Harley, J. M., Bouchet, F., Papaioannou, N., Carter, C., Trevors, G., Feyzi-Beghnagh, R., Azevedo, R., Landis, R. S. (2014). *Assessing Learning with MetaTutor, a Multi-Agent Hypermedia Learning Environment.* Paper to be presented at a symposium on Innovative Practices for Assessing Learning in Computer Based Learning Environments at the 2014 annual meeting of the American Educational Research Association, Philadelphia, PA.

Assessing Learning with MetaTutor, a Multi-Agent Hypermedia Learning Environment

Jason M. Harley[1] François Bouchet[2], Niki Papaioannou[3], Cassia Carter[3], Gregory Trevors[1], Reza Feyzi Behnagh[1], Roger Azevedo[4] and Ronald Landis[3]

[1]McGill University, Montréal, Québec, Canada
[2]Sorbonne Universités, UPMC Univ Paris 06, Paris, France
[3]Illinois Institute of Technology, Chicago, IL, USA
[4]North Carolina State University, Raleigh, North Carolina, USA

Corresponding author: {jason.harley@mail.mcgill.ca}

Abstract. In this paper we discuss the ways in which assessments of learning were evaluated, designed, and implemented in MetaTutor (a multi-agent, hypermedia learning environment about several human body systems; Azevedo et al., 2012, 2013). We also share the lessons that we have learned from assessing learning with MetaTutor across three different universities and three studies ($N = 336$). Techniques and considerations for assessing learning are shared, including pedagogical and psychometric properties of items. Results revealed significant differences in pre to posttest learning across all studies and that changes to the content and assessments of learning lead to lower scores, in particular, on the pretest where scores had previously been high and limited variance. The methods and results of this paper can be used to motivate other researchers who use CBLEs to examine and improve the psychometric and pedagogical features of their learning assessments.

Introduction

Computer-based learning environments (CBLEs) are complex systems capable of a myriad of advanced functions, including measuring, tracking, and in some cases, adapting to learners based on information such as their learning assessments (e.g., pre-test scores), experience of positive emotions, and use of effective self-regulatory skills (Azevedo et al., 2013; Kinnebrew, et al., 2013; Lajoie et al., 2013; Shute & Kim, 2013). However, as CBLEs have become increasingly more sophisticated the methodological reporting and empirical scrutiny regarding their assessments of learning have commonly become a secondary consideration despite opportunities for innovation (Azevedo & Aleven, 2013; Mislevy, 2010). Online assessments of learning, in particular, provide valuable opportunities for using formative assessments of learning through the administration of quizzes. The administration and scoring of digital quizzes in CBLEs is easier than paper-based versions and the information from them can be used to customize participants' learning sessions. For example, CBLEs can use quiz score information to provide immediate feedback about learners' performance and to make suggestions for next steps (e.g., re-read the page, continue). CBLEs also provide unique opportunities for enhancing traditional assessment tools, such as quizzes, through the use of open-learner models (OLMs). OLMs provide learners with a visual representation of their progress which can be generated through self-reporting, quiz results, or information from the system logs (or a combination of) (Fezyi-Behnagh et al., 2013). CBLEs capacity to log information opens up new options for assessing learning subtly through the use of algorithms that identify learner-system interaction parameters that correspond to different types of learning and psychological processes (e.g., creativity; Shut & Kim, 2013).

The purposes of this paper are to discuss the ways in which assessments of learning are iteratively evaluated, designed, and implemented in MetaTutor (a multi-agent, hypermedia learning environment about several human body systems; Azevedo et al., 2012, 2013) and to share the lessons that we have learned from assessing learning with it across three different universities and three studies. Techniques and considerations for assessing learning are shared, including pedagogical and psychometric properties of items. Results comparing changes in pre to posttest learning are also presented and the effectiveness of content and assessment changes as well as limitations and future directions are discussed.

Methods and Data Sources

This section has been structured to provide details about the participants this paper's analyses drew upon as well as the learning environment, MetaTutor (Azevedo et al., 2013), before describing the different methods used to measure learning. The Experimental Procedure describes the context and process through which the data were collected.

Participants

Three hundred and thirty-six (N= 336; 197 female, 94 male) undergraduate students from three North American universities participated in this study. The universities included a large mid-South US public university (University A; N = 101), a large, Canadian, public university (University B; N = 42), and a small, mid-West, private university (University C; N = 54). Participants were randomly assigned to one of the experimental conditions tested (described below). This paper describes results from three studies using MetaTutor. One hundred and fifty-two participants (N = 152; 79 female, 27 male; 101 from University A and 51 from University B) participated in Study 1. One hundred and twenty four (N = 124; 81 female, 42 male; 82 from University B and 42 from University C) participated in Study 2. Sixty participants participated in Study 3 (N = 60; 36 female, 24 male; 12 from University C and 48 from University B.

MetaTutor

MetaTutor (Azevedo et al., 2012, 2013) is a multi-agent Intelligent Tutoring System (ITS) and hypermedia learning environment which consists of 38-48 pages of text and static diagrams organized by a table of contents displayed in the left pane of the environment (content varies by version/study of MetaTutor). The versions of MetaTutor used in this experiment were comprised of material on the human circulatory system. Participants were instructed to learn this material as they interacted with four embedded, pedagogical agents (PAs). The four PAs' instructional scaffolding varied depending on the experimental condition that learners were assigned to. Aside from PA scaffolding, the conditions within each study were identical. In the prompt and feedback condition (PF) condition, learners were prompted by the PAs to use specific self-regulatory processes (e.g., to metacognitively monitor their emerging understanding of the topic or deploy a specific cognitive learning strategy such as re-reading or coordinating informational sources), and were given feedback about their use of those processes. In the control (C) condition, participants did not receive prompts or feedback and could only perform these self-regulatory processes on their own initiative by clicking on the buttons on a palette in the interface. In addition to the PF and C conditions, participants in Study 1 were also randomly assigned to a prompt-only (PO) condition, which was similar to the PF condition with the exception that no feedback was provided to students with regard to the quality of their responses (e.g., accuracy of their metacognitive judgments, scores on quizzes). Table 1 provides a summary of differences between PA behavior in the PF and C conditions. Differences between these conditions in Studies 2 and 3 exist, primarily through meta rules where the frequency of scaffolding was adapted based on participants effective use of SRL behaviors. Describing the changes in more depth is, however, beyond the scope of this paper (for more information see Bouchet, Harley, and Azevedo, 2013)

Table 1

Summary of PA Behaviour by Condition

Learning Context	Summary of PA Behaviour	
	Prompt and Feedback Condition	Control Condition
Sub goal setting	Pam engages in interactive dialogue with student in order to relate (e.g., through counter-proposals) their proposed sub goal to one of the seven ideal sub goals. If proposal is un-related, Pam provides learner with additional opportunities. Pam provides feedback regarding appropriateness of proposed sub goal.	Pam counter-proposes one of the seven ideal sub goals if the learner doesn't initially propose one. No feedback is given regarding the appropriateness of their proposal.
Self-regulated Learning Behaviour	Prompts: Sam, Mary, and Pam provide prompts to students to engage in SRL behaviours, such as making summaries, content evaluations, and activating prior knowledge.	Prompts: Sam, Mary, and Pam do not provide any SRL prompts during the learning session after the sub goal setting phase.
	Feedback: Sam and Mary provide feedback regarding students' use of these processes. For example, whether students have written an appropriate summary, or whether Mary agrees that the content of a page is relevant to the sub goal learners are currently working on. Pam does not evaluate learners' prior knowledge activation. Sam and Mary provide feedback on learners' use of SRL irrespective of whether they are prompted by the agent or self-initiated (through the SRL palette) by the learner.	Feedback: Sam and Mary do not provide any feedback.
Administration assessments	Mary administers page and sub goal quizzes to participants. She provides learners with feedback regarding how they did on these assessments to help them make decisions regarding whether to move on or revisit the learning material.	Mary administers page and sub goal quizzes to participants. No feedback is provided.

Experimental Procedure

The experiment consisted of two sessions (session 1 lasted approximately 30 minutes and session 2 lasted approximately three hours), which were separated by a maximum of three days or a minimum of one hour to avoid participant fatigue[1]. During

[1] The range of time existed to balance flexibility and therefore minimize attrition, as well as reduce the likelihood that participants would be exposed to new information on the circulatory system that could affect their performance on the posttest.

session 1, participants read and signed an informed consent form and were administered several self-report questionnaires (e.g., demographics, AEQ-trait emotions) and a pretest on the circulatory system. All participants completed the session individually on a desktop computer with a Core 2 Duo 2.80 GHz processor, 2GB of RAM and Windows XP, using a 17” monitor with a 1024x768 resolution. Participants were compensated up to $10 at the end of session 1 ($10/hr.) or given course credit.

During session 2, participants were shown an introductory video presenting information regarding the MetaTutor learning environment, including training on how to use different components and navigate through the system. Each participant was given a period of time to learn as much as possible about the circulatory system using MetaTutor. In Study 1 participants had two hours to learn as much as possible about three sub goals. In Study 2 they had one hour to learn as much as possible about two sub goals. Participants had 90 min. to learn as much as possible about two sub goals in Study 3. Halfway through the learning session, participants were given the opportunity to take a five-minute break, though not all participants chose to take it. At the end of session 2, participants were asked to complete a posttest. Lastly, all participants were debriefed and received up to $30 ($10/hr) for participating in the second session.

Learning Measures

Six multiple-choice questions were developed for each of MetaTutor’s pages. Pre/Posttest, page quiz, and subgoal quiz questions were drawn from this bank.

Page quizzes. Depending on the experimental condition, participants could be prompted or elect to take a three-item page quiz to assess their understanding of page-specific content.

Sub goal quizzes. Participants set sub goals (2-3, depending on the study) at the beginning of the learning session to help guide their learning session. Sub goal quizzes contained ten items and were designed to assess whether they had sufficiently mastered the content pertaining to each of the sub goals they had established for learning. These items were drawn from the total bank of items that pertained to the pages relevant to a given subgoal.

Pretest and Posttest. Two equivalent 25-item multiple-choice tests on the human circulatory system were used in this study. Participants completed one version of the test as the pretest during session 1 and the other as the posttest during session 2. The order participants received the tests in was counter-balanced. The pretest was designed to assess participants' prior knowledge of the circulatory system and the posttest was designed to assess their knowledge about the circulatory system after interacting with MetaTutor.

Pedagogical item characteristics

Both versions of the pretests included text-based and inferential items. Participants could respond to text-based items using one sentence within a page of content. Inferential items, on the other hand, required participants to go beyond recalling information from the text and demonstrate their understanding through application (e.g., by applying a principle to a different example). As such, the self-report measure assessed learners' cognitive processing at different levels. Using Krathwohl's (2002) revision of

Bloom's Taxonomy, text-based items can be classified as involving cognitive processes related to remembering, such as recognizing and recalling the correct information. In contrast, inferential items involved learners understanding and applying knowledge. Items assessed learners' factual and conceptual knowledge, though generalizations were reserved for inference-related items.

Each of MetaTutor's multiple-choice questions had four foils that corresponded to a target (right) answer, a thematic answer, a near miss, and an unrelated (wrong) answer. Thematic answers were those that were not wrong, but did not contain enough information to properly answer the question. Near miss answers were similar to thematic answers, but may have been too specific and missed important information rather than being too general. One example is the following: Why is the heart muscular? (a) because it must pump blood through the system? [target] (b) because it is connected to arteries and veins [thematic] (c) because it sends blood to the muscles [near miss], and (d) because all body organs are made of muscle [unrelated].

Learning with MetaTutor: Studies 1 and 2

In this section we report on the analyses we ran to examine learning with MetaTutor using pretest and posttest scores.

Results for Study 1

We examined changes in pre (M = 0.68; SD = 0.18) to posttest (M = 0.79; SD = 0.15) learning in Study 1 with a repeated measures ANOVA and observed a main effect of time, $F(1, 146) = 84.31$, $p < .01$, $n^2_p = 0.36$, but no interaction effect between time and experimental group $F(2, 146) = 0.37$, $p > .05$, $n^2_p = 0.01$, nor time and institution, $F(2, 146) = 2.18$, $p > .05$, $n^2_p = 0.02$ on test scores. We failed to find a main effect of experimental group, $F(2, 149) = 0.11$, $p > .05$, $n^2_p = 0.00$, but there was a main effect of university, ($F(1, 145) = 50.12$, $p < .01$, $n^2_p = 0.26$). Students at University B had higher Pretest ($M = 0.80$; $SD = 0.14$) and Postest ($M = 0.89$; $SD = 0.09$) scores than University A (Pretest, $M = 0.61$; $SD = 0.17$; Postest, $M = 0.73$; $SD = 0.15$). Pairwise comparisons using a Bonferroni correction failed to identify any significant differences.

Results for Study 2

In Study 2 we replicated our analyses. We examined changes in pre ($M = 0.76$; $SD = 0.16$) to posttest ($M = 0.84$; $SD = 0.14$) learning in Study 2 with a repeated measures ANOVA and observed a main effect of time, $F(1, 120) = 48.82$, $p < .01$, $n^2_p = 0.29$, but no interaction effect between time and experimental group $F(2, 120) = 0.01$, $p > .05$, $n^2_p = 0.00$ nor time and institution, $F(2, 120) = 0.10$, $p > .05$, $n^2_p = 0.00$, on test scores. We failed to find a main effect of experimental group, $F(2, 149) = 2.95$, $p > .05$, $n^2_p = 0.02$, but there was a main effect of university, ($F(1, 120) = 10.80$, $p < .01$, $n^2_p = 0.08$), on test scores. Students at University B had higher Pretest ($M = 0.79$; $SD = 0.15$) and Postest ($M = 0.87$; $SD = 0.10$) scores than University C (Pretest, $M = 0.70$; $SD = 0.18$; Postest, $M = 0.78$; $SD = 0.18$).

Discussion and Next Steps

The results from our above analyses reveal that students did learn with MetaTutor, but that the experimental condition they were assigned to didn't have a significant effect on their learning. Moreover, pretest scores were high, especially in University B. It is possible that high pretest scores limited the variance and made differences between conditions difficult to identify. We therefore decided to update the content in order to make it more challenging. This involved an examination of the test items as well as the learning material.

Examining and Revising MetaTutor's Content

Extensions and revisions of the pedagogical content

An examination of the content revealed an imbalance in the amount of content (i.e., number of pages) related to each of the seven subgoals, where a range of 4 to 12 pages (out of 38) were related to a given sub goal. As a result, learners had more or less goal-related content to read, depending on the sub goals they set. We therefore identified which sub goals needed additional pages and recruited a biological systems expert to extend the number of pages in MetaTutor from 38 to 48 (creating 10 new pages of pedagogical content on the circulatory system). Table 2 provides a brief overview of descriptive information pre and post the addition of the ten added pages. Balancing the pages related to each of the sub goals better facilitates comparison between subjects who selected different sub goals. Relatedly, the content of the existing pages was slightly shortened or extended, as needed, in order to have a better balance of content per page.

Table 2

Impact of the changes to pedagogical content on pages relevance to the 7 subgoals

	38 pages (Study 2)		48 pages (Study 3)	
	Mean	Std. dev.	Mean	Std. dev.
Number of pages relevant to a sub-goal	7.14	2.85	9.00	1.63
Number of words per page relevant to a sub-goal	1560.90	488.50	1800.90	130.70

In addition to additions and revisions to the text we collected feedback from content experts and lab members on the visual and pedagogical quality of the images and replaced them with higher quality images where appropriate. New, high-quality images were also identified and acquired for the new pages.

Test Items

In order to identify items that needed to be re-written we examined the proportions of learners (from a pooled sample of Studies 1 and 2) that selected each of the four foils for each item. If a target answer was selected by seventy percent or more of learners it was deemed to be too easy and was re-written. Foils relating to the other three options were re-written if they were selected by fewer than three percent of participants, indicating that they were poor distractor items. We therefore sought foils with a selection proportion range between 3-69%. Using these parameters we determined that foils from 94% of pre/posttest questions (55% of all foils) needed to be re-written. Using these same parameters we determined that foils from 70% of all data bank questions (35% of all foils) needed to be re-written (36% overlap with test questions).

All new foils (and items when foil replacement was not sufficient) were either written or reviewed and approved by a graduate student in biology. Two undergraduate students with backgrounds in biology and physiology assisted with the writing and reviewing of items. New items were reviewed by the first and fifth author in order to ensure that they fit with the intended replacement foil (target, thematic, etc.).

In addition to the revisions made to the pre-existing questions, six new questions, three text-based and three inference-based, were written for each of the ten new pages (60 new questions) by the same biological systems expert. The type of question (inference vs. text-based) was reviewed and approved by two PhD students in Educational Psychology, while the content was reviewed by two undergraduate students with expertise in human biology. These questions were then added to the bank of 228 questions used for pre and post test, page and/or sub-goal quizzes.

All of the new questions as well as those formerly used and newly revised (a total of 112) were then pilot tested with 70 new participants from University C who had not previously participated in the MetaTutor experiment or interacted with it prior to answering the questions. The same criteria were applied to these questions and they were deemed suitable for use in MetaTutor Study 3.

At this point, we re-designed the two tests using this bank of 112 new and revised questions, which allowed us to include questions related to the new content and to more accurately balance the number of questions relative to each sub-goal. The two new tests contain 30 questions each, instead of 25 in order to reflect the increase in the amount of the pedagogical content used in the system (cf. Table 3 for more details).

Table 3

Analysis of the questions in the revised tests A and B

		Subgoal ID							Overall
		1	2	3	4	5	6	7	
# Q's associated in test A		6	6	6	5	7	7	6	30[a]
# Q's associated in test B		6	6	6	5	7	7	6	30[a]
% right in test A	*M*	46.8	41.2	41.3	50.5	59.8	50.0	58.9	52.2
	SD	13.1	14.7	16.5	16.7	10.4	11.5	10.9	13.6
% right in test B	*M*	45.7	39.2	44.5	49.7	59.8	54.1	59.3	50.5
	SD	16.9	16.2	16.1	21.2	12.8	11.8	7.3	14.6

a. *This number is different from the sum of the number of questions associated to each sub-goal because questions are often relative to more than one sub-goal.*

Learning with MetaTutor: Study 3

We examined changes in pre (M = 0.61; SD = 0.15) to posttest (M = 0.73; SD = 0.15) learning in Study 3 with a repeated measures ANOVA and observed a main effect of time, $F(1, 56) = 34.70$, $p < .01$, $n^2_p = 0.38$, but no interaction effect between time and experimental group, $F(2, 56) = 0.09$, $p > .05$, $n^2_p = 0.00$, nor time and institution, $F(2, 56) = 0.82$, $p > .05$, $n^2_p = 0.02$ on test scores. We failed to find a main effect of experimental group, $F(2, 56) = 1.00$, $p > .05$, $n^2_p = 0.03$, but there was a main effect of university, ($F (1, 56) =4.47$, $p < .05$, $n^2_p = 0.07$) on test scores. Students at University B had higher Pretest (M = 0.63; SD = 0.14) and Postest (M = 0.74; SD = 0.15) scores than University A (Pretest, M = 0.54; SD = 0.15; Postest, M = 0.69; SD = 0.14).

Comparing Learning Across Studies 1 – 3

We examined changes in pre to posttest learning between Studies and Universities with a repeated measures ANOVA and observed a main effect of time, $F(1, 330) = 150.00$, $p < .01$, $n^2_p = 0.31$, but no interaction effect between time and institution, $F(2, 330) = 1.70$, $p > .05$, $n^2_p = 0.01$, or time and study, $F(2, 330) = 2.14$, $p > .05$, $n^2_p = 0.01$ on test scores. We found a main effect of Study, $F(2, 330) = 21.56$, $p < .01$, $n^2_p = 0.12$, and university, ($F (2, 330) = 33.05$, $p > .01$, $n^2_p = 0.17$) on test scores. Pairwise comparisons using a Bonferroni correction revealed a significant difference between test scores in Study 3 when compared to Studies 1 and 2 ($p < .05$). Table 4 revealed that Pretest and Posttest scores were lower in Study 3 than in either Study 1 or 2.

Table 4

Means and Standard Deviations for Pre and Posttest Scores by University and Study

University	Study	PreTest		PostTest		
		M	*SD*	*M*	*SD*	*N*
C	2	.71	.18	.79	.18	42
	3	.54	.15	.69	.14	12
	Total	.67	.18	.76	.17	54
B	1	.80	.14	.89	.09	51
	2	.79	.15	.87	.10	82
	3	.63	.14	.74	.15	48
	Total	.75	.16	.84	.13	181
A	1	.61	.17	.74	.15	101
	Total	.61	.17	.74	.15	101
Total	1	.68	.18	.79	.15	152
	2	.76	.16	.84	.14	124
	3	.62	.15	.73	.15	60
	Total	.70	.18	.80	.15	336

Conclusions and Discussion

This paper has described the steps we took to examine the psychometric and pedagogical characteristics of the assessments of learning. We used this information to inform a revision of the content and assessments of learning which most notably improved the difficulty of the items and the balance between subgoals and pages. Although the tests items were more challenging for participants from University B and C, we failed to identify a main effect of experimental condition on learning. None-the-less, results across all studies revealed significant differences in pre to posttest learning.

Limitations

One of the limitations of this paper comes from making comparisons between studies in which more changes than content and assessments were made. As mentioned earlier, participants were given different amounts of time and had different numbers of sub goals to learn about over the course of their learning session. Moreover, a variety of new pedagogical agent rules were implemented in Study 3, however (based on the lack of significant differences between experimental conditions) it is possible that they may not have been responsible for much, if any, of the differences regarding learning that were observed between studies. It is also possible that we didn't observe any main effects of experimental condition on learning because of the relatively short time span of the learning session. Learners may not have had enough time to fully benefit from or internalize and implement the SRL strategies the PAs scaffolded in the PF condition (and to a lesser extent, the PO condition in Study 1). A second limitation is the fact that participants were sampled from the general undergraduate student population rather than biology-related programs. This may have had implications with regard to the effort they invested in learning about the content as it was generally unrelated to their academics.

Future Directions

Results from Study 3 are based on a preliminary sample of participants; analyses should be re-run when data collection is complete. Future research with MetaTutor examining learning with a longer learning session may help the PAs' prompts to have an impact. Future research might also examine learners' understanding of the SRL prompts that the PAs are providing to determine if some participants are failing to effectively deploy strategies because they don't understand them as effectively as they could.

Overall, this research is significant because it provides a detailed example of the importance of examining how learning is assessed in CBLEs. More specifically, the methods and results of this presentation can be used to motivate other researchers who use CBLEs to examine and improve the psychometric and pedagogical features of their learning assessments.

Acknowledgements

We would like to thank Anthony Sassi, Valérie Bélanger-Cantara, and Kelsey Anderson for their assistance with content and item revisions. We would also like to thank the following people for their assistance with data collection: Lauren Agnew, Kelsey Anderson, Valérie Bélanger-Cantara, Candice Burkett, Jonathan Burlison, Timothy Carsel, Inderpal Dhillon, Melissa Duffy, Ricardo Erazo, Ashley Fike, Sophie Griscom, Amy Johnson, Lana Karabachian, Nicholas Mudrick, Nicole Pacampara, Alejandra Segura, Victoria Stead, Amber Chauncey-Strain, Michelle Taub, Grace Wang and Wook Yang. The research presented in this paper has been supported by a graduate student fellowship from a Joseph-Armand Bombardier Canadian Graduate Scholarship (CGS) from the Social Science and Humanities Research Council (SSHRC) awarded to the first author and funding from the National Science Foundation (IIS 1008282), The Canada Research Chairs program, and the Social Science and Humanities Research Council awarded to the seventh author.

References

Azevedo, R., & Aleven, V. (Eds.). (2013). *International handbook of metacognition and learning technologies*. Amsterdam, The Netherlands: Springer.

Azevedo, R., Harley, J., Trevors, G., Feyzi-Behnagh, R., Duffy, M., Bouchet, F., & Landis, R.S. (2013). Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. In R. Azevedo & V. Aleven (Eds.), *International handbook of* *metacognition and learning technologies* (pp. 427-449). Amsterdam, The Netherlands: Springer-Verlag.

Azevedo, R., Behnagh, R., Duffy, M., Harley, J., & Trevors, G. (2012). Metacognition and self-regulated learning with advanced learning technologies. In D. Jonassen & S. Land (Eds.), *Theoretical foundations of learning environments* (2nd ed.) (pp. 171-197). Mahwah, NJ: Erlbaum.

Bouchet, F., Harley, J. M., & Azevedo, R. (2013). The impact of different pedagogical agents' adaptive self-regulated prompting strategies with MetaTutor. In C. H. Lane, K. Yacef, J. Mostow, P. Pavik (Eds.), *Lecture Notes in Computer Science: 7926. Artificial Intelligence in Education* (pp. 815-819). Berlin, Heidelberg: Springer-Verlag.

Feyzi-Behnagh, R., Azevedo, A., Legowski, E., Reitmeyer, K., Tseytlin, E., Crowley, R. S. (2013). Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system. *Instructional Science*, 1-23.

Kinnebrew, J. S, Biswas, G., Sulcer, B., & Taylor, R. S. (2013). Investigating self-regulated learning in teachable agent environments. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 451-470). Amsterdam, The Netherlands: Springer-Verlag.

Lajoie, S. P. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from Avionics and Medicine. In K. A. Ericsson (Ed.), *The development of professional performance: Approaches to objective measurement and designed learning environments* (pp. 61-83). New York, NY: Cambridge University Press.

Mislevy, R,J. (2010). Some implications of cognitive psychology for educational assessment. *Research Papers in Education*, *25*, 253-270.

Shute, V. J., & Kim, Y. J. (2013). Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology (4th Edition)*. New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.