An Investigation of Accuracy of Metacognitive Judgments during Learning with an Intelligent Multi-Agent Hypermedia Environment

Reza Feyzi-Behnagh (reza.feyzibehnagh@mail.mcgill.ca)
Zohreh Khezri (zohreh.khezri@mail.mcgill.ca)
Roger Azevedo (roger.azevedo@mcgill.ca)
Department of Educational and Counselling Psychology
Laboratory for Study of Metacognition and Advanced Learning Technologies
McGill University
3700 McTavish Street, Montréal, Québec, H3A 1Y2, Canada

Abstract

Successful learning with advanced learning technologies is based on the premise that students adaptively regulate their cognitive and metacognitive processes. However, research suggests that students are rather dysregulated in their learning. One major source of dysregulation is based on inaccurate metacognitive judgments made during learning. This study investigated learners' accuracy and confidence in metacognitive judgments made in the context of learning about the human circulatory system with MetaTutor, a multi-agent intelligent hypermedia learning system. 83 college students took part in this study, and their interactions within MetaTutor in the two-hour learning session provided data for this study. In general, the results revealed that learners were overconfident to differing degrees in ratings of their judgments of learning (JOLs) and feelings of knowing (FOKs). It was also found that receiving timely prompts and adaptive feedback from the artificial agent in MetaTutor improved the accuracy of metacognitive judgments. Learners in the Prompt and Feedback condition (PF) were overconfident to a lesser degree than those in other conditions (Prompt Only [PO] and Control). Finally, one-way ANOVA and Tukey post-hoc results indicated that learners who received prompts and feedback attained significantly ($p < .05$) better learning efficiency scores than learners in Control and Prompt Only conditions.

Keywords: Metacognitive Judgments; Hypermedia; JOL; FOK; Accuracy; Multi-Agent Learning Environment.

Objectives and Theoretical Framework

Successful learning with advanced learning technologies is based on the premise that learners adaptively regulate their cognitive process based on accurate metacognitive judgments during learning (Azevedo, Moos, Johnson, & Chauncey, 2010). However, there is ample empirical evidence to suggest that learners usually do not regulate key cognitive, metacognitive, affective, and motivation processes during learning with advanced learning technologies such as multi-agent environments (Azevedo et al., in press; Biswas et al., 2010; Graesser & McNamara, 2010; McQuiggan & Lester, 2009; White et al., 2009). In other words, learners typically do not deploy effective learning strategies, modify and update internal cognitive standards, correct behavior based on feedback and scaffolding from the learning system or a tutor, metacognitively monitor their use of strategies or make accurate metacognitive judgments. For example, students' failure to metacognitively monitor their learning, make accurate metacognitive judgments, and deploy regulatory processes are detrimental and can negatively impact their learning. One approach to address this issue is to develop multi-agent learning environments that embody artificial pedagogical agents that are designed to model, scaffold, and foster students' metacognitive processes during learning (see Azevedo et al., in press; 2010; Leelwaong & Biswas, 2008; Schwartz et al., 2009; White et al., 2009).

The goal of this study was to investigate the effects of a multi-agent hypermedia learning environment, MetaTutor, on the accuracy of learners' metacognitive judgments during their learning of the human circulatory system. The metacognitive judgments investigated in this study included Judgments of Learning (JOLs) and Feelings of Knowing (FOKs), which were either prompted by one of MetaTutor's four pedagogical agents or initiated by the students themselves using an SRL (Self Regulated Learning) palette available to them during the learning session. Research by the MetaTutor team has revealed key self-regulatory processes, related to planning, metacognitive monitoring, learning strategies, and methods of handling cognitive task demands, which are deployed by students while learning about complex science topics (see Azevedo & Witherspoon, 2009). One of the main objectives of the MetaTutor project has been to test the effectiveness of pedagogical agents as external regulatory agents in scaffolding students' learning. Pedagogical agents have the potential to provide students with information that will help them become strategic, motivated, and independent learners. One of the areas where pedagogical agents can help students to better regulate their learning is by improving the accuracy of metacognitive judgments, such as JOLs and FOKs. Nelson (1996) argued that, metacognitive judgments are notoriously inaccurate most of the time. He defined the accuracy of metacognitive judgments in terms of the correlation between the respective metacognitive judgments and a subsequent performance score. When a learner's metacognitive judgment rating and performance score correlate closely, they are well 'calibrated'. Lack of confidence as well as overconfidence not justified by one's performance can threaten short- and

long-term learning outcomes of the task. As noted by Boekaerts and Rozendaal (2010), if left unattended, over-confidence or under-confidence in one's skills and knowledge may spread to the domain and may eventually become a personality trait. Winne (2010) argued that the lack of accuracy in metacognitive judgments can be due to a shortage of cognitive resources, specifically working memory and attentional resources, which might be already in use by other cognitive or metacognitive processes, like managing progress toward goals. One of the purposes of using pedagogical agents to provide students with feedback on their performance and correctness of metacognitive judgments is to improve the accuracy of these judgments, because if learners can judge what material they have learned well and what they have not, they can focus their attention on the poorly-learned information, else if their judgments are inaccurate, they cannot successfully guide their learning (Dunlosky & Lipko, 2007), and deploy remedial strategies (e.g. re-reading and taking notes). Although outcome feedback on metacognitive judgments increases students' accuracy, a primary role of feedback in calibration is to change the learners' level of confidence (Stone, 2000). Stone (2000) argues that external feedback, from a teacher or an external agent, can influence how a task is assessed, and will lead to the improvement of students' internal feedback loop as well as their self-regulation of learning. There are two major methods for assessing accuracy – *relative accuracy* and *absolute accuracy*. Dunlosky & Lipko (2007) define relative accuracy as the degree to which one's judgments correlate with his/her own test performance. Absolute accuracy is also defined as whether a person's judgments are over- or under-confident. In this study, we report several analyses of student-initiated and system-initiated metacognitive judgments across three experimental conditions during a two-hour learning session with MetaTutor.

Method

Participants

A total of eighty-three (N=83) participants (70% females) drawn from the two large colleges located in large metropolitan areas took part in this study. They each received $40 for completing the two-day experiment. The participants were randomly assigned to one of three conditions: Prompt and Feedback (PF), Prompt Only (PO), and Control. The PF condition received timely prompts from the pedagogical agents in the learning environment to use different SRL processes and received feedback regarding their performance on the deployment of the metacognitive processes. The PO condition received the same prompts, but no feedback was provided on their performance. Finally, the control group received no prompts and they were free to learn without help from agents in MetaTutor.

MetaTutor, Apparatuses, and Materials

MetaTutor included 41 pages of text and diagrams, designed to detect, model, trace and foster students' self-regulated learning about complex science topics like the human circulatory, digestive and nervous system (Azevedo & Witherspoon, 2009; Azevedo, Johnson, Chauncey, & Graesser, 2011) (See Figure 1). The content for the learning

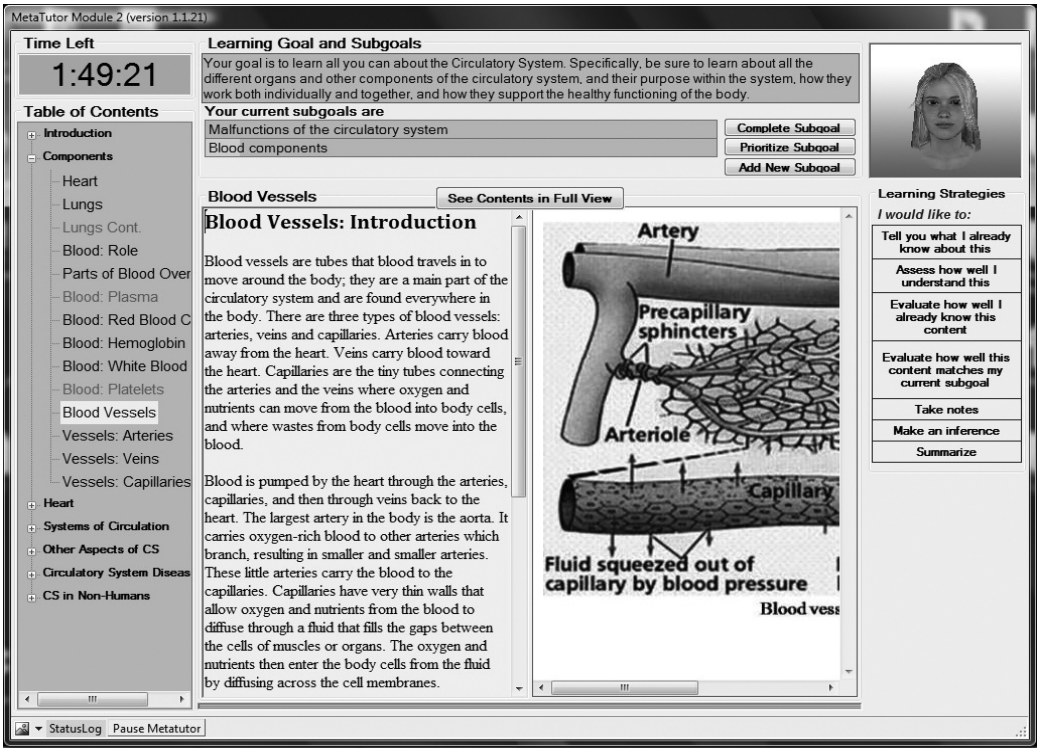

Figure 1. MetaTutor screenshot

session was material on human circulatory system. Several apparatuses were used to collect data during the learning session, including a Tobii T60 LCD remote eye–tracker, used to collect the eye-tracking data; a digital microphone for the concurrent think-aloud protocols; system-generated log files; and a high-definition digital video camera used to collect participants' facial expressions to analyze their emotions during learning.

Experimental Procedure

On the first day of the experiment, participants took a test which measured their knowledge of SRL processes, as well as a pre-test examining their prior knowledge about the human circulatory system. On the second day of the experiment, learners started by setting three sub-goals for their learning at the beginning of the learning session. During their interaction with the learning environment, four computerized pedagogical agents (Gavin the Guide, Mary the Monitor, Pam the Planner, and Sam the Strategizer) helped participants interact with the environment, helped them plan, monitor, use appropriate strategies, and provided timely prompts and appropriate feedback (only in PF condition). The students were also free to choose SRL processes from an SRL palette in the environment interface, which included buttons for initiation of different planning, monitoring and control processes (See Figure 2). The assessments used in the system were a pretest and a posttest, each comprised 25 multiple-choice items. Posttest questions included text-based and inferential questions. In addition to the pretest and posttest, throughout the learning session, the participants were tested with short quizzes after they made a judgment of learning (JOL), feeling of knowing (FOK), and sub-goal completion. The results of the short quizzes in the PF condition led the system's subsequent behavior, and helped the system to provide the participants with proper adaptive feedback.

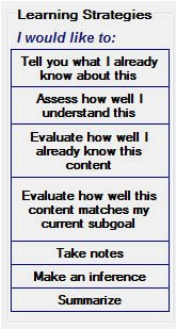

Figure 2. SRL palette in MetaTutor interface

The learners were given two hours to learn about the human circulatory system using MetaTutor, and had the opportunity to take a short five-minute break after the first half of the session. After the two-hour learning session, the participants had 20 minutes to complete the posttest on the material they had learned, and finally they were paid and debriefed at the end of the experiment.

MetaTutor was designed to collect and record all participant interactions with the learning environment and upload these interactions into a log-file which was created for each participant. Specifically, we examined and extracted data for two types of metacognitive judgments, namely Judgments of Learning (JOLs) and Feelings of Knowing (FOKs). During the learning session, all participants had the option of clicking on an SRL palette as a behavioral indication that they were about to deploy different cognitive strategies like summarization or make a metacognitive judgment regarding their performance (*User-initiated* SRL process). Strategies and metacognitive judgments were also prompted at appropriate times by the pedagogical agents (in PO and PF conditions) to scaffold learners' self-regulation (*System-initiated* SRL process). When the participants clicked on the SRL palette or were prompted by Mary to make a judgment of learning (JOL), they were asked to indicate how well they understood the content they had just read on a 6-point Likert scale, ranging from one (*I strongly believe I do not understand this content*) to six (*I strongly believe I do understand this content*). Additionally, the participants could click on a button related to FOK and assess how well they already knew the content they were reading, on a 6-point Likert scale. After making JOLs and FOKs, the participants were asked to take a short quiz and answer three questions to assess the accuracy of their judgment. In the PF condition, participants also received feedback on their performance on the quiz and the accuracy of their JOL or FOK. All metacognitive judgments and quiz scores were recorded in individual log-files along with other interactions with MetaTutor.

Data Analysis

The data analyzed in this paper are part of a major study, which were extracted from the log-files created by MetaTutor for each of the 83 participants across the three conditions (Prompt Only, Prompt & Feedback, and Control). For the purpose of this study, the values of the JOLs and FOKs together with the quiz scores following these metacognitive judgments were extracted by analyzing each participant's log-file. Time stamps in milliseconds were also extracted for FOKs and JOLs, together with data on whether these metacognitive judgments were system-initiated or user-initiated. We then coded the rating values of FOKs and JOLs, which were made on a 6-point Likert scale, into two classes, namely, FOK +, FOK -, JOL + and JOL -, in a way that ratings of 1 to 3 were coded into "+" valence, and ratings of 4 to 6 were coded into "-" valence.

The learning efficiency score was calculated by dividing the raw posttest score by learning time in minutes, which was the time spent on learning content (Faw & Waller, 1976; Simons, 1983). This calculation was performed to account for the amount of time students in both experimental conditions spent on the actual content of the circulatory system by subtracting the time they spent interacting with the agents. Interactions with the agents,

taking quizzes, writing summaries, and other instances when the instructional content was not visible, were not included in learning time. Pretest and posttest scores were also recorded for each participant in the experiment.

We used two measures of *bias* and the *Goodman-Kruskal Gamma* correlation to describe the degree to which judgments of learning and feelings of knowing correlated with performance. The calculations were done based on a two-by-two contingency table created by comparing the JOL and FOK ratings (+ and -) with the performance on the subsequent quiz (low vs. high). Bias score was calculated as the difference between the proportion of high JOLs and FOKs (+) to the relative performance (total correct JOLs/FOKs divided by the total numbers of JOLs/FOKs). Bias scores greater than zero indicate over-confidence, scores less than zero suggest under-confidence, and zero indicates perfect accuracy of confidence and performance. The Goodman-Kruskal gamma correlation is a measure of relative accuracy of performance outcomes according to the confidence judgments made by participants (Dunlosky & Metcalfe, 2009; Schraw, 2009). Gamma indicates the trend in judgments relative to the trend in performance scores. Gamma scores close to +1.0 indicate perfect correlation between JOL/FOK ratings and performance on the subsequent quiz. Bias scores and Gamma correlation were calculated for JOLs and FOKs made by the participants across the three groups in the study.

In order to check for the degree of accuracy of FOK and JOL ratings made by participants with regard to their performance, and investigating any *under-* and *over-estimations*, we tallied the number of agreements between FOK and JOL ratings and quiz performance on a 3-by-3 contingency table (FOK or JOL ratings from 1 to 3 by performance on quiz from 1 to 3). The JOL/FOK ratings made by participants in the MetaTutor environment were initially on a 6-point Likert scale, but we decided to transform the 6-point Likert ratings into 3-points, so that better comparison can be made with performance on a 3-item quiz. This way, we would have a symmetrical contingency table, and can investigate the accuracy of ratings with regards to the performance on the subsequent quiz. The frequencies and percentages of accurate judgments, under- and over-confidence in FOKs and JOLs were obtained across the three experimental groups.

Results and Discussion

Learning Efficiency and Time on Content

The comparison of total time spent on task indicated a significant difference among the three conditions, F (2, 80) = 30.55, $p < .05$, $\eta_p^2 = .045$. Tukey-HSD post hoc analyses revealed that all three groups significantly differ from each other in total time on task. An analysis of variance (ANOVA) on the learning efficiency scores indicated a significant effect of learning condition on learners *learning efficiency* (F [2, 80] = 5.538, $p < .01$, $\eta_p^2 = .122$). Tukey-HSD post-hoc comparisons revealed that the Prompt and Feedback (PF) condition significantly outperformed the Control condition ($p < .05$). A marginal difference was found between the PO and Control conditions ($p = .052$). No significant difference was observed between PF and PO conditions in learning efficiency scores. *Learning time* was calculated by summing the amount of time spent viewing the instructional content, including pages and images. A one-way ANOVA indicated a significant difference among the groups in learning time, F (2, 80) = 30.541, $p < .001$. Tukey-HSD post-hoc analyses indicated that the Control group had a longer total learning time (M = 86.39 minutes, SD = 13.54) compared to both the PO condition (M = 68.51, SD = 14.20) and the PF condition (M = 58.93, SD = 11.74), $p < .001$. Additionally, the PO condition had a significantly longer learning time compared to the PF condition, $p < .05$. These findings indicate that receiving agent prompts to deploy SRL processes and receiving subsequent adaptive feedback improves learning, as indicated by learning efficiency scores.

Metacognitive Judgments

In order to compare the system-initiated (prompts) vs. user-initiated (clicks on the button on the SRL palette) JOLs and FOKs (+ and -) in the two experimental conditions (PO and PF), 2 x 2 chi square contingency table analyses were conducted. These analyses do not include the Control group since participants in this condition do not receive prompts by the system to deploy SRL processes. The results indicated that there is a significant difference in the distribution of user- vs. system-initiated metacognitive processes across both experimental conditions ($p < .05$). This indicates that learners make fewer JOLs by clicking on the SRL palette when the system does not prompt them to make a JOL. Also, more positive than negative FOKs and JOLs are observed in all conditions. Analyzing the accuracy of these metacognitive judgments might shed light on how calibrated the students were while making the judgments. A summary of chi-square results is presented in Table 1.

Table 1. Frequencies of $\chi 2$ analysis of user- vs. system-initiated JOLs and FOKs by valence

Cond.	Initiation	JOL+	JOL-	χ^2	p
PF	User	66	19	5.207	0.022*
	System	68	41		
PO	User	29	1	4.743	0.029*
	System	114	28		
Control	User	72	11		
	System	0	0		
		FOK+	FOK-	χ^2	p
PF	User	46	6	23.57	0.000*
	System	22	29		
PO	User	45	6	6.534	0.010*
	System	25	13		
Control	User	44	11		
	System	0	0		

Measures of Accuracy

In order to calculate the accuracy of metacognitive judgments (agreement between judgment and performance), we used Goodman-Kruskal Gamma, which is a measure of correlation and is based on the difference between concordant and discordant pairs. Under statistical independence, Gamma will be 0, which means there is no correlation between judgment and performance. The value of Gamma ranges from -1 to +1. In this study, whenever there is an agreement between an FOK or JOL rating (+ or -) and the corresponding subsequent quiz score, it is considered a concordance. The results of Gamma calculation indicate the degree of association between FOK or JOL ratings and performance on the subsequent quiz, in other words, the agreement of judgments or ratings with performance (see Table 2). As illustrated in Table 2, there is a significant correlation between JOL judgments and performance in both the PO and PF conditions ($p < .05$). Specifically, in the case of JOL in the PO group, a strong agreement can be observed between ratings and performance ($G = .638$, $p < .001$). Nelson and Dunlosky (1991) reported an average value of Gamma for immediate metacognitive judgments from +.09 to +.48, and quote other similar studies finding average Gammas of +.33.

Table 2. Gamma and Bias score summary table

	Gamma Correlation				Bias Score	
	FOK	Sig	JOL	Sig	FOK	JOL
PO	.296	.262	.638	.001*	.15	.11
PF	.184	.051	.145	.047*	.009	.06
Ctrl	-.256	.502	-.20	.594	.18	.10

$p < .05$

The Gamma values obtained in this study for FOKs and JOLs are in approximately the same range as those found by other researchers in similar studies. The medium and low Gamma correlations obtained here indicate low accuracy of JOLs and FOKs made by learners in different conditions. Better accuracies are observed for PO condition, which might be related to the fact that in the absence of agent feedback, participants had to become more independent metacognitively, and monitor their learning more accurately.

In order to investigate the degree of over- and under-confidence, bias scores were calculated. Bias (Kelemen, Frost & Weaver, 2000) is a measure of overall degree to which confidence matches performance. Bias scores greater than zero indicate over-confidence and bias scores less than zero show under-confidence. As can be seen in Table 2, participants in all three conditions were over-confident in their FOK and JOL ratings, to differing degrees. This is in line with findings of Lichtenstein and Fischhoff (1977), where they argued that the most common bias observed in metacognitive judgments is over-confidence. The bias scores for JOLs and FOKs in the PF condition are very small, which corroborates the argument made by Sharp and colleagues (1988) that when learners are exposed to performance feedback, their confidence judgments improved across sessions to a greater extent than in other conditions. As explained in the Procedure section above, in order to investigate the degree to which participants were accurate, over- or under-confident in judgments of their performance on a subsequent quiz, we investigated the frequency of correct and incorrect judgments. The results indicated that in the case of JOLs, participants were accurate about less than half the time (34.8, 41.7, and 44.5 percent for PO, PF and Control conditions, respectively).

Table 3.Percentage of confidence and accuracy of JOLs and FOKs

		Under-confidence	Over-confidence	Accurate Judgment
JOL	PO	16.27 %	48.83 %	34.88 %
	PF	22.16 %	36.08 %	41.75 %
	Ctrl	16.86 %	38.55 %	44.57 %
FOK	PO	21.34 %	55.05 %	23.59 %
	PO	30.09 %	33.98 %	35.92 %
	Ctrl	18.18 %	50.90 %	30.90 %

With regards to the accuracy of FOKs, the participants were even less accurate in comparison to when they made judgments of their learning (23.5, 35.9, and 30.9 percent for PO, PF and Control conditions, respectively). A summary of findings about confidence and accuracy in JOLs and FOKs is displayed in Table 3.

Conclusions and Future Directions

One of the goals of this study was to investigate the accuracy of metacognitive judgments (FOKs and JOLs) made by students while learning with a hypermedia multi-agent environment. The findings were generally in line with results of previous studies (e.g., Nelson & Dunlosky, 1991; Lichtenstein & Fischhoff, 1977) on accuracy of learners' metacognitive judgments, in terms of the magnitude of Gamma correlations between judgments and performance, and the general bias learners have towards overconfidence in making JOLs and FOKs. We also found that learners receiving prompts and feedback (i.e., those in the PF condition) from pedagogical agents were less overconfident in their JOLs and FOKs than learners in other conditions. This provides support for the effectiveness of prompts and adaptive feedback on improving learners' calibration in their metacognitive judgments. The findings also have implications for the design of intelligent and adaptive computer-based learning environments to help students self-regulate their learning in a better way and become more calibrated in their metacognitive judgments, which will lead to improved learning.

The data for the current study was obtained from log-files generated by MetaTutor, which contained a detailed record of learners' interactions with the learning environment. However, the use of on-line trace methodologies like eye-tracking data and concurrent think-alouds could provide

additional evidence about the nature of these metacognitive processes. Another avenue for future research is the investigation of the accuracy of delayed-JOLs in the context of multi-agent learning environments. The delayed-judgment of learning effect has been studied by a number of researchers, and they have found that delayed judgments are significantly more accurate than immediate judgments (Thiede et al., 2005; Veenman et al., 2006).

Acknowledgments

Funding supporting this study was provided by the National Science Foundation (DRL# 0633918 and DRL #1008282). The authors would like to thank Amy Johnson, Candice Burkett, and Amber Chauncey Strain for data collection.

References

Azevedo, R., Cromley, J. G., Moos, D. C., Greene, J. A., & Winters, F. I. (in press). Adaptive content and process scaffolding: the key to facilitating students' self-regulated learning with hypermedia. *Psychology Science Quarterly*.

Azevedo R., Johnson A., Chauncey A., & Graesser A. (2011). Use of hypermedia to convey and assess self-regulated learning. In B. Zimmerman and D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 102-121). New York: Routledge

Azevedo, R., Moos, D., Johnson, A. M., & Chauncey, A. D. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: issues and challenges. *Educational Psychologist, 45*(4), 1-14.

Azevedo, R. & Witherspoon, A. M. (2009). Self-regulated learning with hypermedia. In D. J. Hacker, J. Dunlosky, and A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 319-339). Mahwah, NJ: Routledge.

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning, 5*(2), 123-152.

Boekaerts, M. & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Leaning and Instruction, 20*, *372-382*.

Dunlosky, J. & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16 (4),* 228-232.

Dunlosky, J. & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage.

Faw, H. W., & Waller, T. G. (1976). Mathemagenic behaviors and efficiency in learning from prose materials: Review, critique, and recommendations. *Review of Educational Research, 46*(4), 691-720.

Graesser, A. & McNamara, D. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist, 45*(2), 234-244.

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidnce against a general metacognitive ability. *Memory & Cognition, 28* (1), 92-107.

Leelawong, K. & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of AI in Education, 18*, 181-208.

Lichtenstein, S. & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behvaior and Human Performance, 20*, 159-183.

McQuiggan, S. W. & Lester, J. C., (2007). Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies. 65*, 348-360.

Nelson, T. O. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2 (4),* 267-270.

Nelson, T. O. (1996). Consciousness and metacognition. *American psychologist, 51*, 250-256.

Schwartz, D. L., Chase, C., Chin, D. B., Oppezzo, M., Kwong, H., Okita, S., Biswas, G., Roscoe, R., Hogyeong, J., & Wagster, J. (2009). Interactive metacognition: monitoring and regulating a teachable agent. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Handbook of metacognition in education* (pp. 340-358). New York: Routledge.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 33-45.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes, 42*, 271-283.

Simons, P. R. J. (1983). How we should control time on task – Or should we? *Instructional Science, 11*, 357-372.

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12 (4),* 437-475.

Thiede, K. W., Dunlosky, J., Griffin, D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*, 1267-1280.

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

White, B., Frederiksen, J., & Collins, A. (2009). The interplay of scientific inquiry and metacognition: More than a marriage of convenience. In D. J. Hacker, J. Dunlosky, and A. C. Graesser (Eds.) *Handbook of metacognition in education* (pp. 175-205). New York: Routledge.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist, 45*, 267-276.