = 3 of the 21). The mean final recall values as a function of pre-JOL recall are presented in Figure 1. Overall, when pre-JOL recall was correct, mean final recall was .96. When pre-JOL recall was incorrect, mean final recall was only .11.

Preschoolers may not have understood that pre-JOL recall was a good predictor of final recall, more specifically, that incorrect pre-JOL recall of an item was a good indicator that final recall of the item would also be incorrect. Most tended to ignore this information when making delayed JOLs (see Figure 2). When the children failed to recall a name or recalled an incorrect name, their likelihood of judging that they would recall the correct name later was .69, which not only exceeded the actual likelihood of their recalling the name later (.11), but was also significantly greater than .50, t(28) = 2.65, p < .05. Thus, when they did not recall the correct name at pre-JOL recall, the preschoolers as a group were more likely to say they would recall it later than not. Keep in mind, however, that the .69 figure is essentially the average of 1.00 from the 14 children who predicted recalling every animal's name and .39 from the remaining 15 children. The latter children's likelihood of saying they would recall such names later



Figure 2. Mean JOL magnitude as a function of pre-JOL recall in Experiment 1.

did not differ from chance, t(14) = 1.29, p > .10. Of the 13 preschoolers who did not show a response bias in their delayed JOLs, some may have been influenced by pre-JOL recall. Immediate pre-JOL recall exceeded delayed pre-JOL recall in all 13, and ten mirrored this pattern in their recall predictions (i.e., immediate > delayed) and only two showed the opposite pattern (immediate < delayed). This difference (10 vs. 2) was significant, $\chi^2(1, N = 12) = 5.33$, p < .05.

Was pre-JOL recall of delayed items correlated with JOLs for delayed items among the 13 preschoolers who did not always say yes? Three did not pre-recall a single name, and so *G* could not be computed for them. Of the ten who pre-recalled at least one name, eight predicted perfect final recall of the name(s) that they pre-recalled (M = 1.00) and predicted less than perfect final recall of the name(s) that they failed to pre-recall (M = .36), whereas two showed the opposite pattern (Ms = 0.00 and .77, respectively). The chance probability that eight out of ten would show a pattern consistent with pre-JOL recall just missed the one-tailed cutoff, p = .055. Statistical power is a problem here because the sample is small. Most likely, the delayed JOLs of six-to-eight preschoolers were influenced by pre-JOL.

Although some preschoolers may have used pre-JOL recall as a cue to future recall, the majority showed no evidence that they understood that failure to recall an animal's name meant that they were likely to fail to recall it later. This result is consistent with Lipko et al's (2009) findings that most preschool-age children's predictions about their memory are highly overconfident and that they do not understand that past memory performance tends to be a reliable indicator of future memory performance.

For delayed JOLs, the mean gamma between pre-JOL recall and final recall could be computed for all of the 39 3rd graders. As with preschoolers, pre-JOL recall and final recall were significantly correlated in the delayed condition (mean G = .99). When pre-JOL recall was correct, mean final recall was .95. When pre-JOL recall was incorrect, mean final recall was only .03. The 3rd graders also showed a better understanding of the importance of pre-JOL recall than preschoolers. Their delayed JOLs were significantly lower when pre-JOL recall was incorrect (M = .38) than when it was correct (M = .94), t(38) = 13.64, p < .001. This result provides evidence that the 3rd graders were using pre-JOL recall as a cue when making delayed JOLs.

The 3rd graders' reliance on pre-JOL recall as a cue provides an explanation for why their JOLs were higher in the immediate condition (M = .88) than in the delayed

condition (M = .66), t (38) = 8.33, p < .001. When a 3rd grader correctly retrieved the name during pre-JOL recall, his or her likelihood of predicting final recall of a name was quite high (M = .94 in delayed and .88 in immediate). However, the child's likelihood of predicting later recall of a name was lower if he or she did not correctly recall it during pre-JOL recall (M = .38), which only occurred in the delayed condition. Mean final recall of items that were not correctly recalled at pre-JOL recall was .03, which is lower than the predicted recall of those items, t(38) = 8.11, p < .001.

The 3rd graders' reliance on pre-JOL recall as a cue also provides an explanation for why their JOLs were less accurate in the immediate condition than in the delayed condition (as already reported, mean G = .31 and .93, respectively). For items that were successfully pre-recalled, predicted recall exceeded final recall in the immediate condition (.88 vs. .60), t(38) = 7.88, p < .001, but not in the delayed condition (.94 vs. .95). Note that although they overestimated the reliability of correct immediate pre-JOL as a predictor of final recall, they did show some sensitivity to the fact that it was not as reliable a predictor as correct delayed pre-JOL recall (i.e., the difference between .88 and .94 was significant, t(38) = 3.00, p < .01.). It is possible that they had a more accurate understanding of the differential reliability of pre-JOL recall, but used pre-JOL recall to guide their immediate JOLs because they could not identify a more reliable cue.

2.3.3 Judgment Magnitude and Final Recall as a Function of Item Relatedness

Another factor that has been shown to influence JOLs is item relatedness. The mean final recall values as a function of item relatedness are presented in Figure 3. Only



Figure 3. Mean final recall as a function of item relatedness in Experiment 1.

data from the delayed condition are discussed because children in the immediate condition obviously did not discriminate between related and unrelated items given their strong bias to say they would recall all items. Preschoolers' final recall of related items (M = .52) was significantly better than their final recall of unrelated items (M = .15), t(28)= 5.50, p < .001. Unlike adults and older children, preschoolers were rather insensitive to the influence of item relatedness on memory (see Figure 4). Overall, their JOLs for related items (M = .78) were not significantly different from their JOLs for unrelated items (M = .70), t(28) = 1.57, p = .13. Even the preschoolers who distinguished between items in their delayed JOLs did not show a tendency to use item relatedness as a cue



Figure 4. Mean JOL magnitude as a function of item relatedness in Experiment 1.

when making their predictions. Although they were more sensitive to this cue than the preschoolers who predicted they would recall every item, their JOLs for related items (M = .67) were not significantly different from their JOLs for unrelated items (M = .49), t(12) = 1.62, p = .13. Power is a problem for this analysis because only 13 preschoolers made differential delayed JOLs.

In the 3rd graders, item relatedness influenced the level of final recall, M = .73 for related items compared to .22 for unrelated ones, t(38) = 11.76, p < .001. Unlike the preschoolers, they gave significantly higher delayed JOLs to related items (M = .88) than to unrelated items (M = .43), t(38) = 11.10, p < .001. This result is consistent with the

proposal that 3rd graders use item relatedness, or some factor associated with it (e.g., ease of learning) (Koriat, Ackerman, Lockl, & Schneider, 2009) when making delayed JOLs.

The use of pre-JOL recall as a cue for making JOLs did not completely account for the effect of item relatedness on delayed JOLs, however. For items that were correctly recalled at pre-JOL recall, JOLs for related items (M = .94) did not differ from JOLs for unrelated items (M = .95), t(26) < 1. For items that were not correctly recalled at pre-JOL recall, JOLs were higher for related items (M = .70) than unrelated items (M =.26), t(27) = 4.85, p < .001. Relatedness seemed to influence JOLs for items that were not correctly recalled at pre-JOL recall. One possible explanation for this effect is that children were more likely to make commission errors at pre-JOL recall for animals with related names and omission errors for animals with unrelated names. In other words, it may have been easier for children to generate an incorrect related name at pre-JOL recall than an incorrect unrelated name.

CHAPTER 3

EXPERIMENT 2

3.1 INTRODUCTION

The results of Experiment 1 suggest that most preschoolers cannot make accurate delayed JOLs. More specifically, they do not seem to consider past performance on a memory task when making predictions about future performance on that task. However, an alternative explanation for these results is possible. About half of the preschoolers in the first experiment showed a "yes" response bias. That is, about half said that they would recall all of the animals' names at final recall. If the procedure for eliciting JOLs were changed so as to reduce the influence of such a response bias, preschoolers might make more accurate delayed JOLs than in Experiment 1. The main goal of Experiment 2 was to examine this possibility.

In Experiment 2, a two-item forced choice paradigm was added to examine the accuracy of preschoolers' delayed metacognitive judgments. Asking the children to make a choice between two items should reduce the influence of a "yes" response bias because the procedure does not allow the child to say "yes" to both items. The general procedure was the same as in Experiment 1, except for a few changes. First, Experiment 2 included only a delayed group because its primary goal was to explore whether preschoolers can make accurate JOLs under any condition. Second, the children were

taught names for ten stuffed animals instead of six. The four additional animals were included so that there were a sufficient number of observations in the forced choice task. Third, due to the larger number of names the children had to learn, a second study session was added to insure recall was not on the performance floor. This additional study session included recall and feedback. In this session, the experimenter asked the children to name each of the animals. They were told whether their response was correct or incorrect and then the experimenter said the correct name.

The final and most important change was the addition of the forced choice task. For this task, the children were shown two animals at a time. As each pair was presented, the experimenter asked which animal's name they would be more likely to recall later. The key trials were ones in which an animal whose name was correctly recalled was paired with either an animal whose name was incorrectly recalled or an animal for whom no name was recalled. A "yes" response bias would not affect the children's choices in this task because they had to choose one of the animals. Therefore, in the forced choice test, preschoolers might show a greater ability to distinguish between well-learned items and less well-learned items when making predictions about future performance on a recall task.

3.2 METHOD

3.2.1 Participants

Twenty-five preschoolers (M age = 4-10; range = 3-11 to 5-10) were recruited from middle class, suburban schools in Northeast Ohio. There were 12 girls and 13 boys.

None of the children participated in the first experiment. Each child received a small prize for participating.

3.2.2 Materials and procedure

Children were tested individually at their preschools. They were taught names for ten different animals (see Appendix A). As in Experiment 1, half of the names were associated with the animal and half were not. All animal names were learned in a single block. Children only made delayed JOLs.

Each animal was presented to the child one at a time. The experimenter asked the child to say what type of animal it was. Next, the experimenter said the animals name and explained the association if the name was related to the animal. The child was asked to repeat each name to insure that he or she heard it correctly. After all animals were presented, the children completed a second study session, which included recall and feedback. The experimenter asked the child to name each animal. If the name was correct, the experimenter told the child that he was right and then repeated the name (e.g., "That's right! The frog's name is Hoppy!"). If the name was incorrect or if the child did not recall a name, the experimenter simply reminded the child of the correct name (e.g., "The pig's name is Brian."). After the two study sessions there was a two minute delay, which was filled with an age appropriate distractor task (i.e., coloring). Procedures for pre-JOL recall and the JOLs themselves were the same as in Experiment 1. For each JOL, the child was asked to recall the animal's name and whether or not he would remember it later.

After all JOLs were made, the child answered five forced choice questions. Two animals were shown to the child at a time. The decisions about which animals were paired together were based on pre-JOL recall. The experimenter began by pairing animals whose names were correctly recalled with animals for whom no name was recalled. Next, animals whose names were correctly recalled were paired with animals whose names were incorrectly recalled. The remaining animals were randomly paired together. Both animals in a pair were always presented at the same time. Also, the placement of the animals (i.e., whether the animal whose name was correctly recalled was to the child's left or right) was counterbalanced. The experimenter asked the child which animal's name he would be more likely to recall later. Lastly, the child completed a final recall test. The animals were presented in the same order for the two study sessions, the JOLs, and final recall.

3.3. RESULTS AND DISCUSSION

3.3.1 Judgment Accuracy

Fewer children had response biases in the current experiment than in Experiment 1. Only 5 of the 25 children said they would remember all of the animal's names, compared to 14 of 29 in Experiment 1, $\chi^2 = 4.71$, p < .05.

Gamma between JOLs and final recall could be computed for 15 of the 25 children. (In addition to those showing a positive response bias, three showed a negative response bias, and one other child did not generate any names in final recall.) In contrast

to the bimodal distribution of *G* scores in Experiment 1 (six 1.0 and four -1.0), G scores in the current experiment were negatively skewed (eight 1.0, three other positive, three other negative, and one -1.0). Also in contrast to Experiment 1, mean *G* (.55) was significantly greater than zero, *t* (14) = 3.13, *p* < .01. The number of children whose G scores were positive (11) was approximately the same as the number who were either response-biased or made inaccurate JOLs (13), whereas in Experiment 1, there were fewer who fit the first description (6) than the second (20), $\chi^2(1, N = 26) = 7.54$, *p* < .01. This difference between experiments approached the conventional level of significance, $\chi^2(1, N = 50) = 2.88$, *p* = .09.

3.3.2 Judgment Magnitude and Final Recall as a Function of Pre-JOL Recall

Gamma between pre-JOL recall and final recall could be computed for only 19 of the 25 children because six did not pre-recall any names. In the group of 19, pre-JOL recall and final recall were perfectly correlated (mean G = 1.00). When pre-JOL recall was correct, mean final recall was .89, and when it was incorrect, mean final recall was only .04. Thus, as in Experiment 1, pre-JOL recall was a highly diagnostic cue for final recall even for children who had predicted recalling every name (N = 5 of the 19) or whose JOLs were differentiated, but inaccurate (N = 3 of the 19).

There was evidence that the children who made differential JOLs (those who predicted recalling some names, but not others) were influenced by pre-JOL recall. Of these 16 preschoolers, gamma between pre-JOL recall and JOL could be computed for 14. Two did not pre-recall any names. Eleven of the 14 had perfect *G* scores of 1.0, and

only two had negative scores. The proportion who had positive G scores in the current experiment was comparable to that in Experiment 1 (8 of 10). Also similar to Experiment 1, each child who had a positive G score predicted recalling a higher proportion of the name(s) that they had pre-recalled correctly (M = .97) than of the name(s) that they had not pre-recalled correctly (M = .35).

For the entire sample of children, mean rate of predicting final recall of names that had not been pre-recalled correctly was .48, which did not differ from chance.

The higher rate found in Experiment 1 (M = .69) was entirely due to the greater proportion of children in Experiment 1 who predicted recalling every animal's name.

3.3.3 Judgment Magnitude and Final Recall as a Function of Item Relatedness

Not surprisingly, final recall of related names (M = .43) was higher than final recall of unrelated names (M = .14), t(24) = 4.21, p < .001. Recall of related names (M = .34) was also higher than recall of unrelated names (M = .06) at the second study session, t(23) < 1, or at pre-JOL recall (M = .46 and .13, respectively), t(24) = 4.39, p < .001. As in Experiment 1, preschoolers' JOLs for related names (M = .63) were not significantly different from their JOLs for unrelated names (M = .54), t(24) = 1.66, p = .11. Even among the 16 preschoolers who did not show a response bias, JOLs for related names (M = .61) were not significantly different from unrelated names (M = .61) were not significantly different from JOLs for unrelated names (M = .48), t(15) = 1.55, p = .14.

Due to small sample size, power is a problem with this last analysis. It was also a problem for this same kind of analysis in Experiment 1. If the results of the two

experiments are combined in a meta-analysis by the Stouffer method (Rosenthal, 1991), the preschoolers who made differential delayed JOLs (N = 29) predicted recall of more related names (M = .64) than unrelated items (M = .48), z = 2.11, p < .05. It is important to keep in mind, however, that the other 25 preschoolers in the two experiments made undifferentiated JOLs (usually predicting recall of every name), and so did not show sensitivity to item relatedness.

3.3.4 Forced choice questions

The forced choice test was included to explore whether those preschoolers who showed a response bias when making JOLs nevertheless realized that they would be more likely to recall a name that had been recalled correctly at the time of judgment (correct pre-JOL recall) than a name that had not. Forced choice performance could be evaluated for only 19 preschoolers because six did not pre-recall any names correctly. The tendency to select a toy whose name had been pre-recalled correctly over a toy whose name had been pre-recalled incorrectly was strong, M = 0.83 (SD = .19), t(18) = 7.48, p <.001. This tendency was especially strong when the incorrect choice was a toy for which the child had failed to generate a name at pre-recall (an omission error), M = .91 (N who had at least one trial of this type = 16), as opposed to a toy the child had named incorrectly at pre-recall (a commission error), M = .71 (N = 10).

Did the children who showed a response bias when making JOLs show sensitivity to a name's pre-recall status in their forced-choice responses? Forced choice performance of only five of the response-biased children could be evaluated because the other four did not pre-recall any names correctly. The five children's mean rate of selecting the toy whose name had been pre-recalled correctly was .69 (SD = .28), which was not significantly greater than chance (.50), t(4) = 1.72, p = .16. Power is clearly a problem for this analysis. Only three of the response-biased children received trials in which a correctly pre-recalled item was pitted against an omission error. All three favored the correctly pre-recalled item on these trials (M = .89, SD = .19), but this sample is too small to evaluate with inferential statistics.

Two other children's JOLs were negatively correlated with pre-JOL recall. Thus, there were seven children whose JOLs were not positively influenced by the pre-recall status of the names and who received at least one forced choice test trial. This group's rate of selecting a toy whose name had been pre-recalled correctly over a toy whose name had been pre-recalled correctly over a toy whose name had been pre-recalled chance, t(6) = 2.76, p < .05. This result supports the conclusion that preschoolers show greater sensitivity to an item's pre-recall status when making forced-choice JOLs than when making yes-no JOLs.

CHAPTER 4

EXPERIMENT 3

4.1 INTRODUCTION

In Experiment 1, most preschoolers were extremely overconfident, even when making their JOLs after a delay. Fourteen of the 29 participants even said that they would recall all of the trained names. Among the 13 preschoolers who made differential JOLs, there was no evidence that these judgments were accurate, although there was suggestive evidence that they were influenced by whether or not they recalled a name at the time of judgment. In Experiment 2, fewer preschoolers' delayed JOLs were extremely overconfident. Only 5 of the 25 said they would recall every trained name. Among the remaining preschoolers, there was evidence that their JOLs were accurate (mean gamma = .55) and influenced by pre-JOL recall. It is unclear why more preschoolers showed an ability to make accurate delayed JOLs in Experiment 2 than in Experiment 1.

One possible explanation for this discrepancy has to do with a methodological difference between the two experiments. Experiment 1 included a single study session in which the experimenter said the animal's name and the child repeated it. In Experiment 2, a second study session was added to improve final recall. The children were asked to recall the animal's name and then were given feedback concerning the correctness of

their answer. The experimenter also repeated the animal's correct name. Although the extra study session was only intended to improve final recall, it may have also promoted more accurate metacognitive monitoring.

Previous research has shown that practice (i.e., repeatedly studying the same list of items) can lead to improved accuracy of JOLs in both adults (e.g., Koriat, 1997) and children (Koriat and Shitzer-Reichert, 2002). Koriat and Shitzer-Reichert (2002) conducted two experiments examining the influence of practice on JOL accuracy in children. As described above, 2nd and 4th graders learned a list of word pairs. The children studied and made a JOL for each word pair. The study session was followed by a recall test, in which the children had to say the target word when presented with the stimulus. To examine the effects of practice, this study-test cycle was repeated three more times. In their first experiment, they also examined the effects of feedback on JOLs. The procedure was the same for all children, except half of the children heard a sound when they recalled an item incorrectly on the recall tests. In the second experiment, they examined the effects of timing and the type of cue used for the JOLs. Half of the JOLs were made immediately after studying a word pair and the other half were made after all word pairs had been studied. Also, when making JOLs, half of the children were shown the complete stimulus-response pair. The remaining children were only shown the stimulus.

In Experiment 1, feedback had little effect and was not discussed. However, practice led to an increase in both JOLs and recall. More importantly, the accuracy of JOLs also increased with practice. Practice also led to more accurate JOLs in Experiment 2, except in the delayed, stimulus alone-condition. The accuracy of JOLs in the delayed, stimulus-alone condition was high at the first study-test cycle and remained high throughout the experiment. During the third study-test cycle, the accuracy of JOLs in the other conditions reached the same level of accuracy as the JOLs in the delayed, stimulus-alone condition. Koriat and Shitzer-Reichert (2002) concluded that practice and delay provide two different ways to improve the accuracy of JOLs.

Although the children did not have extra practice making JOLs, it is possible that the additional study session with recall can explain the improved accuracy of delayed JOLs in Experiment 2 of the current study. Koriat and Shitzer-Reichert's (2002) results showed that additional study-test blocks can improve the accuracy of JOLs. Specifically, the additional recall tests may have been responsible for this improvement. As shown in the current experiments, success on a recall test is a good predictor of future recall. Multiple retrieval attempts may make this cue more salient. In Experiment 2 of the current study, failing to recall an item twice (i.e., at the second study session and pre-JOL recall) may have helped children realize that it is unlikely they would recall it in the future. One failed attempt may not have been enough to make retrieval success a salient cue for preschoolers.

In Experiment 1, failing to recall the trained names may have helped some preschoolers to make accurate delayed JOLs. There was a trend for the relation between delayed JOLs and pre-JOL recall to be stronger when these were made after having made immediate JOLs (and been tested for their accuracy), but not before. The gamma correlation between delayed JOLs and pre-JOL recall was +1.0 for six children and -1.0 for one child when delayed JOLs were made second, p = .06, but +1.0 for two and -1.0 for one when delayed JOLs were made first, p > .10. The experience of failing to generate most of the immediate items in the final recall test, despite having predicted being able to recall of all of them, may have led preschoolers to doubt, when making delayed JOLs, that they would ultimately recall the names they were failing to recall at the time of the judgment.

The goal of Experiment 3 was to explore whether an additional study session with recall can account for the improved accuracy of delayed JOLs in Experiment 2. Half of the participants completed a single study session with no recall as in Experiment 1. The other half completed two study sessions. The additional study session included recall and feedback as in Experiment 2. It was predicted that children with an additional recall attempt and feedback would make more accurate JOLs than the children with a single study session. Preschoolers may be able to accurately monitor their memory if their judgments are made after a delay and they have more than one retrieval attempt.

4.2 METHOD

4.2.1 Participants

Sixty preschoolers (M age = 5-1; range = 3-10 to 6-6) were recruited from middle class, suburban schools in Northeast Ohio. There were 33 girls and 27 boys. None of the children participated in either of the first two experiments. Each child received a small prize (e.g., stuffed animal or puzzle) for participating.

4.2.2 Materials and procedure

The children were tested individually at their preschools. All children were taught names for eight animals (see Appendix A). As in the previous experiments, half of the animal's names were related to the animal in some way and half were unrelated. To examine the effects of having an extra study session with recall, the children were split into two groups. Twenty-nine children received only one study session as in Experiment 1 (i.e., no practice group). Thirty-one children received a second study session including recall and feedback as in Experiment 2 (i.e., practice group).

The materials and procedures used in the no practice group were identical to those in the delayed condition of Experiment 1, except for two changes. First, after all of the JOLs were made, the children completed the forced choice task used in Experiment 2. Second, as described above, the children were taught names for eight animals. Training six items, as in Experiment 1, was judged to risk the possibility that some children would not get enough trials in the forced choice task (i.e, ones that pitted a toy that had been pre-recalled correctly against a toy that had been pre-recalled incorrectly). Training ten items, as in Experiment 2, was judged to be too many for the children who received only one study session. In Experiment 2, mean recall after one study session was rather low (.20).

Children in the no practice group completed a single study session, in which the experimenter presented each animal individually. They were asked to say the type of animal and then the experimenter stated the animals' name. As in the previous experiments, if the name was related to the animal, the experimenter explicitly pointed

out the association (e.g., "An octopus has a lot of legs. So his name is Legs.") To insure that the children heard the names correctly and were paying attention, they were asked to repeat the animal's names. The study session was followed by a two-minute delay that was filled with an age appropriate task (i.e., coloring). Next, children attempted pre-JOL recall and made a JOL for each animal. Next, the children were asked four forced choice questions. The procedures of the forced choice task were the same as those used in Experiment 2. Immediately following the forced choice task, the children completed a final recall task.

The materials and procedures used in the practice group were identical to those used in Experiment 2, except the children only learned the names of eight animals. The first study session was the same as in the no practice group just described. Immediately after the first study session, children in this group completed a second study session which included recall and feedback. Each animal was presented individually for a second time. The children were asked to recall the animal's names and then the experimenter provided feedback. The children were told whether their answer was correct or incorrect and then the experimenter repeated the correct name for the animal. The remaining procedures were identical to those used in the no practice group described above, which included a two-minute delay, pre-JOL recall, JOLs, the forced choice task, and final recall.

4.3. RESULTS AND DISCUSSION

4.3.1 Judgment Accuracy

The hypothesis that fewer children would show a response bias in the practice group than in the no practice group was not supported. Nine of the 31 children in the practice group (29%) and ten of the 29 children in the no practice group (34%) showed a response bias. In every case but one, the bias was to predict recalling all eight animal names. The proportions of children showing response biases were closer to the proportion in Experiment 2 (36%), in which children received additional practice, than in Experiment 1 (55%), in which children did not receive additional practice. Overall, the proportion of children showing a response bias in the current experiment (32%) was significantly less than the proportion who did so in Experiment 1, $\chi^2(1, N = 89) = 4.53$, *p* < .05.

Gamma between JOLs and final recall could be computed for 20 children in the practice group and 16 children in the no practice group. (In addition to the children with response biases, five others generated no correct names in final recall.) As in Experiment 2, the distributions were highly negatively skewed, with the majority of children obtaining perfect G scores (i.e., 1.0). In the practice group, 17 children had perfect G scores and the remaining three had positive ones (overall M = .94). In the no practice group, 13 had perfect G scores, one had a positive score, and two had negative scores (overall M = .73). Although the difference between groups was in the expected direction, it was not significant. Although there were relatively more children with accurate JOLs

(positive G scores) compared to inaccurate ones (response biases or negative G scores) in the practice group (20 vs. 9) than in the no practice group (14 vs. 12), this difference was also not significant, $\chi^2(1, N = 55) = 1.33$, p > .10.

Gamma is a measure of relative accuracy and does not provide information about overestimation or underestimation (Dunlosky & Metcalfe, 2009; Nelson, 1996). Thus, if misses (predict failure to recall the correct name, but actually recall it) were rare, two groups could have comparable gammas even if they differed in false alarms (predict recall of correct name, but do not actually recall it). In the current experiment, misses were rare in both the practice and no practice groups. Of the children for whom G could be computed (who will be the groups compared in the remaining analyses in this paragraph), only six of 20 in the practice group and three of 16 in the no practice group produced a miss, p > .10. Mean rates of misses were .11 and .11 in each group. In contrast, 15 of 20 children in the practice group and 14 of 16 children in the no practice group produced a false alarm, p > .10. Rates of false alarms were lower in the practice group (M = .29, SD = .22) than in the no practice group (M = .48, SD = .26), t(34) = 2.40, p < .05.

This last result supports the hypothesis that experience making delayed JOLs and receiving feedback about their accuracy helps preschool-age children to improve the accuracy of their JOLs. Their tendency to be overconfident about items that they ultimately failed to recall correctly decreased relative to children who did not get such experience. Their tendency to be underconfident about items that they ultimately

succeeded in recalling, which was a rather minimal tendency to begin with, was not affected.

4.3.2 Judgment Magnitude and Final Recall as a Function of Pre-JOL Recall

Gamma could be computed between pre-JOL recall and final recall for 28 children in the practice group and 26 children in the no practice group. As in the previous experiments, these measures were highly correlated (mean G = .92 and 1.00 in the practice and no practice groups, respectively), indicating that pre-JOL recall was a diagnostic cue that the children should have used when making JOLs. When pre-JOL recall was correct, mean final recall was .86 in the no practice group and .88 in the practice group. When pre-JOL recall was incorrect, mean final recall was only .02 in the no practice group and .11 in the practice group.

Due to the extra study session with recall and feedback, it was hypothesized that children in the practice condition would be more likely than children in the no practice condition to use pre-JOL recall as a cue when making their JOLs. Gamma between pre-JOL recall and JOLs could be computed for 18 children in the no practice group and 21 children in the practice group. The mean gammas were .75 and .88 in the no practice group and the practice group, respectively. As in Experiment 2, these distributions were highly skewed. Fifteen of the 18 children in the no practice group had *G* scores of 1.0, and two had negative scores. In the practice group, 16 of the 21 children had perfect *G* score, and none had negative scores. These distributions were not significantly different, $\chi^2(1) = .303$.

Even though gammas between pre-JOL recall and JOLs were high in both conditions, it was possible that they were high primarily due to the rarity of misses (where a name is pre-recalled correctly, but the child predicts failure to get it correct in final recall). A child needs to produce at least one miss to receive a gamma of less than 1.0. When they correctly recalled an animal's name at pre-JOL recall, mean JOLs in the practice and no practice groups did not differ (M = .90 and .94, respectively), t(56) < 1. However, when they failed to recall an animal's name at pre-JOL recall, mean JOLs in the practice groups were significantly lower than mean JOLs in the no practice group (M = .49 and .69, respectively), t(57) = 2.19, p < .05. In other words, when they did not correctly recall an item at pre-JOL recall, the children in the practice condition were less likely to predict that they would recall it later.

4.3.3 Judgment Magnitude and Final Recall as a Function of Item Relatedness

Overall, final recall of related items (M = .60) was higher than final recall of unrelated items (M = .21), t(59) = 10.66, p < .001. In contrast to the previous experiments, preschoolers gave higher JOLs to related items (M = .77) than to unrelated items (M =.66), t(58) = 2.12, p < .05. The remaining analyses in this section will only include children who did not have a response bias (N = 41) because those with a response bias did not differentiate between items in their judgments. A 2 (Condition: practice vs. no practice) X 2 (Item type: related vs. unrelated) mixed analysis of variance of judgment magnitude yielded a significant main effect of item type, F(1,39) = 27.33, p < .001. Among these children, judgments were higher for related items (M = .77) than for unrelated items (M = .45). There was no evidence that the practice group was more sensitive to item relatedness than the no practice group. That is, the group by item type interaction was not significant, F(1,39) = 1.57, p > .20.

A 2 (Group) x 2 (Item type) mixed analysis of variance was also conducted on final recall with children who did not have a response bias. There was a significant main effect of item type, F (1, 39) = 69.98, p < .001. The children recalled more related items (M = .57) than unrelated items (M = .17). Practice did not have a significant effect on the final recall of items and did not interact with item relatedness.

4.3.4 Forced choice questions

Importantly, as in Experiment 2, the forced choice task provided a way to explore whether children who showed a response bias when making JOLs knew that they would be more likely to recall a name that had been correctly recalled at pre-JOL recall than a name that had not been recalled. Performance on this task could be evaluated for 56 children. One child failed to complete the task, one child correctly pre-recalled all of the names, and two children did not correctly recall any of the names. Overall, children performed well on the forced choice task. They were able to distinguish between names that were recalled at pre-JOL recall and ones that were not. When an animal whose name was correctly recalled at pre-JOL recall was paired with an animal whose name was not successfully recalled, the mean rate of choosing the animal whose name had been correctly recalled was .87, which was significantly greater than chance, t(55) = 11.87, p < .001. The tendency to choose the animal whose name was correctly pre-recalled was

stronger when the child failed to generate a name at pre-recall (omission error), M = .91 (*N* who had at least one trial of this type = 51), than when the child generated an incorrect name (commission error), M = .67 (N = 21).

Even children who showed a response bias when making JOLs were sensitive to whether they recalled a name at pre-JOL recall when completing the forced choice task. Forced choice performance could be evaluated for seventeen of the response-biased children. One child failed to complete the forced choice task and another child correctly pre-recalled all of the names. The response-biased children's mean rate of selecting the animal whose name had been correctly pre-recalled was .79, which was significantly greater than chance, t(16) = 3.93, p = .001. Not surprisingly, the tendency to choose the animal whose name was correctly recalled at pre-JOL recall was stronger when the child failed to generate a name at pre-recall (omission error), M = .89 (N who had at least one trial of this type = 14), than when the children generated an incorrect name at pre-recall (commission error), M = .60 (N = 9). Performance on the forced-choice task shows that even response-biased children can distinguish between items that were successfully recalled and items there were not successfully recalled at pre-JOL recall, and they can use this information to make comparative judgments about which items will be recalled in the future.

CHAPTER 5

GENERAL DISCUSSION

The main goal of the current experiments was to explore whether preschoolers can make accurate item-by-item JOLs. In Experiment 1, when judgments were made immediately after a name had been taught, nearly every child predicted subsequent recall of every name. When judgments were made after a delay, however, only about half of the children showed this response tendency. Despite this reduction in overconfidence, the delayed JOLs of those who predicted one or more recall failures were still overconfident and were not correlated with final recall. The vast majority showed no evidence that they understood that failure to recall an animal's name at pre-JOL recall meant that is was unlikely that they would recall the name later. Consistent with previous research, 3rd graders were able to make accurate JOLs in the delayed condition. They used pre-JOL recall and item relatedness to help make accurate judgments.

Experiments 2 and 3 focused only on delayed JOLs in preschoolers. In Experiment 2, 80% of the children predicted at least one recall failure, and they were less likely to recall these items than ones for which they had made positive JOLs. In the forced choice task, even children who made undifferentiated or inaccurate delayed JOLs were more likely to say they would recall a name that had been correctly recalled in pre-JOL recall than a name that had not been correctly recalled. Thus, the results of Experiment 2 supported the hypothesis that children would use pre-JOL recall as a cue when making delayed JOLs. There was no support for the hypothesis that item relatedness would also influence these judgments.

The goal of Experiment 3 was to explore whether an additional practice trial with recall and feedback was responsible for the improved accuracy of JOLs in Experiment 2. The performance of a group with an additional practice trial was compared to the performance of a group without an additional practice trial. The tendency to predict perfect recall was less prevalent than in Experiment 1, but did not differ by group. Also, the JOLs of those who predicted at least one recall failure were more accurate than in Experiment 1, but the average gamma correlation between JOLs and final recall did not vary by group. When a name was correctly recalled in the pre-JOL phase, nearly every child in each group almost always predicted that they would recall the name later. However, children in the practice group were less likely to say they would recall an item that was not correctly recalled at pre-JOL recall than were children in the no practice group. The additional practice trial increased children's sensitivity to the reliability of recall failure as a predictor of a similar recall failure in the future.

As in the previous experiments, final recall of related items was higher than final recall of unrelated items. In contrast to previous experiments, JOLs in Experiment 3 were higher for related items than for unrelated items. As in Experiment 2, children did very well on the forced choice task. Even those whose yes-no JOLs had been undifferentiated or inaccurate were more likely to say they would recall a name that they

47

had gotten correct during pre-JOL recall than a name that they could not produce during pre-JOL recall..

Overall, the number of children with accurate JOLs compared to inaccurate ones was greater in Experiment 3 (34 vs. 21) than in Experiment 1 (6 vs. 20), $\chi^2(1, N = 81) =$ 10.60, p < .005, but not in Experiment 2 (11 vs. 13), $\chi^2(1, N = 79) = 1.74$, p > .10. A comparison of just the practice condition of Experiment 3 to Experiment 2 approached the conventional level of significance, $\chi^2(1, N = 53) = 2.89$, p = .09.

It is not clear why children in Experiment 1 did worse than children in Experiments 2 and 3. One possibility is that it may have helped to get four extra training items in Experiment 2 and two extra training items in Experiment 3. These trials could have provided an additional opportunity for children to realize that that they could not recall most of the items in pre-judgment recall and that it would be difficult to remember all of the names, especially the unrelated ones.

Another possibility is that age or gender differences can account for the poorer performance in Experiment 1. The delayed judgments of children in the no practice group of Experiment 3 are directly comparable to those of the children in Experiment 1 who made a block of delayed judgments before making immediate judgments. Neither group had any practice making JOLs. The mean age of children in the no practice group of Experiment 3 was 61 months, but the mean age of children in the delayed first group in Experiment 1 was only 57 months. Also, the ratio of boys to girls was different between the two experiments. There was an equal distribution of girls and boys in the no practice group of Experiment 3 (15 boys, 14 girls), but there were more girls than boys in the delayed first group of Experiment 1 (5 boys, 9 girls). Both of these factors may have contributed to the differences in performance between Experiment 1 and Experiment 3.

Although performance in Experiments 2 and 3 was better than in Experiment 1, it was far from perfect. Even in the practice condition, about 30% of the children still predicted perfect recall. Approximately another 20% predicted recall failure, but their predictions were either uniformly negative or inaccurate (i.e., negatively correlated with performance). In addition, the average rate of false alarms was substantial.

Future research should explore why when preschoolers fail to recall a name, they show such a strong tendency to predict that they will recall it later. Although they may truly believe that they will recall the name, another possibility is that they just like saying "yes" or think that this response is what the experimenter wants from them. To explore this option, children could be asked if they will forget an animal's name later rather than if they will remember the animal's name. If children continue to say "yes" to all of the items, this would suggest that they simply have a hard time inhibiting the "yes" response and do not actually believe they will remember all of the animal's names.

Although many of the preschoolers said they would remember all of the items, the majority in Experiments 2 and 3 made fairly accurate delayed JOLs and showed that they understood that current performance on a memory task predicts future performance. These results are consistent with Schneider et al.'s (2001) finding that Kindergarteners could make accurate item-by-item delayed JOLs. In contrast, Lipko et al. (2009) found that preschoolers were unable to make accurate global judgments. They had children study 10 pictures of familiar objects and then asked them how many out of ten they

would remember later. The children's judgments remained overconfident across five trials, even though they were able to accurately recall how many items they recalled on previous trials.

There are several differences between the paradigm used by Lipko et al. (2009) and the paradigm used in the current experiments that may help explain the difference in results. First, the children in Lipko et al.'s (2009) studies made global judgments. In other words, they had to make judgments about all of the items at once rather than one at a time. As compared to the item-by-item judgments examined in the present research, global judgments may be more susceptible to overconfidence because they require children to consider a greater amount of information. Children have limited working memory capacities, which might make it more difficult for them to make global judgments. Also, the preschoolers in Lipko et al.'s (2009) studies were not asked to attempt pre-judgment recall before making their judgments. Pre-judgment recall might have helped the children realize how many items they could actually recall. However, even if the preschoolers had attempted pre-judgment recall, it might have been too difficult for them to keep track of how many items they were able to recall and then to use that information to make a global judgment. Based on the results of the current experiments, failure to understand that the past predicts the future is not responsible for overconfidence in Lipko et al.'s (2009) paradigm.

It would be interesting to explore whether the results of the current experiments would generalize to other types of information. Specifically, future research should examine whether young children can make accurate JOLs about more naturally acquired information, such as word meanings. Research shows that two- and three-year-olds will often say they know the names for unfamiliar objects, such as an egg slicer, even though they cannot provide a label for those objects (Lipowski & Merriman, in press; Merriman & Marazita, 2004). In contrast, four-year-olds are able to make accurate judgments about whether they know the names for familiar and unfamiliar objects. It may be possible to improve the performance of young preschoolers on this type of task by having them attempt pre-judgment recall. One group of children could simply make judgments about whether they know the names for familiar and unfamiliar objects. A second group could be asked to try to recall an object's name before making a judgment about whether they know the names for familiar and unfamiliar objects. A second group could be asked to try to recall an object. Pre-judgment recall may help young preschoolers realize that they do not know the names for unfamiliar objects.

The benefit of practice found for JOLs in Experiment 3 may also extend to children's awareness of their knowledge of language. A practice group could make judgments about whether they know the names for several objects in a set. Later, the children could be asked to name the objects and then make judgments about whether they know the names for objects from a different set. A control group could just make judgments for objects in the latter set. The experience of being unable to generate names for objects that they had identified as having known names may make them less willing to identify other unfamiliar objects as having known names.

Children might also show a greater awareness of their name knowledge in a forced-choice paradigm like the one used in Experiments 2 and 3, compared to the yes-no question paradigm used by Merriman and colleagues. For example, children could be

asked whether they would be more likely to come up with the correct name for a familiar object (e.g., book) or for an unfamiliar object (e.g., egg slicer). Based on the results of the current experiments, at least some preschool children should understand that their current knowledge state is a good predictor of what they will know in the future. In general, it may be easier for preschool children to make judgments about whether they know naturally-acquired object labels because these labels may be better established in their long term memories than names that they have just been taught for stuffed animals.

Future research can also further examine the role of pre-judgment recall when making JOLs. In the current experiments, children were asked to recall the name of an animal before making a JOL for that item. This should have made pre-JOL recall a very salient cue that children could use when making JOLs. The results suggest that some preschoolers used this information when making their JOLs. However, the current experiments do not provide information about whether preschoolers naturally attempt to recall a piece of target information before they make a metacognitive judgment about it. As discussed above, research shows that adults' JOLs are not affected by whether or not they are explicitly asked to try to recall a target word before making a JOL (Nelson, Narens, & Dunlosky, 2004). This result suggests that adults naturally attempt to recall a piece of information before making a JOL about it. In the current experiment, preschoolers often did not use pre-JOL recall as a cue, even though this information was made salient. Therefore, it is highly unlikely that preschoolers would automatically attempt to recall an item before making a judgment about it. Future research should examine whether young children naturally attempt recall before making JOLs.

Finally, it would be interesting to see if children who can make accurate metacognitive judgments can apply this knowledge to metacognitive control. One of the reasons why metacognitive monitoring is so important is that it provides valuable information that people could possibly use when making decisions about how to allocate study time. Preschoolers have never engaged in self-directed study. Therefore, they may not be able to make effective decisions about which items should be studied further. However, in Experiments 2 and 3, preschoolers were very good at predicting which of two names they would be more likely to recall later in the forced choice task. It may be possible to help preschoolers use this ability when making judgments about which items they should study again. For example, as in the forced choice task used in the current experiments, children could be shown one animal whose name was correctly recalled at pre-JOL recall and one animal whose name was not successfully pre-recalled. Then the experimenter could ask which animal the child would like to study again. It is possible that preschoolers can use their predictions of which items they will be more likely to remember later to make decisions about which items should be studied further. Importantly, future research can help explore the relationship between metacognitive monitoring and metacognitive control in young children.

The current research explored whether preschool children can make accurate JOLs. Although some children said they would remember all of the animal's names, many preschoolers were able to make accurate yes-no JOLs after a delay. Practice retrieving the names helped reduce overconfidence. Performance on the forced-choice task showed that preschoolers can distinguish between well-learned and less well-learned items. The current experiments provide evidence that most preschool children understand that current performance on a memory task predicts future performance on that task.

REFERENCES

- Atkinson, R. C., & Shiffrin. R. (1968). Human memory: A proposed system and its control processes. In K. Spence & J. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89-195). New York: Academic Press.
- Dunlosky, J., Hertzog, C., Kennedy, M. R. F., & Thiede, K. W. (2005). The selfmonitoring approach for effective learning. *International Journal of Cognitive Technology*, 10, 4-11.
- Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic-extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1180-1191.

Dunlosky, J, & Metcalfe, J. (2009). Metacognition. Beverly Hills, CA: SAGE

- Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, 20, 373-380.
- 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., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy: Insights from the processes underlying judgments of learning in children. In PChambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, function, and use* (pp. 1-18). Dordrecht, Netherlands: Kluwer Academic.
- 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.
- Marazita, J. M., & Merriman, W. E. (2004). Young children's judgment of whether they know names of objects: The metalinguistic ability it reflects and the processes it involves. *Journal of Memory and Language*, 51, 458-472.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of the feelingof-knowing predictions. *Psychological Bulletin*, *95*, 109-133.
- Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. Comments on Schraw (1995). *Applied Cognitive Psychology, 10,* 257-260.
- 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.
- Roediger, H. L. III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-210.

- Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, 137(1), 131-148.
- 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.
- Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, 3, 315-316.

APPENDIX

Animals and Names Used as Stimuli

<u>Animal</u>	Related Name	Unrelated Name
Lion ^{abcd}	Roary	Jim
Cat ^{abcd}	Stripes	Molly
Ladybug ^{abcd}	Dotty	Anna
Frog ^{abcd}	Норру	Tom
Octopus ^{abcd}	Legs	Steve
Cow ^{abcd}	Moo Moo	Betty
Elephant ^{abcd}	Peanut	Suzy
Pig ^{abcd}	Oinky	Brian
Dolphin ^{abc}	Swimmer	Mary
Bear ^{abc}	Teddy	Emily
Dog ^{ab}	Droopy	Bob
Turkey ^{ab}	Gobbler	George
Mouse ^b	Squeaky	Laura
Monkey ^b	Banana	Katie
Owl^b	Hootie	Chris
Lamb ^b	Wooly	Sarah
Cheetah ^b	Speedy	Eric

Mark

- ^a Experiment 1- preschoolers
- ^b Experiment 1- 3rd graders
- ^c Experiment 2
- ^d Experiment 3