Annual Report to the Institute of Education Sciences
June 30, 2007

iSTART: Interactive Strategy Training for Active Reading and Thinking
(R305G040046)

Danielle S. McNamara (PI)

University of Memphis

INTRODUCTION

Many students have difficulty understanding what they read, particularly the challenging textbooks they encounter in their academic courses (Bowen, 1999; Snow, 2002). Such texts can be better understood by teaching students how to use active reading strategies that enhance comprehension. In this project, we focus on teaching high school and college students just such reading strategies through an automated strategy trainer called iSTART (Interactive Strategy Trainer for Active Reading and Thinking, McNamara, Levinstein & Boonthum, 2004).

To date, numerous experiments assessing the efficacy of iSTART have been conducted with over 1,000 middle school, high school, and college students. The convergence of findings suggests that iSTART is effective in helping students use strategies to learn from texts, and enhances comprehension, particularly among low-knowledge readers (Magliano, Todaro, Millis, Wiemer-Hastings, Kim, & McNamara, 2005; O'Reilly, Sinclair, & McNamara, 2004a, 2004b). The goal of the current project is to expand iSTART so that it can be used more effectively and efficiently in high school classrooms.

iSTART: Automated Strategy Training

iSTART is a web-based computer program that uses automated agents to provide reading strategy training. iSTART currently incorporates theoretically motivated Self-Explanation Reading Training (SERT; McNamara, 2004; McNamara & Scott, 1999), which teaches students to self-explain science texts by using active reading strategies that facilitate and enhance comprehension, such as paraphrasing, elaborative inferences and bridging inferences (Gernsbacher & Hargreaves, 1988; Pressley, Wood, Woloshyn, Martin, King, & Menke, 1992; Rosenshine & Meister, 1994).

iSTART has three modules: *Introduction* **(**students watch the teacher-agent explain the reading strategies to two student-agents); *Demonstration* (students are quizzed on various aspects of the SERT strategies); and *Practice* **(**students practice generating typed self-explanations while the program provides feedback on performance). The practice section incorporates feedback that is adaptive to the level of student performance.

Empirical studies with high school and college students suggest that iSTART improves both reading comprehension and the quality of self-explanation during the process of reading (e.g., McNamara, O'Reilly, Best, & Ozuru, 2006; O'Reilly et al, 2004a, 2004b). The current IES grant aims to scale-up iSTART so that it can be more widely accessible to users of different abilities.

iSTART: The System We Envision

iSTART targets the kind of active reading strategies that research has shown to be characteristic of skilled, successful readers (e.g., Bereiter & Bird, 1985). These reading strategies are particularly important for understanding textbooks because they often include unfamiliar, challenging material. However, teachers are often inadequately prepared to provide reading strategy training to the students in their classrooms (see, Snow, 2002). While they may be well versed in their particular domain, it is frequently the case that they have little training with regard to reading instruction, or how to incorporate reading strategy training into their classrooms (Baker, 1996; Snow, 2002). This impediment is augmented by the wide range of students' needs (Cornoldi & Oakhill, 1996) – students each have particular reading deficiencies and learn reading strategies at different rates. Teachers clearly do not have time to give individualized reading strategy training to each of their students.

iSTART offers a solution to this educational problem by providing automated reading strategy training that is adaptable to students' needs and rates of progress while taking into consideration the teacher's course demands. However, to fulfill such an active role in classrooms, the current version of iSTART must be augmented in two principle ways: (1) by improving the ability of the system to adaptively respond to students' and teachers' needs and, (2) by creating a teacher interface that allows teachers to integrate iSTART into their classrooms.

Incorporating a computerized tutor into a classroom is not as simple as merely giving the system to the teacher and expecting that it will be used consistently or successfully. There are many constraints that must be met in order for the system to be integrated into classroom in an effective manner. First, the teacher must understand the need for reading strategy training, and be receptive to intelligent tutoring systems. Second, the program must be easy for the teacher to use, and the system should also include a component that handles the teachers' questions about the program and data on the students' progress. Thus, a subcomponent called the teacher interface must be developed to facilitate teachers' use of iSTART in the classroom. Third, the number of course topics and range of text difficulty covered during the practice sessions need to be increased to make the system applicable to a wide variety of educational topics and students with varying levels of ability.

We have completed year three of the three-year funding period. We will complete our research during the up-coming fourth year (with a no-cost extension). The summary below briefly describes our third year of progress toward our four objectives. Following this are more detailed summaries of the project goals and sub-goals and our progress toward accomplishing our goals. Overall, we have made excellent progress in achieving the goals of our grant proposal.

SUMMARY OF PROGRESS

Objective 1: Text Domains. Our first goal is to increase the number of texts used in the self-explanation practice module. By incorporating history and narrative texts into the existing practice texts (currently of the domain of science), students with varying ability levels will be able to use iSTART across three curriculum domains. We have collected a corpus of 180 texts, aligned with the national education standards from different genres. The iSTART system has been expanded to be able to evaluate self-explanations in the additional domains of history and literature. Additionally, we are improving the self-explanation evaluation model by including a fuzzy-logic inference engine (Bellissens & McNamara, 2006). We conducted an experiment in which high school and college students were trained to self-explain with iSTART using the standard science texts. However, in addition to being assessed on the quality of their self-explanations on science texts, students were also required to generate self-explanations for *history* and *literature* texts. The data from this study will help us to evaluate the effects of iSTART training on students' ability to self-explain texts across different subject domains. In addition, it will provide valuable data for the self-explanation evaluation algorithm, allowing better calibration which will result in more precise and accurate feedback to the trainee. We are currently analyzing this data and preparing several papers for publication.

Objective 2: Adaptability to the User. Our second goal is to match the type and difficulty level of the strategy training to the ability level of the reader; that is, tailoring the training to each student's background knowledge and reading ability. Matching text to readers primarily consists in knowing how readers self-explain after certain kinds of discourse stimuli, and in tuning the practice according to what we know about the relationship between text structure and student self-explanation. The basic principle here was to find a way to define those textual stimuli that could influence the reader's self-explanation and interact with reader's characteristics such as reading skills and prior knowledge. We have conducted several experiments to address this issue.

Objective 3: Responsiveness to Strategy Deficits. Our third goal is to expand the training modules so that we can match the reading strategy training to the level of the reader. For example, during initial training, a low-ability reader should master lower-level strategies, such as paraphrasing, whereas a high-ability reader should focus on learning higher-level strategies, such as constructing inferences through bridging inferences or elaboration. As such, we will therefore refine the reading strategy instruction so that readers are taught strategies at the appropriate level. We have conducted an experiment that examines this issue. We have also created pretest assessments of students' prior domain knowledge in science, history, and literature. Prototypes of five new strategy modules were created: *What is Reading* teaches basic concepts about reading, *Previewing* teaches the students to look at the main parts of the text and get an idea of what a text is about *before* actually reading the text; *Paraphrasing* provides more practice on paraphrasing, particularly for less skilled readers; *Question Asking* teaches students to ask deeper level questions; and *What's the Main Point* teaches students to concisely convey the main points of the text. In addition to these new modules, we have also created a more adaptive version of the question manager underlying our Demonstration module. The modifications to this module make it possible to use a range of texts with different difficulty levels (see Subgoal 2.1) and also adjust the difficulty level of the questioning according to the student performance.

Objective 4: Teacher Interface. Our goal is to create an interface that enables teachers to use iSTART effectively. The interface serves three purposes: (1) it provides necessary information and training for teachers to use iSTART, (2) it allows some degree of freedom for teachers to adjust the training program by selecting texts of their choice, and when they will be read and (3) it gauges students' progress in the training program. We have developed the first versions of most of the components of the teacher interface. One part of the *Training Organizer and Management Module* (TOM), allows teachers to sign-up their students by electronic copy and paste from their class rosters, add students one at a time, transfer students to other sections, and keep their iSTART rosters synchronized with their class rosters. Also, the scheduling and policy features allow teachers to set up iSTART and integrate its library of readings with their classroom lessons by topic or specific reading while maintaining adaptation to the abilities of the student. As a part of the *Teacher Instruction Module* (TIM), we have created a version of the iSTART trainer that provides a "hands-on" tour of the modules. We have also prepared a set of web pages that provide information to help the teachers understand the theoretical aspects underlying reading strategies. Lastly, a teachers' needs assessment was performed in which the Center for Research in Educational Policy (CREP) administered surveys to teachers from five high schools.

PROJECT GOALS AND PROGRESS REPORT

Objective 1: Text Domains

Our first objective is to increase the number of texts used in the self-explanation practice so that iSTART can be used across curriculum domains by students. Increasing text domains used in the practice session contributes to increasing the systems' ability to adapt to students with varying ability levels. The current version of iSTART provides self-explanation practice through two science texts that focus on the development of thunderstorms and the origin of coal. While this has been sufficient to conduct our initial evaluations of the system, it is not sufficient for our goal to take iSTART to a larger scale for the following reasons. First, increasing the number of texts is essential in order for struggling students to receive a greater amount of practice than is currently possible. Although practice with two texts provides substantial benefits, our research has shown that this is not enough practice for the students most in need of training. Second, our experiences with students and teachers over the past few years have made it evident that the reading strategy training offered by iSTART should be available in additional domains. For instance, in our previous research teachers have reported to us that even though the students were trained to use self-explanation in their science classes, the students would frequently also use the new strategies for their courses in history and literature.

Expanding the number of topics covered by iSTART is beneficial for both applied and theoretical reasons. First, teachers will be more receptive to using iSTART in their classrooms if topic-relevant material (or better yet, selections from their course textbook) is available for the students to read using the iSTART system. In other words, iSTART will be better used by teachers if it allows coverage of required content while their students are learning the reading strategies. In theoretical terms, increasing the number of topics covered will also contribute to our understanding of reading processes. It will both afford us the opportunity to examine how different strategies are used and contribute to comprehension across different domains.

Our research plans to meet Objective 1 are represented as subgoals in Table 1. The table indicates progress made with each subgoal and the period in which we expect to complete the work.

Table 1: Subgoals for Objective 1

Subgoal	Description of activity	Progress	Time frame
1.1	Create a library of digitalized texts	Completed	Years 1 and 2
1.2	Create a self-explanation evaluation system t easily adaptable to a wide range of texts	Completed	Years 1 and 2
1.3	Evaluations of the feedback system using self-explanations from narrative text (Study 1)	Completed	Years 1 and 2
1.4	Experimental study involving collection of self-explanations for science, history and narrative texts	Completed	Years 1 and 2

Subgoal 1.1: This subgoal involves building text libraries for iSTART practice by increasing the number of texts available in the self-explanation practice so that iSTART can be used more extensively both within curriculum domains (e.g., different topic within science) and across curriculum domains (e.g., science, history, and literature). In order to facilitate the widespread use of iSTART, the selection of the text passages was aligned with the associated national education standards. We are continuing to collect these materials and evaluate them in terms of their difficulty levels (see Subgoal 2.1 for more information). So far we have collected more than 100 science texts, 50 history texts and 30 narrative texts. With respect to science texts, focus is given on biology and physical science texts, but other domains such as earth science and life science texts are also included. Further, these texts are catalogued in terms of the specific area of national or state (i.e., Tennessee) standard each text covers and various text features that are assumed to affect comprehension difficulty (i.e., vocabulary, sentence length, argument overlap, etc.). History and literature texts have been collected from various sources including electronic texts located in internet sites. In collecting these texts, care was taken to make sure that that content is truthful (in case of history texts) and the texts cover a variety of topics from world to North American history. Hence, overall we have a relatively comprehensive text library that is organized in terms of domain, topic, state/national standards it covers, and various aspects for text difficulty to accommodate the various demands of training situations present high school.

Subgoal 1.2: This subgoal involves the refinement of a self-explanation evaluation system. This system gives rapid evaluations and responses to student self-explanations. It can also easily be adapted to a wide range of texts. The self-explanation evaluation algorithms must be easily applied to a wide variety of texts. Our first version of an algorithm to achieve this goal is described in McNamara, Boonthum, Levinstein, and Millis (2006). We have created an evaluation system that can be automatically applied to virtually any text and is highly predictive of self-explanation quality. This revised system was used in the experiments conducted during the 2004-2005 school year. Informal examinations of the accuracy of the system's evaluations of self-explanations over a variety of texts indicate that the system performs quite well. After we completed the human coding of the protocols, which was necessary for a systematic evaluation of the system, and in the continuity of the current model, we decided to improve the model by including a fuzzy inference engine (Bellissens & McNamara, 2006). The fuzzy inference engine is able to abstract rules from human ratings to determine the quality of the self-explanation that the students produce. We are enthusiastic about the preliminary results we obtained because the protocols we used involved a wide variety of texts and participants. However, the expected improvements in performance evaluation with the application of this algorithm did not show up in more extensive analyses for several reasons. Specifically, we needed to find new variables that described the relationship between a self-explanation and the previous text. We proposed to use *Entailer* fill this gap (McCarthy, Rus, Crossley, Bigham, Graesser, & McNamara, 2007).

The purpose of Entailer is to evaluate the degree to which a self-explanation is entailed by a source text (Rus, McCarthy, Lintean, Graesser, & McNamara, 2007). Entailer uses minimal knowledge resources and delivers high performance compared to similar systems. The approach encompasses lexico-syntactic information, negation handling, and synonymy and antonymy embedded in a thesaurus (WordNet; Miller, 1995). Entailer functions by having each pair of text fragments (assigned as text [T] and hypothesis [H]) mapped into two graphs, one for T and one for H, with nodes representing main concepts and links indicating syntactic dependencies among concepts as encoded in T and H, respectively. An entailment score, entail(T,H), is then computed quantifying the degree to which the T-graph subsumes the H-graph (see Rus, McCarthy, &

Graesser, 2006, for a full discussion). The score is the weighted sum of one lexical and one syntactic component. The lexical component expresses the degree of subsumption between H and T at word level (i.e., vertex-level), while the syntactic component work does the same thing at syntactic-relationship level (i.e., edge-level). The derived Entailer score is so defined as to be non-reflexive, such that entail (T,H) does not entail (H,T).

We have also created three semantic spaces for the LSA algorithms that are part of the system. LSA is a statistical model of semantic memory. Theoretically, LSA uses induction processes to transform specific co-occurrences between words and contexts in a large corpus of text into abstract relations, called similarities, between words of the corpus. As one can use LSA to evaluate similarities between words, between words and passages, or between two passages, LSA has demonstrated its usefulness in estimating the quality of a text production in comparison to the text segment to which it refers (Graesser, Wiemer-Hastings, Wiemer-Hastings, Harter, Person, & the TRG, 2000; Landauer, Laham, & Foltz, 2003; McNamara, Boothum, Levinstein, & Millis, 2007; Magliano, Weimer-Hastings, Millis, Muñoz, & McNamara, 2002; Wolfe & Goldman, 2003).

The semantic spaces (or large corpora of text used in LSA) contribute to evaluations by providing a semantic comparison of the self-explanation to the target text. We have created LSA spaces for narrative and science texts, and a mix of those two. Evaluations of the narrative space have indicated that it does not perform well (i.e., it does not produce results equivalent to or as good as other spaces). As a result, we have decided to use Science and General TASA spaces in iSTART evaluations. Hence, we use those two spaces not only to evaluating the self-explanation in practice module, but also to estimate difference of self-explanation quality in pretest and posttest – before and after the iSTART trainer – explained by training.

Subgoal 1.3: This subgoal requires coding self-explanation protocols across a wide variety of texts to support the revision and evaluation of the system's algorithms. In the first year, we completed a preliminary study in which high school students self-explained four types of texts (science, literary analysis, narrative, and wellness) after receiving iSTART training with science texts. In the second year, the protocols from this study were rated by human raters in terms of strategy use and quality. These ratings have been compared to the judgments made by the current system (McNamara et al., 2006; McNamara et al., 2004; Millis, Kim, Todaro, et al., 2004) and the fuzzy system (Bellissens & McNamara, 2006). The match between trained raters' judgments and the current system's judgments (in terms of d-primes and Cohen's Kappa) indicates the appropriateness of the judgments made by the system, although the match was better with the fuzzy system. We have completed the ratings of the protocols collected in the transfer experiment described in goal sub-goal 1.4. Unfortunately, with those protocols, the fuzzy logic system does not produce better results than the former system without the fuzzy engine. As a result, the former system will most likely be combined with Entailer (McCarthy et al., 2007) for the final version of iSTART. Entailer offers a number of advantages over current text relatedness measures such as LSA and overlap indices. First, Entailer addresses both syntactical relations and negation, which is not the case with LSA. Second, Entailer addresses asymmetrical issues by evaluating text non-reflexively, so $\text{entscore}(H,T) \neq \text{entscore}(T,H)$. As such, the evaluation of a self-explanation to a source text will be different from the evaluation of the stimulus to the response. Third, Entailer handles negations; hence, it offers the opportunity of providing more accurate feedback. Currently, Entailer is not equipped to handle problems such as misspellings and typos any more than other text relatedness measures. We are working on the implementation of Entailer in a complete feedback system.

Subgoal 1.4: This subgoal involves the testing of the transfer of self-explanation training on science texts to texts from three domains: science, history, and literature. This experiment also allowed us to collect additional self-explanation protocols from history and narrative domains.

We investigated the effectiveness of iSTART in facilitating the production of quality self-explanation of non-expository text. The experiment was a within-subjects design with genre (science, literature, history) and time of test (pre, post) as within-subjects variables. The six texts (2 x 3 genres) were selected and equated on the number of words, Flesch-Kincaid Grade level, and argument overlap. During the pretest, students typed SEs for eight target sentences within each of three counterbalanced texts (science, literature, history) presented in a random order. During the iSTART training, students progressed through the three main sections of the program: *Introduction*, *Demonstration*, and *Practice*. During the posttest, students typed SEs for eight sentences for each of the remaining three texts (science, literature, history). The quality of students' SEs was evaluated by the iSTART algorithm (McNamara, Boonthum, Levinstein, & Millis, 2007) which rates the quality of SEs on a 0-3 point scale, with 3 being the best score.

The experiment included 272 students (77 high school, 195 college). These data have been collected and are currently being analyzed. Our initial pilot data (69 college students) have been analyzed and were presented at this year's Society of Text and Discourse conference (O'Reilly, Taylor, Duran, & McNamara, 2006). The participants were 66 undergraduate students who self-explained text from three genres: science, history, and literature. The quality of their self-explanations was evaluated using the iSTART algorithm. As can be seen in Table 2, self-explanation quality increased from pretest to posttest for each genre. These differences were significant, science: $t(65) = 2.28$, $SD = .63$, $p < .05$; history: $t(65) = 2.03$, $SD = .70$, $p < .05$; literature: $t(65) = 3.64$, $SD = .56$, $p < .01$. Thus, our preliminary evidence for college students showed evidence that iSTART training transfers to other text genres.

Table 2. Pilot study: Pretest-posttest self-explanation quality.

Genre	Pretest	Posttest
Science	1.80	1.98
History	1.90	2.08
Literature	1.95	2.20

We sought to replicate this finding for college students and compare these findings to a control, no-training group. During the training session, participants in the control group took a tutorial in web design using the same texts as those read in the experimental training session. There were 58 participants in the iSTART condition and 61 in the control condition. See Table 3 for the self explanation scores for each group and genre. A 2 (training group: iSTART vs. control) x 2 (testing occasion: Pretest vs. Posttest) x 3 (genre: science vs. history vs. literature) ANOVA revealed an interaction between training group and testing occasion, $F(1, 117) = 4.53$, $MSE = .251$, $p < .05$. Follow-up *t*-tests indicated that self-explanation quality decreased from pretest to posttest for the control group but (*M* pretest = 1.85; *M* posttest = 1.72), but there was no change for the iSTART group (*M* pretest = 2.14; *M* posttest = 2.17). There was also a main effect of genre, $F(1, 234) = 13.18$, $MSE = .120$, $p < .001$, showing that overall self explanation quality was lower for science (*M* = 1.88) than the other genres (*M* history = 2.02; *M* literature = 2.02). Planned comparisons were conducted to test the transfer of training to the other text genres. For the iSTART group, these comparisons showed no evidence of transfer except for a trend for history texts ($p = .11$). As a comparison, self explanation quality decreased from pretest to posttest for the control condition in all three genres (science: $t(60) = 1.95$, $SD = .49$, $p = .056$; history: $t(60) = 2.05$, $SD = .52$, $p < .05$; literature: $t(65) = 2.31$, $SD = .48$, $p < .05$. This study did not replicate the transfer effects found in the pilot study. However, it should be noted that this may have been due to differences in the population given that performance actually dropped in the control condition from pretest to posttest. In this perspective, iSTART prevented a drop in performance and slightly increased self explanation quality for history texts (although this was only a trend).

Table 3. iSTART vs. Control: Pretest-posttest self-explanation quality.

Group	Genre	Pretest	Posttest
iSTART	Science	2.06	2.05
	History	2.14	2.28
	Literature	2.22	2.18
Control	Science	1.76	1.64
	History	1.89	1.75
	Literature	1.91	1.77

We also tested the effects of iSTART on transfer in a high-school population. All participants received iSTART training. The mean self explanation quality for each genre and condition are presented in Table 4. A 2 (testing occasion: Pretest vs. Posttest) x 3 (genre: science vs. history vs. literature) ANOVA revealed a main effect of genre, $F(1, 234) = 10.95$, $MSE = .114$, $p < .001$, showing that overall self explanation quality was lower for science (*M* = 1.89) than the other genres (*M* history = 2.07; *M* literature = 2.03). Additionally, there was a marginally significant interaction between testing occasion and genre, $F(1, 138) = 2.50$, $MSE = .122$, $p = .086$. Planned comparisons were conducted to test the transfer of training to the other text genres. There was a marginal increase in self explanation quality for history texts, $t(69) = 1.75$, $SD = .63$, $p = .084$, no change for literature texts, $p = .82$. Additionally, there was a significant increase in quality for science texts, $t(69) = 2.12$, $SD = .62$, $p < .05$. These data suggest that training did not transfer to literature texts for high-school students, but did have a marginal transfer for history texts.

Table 4. High-school students: Pretest-posttest self-explanation quality.

Genre	Pretest	Posttest
Science	1.81	1.97
History	2.01	2.14
Literature	2.04	2.02

We were further interested in what pretest measures would predict transfer of training after iSTART. A transfer score was created for the literature genre and the history genre by subtracting posttest self explanation quality scores from pretest scores; the higher the resulting score, the more successful the transfer. Table 5 presents a correlation matrix among the two transfer scores and the pretest measures (described in Subgoal 3.1).

Table 5. Correlations among literature and history transfer scores and pretest measures.

Transfer score	Gates Reading Score	PK Literature	PK History	MSI Strategy Knowledge	Good reading goals	Epistemology of Social World	Epistemology of Physical World
Literature	.07	.07	-.06	-.01	-.06	-.12	-.05
History	.19	.18	.24*	.15	.11	.14	-.02

* $p < .05$

These data show that the only the history genre transfer score significantly correlated with a pretest measure, i.e., knowledge of history. Albeit significant, it is a weak correlation. Nonetheless, it is possible that having knowledge of history helped students apply self explanation strategies learned while comprehending science texts to the history domain. Similar to science texts, one main purpose of history texts is to inform the reader. It may be that this similarity may facilitate transfer of reading strategies, as long as there is sufficient knowledge of that domain. Literature texts mainly tell a story, so it is possible that different self explanation strategies are more appropriate for literature than science or history texts.

Overall, these experiments provide some evidence that self explanation training in the science domain may transfer to other domains, with the exception of literature. Nonetheless, the evidence on the whole suggests that strategies may be domain specific; students should practice using reading strategies within the target domain. Thus, we do need to provide practice texts that cover a range of domains.

Objective 2: Adaptability to the User

The second objective is to make the iSTART system more adaptable to a wider range of students. A major step toward this goal is increasing the range of difficulty of the training texts. The previous training texts were targeted at high school students. However, for some readers, these texts may have not been challenging enough to stimulate processing at the zone of proximal development (Brown, Ellery, & Campione, 1998; Hung, 2001; Luckin, 2001; Murray & Arroyo, 2002, 2004; Vygotsky, 1978). Our past studies with high school students indicated that students varied greatly in terms of their reading skills, including their ability to self-explain texts using reading strategies (Best, Ozuru, & McNamara, 2004; McNamara et al., 2006). Whereas some advanced students can use elaboration based on general world knowledge and personal experiences, many students often struggle with producing paraphrases when a target sentence is long, complex, and/or contains unfamiliar words. Objective 2 is aimed at providing practice that adapts to students' current ability level.

Students' ability levels have been measured through a collection of assessment instruments (e.g., prior domain knowledge, reading skills, learning goals, etc.). We acquired texts at varying levels of difficulty (e.g., junior high school, high school and Freshmen College) based on national educational standards. We also computed the difficulty level of the texts that students self-explained by defining some cohesion characteristics of the texts (see Objective 2.1) and we tested successfully the relevance of that computation. In addition, we have analyzed texts and data from study 2 (see Objective 1.4) in order to investigate the relationship between text difficulty (in the domains of science, history, and literature) and self-explanation quality. That analysis allowed us to define within the texts, three categories of sentences as a function of their dependency to previous sentences in a text, and the level of dependency can predict the quality of self-explanation. As a result, a sentence dependency program was implemented into iSTART to choose target sentences as a function of trainees' needs. Finally, an experiment in which we varied text difficulty and sentence dependency has been conducted. We are currently coding the participants' verbal protocols.

Our research plans and progress for meeting Objective 2 are represented as subgoals in Table 6.

Table 6: Subgoals for Objective 2

Subgoal	Description of activity	Progress	Time frame
2.1	Increase range of text difficulty	Completed	Years 1 and 2
2.2	Matching the text to the reader (Study 3)	Completed	Years 2 and 3

Subgoal 2.1: Increase range of difficulty. This subgoal involves increasing the range of text difficulty and finding a method to determine the relative difficulty of group of texts according to educational standards but also to indices of textual cohesion and vocabulary accessibility. We defined difficulty according to two dimensions. The first one is the *difficulty between texts*, which was calculated using a multifactor index that included text cohesion and word difficulty. The between-text difficulty indices allowed us to categorize texts, from the same domain, as difficult or easy (Duran, Bellissens, Taylor, & McNamara, 2007).

The second dimension is the *difficulty within texts*, that is, the degree to which each text sentence is dependent or independent of previous sentences (i.e., how interrelated they are), which can influence inference generation and the quality of self-explanations as a function of readers' characteristics (Bellissens, Jeuniaux, Duran, & McNamara, 2007). The notion of within-text difficulty is also intended to help us to determine which sentences should be used as target sentences for self-explanation to maximize the benefit of training for students with different levels of ability.

Difficulty between texts. The texts contained in the text library described earlier have been analyzed and indexed in terms of various relevant text features that affect their difficulty level. We have acquired science texts at varying levels of difficulty (e.g., junior high school, high school, and college), based on national education standards. We have determined which of the many Coh-Metrix indices of text characteristics (Graesser McNamara, Louwerse & Cai, 2004; McNamara, Louwerse, & Graesser, 2002) are indicative of difficulty. Analyses were conducted within a representative sample of the larger text library. This total sample consisted of 60 texts (M=400 wds), extracted from 23 science textbooks. Twenty subtopics were covered including 10 *physical science* and 10 *life science* subtopics. Each subtopic included an excerpt from junior high-school, high-school, and freshman-college levels.

For each text, we used Coh-Metrix to compute 30 referential overlap and 24 vocabulary accessibility indices. *Referential overlap* indices represent a factor of cohesion that approximates conceptual redundancy and relatedness between sentences. High values of referential overlap can assist readers in integrating content into a coherent mental representation of the text (McNamara & Kintsch, 1996). *Vocabulary accessibility* indices represent a factor of semantic information that varies in word familiarity, ambiguity, and abstractness. These word characteristics are important in influencing the activation of concepts from memory while reading (Freebody & Anderson, 1983).

The large number of Coh-Metrix indices were then reduced to six indices, three for each group, by a Principle Component Analysis (PCA), with varimax rotation. This resulted in referential overlap indices that represent content word overlap in adjacent sentences, and vocabulary accessibility indices that represent the average standard deviation of word concreteness and imageability.

To identify our hypothesized difficult and easy-texts, we evaluated the texts in an unsupervised cluster analysis. A text was considered *difficult* if it was assigned to a cluster with the lowest mean scores for referential overlap and vocabulary accessibility. Conversely, a text was considered *easy* if it was assigned to clusters with the highest mean. Thus, we selected eight texts, representative of four *life science* topics, with each topic containing a difficult and easy version.

To validate the identified difficult and easy texts, 24 participants read four texts (two easy, two difficult). When the text had been read, participants were instructed to type their recall. The dependent variables of recall and reading time per sentence were recorded.

The mean number of words recalled for the difficult and easy texts was then compared. A one-way within-subjects ANOVA demonstrated a significant effect for type of text (difficult/easy), $F(1,23) = 41.80$, $p < .001$, $\eta^2 = .655$. Participants recalled more from the easy texts than from the difficult texts. The mean number of idea units recalled for easy texts was 12.09 (SD = 4.55) and the mean number of idea units recalled for difficult texts was 8.37 (SD = 2.92).

We also used LSA to evaluate the contextual similarity between the free recall and the text from which the recall was generated. This allowed us to asses the qualitative differences of difficult and easy texts. The LSA values for each of the four recalled texts were submitted to a one-way within-subjects ANOVA. There was a significant effect for type of text (difficult vs. easy), $F(1,23) = 13.19$, $p < .001$, $\eta^2 = .528$. Participants' recall was more contextually similar to the text for the easy texts than for the difficult texts (see Table 2). The mean LSA value for the overlap between easy texts and recall was .78 (SD = .07) and the mean LSA value for the difficult-texts and recall was .69 (SD = .11).

To assess the effect of reading time, sentences were first normalized for differences in length by dividing reading time by number of words. Reading times of +/- 2 SD (< 2% of data) were removed. A one-way within-subject ANOVA demonstrated a statistically significant effect when reading times were normalized by number of words $F(1,23) = 8.09$, $p < .01$, $\eta^2 = .269$. This result suggests that participants spend more time reading difficult texts. The mean reading time per word (ms) for easy-texts was 363.23 (SD = 19.74) and the mean reading time per word (ms) for difficult texts was 426.63 (SD = 34.25).

This initial investigation contributes to Subgoal 2.1 by demonstrating that Coh-Metrix indices can significantly identify texts that have unique influences on human comprehension. Further, the described method is a theoretically grounded technique that can be used to increase the range of text difficulty. We are currently in the process of implementing the indices into an algorithm for automatically identifying the difficulty of any text, including our complete text library.

Difficulty within text. A large body of research has addressed how the linguistic representation of a text guides the formation of bridging or elaborative inferences during text comprehension (Singer & Remillard, 2004). Some authors have also demonstrated that even though poor readers were not able to generate elaborative inferences automatically during reading they could encode them to use later (Murray & Burke, 2003; Schmalhofer, McDaniel & Keefe, 2002). It results that the relationship between trainee characteristics and self-explanation quality is complex and probably depends on how text cohesion structure is conceptualized. For instance, viewing text cohesion structure either as content word overlap between sentences versus as causal links relations between sentences, leads to different expectations about text comprehension and explanation. Indeed, an implicit causal link device, such as *causal verbs*, demands more knowledge to be recognized, or to be processed, than do *causal particles* (Sanders & Noordman, 2000) or content word overlap (Kintsch, 1998). The former depends on knowledge and the latter is an obvious cohesion device for many readers.

Hence, we designed a model that computed text cohesion structure as a function of different indices of textual cohesion (content words overlap, semantic similarity, causal link). The output of the model was the dependency levels of text sentences. According to the model, if a sentence shares many content words with previous sentences in the text, that sentence is considered as dependent, and as independent if the sentence does not share content words with previous sentences. The same reasoning has been applied to semantic similarity and causal links indices between sentences.

We applied the model to materials and data from Study 2. Two experts scored 3696 self-explanations from the study 2. They judged whether participants added information in their self-explanations in comparison to the target sentence, and whether added information came from the text itself (bridging inference) or from information outside of the text (elaborative inferences). Of the 3696 self-explanations; 9% contained both bridging and elaborative inferences, 49% consisted of only paraphrasing or repeating the target sentence, 21 % contained bridging inferences but no elaborative inferences, and 21% contained elaborative inferences but no bridging inferences. The model was applied to the six texts used in the study 2.

The model defined the dependency of the eight target sentences per text, as a function of 1) content word overlap, then 2) semantic similarity, and finally 3) causal links. For each type of dependency, we conducted a two-way within-subjects ANOVA, including sentence dependency (Dependent, Intermediate, Independent) and the type of generated inference (Bridging, Elaborations), with the number of generated inferences as the dependent variable.

As a main result, sentence dependency had a significant effect on the type of inferences generated. Indeed, more elaborative than bridging inferences were generated when the target sentence was independent and fewer when the target sentence was dependent. That was true with a) content word overlap cohesion, $F(2, 152) = 12.35$, $p<.01$, b) semantic similarity cohesion, $F(2, 152) = 3.16$, $p < .05$, and c) causal link cohesion, $F(2, 152) = 24.85$, $p < .01$.

As an intriguing result, prior knowledge significantly interacted with Sentence dependency and Inference type only when the dependency was defined on a causal link basis. We used a mixed model with Sentence dependency and Inference type as within-subjects factors, and Prior knowledge as a between-subjects factor. The three-way interaction including Sentence dependency, Inference type, and Prior knowledge was significant: $F(4, 148) = 4.29$, $p < .01$. In a separate Sentence dependency by Inference type analysis, high-knowledge participants made significantly more Bridging inferences with Dependent than with Independent sentences, and more Elaborations with Independent than with Dependent sentence, $F(2,44) = 27.50$, $p < .01$. The same trend was found for Intermediate knowledge participants. However, Low knowledge participants generated the same number of Elaborative and Bridging inferences after an Independent target sentence, .21 vs. .22, respectively, in a separate analysis including only the Low knowledge participants, the interaction between Sentence dependency and Inference type was not significant, $F(2, 40) < 1$.

As a result, Sentence dependency is correlated to self-explanation quality. Dependent target sentences led mostly to Bridging inferences within the self-explanations and Independent sentences led mostly to Elaborative inferences within the self-explanations. Moreover, only knowledgeable readers were able to process implicit textual causal cohesion, as measured by our model.

Subgoal 2.2 (Study 3): The past year, we have addressed the question of matching text to reader through a study that included 74 college students that vary in terms of their reading ability and domain knowledge. In that study, the difficulty *between* experimental texts was manipulated, as describe above, to be at two different levels – low and high (see subgoal 2.1). The difficulty within the text was manipulated as well: each experimental text included 9 target sentences to be self-explained. Participants were asked to read 4 texts and self-explain 9 target sentences per text. Three of the target sentences were dependent, three were independent, and three were of an intermediate level of dependency. Seventy-four college students from the University of Memphis participated in the experiment for credits. We are still coding the protocols of that experiment.

We expect to observe an interactive effect such that low-ability readers will show greater inference generation (bridging and elaborative inferences) during high-cohesion text reading, and high-ability readers will show greater improvement using the low-cohesion text (see e.g., McNamara, Kintsch, Songer, & Kintsch, 1996). In addition, according to the theory of the zone of proximal development, the greatest level of improvement should occur when participants read a text that they find to be moderately difficult given their ability level. Hence, we predict a greater inference generation by self-explaining independent target sentence presented in Easy texts.

We are currently processing the newly collected data to determine the relationship between text difficulty and self-explanation quality. The analysis is scheduled to be completed by this summer.

Objective 3: Responsiveness to Strategy Deficits

The third objective is to improve the adaptivity of the system through an improved student model and the inclusion of additional strategy modules. This objective will be reached through several means. First, there will be a refined battery of individual difference measures. This will help inform the second subgoal of developing an improved student model. Third, additional strategy modules focusing on the less skilled reader will be developed. During this process, the system components will be assessed in terms of their usability. Lastly, the expanded iSTART systems' educational efficacy will be evaluated. The progress made in each of these three areas is reported in Table 7 below.

Table 7: Subgoals for Objective 3

Subgoal	Description of activity	Progress	Time frame
3.1	Refine battery of individual difference measures	In Progress	Years 1 -3
3.2	Creation and revision of student model of strategy needs	In Progress	Years 2 - 3
3.3	Expansion of strategies included in the system	In Progress	Years 2 -3
3.4	Assess system usability (Study 4)	In Progress	Years 1 - 3

Subgoal 3.1: This subgoal involves the development of improved individual difference measures. In previous research we have assessed students' competencies and aptitudes using a large battery of pretests prior to training using iSTART. The pretest measures included general reading skill (e.g., Gates-MacGinitie reading test), metacognitive reading strategy knowledge (Schmitt, 1990), metacognitive skills (e.g., Mokhtari & Reichard, 2002), general science knowledge, and specific science vocabulary knowledge. Over the past few years, we have collected data on thousands of participants that used our existing battery of pretest assessments. This large data set enabled us to evaluate our measures' ability to predict students' reading comprehension and learning gains. Our efforts have focused on evaluating the validity of our pretest measures in terms of predicting students' performance on science passage comprehension and self-explanation quality for science texts that were read both before and after training using the iSTART system.

While it is feasible to conduct two-hour pretest sessions in the laboratory, it is not so feasible to expect that teachers would be able to do so as well. Therefore, our goal is to maximize the predictive power of the pretest assessments while reducing the time necessary to administer these tests. Thus far, we have evaluated each pretest measure and constituent questions in terms of whether they predict students' comprehension performance and self-explanation quality on a variety of science texts. Items (or entire measures) that showed little or no predictive validity were eliminated.

We have made significant progress in refining the battery of individual difference measures. This included the development of a more differentiated pretest that better assesses students' pre-existing knowledge on science, history and literature domain, as well as the addition of new measures of learning strategy, epistemology, and motivational factors.

The development of the prior knowledge questions occurred in the following four phases. *Phase 1:* We analyzed the data from previous studies to select the most predictive items. The selection process consisted of two stages. First, we eliminated questions whose difficulty level (i.e., proportion correct) was either above 60% or below 30%. Second, we correlated each question with several different individual difference measures (e.g., reading skill, metacognitive knowledge) and also with student's performance on several comprehension tests. Questions with low correlations to these measures were eliminated. *Phase 2:* Additional questions were taken from a test bank at Oswego City School District, Regents Exam Prep Center. This resource is an online facility that helps high school students prepare for the New York state standards subject domain tests. Questions were selected based upon moderate difficulty and topics that were not covered by our existing measures. *Phase 3:* Questions were generated by two of the posttest doctoral fellows to fulfill our needs that were not met by existing assessments. These were developed by sampling topics covered in high school textbooks and brief study guides. We developed near-miss, thematic, and unrelated question distracters based on the guidelines of Graesser and Wisher (2001). *Phase 4:* A total of 55 multiple choice prior knowledge questions (18 science, 18 history, and 19 literature) were piloted with 15 undergraduate students to examine the item performance. Thirty questions (10 each for each domain) were selected based on the item performance such that no items selected had either a ceiling (> .9) or floor effect (< .25 = chance level performance). These 30 items were used in the transfer experiment conducted in the fall of 2005, and the spring of 2006. Overall item characteristics indicated that the items in each domain have reasonable variation in item difficulty without exhibiting ceiling or floor effect.

A second strategy for improving the individual difference measures was to include additional measures that have previously been found to influence learning. The new instruments were designed to measure students' learning goals, epistemological beliefs, and topic interest.

Interest. There is evidence indicating that students' engagement and learning from reading texts is influenced by their interest in the topic under study (e.g., Alexander, 1992), their learning goals (e.g., Taylor, 2004), and their epistemological beliefs (e.g., Daneman & Hannon, 2001). Preliminary analyses indicate that students' interest in the domain of science is significantly correlated with their ability to self-explain science texts prior to training, even after removing the effect of science prior knowledge and reading ability, thus indicating a unique contribution of domain interest to students' ability to self-explanation science texts (Taylor & McNamara, 2006). In addition, strategy training via the iSTART system resulted in a significant improvement in self-reported understanding of the science texts and self-explanation quality. At pretest, the measures of students' understanding were significantly correlated with their measure of interest in specific texts. However, after training, these measures were no longer significantly correlated. Hence, overall, iSTART training appears to improve readers' understanding while making it less dependent upon the learner being interested in the specific text (Taylor, Sinclair, & McNamara 2006). Similar analyses on the history and literature domains are under way.

Learning Goals. There are two primary learning goals used by students – maximizing learning or minimizing effort. Students seeking to maximize their learning are often required to exert greater effort and engage in what Bereiter and Scardamalia (1989) refer to as *intentional learning*. In contrast, some students do not seek to maximize their learning, but instead merely seek to minimize the amount of mental effort expended, taking an *effort minimization* approach. A nine-question measure of students' learning goals was introduced. Analyses of this new measure are in progress.

Epistemology. Another new measure that was introduced assessed the students' epistemology or their understanding of the nature of knowledge and knowing. Adapted from the work by Kuhn (Kuhn, Cheney, & Weinstock, 2000; Kuhn & Weinstock, 2002), this assessment helped to determine participant's epistemological level of understanding, holding that beliefs fall into one of three categories: (1) Always either true or false, (2) Relativist and merely a matter of opinion, or (3) Judgments requiring the backing of evidence. Analyses of this new measure are also underway.

Comprehension monitoring. One of the key individual difference measures in reading strategy training is metacognition, or more specifically the ability to monitor reading comprehension. In the past many researchers, including ourselves, attempted to measure comprehension monitoring ability using some form of self-report, assessing the whether students actually know "what to do" or "how to do it" when they need to monitor comprehension. But this approach has a number of problems including response bias. We began to pursue the possibility of measuring comprehension monitoring ability using a form of on-line procedure in which readers' sensitivity to their own reading time is measured. As the first step, we conducted an experiment that examined whether and to what extent readers are aware of their on-line processing difficulty by measuring the relations among sentence difficulty, reading time, and participants' subjective judgment of the difficulty level of the sentence.

We have analyzed part of the data, and this finding is published as a proceeding for the Cognitive Science Society 2007 (Kurby, Ozuru, & McNamara, in press). In this report we briefly describe the main findings of the experiment. Thirty-two undergraduate students read a psychology text on brain anatomy in one of the three reading conditions (i.e., read only, judgment of learning, or sentence difficulty judgment). The text was comprised of 62 sentences with sentence lengths varying from 4 to 49 words per sentence. The Flesch Reading Ease for each sentence was calculated using Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004). The scores for the 62 sentences varied from 0 to 96.7. The frequency distribution of sentences with different reading ease levels is presented in Table 8.

Table 8. Frequency distribution of sentences as a function of reading ease level

Range of Flesch Reading Ease	Number of sentences
0 -19	6
20 -39	18
40 -59	23
60 -79	13
80 -100	2

Prior to the main task, participants' reading ability was measured with Gates MacGinitie reading ability test. In the main task, participants read the text one sentence at a time. The *Read Only* condition participants were instructed to read the text carefully one sentence at a time. The judgment of learning condition (*JOL*) participants were asked to indicate how well they would answer a question based on each sentence using a 4-point scale (1=likely to be wrong, 4=likely to be correct). The sentence difficulty judgment condition (*JOD*) participants were asked to indicate their subjective difficulty estimate of each sentence using a 4-point scale (1=very easy, 4=very difficult) immediately after reading each sentence. Reading time of each sentence and the judgment responses for the judgment conditions were recorded. Immediately after reading the text and performing the judgment task, participants answered the multiple choice questions in the absence of the text.

In this report, we describe the analysis of the relationship between judgments and reading time and objective sentence difficulty. Table 9 below presents the correlation between the reader's reading time and judgments as a function of reading ability level measured by the Gates MacGinitie test. A positive value means that reading time increases (i.e., slows down) as judgments increase. The *t*-tests revealed that only the high skilled readers showed significant correlations between reading time and judgments (JOL-Low: $t(5) < 1$; JOL-High: $t(4) = 3.85$, $p < .05$; JOD-Low: $t(3) < 1$; JOD-High: $t(5) = 2.92$, $p < .05$). These results indicate that there are systematic differences between high and low skilled readers in terms of their ability to monitor processing fluency.

Table 9. Mean Beta Weights for each Group and Skill.

Condition	Reading Skill	Beta Weight	Beta Weight *SE*
JOL	Low	.04	.05
	High	.21	.06
JOD	Low	.08	.11
	High	.27	.09

However, this finding does not rule out the possibility that metacognitive judgments may be more directly predicted by text difficulty rather than processing speed (i.e., reading time) itself. That is, readers may be making judgments based on objective difficulty of the sentence, and the correlation is just a byproduct of this judgment. To address this issue, we conducted item-based analyses using hierarchical linear regressions on the mean judgments per sentence to investigate the contribution of reading time (RT) to judgments over and above a measure of sentence difficulty (Flesch Reading Ease: FRE). Specifically, we calculated average judgments for each sentence across participants separately for high and low skilled readers in each condition (JOL and JOD). Then, using mean judgment as the criterion variable, we performed hierarchical linear regressions using average reading time of each sentence and FRE of each sentence as the predictor variables. Flesch Reading Ease (FRE) was entered in the first step and average reading time per word (RT) was entered in the second step. We were interested in the ΔR^2 for reading time for each analysis. A separate analysis was run for each of the four groups (condition x reading skill). The total R^2 and ΔR^2 for each step and predictor (FRE, RT) for each analysis are presented in Table 10.

The first column of Table 10 (i.e., step 1) indicates that our measure of sentence difficulty was a significant predictor of judgments for all groups except for the JOL-Low skilled group. This result shows that when objective text difficulty is entered in alone, it accounted for a significant amount of variance in participants' judgment ratings of the sentence. However, looking at the results reported in the third column (i.e., ΔR^2), we see that the skilled readers in the JOD condition based their judgments on their reading time over and above text difficulty. These readers appear to base their judgments on both text difficulty and on their assessment of their processing difficulty at the time of reading. This suggests that at least skilled readers take into consideration the subjective difficulty of their text processing (i.e., reading time) in addition to the text features when making sentence difficulty judgments. In contrast, skilled readers' JOLs were predicted by objective text difficulty alone. Finally, less skilled readers' judgments, in particular in the JOL condition, were related to neither sentence difficulty nor reading time in a systematic way. Overall, the experiment indicated that there is systematic variability in people's ability to monitor their own processing fluency when reading complex sentences, and indicates that measurement of people's ability to judge processing fluency based on reading time, taking into account text difficulty, might be a productive approach to measuring individual differences in metacognition.

Table 10. Total R^2 and ΔR^2 for each step, condition, and reading skill

Condition	Reading Skill	R^2		ΔR^2
		FRE (step 1)	RT (step 2)	
JOL	Low	.000	.000	.000
	High	.230*	.232*	.002
JOD	Low	.068*	.070*	.002
	High	.159*	.245*	.086*

* $p < .05$

RSAT. We have begun to include an assessment tool called the Reading Strategy Assessment Tool (RSAT: Magliano & Millis, 2004) which is funded by the IES, as a method to assess students' reading comprehension ability. In this tool, participants read texts on a computer, and either answer a question or report the content of their on-going thought-process at selected sentences in the context of what they've understood so far about the text. This assessment technique, hence, is based on a think aloud paradigm. This tool provides an assessment of a participant's online comprehension by providing an insight into contents of working memory related to comprehension. In addition, the questions provide an assessment of the participant's comprehension of the text. Each think-aloud protocol is assessed in terms of the presence of paraphrases, bridging inferences, and elaborations. For paraphrases, it counts the number of words in each protocol that overlaps with the current sentence. For bridging inferences, it counts the number of words in each protocol that overlap with the immediate prior sentence (local bridge), and counts the number of words that overlap to prior text sentences beyond the immediate prior sentence (distal bridge). For elaboration, it computes two scores which are then added. It counts the number of words in each protocol that do not match any word in the previous text and current sentence, and counts the number of words in each protocol that match words in sentences not yet read. The answers to the questions are evaluated based on the number of words in each answer that overlap with words in the ideal answer. This tool then provides five scores: 1) a paraphrase score, 2) a local bridging score, 3) a distal bridging score, 4) an elaboration score, and 5) a score of question answering performance. These scores can be used as proxy measures of the quality of comprehension process. We are conducting further investigations both independently and in collaboration with the Magliano and Millis lab to fine tune the use of this tool as an individual difference measure.

Subgoal 3.2: This subgoal involves the creation and use of an improved student model of strategy needs. The model will be based on data gathered from the pretest battery and from the student interactions in the course of training. We envision a dynamic student model that makes use of pre-assessment data to initiate training and then is continuously updated as a function of training performance data, such as response times and question answering accuracy (e.g., strategy quiz performance). For example, in the introduction, the student is given a brief quiz after each strategy is introduced. In the demonstration section, we learn which strategies the student is able to, or prefers to identify. In the practice section, we learn how well the student is able to explain the texts and to identify strategies that are used. As this information is added to the student model, it will be possible to predict which avenues of iSTART training are most valuable as the student progresses.

We intend to make use of theoretical guidelines as well as empirical data to direct training. Both theory and data will guide options such as what level of scaffolding to provide, what strategies to emphasize, the number of examples to provide, and the amount of practice to offer. Constructing and revising a model of the student throughout training will allow us to map the characteristics of the student to an appropriate program of training. To this end, we will use data collected from previous experiments to examine the link between reader aptitudes (pretest assessments mentioned above) and the success with which students learn and use strategies incorporated in iSTART (e.g., paraphrasing and bridging). Specifically, we aim to identify how students' level of reading skill and prior topic knowledge is associated with students' performance on the iSTART module tasks (quizzes incorporated in the Introduction module) and the use of strategies (self-explanation quality rating generated in the practice module). The overarching idea is to determine which reading strategies are appropriate for students of differing ability levels. For example, students with poor reading skills may demonstrate a poor knowledge of higher-level strategies (e.g., elaboration and bridging) and thus benefit more from learning the lower-level strategies (e.g., paraphrasing). Conversely, students with high reading skills may demonstrate a better grasp of the higher-level strategies and thus benefit from focusing solely on higher-level reading strategy training.

We have conducted a series of analyses to gain an understanding of possible student models that represent current ability and potential to effectively comprehend the various text materials. One of the analyses examined the profiles that are indicative of relatively good and poor self-explainers. The other analyses examined the profiles that are indicative of learning potential; that is, whether a student is likely to improve ability to self-explain based on iSTART training.

Profile analyses serve to reveal patterns of results that are indicative of different groups of participants. That is, the goal of the analyses is to obtain an overall "picture" of reader characteristics that can be used to determine the nature of training to-be-provided to students. Profile analysis involves between-group multivariate analyses which compare performance on multiple measures within and between groups. These analyses consider each measure as a separate DV.

A profile analysis included three tests. The first is a test for flatness of the profiles. That is, does performance differ among the different measures? A main effect of measure-type would indicate that the scores are not "flat." In other words, presence of a significant main effect indicates that performance averaged across groups differs significantly between different measures. For example, participants' scores on the reading ability test might be higher relative to their scores on self-explanation ability. As such this analysis helps us gain an understanding of the overall landscape of student features. This analysis requires that all measures are on the same scale. As such, all measures were proportioned by dividing mean performance of each measure by the highest possible score. This created a score that ranged from 0 to 1 for each measure. For every measure, a score of 1 means perfect performance and 0 means completely inaccurate performance.

The second test examines the differences in levels in terms of all the dependent variables. That is, do the different groups of subjects score differently on the dependent variables overall (i.e., collapsed across all measures)? Hence, a significant effect in this analysis would mean that the dependent variables included in the analysis effectively differentiate two groups of participants (high self-explainer versus low self-explainer; and high improver versus low improver).

The third test examines parallelism. That is, it examines whether or not the profiles for the different groups are parallel to each other. If they are parallel, then that means their profiles are the same. If they are not parallel, then they are different. Parallelism, or non-parallelism, is revealed by the multivariate interaction between group (i.e., high vs. low self-explainer) and type of dependent variables. Hence, presence of a significant interaction indicates whether high self-explainer scores are higher than low self-explainer depends on a specific measure, indicating that profile is not parallel.

In sum, we are interested in: 1) a main effect of dependent measure (flatness); 2) main effect of the level (whether high ability and low ability participants significantly differ from each other); and 3) parallelism – or lack of interactions (whether the difference between high and low ability participants varies as a function of the dependent variables)

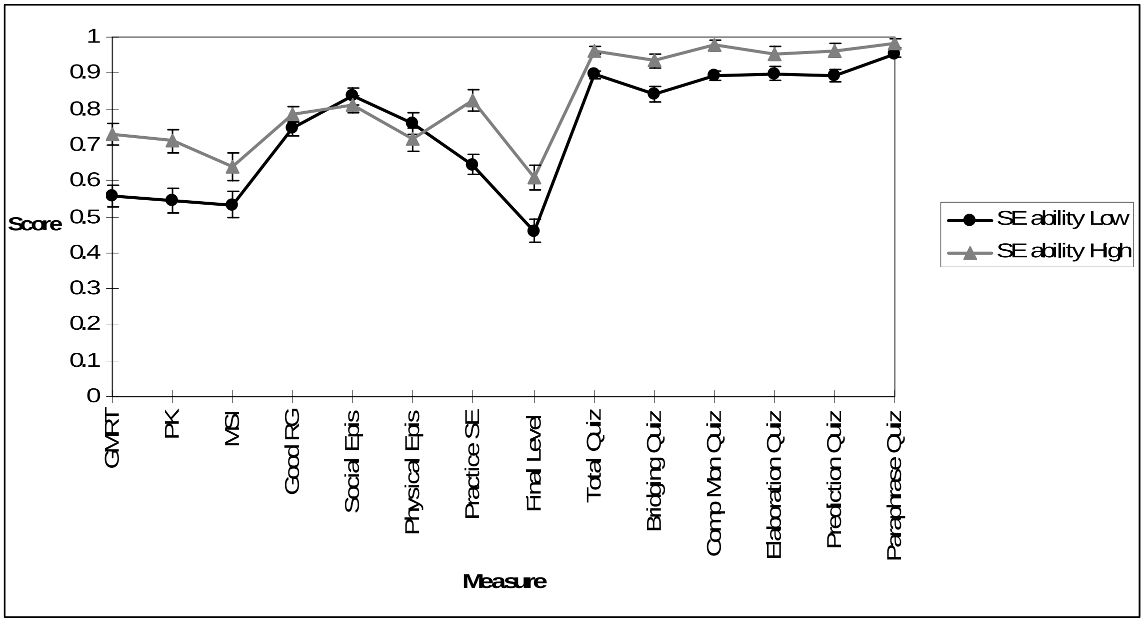
There are potentially many measures that are useful in representing high/low self-explainers or high/low improvers. The dependent variables used in the present analyses are listed in Table 11 below. We divided the high school participants (n = 70) in the transfer study into high and low groups using the following technique. First, a grouping variable was created by performing a median split of pretest self-explanation performance (science genre). Those above the median were classified as high explainers, those below the median were classified as low explainers, and those directly on the median were not included in the analysis (10 participants). For this first profile analysis we examined measures associated with different reader characteristics, and measures that interacted with iSTART (see Table 11).

As mentioned above, this first profile analysis examined the difference in profiles between high self-explainers and low self-explainers as determined by pretest measures. A median split was performed on the pretest self explanation tests and those above the median were labeled High ability self explainers (31 students) and those below the median were labeled Low ability self explainers (29 students). Those directly on the median were removed from the analyses (10 students). Figure 1 presents the different profiles for each reader type.

Table 11. Measures included in the profile analyses of self explanation quality and responsiveness to iSTART (with descriptions).

Measure	Type	Description
GMRT	Reader characteristic	Score on Gates-MacGinitie
PK	Reader characteristic	Score on prior knowledge test
MSI	Reader characteristic	Score on Metacognitive Strategies Index
Good RG	Reader characteristic	Score on reading goals questionnaire (Good goals)
Social Epis	Reader characteristic	Score on social epistemology
Physical Epis	Reader characteristic	Score on physical epistemology
Practice SE	iSTART measure	Self-explanation quality in practice module
Final Level	iSTART measure	Student level variable at end of demonstration module
Total Quiz	iSTART measure	Total quiz performance in Introduction module
Bridging Quiz	iSTART measure	Quiz performance for bridging inference questions in Introduction module
Comp Mon Quiz	iSTART measure	Quiz performance for comprehension monitoring questions in Introduction module
Elaboration Quiz	iSTART measure	Quiz performance for elaboration inference questions in Introduction module
Prediction Quiz	iSTART measure	Quiz performance for prediction inference questions in Introduction module
Paraphrase Quiz	iSTART measure	Quiz performance for paraphrase questions in Introduction module
Pretest SE score	iSTART measure	Self-explanation quality in pretest

Figure 1. Reader profiles characterizing relatively high and low ability self explainers.



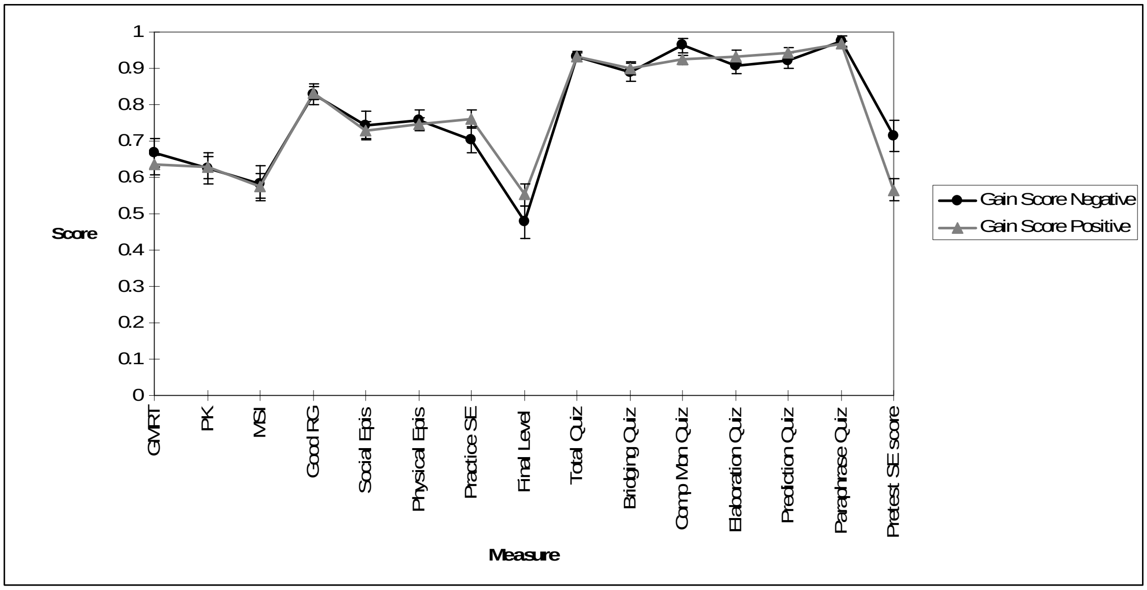
As shown in Figure 1, the relative difficulty of the measures varies. This was confirmed by a significant multivariate effect of dependent variables indicating that the profiles are not flat, $F(47) = 53.63$, $p < .001$. The Figure shows that performance varied widely for the reader characteristic measures, and the two iSTART measures Practice SE and Final Level. Performance across the six quiz scores appear to be flat. There was a significant main effect of levels (high and low ability groups) indicating that overall those with high self explanation ability outscored those with low SE ability, $F(1, 58) = 24.19$, $p < .001$. The test examining parallelism revealed a marginally significant effect, $F(47) = 1.84$, $p = .069$. That is, by and large, the readers show the same profiles. However, the profiles do intersect at several points, suggesting that the difference between high and low self-explainers changes as function of specific dependent variables. Performance did not differ between the groups for three Reader Characteristic measures; Good RG ($p = .17$), Social Epis ($p = .48$), Physical Epis ($p = .33$), and for one iSTART measure; Paraphrase Quiz ($p = .09$). These analyses show that high self explainers generally score higher on most individual differences measures of ability and understand iSTART training itself better as revealed by quiz scores and final level in the demonstration module. However, low and high ability self explainers tend to have similar scores on measures of reading goals and epistemological beliefs.

This second profile analysis examined the profile of high and low improvers. That is whether students who improve after iSTART training can be distinguished from those who do not improve based on the dependent variables. A gain score was computed by subtracting the posttest self explanation quality score for each student from their pretest score. A positive number means that there was a gain after training and a negative number means that there was a decline in performance (i.e., loss) after training. A grouping variable was created based on this score. Those with positive scores were grouped together (i.e., improver 43 students), and those with negative scores were grouped together (i.e., decliner 20 students). Students whose quality stayed exactly the same were not included in the analyses because their group size was too small to include in a profile analysis (5 students). Pretest self explanation quality was included as a measure in this analysis because we are interested in whether or not ones' ability to self explain, measured directly, can distinguish those who gain from training from those who do not. Figure 2 presents the different profiles for each reader type.

As shown in Figure 2, the relative level of score varied as a function of dependent variables, resulting in a significant rejection of profile flatness, $F(49) = 58.89$, $p < .001$. Scores varied for the Reader Characteristics measures and three of the iSTART measures; Practice SE, Final Level, and Pretest SE score. Performance on the quizzes appears rather flat.

The test for levels was not significant, $F(1, 61) < 1$. This shows that those who gain in self explanation quality (improver) do not differ on these different measures of the reader from those who declined (decliner). This suggests that the profiles based on these individual difference scores do not successfully differentiate readers who benefit and readers who do not benefit from iSTART training. However, the test for non-parallelism was significant, $F(49) = 2.74$, $p < .01$, representing a significant interaction between the group (improver vs decliner) and dependent variables, showing that these two groups of readers do show significantly different profiles. As Figure 2 shows, the strongest divergence in the profiles is due to the Pretest SE score measure ($p < .01$). No other differences were significant. This suggests that the best measure of who will have a positive or negative gain after iSTART is their pretest self explanation ability. As the figure indicates, it is apparent that those who gain have significantly lower pre-test self-explanation scores than those who do not gain. One interpretation of this finding may be that the students with high pretest scores have less room to gain. We are currently pursuing this question by modifying the self-explanation scoring system to increase the range in self-explanation score. One criticism of using the pretest self-explanation scores as a predictor of self-explanation gain is that the predictor variable is highly correlated with the outcome variable. That is, the predictor variable is generated from computations using the outcome variable. However, we believe that pretest self-explanation scores can be viewed as a test of ability, and as such we are interested in whether or not training improves performance on that test.

Figure 2. Reader profiles characterizing readers who had positive and negative gains in self explanation quality after training.



Overall these data suggest that pretest individual difference measures strongly distinguish readers with different self explanation ability but that it is difficult to predict gains from these measures. However, a direct measure of self-explanation ability appears to be a good predictor of responsiveness to iSTART. Those with high pretest ability tend to have negative gains from training and those with low pretest ability tend to have positive gains from training.

Subgoal 3.3: This subgoal involves the expansion of available strategy modules. One approach to improving the system's adaptivity is to increase the range of strategy instruction by adding instructional modules covering both lower-level and higher-level reading strategies. These additional modules will augment the system's ability to adapt to the needs of students.

Our research has indicated that some students need more training on how to generate adequate paraphrases before they should attempt to learn higher-level strategies. Although paraphrasing does not lead to deep comprehension, it is a fundamental prerequisite to self-explanation. Thus, we hypothesized that additional support would be needed for the lower ability students. We therefore focused our efforts in the past year on developing two additional modules ('Preparing to Read' and 'Paraphrasing') that teach more fundamental aspects of reading.

*Preparing to Read Module***.** The goal of this module is to help students understand the nature of reading and why training provided by iSTART will help them to become more efficient and successful readers. This module is targeted at less skilled readers who need some additional guidance, beyond that provided by the standard iSTART Introduction section. Although the Introduction may be sufficient for relatively skilled readers who already intuitively know about the processes and effort involved in reading, less-skilled readers may have difficulty grounding the information in their own experience. Specifically, in addition to the original module in which students were introduced to the concept of self-explanation and reading strategies, this module provides students with information about "what is reading for understanding." Incorporation of "reading for understanding" is motivated by the observation that many novice readers (typical high school students) have a belief that reading means just covertly pronouncing words and following text sentence-by-sentence in a serial fashion. They often believe that the information somehow "gets into their head." In order to expose students to a more accurate conception of "reading to understand," this module conceptualizes reading as a series of mental processes or operations that need to result in some form of mental imagery or picture. For example, the module provides an example sentence in which a "A white horse is running in the field." Through the use of this example, students are taught about the notion that "reading to understand" means being able to form some sort of image based on the information stated in the sentence. This idea is in line with current theoretical conceptions of reading comprehension where successful comprehension is conceptualized as a formation of a "mental model." This module is aimed at helping students begin to perceive reading processes as more than just "producing covert sounds corresponding to correct pronunciation of words." The basic format of the module is the same as the introduction of iSTART. Discussion on "what is reading" is developed among the three agents with several specific examples. In addition, with reference to "mental imagery," this module incorporates visual formatting (i.e., image of a running horse) to send a clear message to students about the notion that the goal of reading is to form a mental image of the situation described in the text.

Previewing Module. The goal of this module is to teach the student to get an idea of what the text is going talk about. Hence this is done before they actually start reading the text, which enables them to set a goal of why they are reading. This generally involves a short, focused examination of the text and noting things like titles and subtitles, introductory and concluding or summary paragraphs, review questions, emphasized words and phrases, charts and diagrams. This promotes better comprehension when the text is actually read. In the previous year, a prototype had been developed to explore the conceptual and technical challenges involved in the development of this module. Based on the prototype, we further refined the ideas that needed to be incorporated into the module, and also explored technical limitations. At this stage, we are moving in the direction of developing the module to include three phases with the following features. 1) an introduction with a guided tour in which students move through important parts of the text identified by the system (teacher agent), and complete a preview note (i.e., a table representing an important part of a text) by copying and pasting the important parts of the text (e.g., underlined words, questions, etc.); 2) a guided practice phase in which students select part of the text suggested by the agent in the module and complete a table (labeled) similar to phase 1. In this phase, however, students need to identify the suggested section themselves; 3) independent practice, in which students work with an extended text (a nearly 20 page text that covers an entire chapter in a biology text) and identify important parts, copy and paste these parts into non-labeled cells in order to compose a table that represents all the important ideas in the text, and write a brief remark on each item about how these parts help them understand the text content. Certainly the project to build this type of module is ambitious and requires a lot of effort and work, more than we initially considered. Thus the work would need to continue beyond the current grant period. At the current stage, we have selected a portion of text that will be used in the final practice, and it (i.e., a chapter in a biology text book that includes figures, tables, and photographs) has been successfully incorporated into a computer program for presentation. In addition, we have developed a first draft of the script used in the introduction phase. In this coming year, we plan to tackle the specific challenges such as what kind of feedback we would provide and how the feedback can be computed.

Paraphrasing Module. The goal of this module is to help students to be able to produce quality paraphrases. As mentioned earlier, many high school students have difficulty producing paraphrases. This is problematic because this may mean that they have difficulty understanding the meaning of individual sentences. If readers have considerable difficulty understanding the meaning of individual sentences, they are unlikely to benefit much from training on higher-level reading strategies such as elaboration and bridging inferences. Being able to form a mental model of the meaning of individual sentences is a necessity for higher level text comprehension. For this reason, we developed a new paraphrase module to provide more intensive training on paraphrasing.

The Paraphrasing module has two phases. The first part is an introduction that provides a definition of paraphrasing and an explanation for how paraphrasing can help one better understand texts. The teacher agent and student agents discuss two example sentences and show how to produce a paraphrase.

The second part of the module involves practicing paraphrasing with feedback. There are 12 interactive examples in this phase, and the number of examples that students practice depends on the students' performance. The students are required to read a sentence, and paraphrase the sentence based on their understanding of the target sentence. These 12 target sentences contain multiple, often abstract concepts that are related in somewhat complex ways (e.g., subordination, causal antecedent). Thus, paraphrasing these sentences encourages students to pay attention to the precise relations between these concepts. Students are given two chances, in addition to the initial attempt, to improve their paraphrases to the acceptable level set by the system's algorithms. Hence, when a paraphrase produced in the first attempt is unsatisfactory, the system provides feedback (e.g., that it's too short). Students respond to the feedback by revising their paraphrase. Our hope is, throughout this activity, students become analytical about precise sentence structure and wording so that they try to understand the sentence meaning as accurately as possible. The evaluation algorithm used in the system is similar to the one in the practice module in the current iSTART system. The threshold values that determine whether a protocol is short or too similar have been altered so that the algorithm can appropriately detect a paraphrase which is generally shorter than a full self-explanation, which contains some form of reasoning and inferencing. This algorithm classifies input (i.e., students' paraphrase) into these categories: short, irrelevant, closely similar (SIM1), paraphrase (SIM2), and OK. Student responses assigned values paraphrase and OK are considered acceptable. Based on this classification, the teacher agent provides verbal feedback to the students about the quality of the paraphrase. As mentioned earlier, the paraphrasing module adapts to the students' ability and also their current performance in the module by changing the number of practice examples each student needs to complete based on their ability and performance. In the current version which was used in Experiment 5 (see below), the extensiveness of practice training for each student (i.e., the number of examples) were determined based on these two data: 1) students ability level (high or low ability group) based on the test at the beginning of the experiment; 2) students' actual performance when working with practice examples. Specifically, high ability students work with a minimum of 2 interactive examples, and the low ability group students work with a minimum of 6 interactive examples. The students complete the paraphrase practice only when they finish the minimum number of examples for their ability level and also satisfy one of the following conditions a) producing an acceptable level of paraphrase at the first attempt a minimum of 3 times, and in addition, producing 2 acceptable paraphrases out of 3 previous examples;. b) producing a total of 6 acceptable paraphrases on the first attempt; or c) completing all 12 interactive examples.

An experiment was conducted to test the effectiveness of the paraphrasing module. Our first evaluation of the paraphrase module tested whether or not participants with low ability would benefit from extended practice of paraphrasing, and whether this practice would impact learning gains of high ability students. As such, we used the version of the module that had a minimum of 6 interactive examples for all students in the interactive paraphrase condition. There were two conditions, an iSTART condition and an iSTART with interactive paraphrasing practice condition. We randomly assigned students to either condition constrained by reader ability, which will be described in the following paragraphs.

The experiment proceeded in 4 days/stages. In the first day, participants took a battery of pretest measures. These included a demographics questionnaire, a science prior knowledge measure (PK), the Gates-MacGinitie reading ability test (GMRT: level 7-9), the metacognitive strategies index (MSI; Forget, 1999), and a short form of RSAT. They also read and self-explained 8 target sentences of a text.

Table 12. Correlations among RSAT, MSI, PK, and GMRT.

	RSAT score	MSI	PK
RSAT score	--	--	--
MSI	.31*	--	--
PK	.32*	.18	--
GMRT	.29*	.15	.44**

* *p* < .10
** *p* < .01

Table 13. Mean and standard deviations (in parentheses) for each ability level and condition.

	Measure	Condition	
Ability Level		Interactive paraphrasing	iSTART
Low	RSAT	3.73 (0.75)	3.83 (1.01)
	GMRT	.39 (.09)	.44 (.16)
	PK	.39 (.82)	.33 (.15)
	MSI	.58 (.21)	.46 (.20)
High	RSAT	7.97 (1.16)	7.37 (2.15)
	GMRT	.54 (.14)	.55 (.25)
	PK	.48 (.12)	.47 (.12)
	MSI	.65 (.19)	.65 (.13)

Prior to day 2, participants were randomly assigned to condition with the constraint that there were no between-condition differences in the Gates-MacGinitie, the MSI, the prior knowledge measure, and RSAT. To separate participants into groups, a score based on the RSAT (see above in the individual difference measure section) measures was calculated. We added the RSAT scores for distal bridging, elaboration, and question answering. The bridging and elaboration scores are indicative of the construction of global situation models and the question answering performance measures comprehension. Thus, this score may provide a measure of the construction of accurate situation models. Table 12 provides a correlation matrix among the four individual difference measures. Participants were classified as high ability or low ability based on the RSAT score. Within each ability level, participants were randomly assigned to condition. This resulted in no differences in the four individual differences measures between the groups and skill levels (see Table 13), $F < 1$.

When participants arrived for day 2, they received the training as designated by their condition. The groups both received the Introduction and Demonstration sections on Day 2. The paraphrasing module was embedded in the Introduction section. Participants in the iSTART condition received the standard paraphrasing module, and participants in the interactive paraphrasing condition received the interactive version of the module. On Day 3 participants completed the practice session. On day 4, participants took a self-explanation test and RSAT.

Results

For the first set of analyses, we were interested in the extent to which the different training conditions affected gains in self-explanation quality. We conducted four separate 2 (training condition: interactive paraphrasing vs. iSTART) x 2 (training occasion: pretest vs. posttest) x 2 (Ability: High vs. Low) ANOVAs using median splits on each of the four individual difference measures as the Ability variable. The dependent variable was self-explanation quality as determined by the iSTART algorithm. For ease of exposition, we will discuss the analysis in terms of whether the Ability variable significantly interacted with training condition and the ability level as measured with the performance on GMRT. Table 14 presents the means and standard deviations for each group. Overall, self-explanation quality increased from pretest to posttest (pretest = 1.67; posttest = 1.90), $F(1, 31) = 5.22$, $MSE = .18$, $p < .05$. This variable interacted with training condition, and ability level, $F(1, 31) = 4.31$, $MSE = .18$, $p < .05$. Follow-up tests indicated that for the interactive paraphrase condition, there was a marginally significant increase in self explanation quality from pretest to posttest for the high ability students, $t(7) = 2.02$, $SD = .66$, $p =.083$. For the iSTART condition, there was a significant pretest-posttest increase for the low ability students, $t(9) = 2.44$, $SD = .53$, $p < .05$. Consistent with previous work, we found that iSTART improved self explanation quality for low skilled readers. Practice with paraphrasing in addition to iSTART training, however, did not improve self explanation quality for these readers. Interestingly, this extra training in paraphrasing improved self explanation quality for the high skilled readers.

Table 14. Means and standard deviations (in parentheses) of self-explanation quality for each condition and ability level (GMRT).

	Training Condition	Pretest	Posttest
Ability Level			
Low	Interactive paraphrasing	1.71 (.39)	1.51 (.40)
	iSTART	1.58 (.32)	1.99 (.59)
High	Interactive paraphrasing	1.56 (.77)	2.03 (.46)
	iSTART	1.83 (.46)	2.06 (.48)

In addition, we were interested in the extent to which the different training conditions affected the prevalence of paraphrasing revealed by the think aloud methodology. For this second set of analyses, we conducted planned comparison (*t* tests) on performance on the paraphrasing scores for each condition and ability level (see Table 15). These comparisons showed that the only group that significantly increased in paraphrase ability from pretest to posttest was the high ability students in the interactive paraphrasing condition, $t(7) = 2.28$, $SD = .66$, $p = .056$.

Table 15. Means and standard deviations (in parentheses) of RSAT paraphrase scores for each condition and ability level (GMRT).

	Training Condition	Pretest	Posttest
Ability Level			
Low	Interactive paraphrasing	0.90 (0.63)	1.13 (0.70)
	iSTART	1.10 (0.62)	1.19 (0.53)
High	Interactive paraphrasing	1.07 (0.68)	1.60 (1.03)
	iSTART	0.91 (0.48)	1.09 (0.56)

These results suggest that contrary to expectations, high skilled readers benefited more from extra practice than the less skilled readers. It appears that only the high skilled readers were able to take advantage of the extended paraphrasing practice in a way that enhanced their self explanations. Regular iSTART training may be at less skilled readers' zone of proximal development. As such, adding more practice for these individuals may overly tax their resources and hinder learning. Further work is being conducted to further understand which factors contribute to learning from interactive practice. In addition, we are using the data from the interactive paraphrasing module to fine tune the paraphrase evaluation algorithm.

Question Asking Module. The goal of this module is to help students be able to generate deep-level questions. Here, we make the distinction between a shallow question, which focuses mostly on verification and other short-answer aspects, and a deep-level question, which tends to focus on assertions, judgments, and other questions which require longer answers (Craig, Gholson, Ventura, Graesser, & the TRG, 2000). Such a tool is necessary because (1) previous research on questions asked by students suggests that students ask very few questions (whether shallow or deep-level) during a typical classroom session (Graesser & Person, 1994) and (2) several studies (e.g., King, 1992, 1994) suggest that training in self-questioning may increase students' comprehension of a text.

The Question Asking Module will begin with animated agents discussing the significance of asking questions when reading a text, and then briefly explain how to ask deeper questions (such as those that link ideas within a text or link textual ideas with ideas already present in long-term memory). The module will then enter a "practice" session, where the student will ask questions based on selected target sentences within a science text. The participants' questions will be interspersed with questions asked by another agent. Both the participant and the "asking-agent" will receive feedback from an "instructor-agent." This will allow us to integrate detailed examples of questions and feedback for the asking-agent, while the participant practices and receives more general feedback. Such a "turn taking" style of practice should help scaffold the new strategy for the user. This module is currently under development and will be completed for use in an experiment in the fall of 2007. Technical aspects of the module (e.g., evaluation algorithms) are under development currently.

We are currently conducting an experimental study to gain a better understanding of what types of specific question-asking training is most beneficial for students. In particular, it is necessary to explore the possibility that students with different ability may benefit from training on different types of questions. Although past research on the effect of question asking on comprehension indicated that training on generating deep questions benefited students' comprehension, it is further possible that many students are not ready for deep level questioning. If so, different students may benefit from training on different levels of questions. In order to explore potential interactions between the level of question and individual differences, we are currently conducting an experiment on question asking training that is experimenter-delivered. The data collected in this study will inform us of the directions we should move in terms of the features that need to be incorporated in the question asking module. Further, a Question Asking Module would require substantially different feedback algorithms than those currently employed in the standard iSTART module. To develop an adequate question-detection algorithm, participant protocols are necessary; the question asking experiment will provide us with the data necessary to develop appropriate algorithms.

What's the Point Module. The goal of this module is to help students to find the main ideas in text paragraphs by focusing on their central ideas, themes, and supporting details. The purpose of the module is similar to training systems that teach summarization. However, here the focus will be on finding the main points and distinguishing those from details, rather than writing complete summaries of texts. This module is currently under development. This module includes an introduction phase and a practice module in which the students practice finding the main ideas or themes of passages. The objective of the introduction and practice is to instruct trainees on how to find the point of a paragraph, when it contains a topic sentence and when it does not contain topic sentence. In the latter case, the student is supposed to find ideas that generalize the details in sentences. We are conducting, in summer and fall 2007, an experiment with three purposes: first, we will collect students' one-sentence summaries after reading science text paragraphs; second, we will investigate whether and how students take advantage of the presence or absence of a topic sentence to generalize ideas from one paragraph; second, we will develop an computational algorithm to analyze and rate the summaries.

New Question Manager in Demonstration. Our goal is to reduce the difficulty of the demonstration module and make it more flexible and responsive to the student's level of performance. This initially required a task analysis of what the students needed to do in order to answer questions in the module. Considering factors such as working memory requirements and relevant skills, we have attempted to reduce the complexity of each task by improving the following components.

- Questioning methods: (1) give a set of strategies for selection in a multiple-choice box, (2) tell the student one of the strategies that was used, or (3) tell the student the strategy and remind him/her of the strategy's definition.
- Tool tips: When a multiple-choice dialog box is used, "tool tips" that contain reminders of the strategies' definitions are available to the student. The reminders are activated by clicking on a button next to the strategy choice.
- Focus: The original version required the student to inspect (and parse) the whole self-explanation. We added the capability to direct the student's attention to a portion of the self-explanation.
- Pre-parsing: The parts in the self-explanation where different strategies are used can be highlighted in alternating colors.
- Follow-up response: When the student makes an error, we have the option of telling the student which strategy was used.

Subgoal 3.4: This subgoal involves assessing the system's usability. Our approach to usability analyses will be three pronged, using a combination of information gathered via think-aloud protocols, eye tracking, and task-completion rates. We conducted an investigation that assessed students' eye-tracking movements when using the program. The main findings were that when an agent talks, the participants pay attention to the agent, but they do not look at the agent when the agent is quiet. In addition, participants pay attention to the appropriate agent when the speaking agent refers to that other agent. The results of this eye tracking study show that participants pay attention whenever communicative purposes require attention, but they do not look at the agent when this is not needed (Louwerse et al., 2006).

A second usability study was performed that explored the possibility of providing more user-based control over the iSTART interface by allowing the background color for pages to be alterable. Such a change would provide the user with a more appealing interface from which to interact with the iSTART system. This change must answer two important questions: (1) is there a need for background color change, and (2) will background color change affect performance. To answer these two questions, two experiments were designed. The first experiment assessed the need for color change and was completed during the spring of 2006. The second experiment will determine the effect of background color on iSTART and is planned for the fall of 2007.

Experiment 1 was a preference study for iSTART background colors. In order to provide a manageable set of alternative color schemes, the number of colors examined were constrained to primary (red, yellow, or blue), secondary (green, orange, or violet), tertiary (combinations of one primary and a secondary color of which they were a component color), black, white, or the original color of iSTART. A single snapshot of each module was used for comparison with background colors. Preferences for each of the 15 different background colors were obtained for each iSTART module. Two colors, orange and blue, were preferred across all three modules more than the current iSTART background color. These findings provide the basis for continuation of the study to address whether alterations to background colors will affect performance. Experiment 2 will contain altered iSTART modules allowing learners the option of changing their background color at the beginning of each module. Performance between users using the current background color for iSTART and other colors will be compared.

Objective 4: Teacher Interface

The fourth objective is to provide an effective and user-friendly interface for teachers using iSTART in their classrooms. So far iSTART has only been implemented in classrooms under experimenter control. Scaling up the program means that the classroom teacher would need to have control over how the program is to be used. Our research plans and progress to meet Objective 4 are represented as sub-goals in Table 4.

Table 16 Sub-goals for Objective 4

Sub-goal	Description of activity	Progress	Time frame
4.1	Development of teacher interface	Commenced	Years 1 – 3
4.2	Teachers' needs assessment (Study 6)	Completed	Years 1 – 3
4.3	Assessment of teacher interface (Study 7)	To commence Fall 2007	Year 3
4.4	Assessment of teacher interface in classroom (Study 8)	Preparation commenced	Year 4

Sub-goal 4.1: This sub-goal involves the design of the iSTART user interface to be used by teachers in their classrooms.

Teacher Instruction Module (TIM). In order to maximize the effectiveness of iSTART, we believe that teachers need to have an understanding of relevant underlying psychological theory. We will therefore provide teachers with the basic information about the reading strategies and why they are effective. This information will cover theories of text comprehension, the effect of prior knowledge on reading, and the importance of reading strategies. We have prepared the required material as a set of easily navigated web pages. This approach allows us provide this material in a brief and simple manner. More details can be accessed by following the links in the web pages. This module also provides teachers a "hands-on" tour of the iSTART modules through a special teachers' version of iSTART that allows teachers to explore all modules without being required to complete each one. This enables the teachers to get a better idea of the tasks that their students will experience.

The web pages of the TIM are organized into an easily navigable set that allows teachers to follow ‘next’ and ‘previous’ links linearly through the orientation and also allows them to click on internal links that jump from page to page. They provide discussions of the following topics:

A. Introduction to cognition: the sensory system, passive memory, active memory, and their interaction.
B. Text comprehension: the semantic organization of prior knowledge in long-term memory; the interaction of the textual input with prior knowledge in text comprehension.
C. The iSTART trainer: the introduction, demonstration, and practice modules of the trainer.
D. The iSTART reading strategies: comprehension monitoring, paraphrasing, prediction, elaboration and bridging.

The teachers’ version of the trainer is nearly identical to that of the ordinary trainees’ except that a navigation menu has been added. This menu allows the teacher to exit any module without completing it and to return to an earlier module even if it has been completed. The purpose of this module is not to train teachers in reading strategies but to familiarize them with the training that their students would receive and to allow them to explore the connections between the different parts of the training.

Training Organizer and Manager Module (TOM). The goal of this module is to provide technical and practical support to the teachers so that they can use iSTART effectively. The option of being able to customize the training for different levels of student ability is important for improving the effectiveness and adoption rate of the iSTART system. The teachers will be able to designate specific readings in specified date ranges for different groups of students in the same class or different classes. The *registration and enrollment management* features allow schools to register with iSTART and teachers to sign themselves up, add a class, and register students to their selected class. The *scheduling and policy* features should allow the teachers to set up the curriculum by configuring different modules of iSTART. The program will have default sequences for students based on their assessed ability; however, this module will allow teachers to make special assignments and constrain iSTART’s use of its library to best serve their classes’ needs. We have built a library of texts that encompasses a wide selection of topics and domains from which to choose and have built tools that make it easy to add additional texts. The curriculum policy concerns whether the students take the modules at their own pace, one after the other, or take them on a schedule; whether missed modules can be made up and, if so, whether the current module is taken first or last.

The first version of this module is nearing beta stage and will soon be ready for usability testing. This module includes three sets of features: higher level administration; teacher level registration and enrollment management; and teacher level curriculum management.

Higher level administration provides functionality needed for a future in which iSTART is a scaled-up system serving many schools. It provides for system level administration by an iSTART administrator, the inclusion of new schools with designated principals who may in turn authorize teachers or lead teachers for their schools. In this way school authorities can exercise control over what programs are permitted in their school. An “impersonation” feature permits an administrator, principal or lead teacher to make registration adjustments on behalf of a classroom teacher. All such adjustments are logged.

Individual teachers can create iSTART sections for their classes and enroll students in them. One particularly nice feature is that they can copy and paste many students from a class roll rather than typing them in one at a time. They can also add, drop, or transfer them as may be needed to keep their iSTART roll in synch with their classroom roll. When students are transferred from one teacher's section to another, the receiving teacher may be notified by email. In addition, teachers can create special sections for remedial or advanced work to which students from several existing sections can be added (so that one student can be in more than one section at the same time).

Curriculum management is the most complicated part of this module. The curriculum includes three aspects: iSTART training, special assignments of practice texts, and choice of topics for extended practice. Initially everyone in the class receives training that is adapted to their needs by the iSTART system. Once the training is over, iSTART assigns practice readings to the students according to both their needs and a schedule of topics and subtopics that is chosen by the teacher. In addition, the teacher can schedule specific texts to be used for practice by the class or groups within the class at various points within the term, even during the initial training period.

As an example, a biology teacher may schedule iSTART sessions for her class on a once-a-week basis, estimating that the self-paced pre-testing and training will take students three or four weeks. She would assign a general topic of "biology" for the practice sessions that follow the training, specifying sub-topics of "cells," "plants," "animals," and "environment" for specific date ranges that coordinate with her classroom schedule. For three students who are working on a special project, she would assign a reading on "synapses" for a specific week. When one of those students first logged into iSTART during that week, he or she would first be presented with that reading before any other iSTART training or practice.

Performance Analyzing Tool Module (PAT). The purpose of this module is to provide teachers with a means to monitor students' performance and progress during training. This tool will report the students' pre-training assessment scores, assessments that occur during the iSTART training, and the students' progress through the iSTART curriculum. These reports will help teachers to gauge the students' performance. In addition, by knowing her students' situation in the self-paced curriculum, she can determine when it would be appropriate to introduce special reading texts for groups or for the class. In this Module, the main challenge involves determining the right amount of data to be included in the reports generated for the teachers. Although these reports present summary information, the teacher will also have access to more specific information such as the modules completed, time spent, number of attempts, and performance on each attempt. An interface to access the student's progress on iSTART trainer and pretest has been developed. The reports that can be made available to the teachers are: i) scores on each of the pretests by test, ii) scores on each of the quizzes of the introduction, iii) student level after each of the sentences of the demonstration section, and iv) training modules completed by student. What we intend to include is partially based on teacher interviews.

At this point we have developed the designs for rending both graphs and tables with the desired information in a form that can be easily viewed on a web page without loss of informational context even if the number of students in a class is large. However, two problems remain to be solved. One is the integration of the features of the Performance Analyzing Tool (PAT) with the Training Organizer and Manager (TOM) so that the teachers have the decision support of PAT available when making curricular decisions in TOM. The other is the technical choice of method for transferring raw transactional data collected as the student interacts with iSTART to the summary tables suitable for creating charts, graphs, and spreadsheets. We wish to ensure that the data summarization reliably and timely occurs with minimal disruption of access to the transactional tables.

Sub-goal 4.2 (Study 6): This sub-goal involves assessing the needs of teachers for using iSTART in the classroom. The iSTART system is to be developed in such a way that it enables teachers and students to be able to use the tool with ease and efficiency. As a first step toward assessing teachers' preferences when using the interface, we had previously conducted a pilot study in which teachers that had used iSTART system before provided us with their opinions about the teacher interface and iSTART in general. This information has served to guide our refinement of the iSTART system.

In order to effectively develop iSTART and the teacher interface, it is necessary to evaluate both the general computer technology skills of teachers and the current availability and usability of computer technology in the average public high school. The goal of the current study was to evaluate what high school teachers know about computer technology, the availability of computers, electronic media, and internet technology for teachers, and the extent to which teachers actually utilize the various available technologies.

Principals at five area high schools in the Memphis City School system were selected and recruited by the Center for Research in Educational Policy (CREP) at the University of Memphis. For faculty participation in the iSTART Needs Assessment survey, schools were given a $500 stipend. A total of 127 teachers participated. For the Memphis City Schools the student race/ethnicity percentages for the 2005-2006 school year were as follows: 87% African American, 9% Caucasian, and 4% Other.

A CREP researcher attended after-school faculty meetings at the schools to administer the surveys during March and April, 2006. Surveys were completed during these meetings and were then collected by the researcher. The surveys were anonymous and the data from the surveys are reported in aggregate to preserve the anonymity of respondents.

Three instruments comprised the survey: The Technology Skills Assessment (TSA), Teacher Technology Questionnaire (TTQ), and the iSTART Teacher Questionnaire. The first two are CREP-designed instruments, and the last instrument was designed by iSTART researchers, with input from CREP.

The TSA was developed to assess teachers' perceived technological abilities and knowledge. Teachers rated themselves on a three point scale ('not at all', 'somewhat', 'very easily') with respect to "How easily" they could perform certain tasks. The six task categories included Computer Basics (e.g., 'How easily can you use keyboard commands to cut, copy or delete text?'), Computer Software (e.g., 'How easily can you open and use software programs that are installed on your computer?'), Multimedia Basics (e.g., 'How easily can you import digital video from a camera to a computer?'), Internet Basics (e.g., 'How easily can you use the internet to find help when you have a computer problem?'), Advanced Skills (e.g., 'How easily can you publish information in a variety of media (e.g., printed, monitor display, web-based, video)?'), and Using Technology for Learning (e.g., 'How easily can you use multimedia software to enhance learning experience?'). A sixth section posed questions on the teachers' understanding of site licenses and copyright issues. There were a total of 56 questions in the TSA assessment.

The TTQ focuses on teacher experiences with regard to technology usage, professional development, accessibility of technology resources, and confidence in using technology. Teachers rated statements on a five point scale ('strongly disagree' to 'strongly agree'). Examples of TTQ statements are 'My students have adequate access to up-to-date technology resources' and 'Materials (e.g., software, printer supplies) for classroom use of computers are readily available.' There were a total of 20 technology statements and eight demographics questions (e.g., 'Do you own a home computer?').

The iSTART Teacher Questionnaire focused on teaching strategies, teaching methods, and evaluation techniques. Additionally, teaching demographics were collected. The results of this study will be used to assess the ability of current teachers to utilize available computer technology in the classroom. This information will help in developing an iSTART program that enables teachers and students to use the tool with maximum ease and efficiency.

The Technology Skills Assessment (TSA). Teacher confidence was highest in *Computer Basics* with 93.7% reporting that they could very easily print from a file, while 92.9% said they could proficiently use a mouse or a keyboard. Teachers were less confident in their abilities in *Multimedia Basics*, *Using Technology for Learning, and Advanced Skills.* Some of the areas in which teachers reported feeling least proficient included: importing digital video from a camera to a computer, connecting a computer to a local server, using appropriate digital layout and design, manipulating information in interactive digital environments, and creating an electronic teaching portfolio. The Dimension Summary of the TSA is in Table 17. The item-level data summary is provided in Table 18.

Table 17. Technology Skills Assessment (TSA) Dimension Summary (n=127 Teachers)

#	**Dimension**	**Rating**
1	Advanced Skills	2.00
2	Computer Basics	2.83
3	Internet Basics	2.40
4	Multimedia Basics	2.15
5	Software Basics	2.61
6	Using Technology for Learning	2.11
7	**OVERALL**	2.35

Table 18. Technology Skills Assessment (TSA) Data Summary for Survey Period 1 (2005 – 2006; N = 127)

	% Not at all	% Somewhat	% Very Easily
Computer Basics: How easily can you ...			
Use a spell check tool.	3.1	9.4	87.4
Create basic computer documents (word processed) in a timely manner.	1.6	9.4	89.0
Use help menus for software programs.	5.5	18.9	74.8
Use basic computer terms like mouse, keyboard, hard drive, CD-ROM, and monitor.	0.0	8.7	91.3
Save documents so they can be opened on both a Macintosh and PC.	14.2	8.7	77.2
Create folders on a hard drive or disk.	3.1	15.0	81.1
Save files to specific folders.	3.1	17.3	79.5
Locate and delete unwanted files.	2.4	15.7	81.9
Use keyboard commands to cut, copy, or delete text.	0.8	14.2	85.0
Proficiently use a mouse and keyboard.	0.0	5.5	92.9
Print a document using “Print” from the File menu and/or the toolbar icon.	0.0	6.3	93.7
Software Basics: How easily can you ...			
Use software preview features to check work.	7.9	29.9	62.2
Open and use software programs that are installed on your computer.	2.4	18.1	79.5
Work with and move between two open programs (e.g., Internet and database) to create a product.	9.4	22.0	68.5
Describe the difference between downloading and installing software.	9.4	20.5	69.3
Save documents so they can be opened in a different program (e.g., from Word to Word Perfect).	11.0	22.8	64.6
Install software.	8.7	28.3	63.0
Multimedia Basics: How easily can you ...			
Import digital video from a camera to a computer.	38.6	30.7	29.9
Record and save your voice onto a computer.	52.0	24.4	23.6
Use a scanner to import a photo or document into a computer.	23.6	28.3	48.0
Play a music CD on the computer.	8.7	12.6	78.0
Internet Basics: How easily can you ...			
Connect to the Internet with a modem (phone, cable).	7.9	11.0	81.1
Use Boolean strategies for Internet searches.	42.5	20.5	33.9
Use appropriate software and the Internet to find audio, video, and graphics for lesson plans.	11.0	26.8	62.2
Use the Internet to find help when you have a computer problem.	19.7	27.6	52.8
Determine if information you find on the Internet is accurate and valid.	11.0	27.6	59.8
Evaluate Internet search strategies to determine those that are most efficient.	13.4	28.3	57.5
Determine the usefulness and appropriateness of digital information.	18.1	27.6	53.5
Advanced Skills: How easily can you ...			
Use more advanced computer terms like megahertz, gigabytes, and RAM.	30.7	36.2	33.1
Access information on local area networks (LANs) and wide area networks (WANs).	40.9	24.4	33.1
Use appropriate digital layout and design to meet the needs of defined audiences.	40.9	31.5	26.8
Use appropriate digital layout and design for the selected media (e.g., multimedia, web, print).	38.6	31.5	29.1
Publish information in a variety of media (e.g., printed, monitor display, web-based, video).	30.7	35.4	33.1
Connect a computer to a local server to share files.	45.7	27.6	26.8
Determine if a software program works with an operating system.	38.6	26.8	34.6
Print to a specific printer when connected to a network that has more than one printer.	23.6	23.6	52.0
Use presentation software to share information with specific audiences.	31.5	22.8	45.7
Use Technology for Learning: How easily can you ...			
Use multimedia software to enhance learning experiences.	11.0	41.7	45.7
Use appropriate software (e.g., word processing, graphics, databases, spreadsheets, simulations, and multimedia) to express ideas and solve problems.	15.0	33.1	51.2
Use text and graphics to create and modify solutions to problems.	22.8	31.5	43.3
Use digital audio and video to create and modify solutions to problems.	30.7	39.4	29.1
Use communication tools to participate in group projects.	29.9	31.5	37.8
Manipulate information in interactive digital environments (e.g., simulations, virtual labs, field trips).	38.6	33.1	27.6
Participate in a listserv, chat, and bulletin board session.	32.3	29.1	37.8
Create an electronic teaching portfolio to evaluate your work.	41.7	33.1	23.6
Evaluate electronic portfolio products.	43.3	29.9	25.2
Create technology tools to assess student work (e.g., checklists, timelines, rubrics).	22.8	33.9	40.2
Policy and Ethics: I understand ...			
My school’s acceptable use policy.	4.7	24.4	70.1
The concept of a school site license for software.	9.4	26.0	63.8
How to determine if it is legal to copy a software program or another individual’s electronic work.	13.4	31.5	54.3

Note: Item percentages may not total 100% because of missing input from some respondents

Teacher Technology Questionnaire (TTQ). The Dimension Summary is presented in Table 19, while the item-level data summary is presented in Table 20. Approximately three-fourths of teachers (74.8%) agreed their students could capably use computers at an age-appropriate level. In addition, two-thirds of all teachers (67.7%) reported their own skill levels were sufficient to conduct classes in which students used technology. The same percentage of respondents (67.7%) indicated that teachers in their school were generally supportive of technology integration efforts. The majority of teachers (66.9%) reported that they could meaningfully integrate technology into lessons. The majority of teachers (61.4%) also rated their level of computer ability as good or very good.

On the other hand, only slightly more than one-third (37%) of the responding teachers agreed that their school had a well-developed technology plan. Less than one-half of the respondents agreed that they had adequate administrative support to integrate technology (46.5%) and sufficient materials for classroom use of computers (47.2%). Approximately one-half (49.6%) of the teachers agreed that they routinely integrate technology into their instruction.

Slightly more than one-third of the teachers (35.4%) reported no computers were available for student use in their classrooms. Only 18.1% of teachers indicated having four or more computers available. Thirty-seven percent of teachers (37.0%) disagreed or strongly disagreed that materials for classroom use of computers are readily available. Slightly more than half (56.7%) of teachers indicated that they have received adequate training to incorporate technology into their instruction.

Table 19. Teacher Technology Questionnaire (TTQ)

Dimension Summary		
#	Dimension	Rating
1	Impact on Classroom Instruction	3.58
2	Impact on Students	3.62
3	Overall Support for Technology in the School	3.44
4	Teacher Readiness to Integrate Technology	3.67
5	Technical Support	3.36
6	OVERALL	3.53

Table 20. Teacher Technology Questionnaire (TTQ) Data Summary for Survey Period 1 (2004-2005; N = 127)

Teacher Technology Questionnaire (TTQ) Items	% Strongly Agree and Agree	% Neutral	% Strongly Disagree and Disagree
Indicate the extent to which you agree with each of the following items:			
Most of our school computers are kept in good working condition.	59.8	16.5	22.8
I can readily obtain answers to technology-related questions.	53.5	22.8	21.3
The use of computers has increased the level of student interaction and/or collaboration.	58.3	26.0	14.2
Parents and community members support our school's emphasis on technology.	52.0	30.7	15.0
I know how to meaningfully integrate technology into lessons.	66.9	24.4	7.9
My students have adequate access to up-to-date technology resources.	53.5	13.4	29.1
Materials (e.g., software, printer supplies) for classroom use of computers are readily available.	47.2	13.4	37.0
The integration of technology has positively impacted student learning and achievement.	59.8	23.6	15.0
I am able to align technology use with my district's standards-based curriculum.	55.9	26.8	15.7
Most of my students can capably use computers at an age-appropriate level.	74.8	11.8	11.0
I have received adequate training to incorporate technology into my instruction.	56.7	18.9	22.0
My computer skills are adequate to conduct classes that have students using technology.	67.7	19.7	10.2
Teachers receive adequate administrative support to integrate technology into classroom practices.	46.5	25.2	26.0
My teaching is more student-centered when technology is integrated into the lessons.	59.8	29.9	8.7
Our school has a well-developed technology plan that guides all technology integration efforts.	37.0	33.9	26.0
I routinely integrate the use of technology into my instruction.	49.6	26.0	21.3
Teachers in this school are generally supportive of technology integration efforts.	67.7	24.4	3.9
Technology integration efforts have changed classroom learning activities in a very positive way.	59.8	28.3	7.9
The use of technology has improved the quality of student work.	50.4	31.5	15.7
My teaching is more interactive when technology is integrated into the lessons.	50.4	36.2	11.8
Computers available for student use (%)			
0	35.4		
1	21.3		
2-3	23.6		
4-5	14.2		
6 or more	3.9		

Note: Item percentages may not total 100% because of missing input from some respondents

Table 21. iSTART Teacher Questionnaire Data Summary for Survey Period 1(n=127)

I have used the following reading strategies when teaching during the last 12 months:	*% Unfamiliar Strategy*	*% Never Use*	*% Have Tried Once*	*% Use Sometimes*	*% Use Often*	*% Very Frequently*
PAR Lesson Framework	59.8	24.4	1.6	3.1	3.1	0.8
Textbook Treasure Hunts	32.3	30.7	9.4	11.8	3.9	3.1
Possible Sentences	27.6	24.4	8.7	21.3	7.9	1.6
Two Column Notetaking	11.0	22.0	7.9	33.1	13.4	5.5
Guided Reading Procedure	5.5	5.5	3.9	29.9	33.1	18.9
Remembering	18.1	9.4	2.4	22.0	27.6	12.6
Reading Response Charts	17.3	18.1	5.5	26.8	18.9	5.5
Interactive Notation System for Effective Reading And Thinking (INSERT)	47.2	16.5	7.1	12.6	6.3	3.1
Anticipation Guides	14.2	18.9	8.7	29.1	18.1	4.7
Structured Overviews	11.8	7.9	4.7	29.1	29.1	9.4
Reading Roadmaps	23.6	23.6	8.7	26.0	6.3	5.5
Pattern Guides	37.0	21.3	4.7	19.7	7.1	2.4
3-Level Study Guides	33.9	15.7	7.1	17.3	11.8	4.7
Reciprocal Teaching	20.5	19.7	4.7	30.7	11.8	4.7
Cubing	46.5	22.0	7.1	11.0	3.9	2.4
Semantic Feature Analysis	39.4	22.0	7.9	15.0	8.7	1.6
Speed Reading	18.9	44.9	9.4	14.2	4.7	0.8
Directed Reading Thinking Activity (DRTA)	26.0	15.7	3.9	25.2	11.0	6.3
Concept Guides	15.7	15.7	8.7	32.3	15.0	5.5
Post-Graphic Organizers	7.9	13.4	6.3	33.9	22.0	13.4
Comprehension questions	2.4	5.5	0.0	13.4	43.3	30.7
Reading questions	0.8	2.4	1.6	17.3	37.0	33.9
Paired reading	3.1	26.0	4.7	26.8	20.5	13.4
Subtitle questions	13.4	27.6	3.1	20.5	17.3	7.9
Read/outline	1.6	11.0	8.7	23.6	31.5	18.9

Note: Item percentages may not total 100% because of missing input from some respondents

iSTART Teacher Questionnaire. In the area of reading strategies, many teachers reported that they did not use approximately half of the listed reading strategies. Some of the strategies that were unfamiliar to many teachers were: PAR lesson framework, INSERT, Pattern Guides, Cubing, and Semantic Feature Analysis. Teachers reported most frequently using comprehension questions (74.0% use often or very frequently) and reading questions (70.9% often or very frequently). Despite being unfamiliar with many of the strategies on the survey, nearly three-fourths (72.4% often or very frequently) of teachers indicated they “teach various comprehension strategies to all students throughout the year”.

Regarding teaching activities, teachers reported using lecture most often (63.8% often or very frequently). Other commonly used teaching activities were project reports/presentations, group projects, worksheets, and brainstorming. The most frequently assigned homework was readings from the textbook, and the most common testing method was multiple choice questions. Among the most common teaching practices were: teaching with an overall yearly plan in mind, promoting question/answer classroom interaction, using sources besides the required textbook, giving students the responsibility for learning the material, explaining what students will do next, and using state standards to guide teaching practices.

Sub-goal 4.3 (Study 7): This sub-goal involves assessing the teacher interface. In Year 3, the usability of the system will be tested with six teachers. The teachers will be given a mock set of students, and asked to accomplish a particular set of goals. The goals include, but are not limited to, creating class rosters, setting dates for pre-testing, iSTART training and post-training practice sessions and evaluating performance data sets. Teacher feedback will be used to further iSTART system development.

Subgoal 4.4 (Study 8): This sub-goal involves the in classroom assessment of the teacher interface, and expanded iSTART system. In Year 3, the system will be tested on a small scale with 6 teachers, in approximately 30 classrooms, within a single high school. The teachers will be provided with the necessary training and asked to integrate the system into their classroom. The assessments will include measures of student performance (e.g., pretest/posttest comprehension performance), but it will focus on factors such as teacher classroom practices, the number of reading assignments, reading strategy knowledge, reading strategy instruction, technology experience, and attitude toward technology use in the classroom. This data will be collected via archival data, surveys, focus groups, and classroom observations and will be used to help refine the iSTART system. In addition, teachers’ and students’ use of iSTART will be automatically collected via the web. Using the standard design approach, if problems are detected or if teachers make particular requests, revisions to iSTART will be made during that semester. At the end of the year, survey data will be collected by our outside evaluation team (i.e., CREP) to assess the impact of iSTART on the school and the classroom.

References

Alexander, P. A. (1992). Domain knowledge: Evolving themes and emerging concerns. *Educational Psychologist*, *27*, 33-51.

Baker, L. (1996). Social influences on metacognitive development in reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 331-352). Mahwah, NJ: Erlbaum.

Bellissens, C., Jeuniaux, P., Duran, N. D., & McNamara, D. S. (2007). Towards a textual cohesion model that predicts self-explanation inference generation as a function of text structure and readers' knowledge. In the *Proceedings of the 29th Annual Meeting of the Cognitive Sciences*. Austin, TX: Cognitive Science Society.

Bellissens, C., & McNamara, D. S. (2006, July). Using fuzzy logic to simulate human ratings of self-explanations in iSTART. *Poster presented at the 16th Annual Meeting of the Society for Text & Discourse*, Minneapolis, MN.

Bereiter, C., & Bird, M. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and Instruction*, *2*, 131-156.

Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Bob Glaser* (pp. 361-392). Hillsdale, NJ: Erlbaum.

Best, R., Ozuru, Y., & McNamara, D. S. (2004). Self-explaining science texts: Strategies, knowledge, and reading skill. In Y. B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), *Proceedings of the Sixth International Conference of the Learning Sciences: Embracing Diversity in the Learning Sciences* (pp. 89-96). Mahwah, NJ: Erlbaum.

Bowen, B. A. (1999). Four puzzles in adult literacy: Reflections on the national adult literacy survey. *Journal of Adolescent and Adult Literacy*, *42*, 314-323.

Brown, A. L., Ellery, S., & Campione, J. C. (1998). Creating zones of proximal development electronically. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 341-367). Mahwah, NJ: Erlbaum.

Cornoldi, C., & Oakhill, J. (Eds.). (1996). *Reading comprehension difficulties: Processes and intervention.* Mahwah, NJ: Lawrence Erlbaum Associates.

Craig, S., Gholson B., Ventura, M., Graesser, A. C., & the Tutoring Research Group (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education (Special Issue on Analyzing Educational Dialogue Interaction)*, *11*, 242-253.

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension test such as SAT. *Journal of Experimental Psychology: General, 130*, 208-223

Duran, N. D., Bellissens, C., Taylor, R. S., & McNamara, D. S. (2007). Quantifying text difficulty with automated indices of cohesion and semantics. In the *Proceedings of the 29th Annual Meeting of the Cognitive Sciences*. Austin, TX: Cognitive Science Society.

Forget, M. A. (1999). *Comparative effects of three methods of staff development in content area reading instruction on urban high school teachers*. Unpublished doctoral dissertation, Old Dominion University, Norfolk.

Freebody, P., & R. C. Anderson. (1983). Effects of vocabulary difficulty, text cohesion and schema availability on reading comprehension. *Reading Research Quarterly, 18*, 277-94.

Gernsbacher, M. A., & Hargreaves, D. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*, 699-717.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, & Computers*, *36*, 193-202.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Education Research Journal*, *31,* 104-137.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, 8, 129-148.

Graesser, A. C., Wisher, R. A. (2001). *Question generation as a learning multiplier in distributed learning environments* (Rep. No. A654993). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Hung, D. W. L. (2001). Design principles for web-based learning: Implications for Vygotskian thought. *Educational Technology*, *41*, 33-41.

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Education Research Journal*, *29,* 303-323.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Education Research Journal*, *31,* 338-368.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development*, *15*, 309-328.

Kuhn, D., & Weinstock, M. (2002). What is epistemological thinking and why does it matter? In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 121-144). Mahwah, NJ: Erlbaum.

Kurby, C. A., Ozuru, Y., & McNamara, D. S. (2007). Individual differences in comprehension monitoring ability during reading. In the *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Landauer, T., Laham, D., & Foltz, P. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, New Jersey: Erlbaum.

Louwerse, M.M., McNamara, D.S., Graesser, A.C., Lewis, G. & Zirnstein, M. (2006). An eye for an eye, and for other modalities. In Silva, M. & Cox, A. (Eds.), *Proceedings of the Cognitive Science Workshop "What have eye movements told us so far, and what is next?"* London, University College London.

Luckin, R. (2001). Designing children's software to ensure productive interactivity through collaboration in the zone of proximal development (ZPD). *Information Technology in Childhood Education Annual*, *2001*, 57-85.

Magliano, J. P., & Millis, K. K. (2004). Assessing reading skill with a think-aloud procedure. *Cognition and Instruction, 21,* 251–283.

Magliano, J. P., Todaro, S., Millis, K., Wiemer-Hastings, K., Kim, H. J., & McNamara, D. S. (2005). Changes in reading strategies as a function of reading training: A comparison of live and computerized training. *Journal of Educational Computing Research*, *32*, 185-208.

Magliano, J. P., Weimer-Hastings, K., Millis, K. K., Muñoz, B. D, & McNamara, D. S. (2002). Using Latent Semantic Analysis to assess reader strategies. *Behavior Research Methods, Instruments, & Computers, 34*, 181-188.

McNamara, D. S. (2007). *Reading comprehension strategies: Theories, interventions, and technologies.* Mahwah, NJ: Erlbaum.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*, 1-30.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 227-241). Mahwah, NJ: Erlbaum.

McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247-287.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1-43.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers*, *36*, 222-233.

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict readability and facilitate comprehension* (unpublished technical report). Memphis, TN: University of Memphis.

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research. 34*, 147-171.

McNamara, D.S., & Scott, J.L. (1999). Training reading strategies. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 387-392). Hillsdale, NJ: Erlbaum.

Miller, G.A. (1995). WorldNet: A lexical database. *Communication of the ACM, 38*, 39-41.

Millis, K. K., Kim, H. J., Todaro, S., Magliano, J., Wiemer-Hastings, K., & McNamara, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods, Instruments, & Computers*, *36*, 213–221.

Mokhtari, K., & Reichard, C. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology*, *94*, 249-259.

Murray, T., & Arroyo, I. (2002, June). Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems* (pp. 749-758). Berlin: Springer-Verlag.

Murray, T., & Arroyo, I. (2004, April). The zone of proximal development as a process-oriented metric for instructional software. *Paper presented at the American Educational Research Association 2004 Annual Meeting,* San Diego, CA.

Murray, J. D. & Burke, K.A. (2003). Activation and encoding of predictive inferences: The role of reading skill. *Discourse Processes, 35,* 81–102.

O'Reilly, T. P., Sinclair, G. P., & McNamara, D.S. (2004a). Reading strategy training: Automated versus live. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the*

26th Annual Conference of the Cognitive Science Society (pp. 1059-1064). Mahwah, NJ: Erlbaum.

O'Reilly, T. P., Sinclair, G. P., & McNamara, D. S. (2004b). iSTART: A web-based reading strategy intervention that improves students' science comprehension. In Kinshuk, D. G. Sampson, & P. Isaías (Eds.), *Proceedings of the IADIS International Conference Cognition and Exploratory Learning in Digital Age: CELDA 2004* (pp. 173-180). Lisbon, Portugal: IADIS Press.

O'Reilly, T., Taylor, R. S., Duran, N., & McNamara, D. S. (2006, July). Beyond the science domain: The transfer of reading strategies to non-expository text. *Paper presented at the 16th Annual Meeting of the Society for Text and Discourse,* Minneapolis, MN.

Pressley, M., Wood, E., Woloshyn, V., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychology*, *27,* 91-109.

Rosenshine, B., & Meister, C. (1994). Cognitive strategy instruction in reading. In D. A. Hayes, & S. A. Stahl (Eds.), *Instructional models in reading* (pp.85-107). Hillsdale, NJ: Erlbaum.

Rus, V., McCarthy, P. M., & Graesser, A.C. (2006). Analysis of a textual entailer. In A. Gelbukh (ed.), *Proceedings of the Seventh International Conference of the Computational Linguistics and Intelligent Text Processing* (pp. 287-298). Berlin: Springer-Verlag.

Rus, V., McCarthy, P.M., Lintean, M.C., Graesser, A.C., & McNamara, D.S. (2007). Assessing student self-explanations in an intelligent tutoring system. In D. S. McNamara and G. Trafton (Eds.), Proceedings of the 29th annual conference of the Cognitive Science Society. Austin, Texas: Cognitive Science Society.

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes, 29*, 37-60.

Schmalhofer, F., McDaniel, M. A & Keefe, D. (2002*)* A unified model of predictive and bridging inferences. *Discourse Processes. 33,* 105-132.

Schmitt, M. A. (1990) A questionnaire to measures children's awareness of strategic reading process. *The Reading Teacher*, *49*, 454-461.

Singer, M., & Remillard, G. (2004). Retrieving text inferences: Controlled and automatic influences. *Memory & Cognition, 32*, 1223-1237.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Taylor, R. S. (2004). *Informal science learning: influences of explanatory elaboration and learner control on knowledge acquisition*. Unpublished doctoral dissertation, University of Pittsburgh, PA.

Taylor, R. S., & McNamara, D. S. (2006, July). Interest in the domain of science: Impact on expository science text self-explanation quality. *Paper presented at the 16th Annual Meeting of the Society for Text and Discourse,* Minneapolis, MN.

Taylor, R. S., Sinclair, G. P., & McNamara, D. S. (2006, August). Reducing the constraints of readers' text interest: Impact of strategy training on interest, understanding, and self-explanation quality. *Paper presented at the European Association for Research on Learning and Instruction SIG 2 Biennial Meeting*, Nottingham, UK.

Wolfe, M. B. W., & Goldman S. R. (2003). Use of Latent Semantic Analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments, & Computers*, *35*, 22-31.

Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

2006-2007 Publications

Bellissens, C., Jeuniaux, P., Duran, N., & McNamara, D. (2007). Towards a textual cohesion model that predicts self-explanations inference generation as a function of text structure and readers' knowledge levels. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Boonthum, C., Levinstein, I., & McNamara, D. S. (2007). Evaluating self-explanations in iSTART: Word matching, latent semantic analysis, and topic models. In A. Kao, & S. Poteet (Eds.), *Natural language processing and text mining* (pp. 91-106). London: Springer-Verlag UK.

Crossley, S.A., Dufty, D.F., McCarthy, P.M., & McNamara, D.S. (2007). Toward a new readability: A mixed model approach. In D.S. McNamara and G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Duran, N., Bellissens, C., Taylor, R., & McNamara, D. (2007). Qualifying text difficulty with automated indices of cohesion and semantics. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Kurby, C.A., Ozuru, Y., & McNamara, D.S. (2007). Individual differences in comprehension monitoring ability during reading. *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis.* Mahwah, NJ: Erlbaum.

Levinstein, I.B., Boonthum, C., Pillarisetti, S.P., Bell, C., & McNamara, D.S. (in press). iSTART 2: Improvements for Efficiency and Effectiveness. *Behavior Research Methods.*

Lightman, E.J., McCarthy, P.M., Dufty, D.F., & McNamara, D.S. (2007). Using computation text analysis tool to compare the lyric of suicidal and non-suicidal song-writers. In D.S. McNamara and G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107-136). Mahwah, NJ: Erlbaum.

McCarthy, P.M., & McNamara, D.S. (2007). Are seven words all we need? Recognizing genre at the sub-sentential level. In D.S. McNamara and G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

McNamara, D. S. (Ed.) (2007). *Reading comprehension strategies: Theory, interventions, and technologies*. Mahwah, NJ: Erlbaum.

McNamara, D.S., & the CSEP Lab. (2006). *iSTART: Benefits and effects of extended practice*. (Technical Report). University of Memphis. http://csep.psyc.memphis.edu/mcnamara/pdf/iSTART2006.pdf

McNamara, D.S., & the CSEP Lab. (2006). *iSTART: A classroom study*. (Technical Report). University of Memphis. http://csep.psyc.memphis.edu/mcnamara/pdf/iSTARTweb.pdf

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 227-241). Mahwah, NJ: Erlbaum.

McNamara, D. S., de Vega, M., & O'Reilly, T. (in press). Comprehension skill, inference making, and the role of knowledge. In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes*. Mahwah, NJ: Erlbaum.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers, 36*, 222-233.

McNamara, D. S., & O'Reilly, T. (in press). Theories of comprehension skill: Knowledge and strategies versus capacity and suppression. In F. Columbus (Ed.), *Progress in experimental psychology research*. Hauppauge, NY: Nova Science Publishers, Inc.

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34*, 147-171.

McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 397-421). Mahwah, NJ: Erlbaum.

McNamara, D. S., Ozuru, Y., Best, R., & O’Reilly, T. (2007) The 4-Pronged Comprehension Strategy Framework. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 465-496). Mahwah, NJ: Erlbaum.

Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S., & McNamara, D. S. (2007). Assessing and improving comprehension with Latent Semantic Analysis. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 207-225). Mahwah, NJ: Erlbaum.

Muñoz, B., Magliano, J. P., Sheridan, R., & McNamara, D. S. (2006). Typing versus thinking aloud when reading: Implications for computer-based assessment and training tools. *Behavior Research Methods, Instruments, & Computers, 38*, 211-217.

O'Reilly, T., & McNamara, D. S. (2007) Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43,* 121-152.

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading strategy knowledge on more traditional "High-Stakes" measures of high school students' science achievement. *American Educational Research Journal, 44,* 161 -196.

O'Reilly, T., Taylor, R. S., & McNamara, D. S. (2006). Classroom based reading strategy training: Self-explanation vs. reading control. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1887-1892). Austin, TX: Cognitive Science Society.

Rus, V., McCarthy, P.M., Lintean, M.C., Graesser, A.C., & McNamara, D.S. (2007). Assessing student self-explanations in an intelligent tutoring system. In D. S. McNamara and G. Trafton (Eds.), Proceedings of the 29th annual conference of the Cognitive Science Society. Austin, Texas: Cognitive Science Society.

Taylor, R. S., O'Reilly, T., Rowe, M., & McNamara, D. S. (2006). Improving understanding of science texts: iSTART strategy training vs. web design control task. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 2234-2239). Austin, TX: Cognitive Science Society.

Taylor, R., O'Reilly, T., Sinclair, G., & McNamara, D. S. (2006). Enhancing learning of expository science texts in a remedial reading classroom via iSTART. In S. Barab, K. Hay, & D. Hickey

(Eds.), *Proceedings of the Seventh International Conference of Learning Sciences* (pp. 765-770), Bloomington, Indiana.

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. C. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading sections of the SAT. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 137-172). Mahwah, NJ: Erlbaum.