


Metacognitive Overload!: Positive and Negative Effects of Metacognitive Prompts in an Intelligent Tutoring System

Kathryn S. McCarthy¹  · Aaron D. Likens¹ ·
Amy M. Johnson¹ · Tricia A. Guerrero¹ ·
Danielle S. McNamara¹

Published online: 21 February 2018

© International Artificial Intelligence in Education Society 2018

Abstract Research suggests that promoting metacognitive awareness can increase performance in, and learning from, intelligent tutoring systems (ITSs). The current work examines the effects of two metacognitive prompts within iSTART, a reading comprehension strategy ITS in which students practice writing quality self-explanations. In addition to comparing iSTART practice to a no-training control, those in the iSTART condition ($n = 116$) were randomly assigned to a 2 (performance threshold: off, on) \times 2 (self-assessment: off, on) design. The performance threshold notified students when their average self-explanation score was below an experimenter-set threshold and the self-assessment prompted students to estimate their self-explanation score on the current trial. Students who practiced with iSTART had higher posttest self-explanation scores and inference comprehension scores on a transfer test than students in the no training control, replicating previous benefits for iSTART. However, there were no effects of either metacognitive prompt on these learning outcomes. In-system self-explanation scores indicated that the metacognitive prompts were detrimental to performance relative to standard iSTART practice. This study did not find benefits of metacognitive prompts in enhancing performance during practice or after the completion of training. Such findings support the idea that improving reading comprehension strategies comes from deliberate practice with actionable feedback rather than explicit metacognitive supports.

Keywords Intelligent tutoring systems · Metacognition · Reading comprehension · Log data

✉ Kathryn S. McCarthy
ksmccar1@asu.edu

¹ Arizona State University, Tempe, AZ, USA

Introduction

Metacognition, or “thinking about thinking,” refers to processes related to evaluating what one knows (Flavell 1979). For example, at the close of a challenging class, a student might reflect on how much of the lecture she has actually understood. Later that day, she might make judgments about how long it will take her to review the material and evaluate which specific topics she understands least so that she can spend more time on them. This type of reflection on to-be-learned material is characteristic of *skilled* metacognition.

Metacognition has been shown to relate to a variety of learning outcomes (Hacker et al. 1998; Metcalfe 1996). Unfortunately, students often fail to engage in metacognitive reflection (Pintrich 2000; Varner et al. 2013; Zimmerman 2008; Zimmerman and Schunk 2001) and often inaccurately assess their own learning (Maki 1998). Research in the design of intelligent tutoring systems, or ITSs, has shown that the inclusion of metacognitive prompts can increase performance in the system as well as the efficacy of the training in terms of learning outcomes (Azevedo 2005; Azevedo and Hadwin 2005; Roll et al. 2011).

This study investigates the effects of metacognitive supports in the context of the Interactive Strategy Training for Active Reading and Thinking, or iSTART, an ITS that provides students with instruction and practice to use self-explanation and effective comprehension strategies while reading challenging text (McNamara et al. 2004; McNamara et al. 2006; Snow et al. 2016). Initial work with metacognitive prompts in iSTART indicated that an indirect prompt designed to encourage metacognition improved self-explanation performance during iSTART practice (Snow et al. 2015a). The current study investigates the effects of this prompt and a more direct metacognitive prompt on both in-system performance and post-training learning outcomes.

iSTART

Students often struggle to successfully comprehend the informational texts they encounter in school (NAEP 2015). One method that has been shown to improve comprehension is *self-explanation*, or explaining the meaning of a text to one’s self (Chi et al. 1994). Self-explanation *training* has been shown to further improve reading comprehension (McNamara 2004; McNamara et al. 2006). iSTART is a computer-based version of an effective classroom-based intervention, Self-Explanation Reading Training (SERT; McNamara 2004). In iSTART, students practice writing high quality self-explanations using five comprehension strategies: *comprehension monitoring*, *paraphrasing*, *predicting*, *bridging*, and *elaboration* (McNamara et al. 2004). Comprehension monitoring allows students to evaluate their understanding of the text and to repair comprehension problems by using other strategies (McNamara 2004; McNamara et al. 2007). Paraphrasing, or restating the text in different words, helps readers understand and remember what they have read (Hagaman and Reid 2008). Using prediction, students use prior knowledge and previous text to make inferences about what information may come next in the text (Palinscar and Brown 1984). Bridging refers to when readers connect ideas in the current sentence to ideas from previous parts of the text, whereas elaboration refers to connecting ideas in the sentence to information from prior knowledge (Singer 1988).

In the initial training phase of iSTART, students watch an introductory video on why self-explanation is beneficial for comprehension as well as videos on each of the five strategies. After a summary video, students are transitioned to the non-game-based module, Coached Practice. During Coached Practice, students are prompted to generate self-explanations for various target sentences. A pedagogical agent provides feedback and allows the student to revise this self-explanation (see Fig. 1). iSTART evaluates each self-explanation using a natural language processing (NLP) algorithm (McNamara et al. 2007). This algorithm scores the self-explanation as *poor*, *fair*, *good*, or *great* (internally, this is a numeric score from 0 to 3) and determines which formative feedback messages should be displayed.

After completing one text in Coached Practice, students are directed to the practice environment in which they can play both generative and identification games. They may also revisit the video lessons or Coached Practice. In *generative* games, students read texts and write out their own self-explanations. The same NLP algorithm used in Coached Practice is used to provide a summative score. Students earn more points by writing higher quality self-explanations. In *Showdown*, students' aim to write a better self-explanation than their CPU opponent (Fig. 2). In *Map Conquest*, higher self-explanation scores are awarded more flags to conquer CPU opponents.

In *identification* games, students earn points for correctly identifying the strategy used in an example self-explanation. For instance, in the game *Dungeon Escape*, students must escape from a castle by selecting the correct door. Each door represents a different strategy. Selecting the correct strategy earns points and allows the student to progress to the next room whereas selecting the wrong strategy loses points and eventually wakes the sleeping guard (Fig. 3).

Metacognition

Students generally have poor metacognition and do not readily engage in metacognitive reflection (Pintrich 2000; Varner et al. 2013; Zimmerman 2008; Zimmerman and Schunk 2001). However, better readers demonstrate more comprehension monitoring

The screenshot displays the 'Coached Practice' interface. At the top, a 'Title' field contains the word 'Gravity'. Below this, a text box contains the prompt: 'Gravity: What is gravity? What is the force that causes an object like a car to accelerate down a ramp? You probably know gravity is involved.' To the right of the text box is a small illustration of a male character in a green shirt and white pants. Below the text box, it says 'Turn No: 1/5'. At the bottom left, it shows 'Final points on last self-explanation: 50' and 'You total iSTART points for this text: 50'. A 'NEXT' button is at the bottom center. On the right side, there is a 'History of your Self-Explanations' section with a text entry: 'Sent 1: Yeah, gravity is the force between different objects. I think that the larger the object is, the more gravity pulls other objects towards it. - [50 pts]'. Below this is a 'Self-Explanation:' section with the text: 'Yeah, gravity is the force between different objects. I think that the larger the object is, the more gravity pulls other objects towards it.' A 'Submit Your Self-Explanation' button is below the text. At the bottom right, there is a 'Your Self-Explanation is:' section with a progress bar showing 'Poor', 'Fair', 'Good', and 'Great'. The 'Good' and 'Great' segments are highlighted in green. A 'Scoring Guide' button is at the bottom right.

Fig. 1 Coached Practice



Fig. 2 Showdown

than poor readers (Kirby and Moore 1987; Samuels et al. 2005; Thiede et al. 2017). Prompting metacognition has been shown to improve reading comprehension (see Baker and Beall 2009; Haller et al. 1988; Langenberg 2000; Snow 2002) as well as performance and learning outcomes in ITSs (Azevedo and Hadwin 2005; Azevedo et al. 2016; Graesser and McNamara 2010; Roll et al. 2011).

Methods of prompting metacognition can be categorized in several ways. Prompts can directly or indirectly encourage metacognition (Graesser and McNamara 2010). Direct prompts provide explicit instruction to evaluate knowledge and understanding and can also provide instruction on how to improve metacognitive behaviors. Indirect prompts do not teach monitoring, but encourage use of the metacognitive strategies



Fig. 3 Dungeon Escape

(Bannert et al. 2009). Metacognitive prompts can also be classified as global or local. Global prompts appear at the end of the task and are intended to encourage reflection on overall performance. Local prompts ask students to evaluate their performance on individual trials. Exposure to metacognitive prompts may also encourage learners to engage in similar metacognition in the context of future situations that lack such prompts. Importantly, not all prompts may be appropriate for all tasks. Researchers and educators also need to consider how frequently these prompts are provided. Too few prompts may show little effect, but too many prompts could detract from the actual task.

Metacognition is essential to reading comprehension and a fundamental aspect of self-explanation. Self-explaining is one way of externalizing the comprehension process and it is inherently metacognitive (McNamara and Magliano 2009). The iSTART video lessons provide information about comprehension monitoring, but once in the practice environment, there are no explicit metacognitive prompts. It stands to reason that prompting students to reflect on their performance during practice could improve the efficacy of training. As such, two prompts were developed to encourage metacognition during generative practice (i.e., Map Conquest or Show-down). These prompts encourage students to think about their system performance, which is likely to indirectly increase metacognitive reflection by providing an external evaluation of performance.

Performance Threshold The performance threshold encourages students to consider their performance at the end of each game. The performance threshold is a notification that is triggered when the average self-explanation score for the game is below a given threshold. Self-explanation scores of 0 and 1 indicate that a self-explanation is too short to be of substance, irrelevant to the target sentence and text assigned, or lexically too similar to the target sentence (e.g., a restatement of the target sentence). Scores of 2 or higher reflect that the reader has demonstrated deeper comprehension of the text through the integration of prior knowledge or connection of ideas within the text into their response (Jackson and McNamara 2011). Given that inferencing and integrating are critical for successful comprehension, the performance threshold was set at 2. When the performance threshold is triggered, a pop-up message appears (Fig. 4) and students are directed back to Coached Practice for remediation. As such, this prompt can be considered indirect (no explicit direction to comment on performance) and global (focuses on overall performance rather than performance on individual trials).

Preliminary work by Snow et al. (2015a, b) investigating this performance threshold feature indicated promising results. Students' average self-explanation score increased in the generative game following the notification. This effect was regardless of whether the student chose to stay in Coached Practice or opted to close out of the module early, suggesting that the score increase was due to the notification itself and not simply remedial practice.

Self-Assessment In contrast to the performance threshold, the self-assessment is a direct metacognitive prompt focused at the local level. On certain self-explanations (chosen at random by the researchers), the self-assessment is triggered to have students estimate their score as *poor*, *fair*, *good*, or *great* (reflected numerically as 0–3) and to rate their confidence in that score prior to receiving the score from the system (Fig. 5).

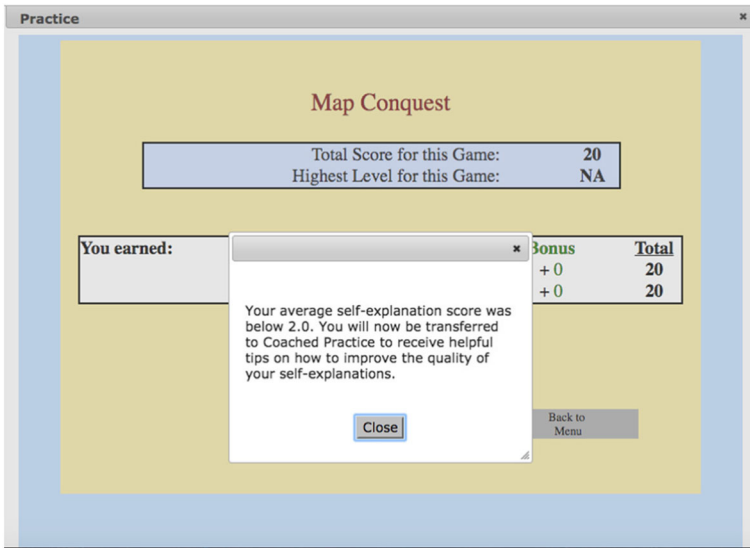


Fig. 4 Performance threshold pop-up notification

Self-assessment prompts have been used to increase metacognitive skill in a variety of educational interventions (see Falchikov and Boud 1989) as well as within ITSs (Gama 2004). In a short study conducted with iSTART, students were prompted to self-assess their performance for two texts. Consistent with previous work on metacognition, they tended to overestimate their performance. However, those with more knowledge related to the topic (i.e. higher prior domain knowledge) were more accurate in their estimates than those with low prior knowledge (Snow et al. 2015a, b).

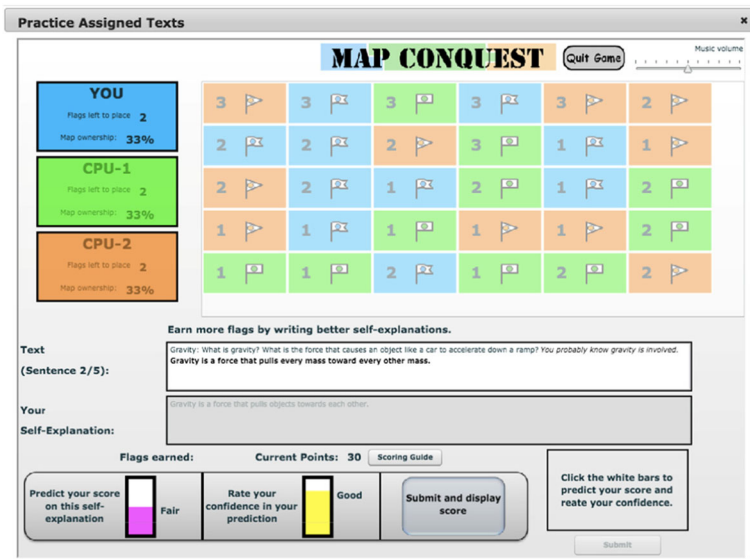


Fig. 5 Self-Assessment Prompt

Current Study

The current study expands on preliminary work (McCarthy et al. 2017) to more thoroughly explore the effects of two metacognitive prompts within iSTART. Previous empirical investigations have looked at the performance threshold and self-assessment prompts independently (Snow et al. 2015a, b), but have not examined their combined effects. Further, these studies have been relatively short-term (1 h) and have focused on how the prompts affect in-system performance. This study addresses these limitations in several ways. First, participants received extended iSTART practice (6 h). Second, we implemented a pretest and posttest that assessed self-explanation and comprehension skill before and after training as well as a more difficult transfer test that did not prompt for self-explanations. Thus, the current study investigates both in-system performance and post-training outcomes. Finally, the study employed a large sample ($n = 116$ receiving iSTART) allowing us to assess both main effects and the interaction of the prompts.

This investigation of metacognitive prompts was conducted as a part of a larger investigation of the effects of extended training in iSTART. Thus, in addition to the 116 participants in the 2(performance threshold: off, on) \times 2(self-assessment: off, on) manipulation, an additional 118 participants completed a pretest and a delayed posttest with no iSTART interaction (i.e., a no training control). Thus, there were two sets of predictions. One related to the overall effect of iSTART and the other specific to the metacognitive prompts embedded within the iSTART condition.

Based on existing work (e.g., McNamara et al. 2007; Jackson and McNamara 2011), it was predicted that the extended practice in iSTART would yield improved self-explanation scores from pretest to posttest. We also predicted this practice would improve comprehension test performance on both the posttest and transfer test and that this benefit would be most evident for inference-based comprehension questions that assess deeper comprehension.

One objective of this study was to examine the extent to which the metacognitive prompts indirectly increase metacognitive reflection by prompting evaluations of performance. If effective, these prompts would be expected to increase performance during training and improve post-training comprehension skill. Comparing the two prompts, self-assessment might be expected to be more effective than the performance threshold because it directly asks students to self-evaluate their performance. However, it does not explicitly compare the students' evaluation to the objective score. In contrast, the performance threshold might be expected to more effectively increase performance because it alerts students that their score is too low and has a tangible consequence of being transitioned to a remedial round of Coached Practice.

Importantly, there were three potential hypotheses regarding how the two metacognitive prompts might interact. Theories of metacognition generally state that as students gain more information about their performance during learning, they are better situated to adapt or change their future learning behaviors and strategies (Mathan and Koedinger 2005; Schraw 1994). Accordingly, it might be hypothesized that students exposed to both metacognitive supports would be best situated to adapt or change their behaviors and strategies, and subsequently show superior performance on the posttest and transfer test. A second hypothesis would be that the benefit of a metacognitive prompt is simply to remind students to consider their performance. If this were the case, combining the two prompts would not provide unique insights for the student relative to

having only one (i.e., having both prompts available would be redundant). A third hypothesis is grounded in theories of skill acquisition (Ericsson et al. 1993; Healy et al. 1993). Such theories would suggest that enhanced performance would come from the development of the skills necessary to complete the task and that this development emerges from extended, deliberate practice, rather than through prompting metacognition. The latter hypothesis predicts an overall effect of iSTART in comparison to the no-training control condition, but no effects of the metacognitive prompts.

Method

Participants

Participants were 234 (147 female, 87 male; $M_{age} = 15.90$) current high school students and recent high school graduates from the southwestern United States. The sample was 48.7% Caucasian, 23.1% Hispanic, 10.7% African American, 8.5% Asian, and 9.0% identified as other ethnicities. Participants were given financial compensation for their participation in the study.

Design

Participants were randomly assigned to either a no-training control condition ($n = 118$) or the iSTART condition ($n = 116$). Within the iSTART condition, participants were randomly assigned to a 2(performance threshold: off, on) \times 2(self-assessment: off, on) between-subjects design yielding four iSTART conditions: threshold only ($n = 28$), self-assessment only ($n = 29$), threshold and self-assessment ($n = 30$), and neither threshold nor self-assessment (iSTART control, $n = 31$).

Materials

Comprehension Tests Two texts were used for the pretest and posttest, Red Blood Cells and Heart Disease. The presentation of these texts was counterbalanced across participants. The texts were approximately 300 words and matched for linguistic difficulty and have been used in other reading comprehension studies (e.g., McNamara et al. 2006; Jackson and McNamara 2011). In each text, participants were prompted to self-explain nine target sentences. After reading, participants answered eight constructed response comprehension questions. These included both *textbase* and *inference* items. Textbase questions are designed to assess shallow comprehension and, thus, have answers that can be found in a single sentence in the text. In contrast, inference questions probe for deeper comprehension as they require the reader to connect information across two or more sentences in the text to derive the answer (See McNamara et al. 1996, for additional information on the construction of these texts and questions).

Transfer Test The transfer test was designed to assess the extent to which students could apply the strategies they had learned to a new context. The transfer text, Plant Growth, was longer (607 words) and more difficult than the pretest/posttest texts, both in terms of its intended audience (college students) and readability indices. Importantly,

participants were not prompted to self-explain while they read the transfer text. After reading, participants completed an 18-item assessment consisting of textbase and inference comprehension questions.

Post-Training Survey After completing iSTART, participants were asked a series of questions regarding their experience with iSTART. These questions assessed their motivation and perceptions of the system.

Procedure

Participants in the iSTART conditions came into the lab for five sessions. In the first session, participants completed a basic demographic questionnaire and a brief prior knowledge test. They then completed the pretest that included the self-explanations and comprehension questions. For the next three sessions (2 h each), participants completed a series of activities within the iSTART system. They first watched the iSTART video lessons and were then transitioned to Coached Practice. After one round of Coached Practice, the participants were allowed to move freely throughout the system for the remainder of the training. All features (lesson videos, Coached Practice, generative games, identification games) were always available. For the appropriate conditions, the performance threshold and self-assessment prompts were triggered during generative games. In the final session, participants completed the post-training survey, the comprehension posttest, and the transfer test.

Those in the no training control condition came into the lab for the pretest session and then returned to the lab after a few days ($M = 3.64$, $SD = .95$) to take the posttest. To ensure equal compensation, participants completed three days of an unrelated task after the posttest.

Scoring

Students' self-explanations generated when reading pretest and posttest passages were scored (0–3) using the iSTART scoring algorithm. Average self-explanation score was calculated by averaging the performance on each of the nine self-explanations. Open-ended comprehension questions for the pretest, posttest, and transfer test, were scored using rubrics developed in previous studies. These rubrics award partial credit (0.0, 0.25, 0.50, 0.75, or 1.0) for incomplete answers (McNamara et al. 2006; Ozuru et al. 2013). Two raters scored 20% of the set and achieved good reliability for all three texts: Heart Disease (Cohen's Kappa = .84), Red Blood Cells (Cohen's Kappa = .85), and Plant Growth (Cohen's Kappa = .80). The same raters then scored the remainder of the constructed responses, maintaining good reliability. Disagreements were settled by a third rater.

Results

Preliminary Results: iSTART Vs. Control

We first report a comparison of the iSTART conditions (i.e., regardless of the type of metacognitive prompt during practice) to the control condition to establish the overall

efficacy of iSTART. These analyses are not the central focus of the current study, but are intended to replicate previous studies conducted with iSTART and to demonstrate the overall effect of training. Subsequent analyses examine the effects of the 2×2 metacognitive manipulations and justify collapsing across training conditions. As will be shown in the following section, these analyses indicate that iSTART increases self-explanation scores from pretest to posttest as well as deep comprehension on a transfer test. Our subsequent target analyses examine the effects of the 2×2 metacognitive manipulations.

Self-Explanation Scores To assess the effect of self-explanation training on posttest self-explanation scores, we conducted a two-level analysis of covariance (ANCOVA) controlling for average self-explanation score at pretest. As shown in Table 1, self-explanation scores were higher for those in the iSTART condition compared to those in the control condition, $F(1, 231) = 29.78, p < .001, \eta^2_p = .11$.

Comprehension Scores A $2(\text{training: iSTART, control}) \times 2(\text{question type: textbase, inference})$ ANCOVA controlling for overall pretest comprehension score indicated a significant main effect of question type, such that participants had higher average scores for textbase items than for inference items, $F(1, 231) = p < .001, \eta^2_p = .10$. There was no main effect of training on posttest comprehension score nor was there a significant interaction, $F_s < 1.00, ns$ (Table 2). These results suggest that there was no effect of iSTART on reading comprehension of grade-appropriate texts when readers were prompted to self-explain.

To investigate the effect of iSTART training on the transfer comprehension test, we conducted a similar $2(\text{training: iSTART, control}) \times 2(\text{question type: textbase, inference})$ ANCOVA controlling for overall pretest comprehension score. This analysis revealed no main effect of training condition, $F < 1.00, ns$, but a significant main effect of question type, $F(1, 231) = 20.21, p < .01, \eta^2_p = .08$, such that students had higher average comprehension scores for the textbase questions than for the inference questions. This was qualified by a significant training by question type interaction, $F(1, 231) = 4.84, p < .01, \eta^2_p = .02$. As shown in Table 2, there was no effect of training for the textbase items, $t(231) = .60, ns$. In contrast, those in the iSTART training condition had higher average scores on the inference items than those in the control condition, $t(231) = 2.30, p < .05$.

To summarize, comparing students who practiced self-explaining with iSTART to a no-training control, iSTART increased the quality of participants' self-explanations at posttest, but had no effect on the immediate comprehension test performance (for which all students were prompted to self-explain). However, in a transfer task in which

Table 1 Means and standard deviations of self-explanation scores at pretest and posttest

	Self-Explanation Scores	
	Pretest	Posttest
iSTART (N = 116)	2.30(.55)	2.43(.47)
Control (N = 118)	2.18(.56)	2.02(.63)

Table 2 Means and standard deviations of comprehension test scores from pretest, posttest, and transfer test as a function of question type

	Comprehension Scores					
	Pretest		Posttest		Transfer Test	
	Textbase <i>M(SD)</i>	Inference <i>M(SD)</i>	Textbase <i>M(SD)</i>	Inference <i>M(SD)</i>	Textbase <i>M(SD)</i>	Inference <i>M(SD)</i>
iSTART (<i>N</i> = 116)	.62(.27)	.40(.22)	.62(.28)	.43(.25)	.36(.21)	.22(.19)
Control (<i>N</i> = 118)	.55(.29)	.36(.22)	.57(.28)	.39(.28)	.34(.20)	.16(.16)

participants were not explicitly prompted to self-explain, those who received training and practice in iSTART yielded deeper comprehension as indicated by higher scores on inference questions.

Metacognitive Prompts

The second set of analyses focused on the 116 participants who completed the 6 h of iSTART lessons and practice and were randomly assigned to the 2(performance threshold) \times 2(self-assessment) metacognitive prompt manipulation.

Post-Training Survey Analysis of the perceptions survey indicated that at least 50% of participants “agreed” or “strongly agreed” with all of the survey questions related to goal setting and the practice environment (50.8%–78.4%). Table 3 shows the average Likert response (out of 5) for each question as a function of the metacognitive prompt condition.

A series of 2 \times 2 ANOVAs indicated that the metacognitive prompt manipulation affected students’ self-reported motivation. Specifically, participants who received both the performance threshold and self-assessment more strongly agreed to the statements: *The feedback during practice was helpful* and *I set goals for myself during practice*.

Bivariate correlations were conducted to assess if strong agreement with either of these statements was related to performance on outcome measures (self-explanation scores and comprehension test scores). Consistent with the iSTART and no-training control comparison, the only significant finding emerged for the inference questions on the transfer test. Stronger agreement with the statement *I set goals for myself during practice* was correlated with average comprehension score on inference items ($r = .22$). All other correlations failed to reach significance.

Self-Explanation Score A 2(threshold: off, on) \times 2(self-assessment: off, on) ANCOVA with pretest self-explanation score as a covariate indicated no main effects nor an interaction, all F s < 1.00 (Table 4).

Comprehension Tests To test the effects of the metacognitive prompts on the open-ended comprehension questions in the posttest, we conducted a 2(threshold: off, on) \times 2(self-assessment; off, on) \times 2(question type: textbase, inference) repeated-measures

Table 3 Average agreement rating as a function of training condition

Item	Control	Threshold only	Self-assessment only	Both	ANOVA
	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	<i>F(p)</i>
I enjoyed using the strategy practice environment	3.33 (1.18)	3.36 (1.06)	3.24 (1.09)	3.55 (0.99)	0.42 (0.74)
The feedback during practice was helpful	3.57 (1.31)	3.46 (1.14)	3.41 (1.09)	4.14 (0.83) **	2.65 (0.05)
The interface had game-like features	3.90 (1.03)	3.64 (1.03)	4.03 (0.94)	4.14 (0.79)	1.44 (0.24)
The environment provided a purpose for my actions	3.63 (1.00)	3.61 (1.13)	3.66 (1.01)	4.10 (0.67)	1.73 (0.17)
I set goals for myself during practice	3.00 (1.20)	3.25 (1.30)	3.38 (1.12)	3.93 (1.00) *	3.39 (0.02)
The visual parts of the environment made practice more enjoyable	3.73 (1.20)	3.18 (1.34)	3.31 (1.17)	3.72 (1.03)	1.66 (0.18)
The objects in the environment were easy to control	3.77 (1.10)	3.64 (1.06)	3.83 (0.93)	4.28 (0.80)	2.28 (0.08)
I wanted to perform well during practice	3.83 (1.05)	3.96 (1.04)	4.21 (0.77)	4.31 (0.71)	1.71 (0.17)
I would use this environment to practice other skills	3.57 (1.31)	3.46 (1.23)	3.38 (1.18)	4.03 (0.87)	1.85 (0.14)

* $p < .05$, ** $p < .01$

ANCOVA. Table 5 shows the average comprehension scores as a function of four training conditions (control, performance threshold only, self-assessment only, both). Performance threshold and self-assessment were between-subjects factors and question type was a within-subjects factor. Pretest comprehension score was entered as a covariate. The only significant result was the main effect of question type, such that participants yielded higher average scores on the textbase items ($M = .62$, $SD = .02$) than the inference items ($M = .43$, $SD = .02$), $F(1, 231) = 6.34$, $p < .05$, $\eta^2_p = .05$. All other effects failed to reach significance (Question Type \times Threshold: $F(1, 110) = 1.92$; all other F s < 1.00).

Table 4 Means and standard deviations of SE score at pretest and posttest as a function of performance threshold and self-assessment conditions

	Self-Assessment Off		Self-Assessment On	
	Pretest	Posttest	Pretest	Posttest
	<i>M(SD)</i>	<i>M(SD)</i>	<i>M(SD)</i>	<i>M(SD)</i>
Threshold Off	2.21(.60)	2.41(.48)	2.43(.48)	2.51(.48)
Threshold On	2.23(.58)	2.44(.40)	2.33(.50)	2.36(.51)

Table 5 Means and standard deviations of comprehension test scores (textbase and inference) as a function of training condition

	Comprehension Scores					
	Pretest		Posttest		Transfer Test	
	Textbase	Inference	Textbase	Inference	Textbase	Inference
	<i>M(SD)</i>	<i>M(SD)</i>	<i>M(SD)</i>	<i>M(SD)</i>	<i>M(SD)</i>	<i>M(SD)</i>
Control	.56(.28)	.36(.25)	.57(.29)	.37(.25)	.35(.19)	.19(.16)
Threshold Only	.55(.29)	.40(.21)	.64(.26)	.41(.23)	.34(.20)	.20(.20)
Self-Assessment Only	.69(.22)	.45(.21)	.65(.26)	.53(.26)	.40(.25)	.26(.23)
Both	.67(.25)	.40(.20)	.62(.30)	.40(.23)	.35(.19)	.22(.16)

A similar $2 \times 2 \times 2$ ANCOVA was conducted with the transfer test scores as the dependent variable. Again, there was a significant effect of question type, $F(1,110) = 6.10, p < .05, \eta^2_p = .05$. Participants produced significantly higher average scores for the textbase items ($M = .36, SD = .02$) than for the inference items ($M = .22, SD = .02$). There were no other significant effects, all $F_s < 1.00$.

In-System Performance

On average, students completed 3.61 ($SD = 3.08$) rounds of Coached Practice (not including the initial training Coached Practice). They played 20.19 ($SD = 13.01$) identification mini-games and 6.32 ($SD = 3.53$) generative practice games.

Self-Assessment Accuracy over Time We used a linear growth model to assess the extent to which self-assessments became more accurate over time and whether accuracy depended on the performance threshold metacognitive prompt. Accuracy was assessed through a measure *discrepancy*, defined as the absolute difference between algorithmic score and self-assessment score. Discrepancy served as a dependent variable in two nested models (see Table 6). Self-explanation number served as the unit of time, t , in both models, where $t = 0, 1, 2, \dots, T$. Time was centered at the first self-explanation so that the intercept reflects the average initial discrepancy without the threshold metacognitive prompt. The nested models are given by

$$Discrepancy_{ii} = \beta_0 + \beta_1 time_{ii} + \beta_{0i} + \beta_{1i} time_{ii} + \epsilon_{ii} \tag{1}$$

and

$$Discrepancy_{ii} = \beta_0 + \beta_1 time_{ii} + \beta_2 threshold_i + \beta_3 (threshold_i \times time_{ii}) + b_{0i} + b_{1i} time_{ii} + \epsilon_{ii} \tag{2}$$

where (1) and (2) represent the unconditional and conditional models, respectively. In addition, i represents the i th participant, where $i = 1, 2, 3, \dots, N$, and ϵ_{ii} is the error in

Table 6 Results from a linear growth model of discrepancy as a function of time and threshold

	Unconditional		Conditional	
	Estimate	SE	Estimate	SE
<i>Fixed effects</i>				
Initial discrepancy, β_0	0.8406***	0.0626	0.8319***	0.0878
Time, β_1	-0.0100*	0.0036	-0.0122*	0.0051
Threshold, β_2	-	-	0.0172	0.1264
Time \times Threshold, β_3	-	-	0.0046	0.0075
<i>Random effects</i>				
Variance initial discrepancy, b_{0i}	0.1527		0.1560	
Variance Time slope, b_{1i}	0.0003		0.0003	
Correlation initial Discrepancy \times Time	-0.5100		-0.5000	
Residual, ϵ_{it}	0.5950		0.5941	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

prediction for each participant at each point in time. Note that (1) simply reflects the linear change in accuracy as a function of time, while (2) examines the effect of the performance threshold over and above changes that occur with practice. Model (2) also examines the extent to which the rate of change in accuracy depends on the performance threshold procedure. Improvement of model fit was assessed via log-likelihood ratio χ^2 test of competing models. Models were fit using the *lme4* package for the R programming language (Bates et al. 2015). Fixed effect p -values were estimated using the *lmerTest* package (Kuznetsova et al. 2016).

The results showed that including performance threshold in the model did not improve model fit over and above the effect of time. However, the results presented in Table 6 do support the conclusion that accuracy improves over the duration of the study. Specifically, the results from (1) suggest that, on average, each additional self-explanation improves accuracy (i.e., reduces discrepancy) by approximately 0.01 points. Model (1) also suggests an inverse relationship between the rate at which accuracy improved (β_{1i}) and initial accuracy (β_0). Participants with larger initial discrepancies increased in accuracy more quickly than did those with lower initial discrepancies, which may reflect an overall regression to the mean.

Performance Threshold To replicate the analyses in the preliminary study conducted by Snow et al. (2015a, b), we assessed performance immediately before and after the performance threshold. Log data were used to identify all generative games in which the average score was less than 2.0, allowing us to compare when the threshold triggered to when the threshold *would have triggered* in the alternate conditions. This indicated that 78 of the 116 participants had at least one average self-explanation score less than 2.0. Though the performance notification could be triggered as many times as necessary, most participants had no more than two instances of an average self-explanation score less than 2.0 (Fig. 6). Thus, we examined only these first two instances. As participants were able to move freely through the system, only 48

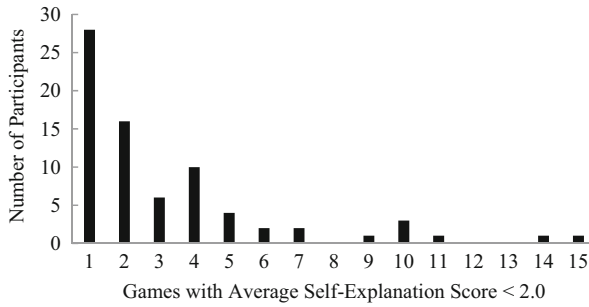


Fig. 6 Frequency of Games with Average Self-Explanation Scores < 2.0

participants across all conditions followed the *generative game, notification, generative game* sequence needed to measure effects. These participants were relatively evenly distributed across the conditions, but it is important to note that the participants in these analyses were not chosen at random, but “self-selected” based on performance. As such, the results should be interpreted accordingly.

We used the average self-explanation score immediately before and after the threshold notification to calculate a *gain score*. For the first instance of notification, the average gain scores in all conditions were positive. Despite the trend observed in Fig. 7, a 2×2 ANOVA indicated no significant effects for performance threshold, $F(1, 47) = 1.92$, *ns*, nor self-assessment, $F < 1.00$. There was also no significant interaction, $F < 1.00$ (Fig. 7).

Fewer participants ($n = 27$) had a second instance of notification. Unlike the first instance, average gain scores were either near zero or negative, indicating that the scores after notification were the same or lower than before the notification. An ANOVA revealed no main effect of performance notification or self-assessment, $F_s < 1.00$, *ns*. There was a significant notification by self-rating interaction indicating that having neither feature or both features did not affect self-explanation score, but that the presence of only one metacognitive feature was detrimental to self-explanation score, $F(1, 26) = 5.46$, $p < .05$, $\eta^2_p = .17$ (Fig. 8).

In sum, there were benefits of iSTART as compared to having no training (i.e., control condition) in terms of post-training self-explanation score and performance on inference questions in a more difficult transfer test in which students were not prompted

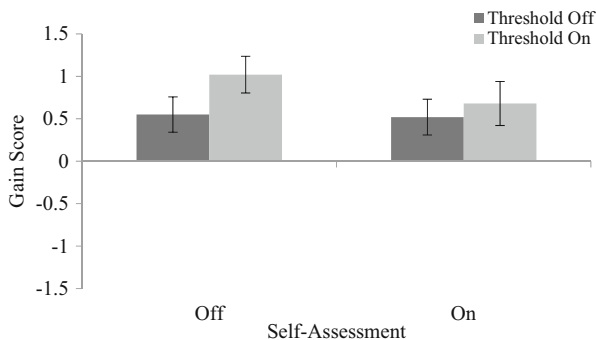


Fig. 7 Average gain score and standard error in first instance of avg. self-explanation score < 2.0 as a function of performance notification and self-assessment

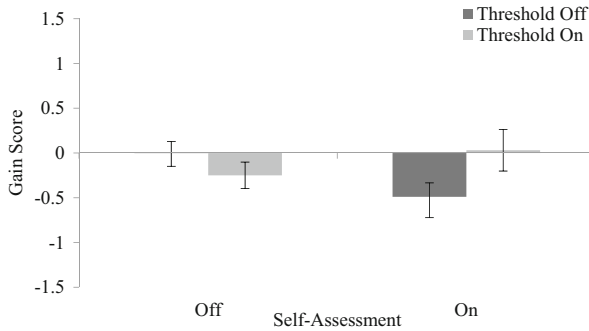


Fig. 8 Average gain score and standard error in second instance of avg. self-explanation score < 2.0 as a function of performance notification and self-assessment

to self-explain. When considering only those who received iSTART practice, the metacognitive prompts yielded positive outcomes in self-reported goal-setting. However, there were no effects of the prompts on the posttest or transfer test. Analysis of the self-assessment data indicated that self-assessments did not become more accurate over time and that there was no effect of the performance threshold. The findings related to the performance threshold notification indicated that the metacognitive prompts did not enhance performance. Indeed, there was some evidence indicating that having only one of the prompts may be detrimental to performance.

Discussion

This study examined the effects of two metacognitive prompts (performance threshold, self-assessment) on performance within iSTART and on post-training learning outcomes. We also explored the more general effect of iSTART as compared to a no training control. Consistent with previous research, iSTART improved high school students' self-explanation quality. Interestingly, comprehension test scores indicated no effect of iSTART on a comparable posttest text for which students were prompted to self-explain. Hence, this study did not replicate previous comprehension gains found for iSTART (e.g., McNamara et al. 2006; Jacovina et al. 2016). Nonetheless, there were significant benefits of iSTART on a more difficult transfer text in which participants were not prompted to self-explain. More specifically, iSTART increased deep comprehension as reflected by higher scores on inference questions. These results suggest that the experience with iSTART encouraged readers to monitor their comprehension and to employ comprehension strategies when encountering difficult texts.

Self-reports revealed that those who received both metacognitive prompts reported that they set goals for their learning and felt the feedback was helpful. However, further investigation indicated that neither prompt affected the self-explanation or comprehension test outcomes. Moreover, analysis of the in-system self-explanation practice revealed that these metacognitive prompts had no effect on the first instance of an average score less than 2.0 and a detrimental effect on performance in the second instance. This interaction in the second instance should be interpreted with caution

given the low sample size. Nonetheless, it is notable that students who are struggling to generate quality self-explanations are those who trigger the performance threshold a second time. As such, these results imply that the metacognitive prompts might be particularly damaging for less skilled readers who are most in need of iSTART.

An important question is how well these findings generalize to other systems and other domains. One possibility is that the lack of effects is idiosyncratic to iSTART. Explaining the text to oneself requires the reader to be continually monitoring comprehension, meaning that the act of self-explanation induces metacognition (McNamara and Magliano 2009). Indeed, in a variety of studies, self-explanation is identified as a type of metacognitive prompt (see Devolder et al. 2012). In ITSs that focus on other skills, metacognitive prompts activate monitoring processes that are not otherwise being recruited. In iSTART, students are already engaged in metacognitive reflection, so the prompts are redundant if not overwhelming.

Alternatively, the inconsistency between these findings and those in the extant body of research could reflect a more fundamental difference across domains. Metacognitive prompting may be more effective in well-defined domains in which students are working to overcome specific misconceptions. The open-ended and ill-defined nature of reading comprehension and self-explanation may be less susceptible to these immediate prompts. For example, students are given feedback about how to improve the quality of their self-explanations, but they are never given explicit information or a specific procedure on how to turn a “good” (2) self-explanation into a “great” (3) one. While scores are based on theoretically-motivated algorithms, the algorithms are essentially a black box to the student. As such, students may be aware that their performance is not adequate, but they may be unsure of how to resolve this discrepancy. Repeatedly asking students to evaluate their performance may be discouraging rather than helpful for students who are already struggling with the task at hand. It is likely that successful implementation of comprehension strategies comes from repeated practice and incremental gains rather than merely “flipping the switch” on a particular process. This is consistent with skill-acquisition theoretical frameworks (Ericsson et al. 1993; Healy et al. 1993), which emphasize the role of deliberate practice.

Ultimately, these results suggest that simply increasing the amount of metacognition is not an effective way of improving self-explanation quality or comprehension. And, based on these findings, we do not intend to include these prompts in future implementations of iSTART and we caution other designers of ITSs to carefully assess the effects of such scaffolding before including it in a learning environment.

Acknowledgements This research was supported in part by IES Grant R305A130124. Opinions, conclusions, or recommendations do not necessarily reflect the views of the Department of Education or IES.

References

- Azevedo, R. (2005). Computer environments as metacognitive tools for enhancing learning. *Educational Psychologist*, 40(4), 193–197.
- Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition—implications for the design of computer-based scaffolds. *Instructional Science*, 33(5), 367–379.

- Azevedo, R., Martin, S. A., Taub, M., Mudrick, N. V., Millar, G. C., & Grafsgaard, J. F. (2016, June). Are pedagogical agents' external regulation effective in fostering learning with intelligent tutoring systems?. In *International Conference on Intelligent Tutoring Systems* (pp. 197–207). Springer, Cham.
- Baker, L., & Beall, L. (2009). Metacognitive processes and reading comprehension. In S.E. Israel, & G.G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 373–388). New York: Routledge.
- Bannert, M., Hildebrand, M., & Mengelkamp, C. (2009). Effects of a metacognitive support device in learning environments. *Computers in Human Behavior*, 25(4), 829–835.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Chi, M. T. H., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.
- Devolder, A., van Braak, J., & Tondeur, J. (2012). Supporting self-regulated learning in computer-based learning environments: Systematic review of effects of scaffolding in the domain of science education. *Journal of Computer Assisted Learning*, 28(6), 557–573.
- Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363–406.
- Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, 59, 395–430.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10), 906–911.
- Gama, C. (2004). Metacognition in interactive learning environments: The reflection assistant model. In J. C. Lester, R. M. Vicario, & F. Paraguac (Eds.), *Proceedings 7th international conference on intelligent tutoring systems* (p. 668e677). Berlin: Springer.
- Graesser, A. C., & McNamara, D. S. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist*, 45, 234–244.
- Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (1998). *Metacognition in educational theory and practice*. New York: Routledge.
- Hagaman, J. L., & Reid, R. (2008). The effects of the paraphrasing strategy on the reading comprehension of middle school students at risk for failure in reading. *Remedial and Special Education*, 29(4), 222–234.
- Haller, E. P., Child, D. A., & Walberg, H. J. (1988). Can comprehension be taught? A quantitative synthesis of “metacognitive” studies. *Educational Researcher*, 17(9), 5–8.
- Healy, A. F., Clawson, D. M., McNamara, D. S., Marmie, W. R., Schneider, V. I., Rickard, T. C., Curtcher, R. J., King, C., Ericsson, K. A., & Bourne, L. E. (1993). The long-term retention of knowledge and skills. *Psychology of Learning and Motivation*, 30, 135–164.
- Jackson, G. T., & McNamara, D. S. (2011). Motivational impacts of a game-based intelligent tutoring system. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th international Florida artificial intelligence research society (FLAIRS) conference* (pp. 519–524). Menlo Park: AAAI Press.
- Jacovina, M. E., Jackson, G. T., Snow, E. L., & McNamara, D. S. (2016). Timing game-based practice in a reading comprehension strategy tutor. In A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Proceedings of the 13th international conference on intelligent tutoring systems (ITS 2016)* (pp. 80–89). Zagreb: Springer.
- Kirby, J. R., & Moore, P. J. (1987). Metacognitive awareness about reading and its relation to reading ability. *Journal of Psychoeducational Assessment*, 5(2), 119–137.
- Kurby, C. A., Magliano, J. P., Dandotkar, S., Woehrle, J., Gilliam, S., & McNamara, D. S. (2012). Changing how students process and comprehend texts with computer-based self-explanation training. *Journal of Educational Computing Research*, 47, 429–459.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2016). lmerTest: Tests in linear mixed effects models. R package version 2.0–32.
- Langenberg, D. N. (2000). *Report of the National Reading Panel: Teaching students to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Bethesda: National Institute of Child Health and Human Development, National Institutes of Health.
- Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Hillsdale: Erlbaum.
- Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist*, 40(4), 257–265.
- McCarthy, K. S., Jacovina, M. E., Snow, E. L., Guerrero, T. A., & McNamara, D. S. (2017). iSTART therefore I understand: But metacognitive supports did not enhance comprehension gains. In B. du Boulay, R.

- Baker, & E. Andre (Eds.), *Proceedings of the 18th international conference on artificial intelligence in education (AIED)* (pp. 201–211). Wuhan: Springer.
- McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38, 1–30.
- McNamara, D. S., & Magliano, J. P. (2009). Self-explanation and metacognition: The dynamics of reading. In J. D. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 60–81). Mahwah: Erlbaum.
- McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of 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. *Behavior Research Methods, Instruments, & Computers*, 36, 222–233.
- 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., 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: Erlbaum.
- Metcalf, J. (1996). Metacognitive processes. In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (pp. 383–411). San Diego: Academic Press.
- NAEP: The Nation's Report Card: Mathematics and Reading at Grade 12. (2015) U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
- Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing text comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology*, 67, 215–227.
- Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1(2), 117–175.
- Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self regulation: Theory, research and applications* (pp. 451–502). San Diego: Academic Press.
- Roll, I., Alevin, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2), 267–280.
- Samuels, S. J., Ediger, K. M., Willcutt, J. R., & Palumbo, T. J. (2005). Role of automaticity in metacognition and literacy instruction. In S. E. Israel, C. C. Block, K. L. Bauserman, & K. Kinnucan-Welsch (Eds.), *Metacognition in literacy learning: Theory, assessment, instruction, and professional development* (pp. 41–59). Mahwah: Lawrence Erlbaum Associates.
- Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology*, 19(2), 143–154.
- Singer, M. (1988). Inferences in reading comprehension. In M. Daneman, G. E. Mackinnon, & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 177–219). San Diego: Academic Press.
- Snow, C. E. (2002). *Reading for understanding: Toward a research and development program in reading comprehension*. Santa Monica: RAND.
- Snow, E. L., Jacovina, M. E., & McNamara, D. S. (2015a). Promoting metacognition within a game-based environment. In A. Mitrovic, F. Verdejo, C. Conati, & N. Heffernan (Eds.), *Proceedings of the 17th international conference on artificial intelligence in education (AIED 2015)* (pp. 864–867). Madrid: Springer.
- Snow, E. L., McNamara, D. S., Jacovina, M. E., Allen, L. K., Johnson, A. M., Perret, C. A., Dai, J., Jackson, G. T., Likens, A. D., Russell, D. G., & Weston, J. L. (2015b). Promoting metacognitive awareness within a game-based intelligent tutoring system. In A. Mitrovic, F. Verdejo, C. Conati, & N. Heffernan (Eds.), *Proceedings of the 17th international conference on artificial intelligence in education (AIED 2015)* (pp. 786–789). Madrid: Springer.
- Snow, E. L., Jacovina, M. E., Jackson, G. T., & McNamara, D. S. (2016). iSTART-2: A reading comprehension and strategy instruction tutor. In D. S. McNamara & S. A. Crossley (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 104–121). Taylor & Francis, Routledge: New York.
- Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2017). How restudy decisions affect overall comprehension for seventh-grade students. *British Journal of Educational Psychology*, 87(4), 590–605.
- Varnier, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research*, 5, 35–59.
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166–183.
- Zimmerman, B. J., & Schunk, D. H. (2001). Reflections on theories of self-regulated learning and academic achievement. *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives*, 2, 289–307.