Instr Sci (2014) 42:159–181
DOI 10.1007/s11251-013-9275-4

Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system

Reza Feyzi-Behnagh · Roger Azevedo · Elizabeth Legowski · Kayse Reitmeyer · Eugene Tseytlin · Rebecca S. Crowley

Received: 8 February 2012 / Accepted: 20 March 2013 / Published online: 11 April 2013

Abstract In this study, we examined the effect of two metacognitive scaffolds on the accuracy of confidence judgments made while diagnosing dermatopathology slides in SlideTutor. Thirty-one ($N = 31$) first- to fourth-year pathology and dermatology residents were randomly assigned to one of the two scaffolding conditions. The cases used in this study were selected from the domain of nodular and diffuse dermatitides. Both groups worked with a version of SlideTutor that provided immediate feedback on their actions for 2 h before proceeding to solve cases in either the *Considering Alternatives* or *Playback* condition. No immediate feedback was provided on actions performed by participants in the scaffolding mode. Measurements included learning gains (pre-test and post-test), as well as metacognitive performance, including Goodman–Kruskal Gamma correlation, bias, and discrimination. Results showed that participants in both conditions improved significantly in terms of their diagnostic scores from pre-test to post-test. More importantly,

R. Feyzi-Behnagh · R. Azevedo
Department of Educational and Counselling Psychology, McGill University, 3700 McTavish St., Montéal, QC H3A 1Y2, Canada
e-mail: reza.feyzibehnagh@mail.mcgill.ca

R. Azevedo
e-mail: Roger.azevedo@mcgill.ca

E. Legowski · K. Reitmeyer · E. Tseytlin · R. S. Crowley (✉)
Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Boulevard, BAUM 423, Rm 523, Pittsburgh, PA 15206-3701, USA
e-mail: crowleyrs@upmc.edu

E. Legowski
e-mail: legoex@upmc.edu

K. Reitmeyer
e-mail: reitmeyerkl@upmc.edu

E. Tseytlin
e-mail: tseytline@upmc.edu

R. S. Crowley
Department of Pathology, University of Pittsburgh, Pittsburgh, PA, USA

participants in the *Considering Alternatives* condition outperformed those in the *Playback* condition in the accuracy of their confidence judgments and the discrimination of the correctness of their assertions while solving cases. The results suggested that presenting participants with their diagnostic decision paths and highlighting correct and incorrect paths helps them to become more metacognitively accurate in their confidence judgments.

Keywords Cognitive debiasing · Computer based education · Intelligent tutoring systems · Metacognitive judgments · Diagnostic reasoning · Metacognition

Introduction

Diagnostic classification is an important part of medical care because a patient's diagnosis is often the main determinant of treatment and prognosis. Clinician overconfidence or underconfidence when evaluating performance on diagnostic tasks can result in sub-optimal patient care. Overconfidence causes the clinician to reach diagnostic closure too quickly before fully considering the evidence in the case, and can result in diagnostic errors (Graber et al. 2005; Voytovich et al. 1985). On the other hand, underconfidence may lead clinicians to pursue unnecessary and inappropriate additional testing and use of consultative services, which may increase the risk of iatrogenic complications (i.e., inadvertently caused by medical treatment or diagnosis procedures), delay appropriate treatment, and unnecessarily inflate the cost of medical care (Berner and Graber 2008; Mann 1993).

Pathologic diagnosis is one type of diagnostic task that requires microscopic examination of human tissues obtained during biopsies and other surgeries, and is of paramount importance in areas such as cancer, skin disease, and liver disease. According to Berner and Graber (2008), the extent of incorrect diagnosis in perceptual specialties such as pathology and radiology, which rely on visual interpretation, typically ranges from 2 to 5 %; however, these rates might be higher in some circumstances. McGinnis et al. (2002) found that in a second review of 5,136 pigmented lesion biopsies (in order to test for melanoma), a change of diagnosis was made in 11 % of cases, 8 % of which changed enough to alter the treatment course. An analysis of more than three hundred pathology malpractice claims filed by patients affected by misdiagnosis indicates that approximately 63 % of claims involved failure to diagnose cancer (skin, breast, and ovarian), resulting in delay in diagnosis or in inappropriate treatment (Troxel 2006). False-negative diagnoses of melanoma, malignant ovarian tumors, and breast biopsies were indicated to be the most common source of malpractice claims against pathologists (Troxel 2006).

Training of pathologists typically requires five or more years, and encompasses both residency training (3–5 years) and advanced fellowship (1–3 years). Training of highly specialized clinicians, such as pathologists, is very difficult for a variety of reasons, including insufficient exposure to infrequently encountered cases, the increased workloads of mentors that limit time for training the next generation of practitioners, and the potential for clinical errors among less-experienced practitioners. Intelligent tutoring systems (ITSs) are one type of computer-based training that could help alleviate these problems by providing a safe environment where residents can practice as frequently as needed and receive tailored feedback and guidance without inadvertently harming patients in the process (Crowley and Gryzbicki 2006). In this study, we investigated the impact of two types of metacognitive scaffolds (playing back the diagnostic decision-making process and considering alternatives) on the diagnostic reasoning of pathology and dermatology residents while solving cases in SlideTutor, an ITS.

ITSs are adaptive and personalized instructional systems designed to emulate the well-known benefits of one-on-one tutoring over other types of instructional methods (Koedinger and Aleven 2007; Shute and Zapata-Rivera 2012; Woolf 2009). ITSs have the potential to accelerate training of novices by providing individualized tutoring in the form of scaffolding and tailored feedback based on a complex interaction between several modules that represent the domain knowledge as well as learner knowledge acquisition and development of expertise. In a one-on-one tutoring situation, the human tutor or the ITS can provide adaptive scaffolding to enhance student learning (see Azevedo and Hadwin 2005; Chi et al. 1994, 2001, 2004; Graesser et al. 1997, 2000; Johnson et al. 2011; Lepper et al. 1997). Studies on the benefits of ITS for student performance have shown that learning improved beyond that achieved through classroom instruction, coming close to what can be achieved with a human tutor (Koedinger and Corbett 2006). ITSs support "leaning-by-doing," and provide individualized support, point out errors, and organize content to cater to the needs of the individual when the teacher has limited time to spend (Corbett et al. 2002; Koedinger and Aleven 2007; VanLehn 2006, 2011). Although there is a great potential for the use of medical ITS, few of these systems have been fully developed (e.g., Azevedo and Lajoie 1998; Clancey 1987; Crowley and Medvedeva 2006; Lajoie and Azevedo 2006; Lajoie 2009; Maries and Kumar 2008; Obradovich et al. 2000; Rogers 1995; Sharples et al. 2000; Smith et al. 1998), a smaller number of which have been empirically evaluated (Crowley et al. 2007; El Saadawi et al. 2008; Woo et al. 2006).

Theories of self-regulated learning

The skill to monitor and control one's cognitive processes, termed self-regulated learning (SRL), is a crucial component of developing expertise in a domain (Winne 2001; Azevedo et al. 2008, Zimmerman 2006). SRL theories attempt to model the ways cognitive, metacognitive, motivational, and emotional processes influence the learning processes. Pintrich (2000) defines SRL as a constructive process where learners set goals based on their prior knowledge and experience and their current learning context. He describes SRL as comprising three main phases: task identification and planning; monitoring and control of learning strategies; and reaction and reflection. Winne and Hadwin's model of SRL (1998, 2008), which is based on the information processing theory, outlines the cognitive processes occurring during learning. The authors propose that learning occurs in four phases: task definition; goal-setting and planning; studying tactics; and adaptations to metacognition. After the learner sets goals, she monitors both her domain knowledge and learning resources (Azevedo and Witherspoon 2009). Next, the learner examines the solution to evaluate its correctness by making a feeling-of-knowing (FOK) judgment, which is defined as the learner's certainty of her actual performance (Azevedo and Witherspoon 2009; Metcalf and Dunlosky 2008).

Different metacognitive scaffolding techniques have been used in ITSs to assist novices in activating prior knowledge, deploying appropriate strategies, and monitoring their learning and the effectiveness of the deployed strategies (Azevedo et al. 2008). Ideally, the computer-based learning environment or ITSs should gradually fade the scaffolding and support based on a dynamic evaluation of the individual learner. We have previously explored the use of ITS on enhancing metacognitive skills when students learn about challenging science topics, such as human biology (Feyzi-Behnagh et al. 2011), and found that traditional cognitive tutoring systems may be limiting as a framework for enhancing metacognition. Azevedo and Hadwin (2005) argue that constant immediate feedback in one-on-one training situations prevents learners from acquiring the required metacognitive

skills in evaluating their problem-solving abilities and becoming aware of their own errors. While the ITS discussed above have focused on global aspects of metacognitive monitoring and control (e.g., modeling, tracing, and scaffolding), none have focused on the types of metacognitive scaffolding provided by the current version of SlideTutor (used in this study).

Metacognitive judgment of clinical accuracy

Substantial prior research reveals that poor calibration of FOK plays a role in medical errors. As Mann (1993) notes, poor calibration can take the form of *underconfidence* in a correct assertion or diagnosis, which could result in a delay of treatment while additional information is sought to confirm the finding, or *overconfidence* in an incorrect assertion or diagnosis, which could result in an incorrect treatment decision because alternative hypotheses are not considered or additional information is not sought. He investigated the confidence levels and calibration of 20 first-year and 27 third-year osteopathic medical students in classifying cardiac dysrhythmia in artificially generated abnormal heart rhythms, and found that medical students were slightly underconfident overall in their diagnoses, and the accuracy and mean confidence level were higher for third-year students. In another study, Freidman et al. (2005) evaluated confidence levels of 72 students, 72 senior medical residents, and 72 faculty internists who provided diagnoses for synopses of 36 diagnostically challenging medical cases, and found mild alignment between participants' correctness and confidence. The misalignment was represented by overconfidence in 41 % of cases for residents, 36 % for faculty, and 25 % for students. A major finding of their study was that there was a positive linear relationship between diagnostic accuracy and participants' clinical experience. Participants' confidence levels also increased linearly with their clinical experience.

Overconfidence is one of many cognitive biases that can affect decision-making processes. According to Croskerry and Norman (2008), one of the possible sources of overconfidence is a *confirmation bias,* in which individuals tend to accept or be overly confident in solutions and conclusions they have reached, instead of considering alternative solutions or looking for disconfirming evidence for their hypotheses. Overconfidence emerges after physicians gain experience and become experts, at which point they solve problems mostly by pattern-recognition processes and recollection of prior similar cases with characteristic features without thinking about differential diagnoses (Berner and Graber 2008). In the culture of medicine, if a physician appears unsure, the uncertainty could be considered a sign of vulnerability and weakness, so physicians learn not to disclose their ambivalence to patients (Katz 1984). One of the solutions offered by Croskerry and Norman (2008) for overcoming overconfidence is to provide prompt feedback on mistakes in the decision-making process, but this kind of feedback is rarely consistently available to physicians. ITSs could provide a method for providing such immediate feedback, and might consequently enhance calibration of accuracy judgments.

Methods of measuring metacognitive judgment accuracy

A variety of methods have been recommended in the metacognition literature for measuring the accuracy, bias, and discrimination of learners' metacognitive judgments (Schraw 2009; Koriat et al. 2002; Nelson 1996). Schraw (2009) categorized the accuracy of metacognitive judgments into two types: absolute and relative accuracy. Absolute accuracy is defined as the measure of the accuracy of a judgment about a specific task, whereas

relative accuracy refers to the measurement of the relationship between multiple judgments and corresponding tasks (Maki et al. 2005). Pearson correlation coefficient or a contingency coefficient (e.g., Gamma) is typically used to measure the relative accuracy of metacognitive judgments (Nelson 1996).

The Goodman–Kruskal Gamma correlation (G) is a measure of relative accuracy, which assesses the relationship between confidence judgments and performance on a criterion task (Goodman and Kruskal 1954; Maki et al. 2005; Nelson 1984). This statistic is calculated as a proportional difference of concordant and discordant pairs. A concordant pair consists of either a positive FOK judgment (i.e., being sure about the correctness of an item identified in a case) and a correct response (i.e., the item being correct), or a negative FOK judgment (i.e., being unsure about the correctness of an item identified in a case) and an incorrect response. On the other hand, a discordant pair denotes a positive FOK judgment and an incorrect response, or a negative FOK judgment and a correct response. *G* correlations range from −1.0 to +1.0, where +1.0 indicates perfect correlation between FOK judgments (sure or unsure) and the actual performance (correct or incorrect), and zero indicates no correlation. It should be noted that measures of relative accuracy, like Gamma correlation, do not measure absolute precision or over- or under-confidence, and the relative accuracy might be quite high while there is low absolute precision (Juslin et al. 1996; Maki et al. 2005; Nietfeld et al. 2006).

Bias assesses the degree of over- or under-confidence of an individual when making a confidence judgment (Schraw 2009). Positive bias scores indicate over-confidence and negative bias scores indicate under-confidence. When confidence perfectly matches performance, the bias score equals zero. Since the bias score can range in negative and positive directions, it indicates the direction and degree of lack of fit between confidence and performance (Schraw 2009). The bias score (Kelemen et al. 2000) is calculated by subtracting the relative performance on all items (total correct items divided by all items) from the proportion of items judged as known (total sure items divided by all items).

Discrimination is the degree to which an individual can distinguish confidence judgments on correct versus incorrect items (Schraw 2009). When an individual is more confident about correct versus incorrect items, there is positive discrimination, whereas when he/she is more confident about incorrect items, there is negative discrimination. According to Schraw (2009), positive discrimination can be interpreted as metacognitive awareness of correct performance because the individual rates higher confidence on correct versus incorrect items.

In this study, we use Gamma, bias, and discrimination measures in order to investigate the effects of the two metacognitive scaffolds (Playback and Considering Alternatives) on the accuracy, bias, and discrimination of metacognitive judgments made by the participants.

Cognitive and metacognitive tutoring in pathology

In previous work, we used think-aloud protocols and a task-analytic approach to explore differences in microscopic diagnosis among novice, intermediate, and expert pathologists (Crowley et al. 2003b). The cognitive model derived from this work was incorporated into SlideTutor, an ITS for teaching visual classification problem-solving to residents and fellows (Crowley et al. 2003a; Crowley and Medvedeva 2006). The central assumption in the architecture of SlideTutor is that cognition is modeled by production rules. We have previously shown that the system produces a four-fold increase in diagnostic accuracy among study participants, and that learning gains are maintained at 1-week retention tests (Crowley et al. 2007). Alternative versions of our tutoring system that target other clinical skills have shown similarly strong learning gains. ReportTutor (Crowley et al. 2005) is a version of SlideTutor that

implements a natural language interface. The system analyzes diagnostic reports written by pathology residents in real-time, and highlights errors they make in their slide review and report writing. Participants who used the system experienced a four-fold improvement in their ability to write accurate and complete diagnostic reports (El Saadawi et al. 2008).

In El Saadawi et al. (2010), we investigated the effect of providing and fading immediate feedback on metacognitive performance of pathology residents in diagnosing pathology cases. The study tested immediate feedback on each student action against a set of metacognitive scaffolds during fading, including immediate feedback on FOK judgments, an inspectable student model, and a pseudo-dialog using a pre-stocked question set. We found that immediate feedback had a significant positive effect on metacognitive performance as well as learning gains. Fading of immediate feedback led to decreased metacognitive performance in terms of accuracy and discrimination of FOK judgments.

The current study investigates the impact of two kinds of metacognitive scaffolds on the accuracy of FOK judgments in SlideTutor. The metacognitive scaffolds used in this study included either showing participants the steps they took in diagnosing a case and then providing them with the correct diagnosis steps (*Playback*), or having them review their decision-making diagrams where their diagnosis steps and the correct diagnosis path were highlighted (*Considering Alternatives*). Furthermore, the study explores whether either of these interventions leads to superior learning.

Research questions

Question 1 What is the impact of Playback versus Considering Alternatives metacognitive scaffolds on participants' cognitive gains during problem solving with a medical ITS?

Question 2 What is the impact of Playback versus Considering Alternatives metacognitive scaffolds on participants' metacognitive judgments (i.e., accuracy of metacognitive judgments, bias, and discrimination) during problem solving with a medical ITS?

Hypotheses

Based on the above research questions and the findings of our previous studies we hypothesize that:

Hypothesis 1 Both metacognitive scaffolding conditions (Playback and Considering Alternatives) will lead to cognitive gains from the pre-test to post-test at the end of the tutoring session.

Hypothesis 2 Participants who are presented with their diagnostic decision paths highlighting correct and incorrect paths (i.e., Considering Alternatives) will become more metacognitively accurate in their metacognitive judgments than participants who are shown a play-by-play of their diagnostic decision-making and the optimum solution path (i.e., Playback).

Methods

Participants

Thirty-one ($N = 31$) participants were recruited from the following institutions: the University of Pittsburgh, Allegheny General Hospital, the University of Pennsylvania, Drexel

University, and Temple University. The participants were pathology and dermatology residents and included nine first-year, nine second-year, five third-year, and eight fourth-year residents. The only exclusion criterion was that participants could not have participated in our previous metacognitive study (El Saadawi et al. 2010). Participants were randomly assigned to one of the two metacognitive scaffolding conditions: *Playback* or *Considering Alternatives*. All participants were volunteers recruited by email, and each received $400 for participation. The study was approved by the University of Pittsburgh Institutional Review Board (IRB Protocol # PRO09030260).

Study design

A repeated-measures study design was used, with metacognitive scaffolds (*Playback* and *Considering Alternatives*) as the between-subjects factor, and metacognitive performance of participants before and after training as the within-subjects factor. The study was conducted over 8 h during 1 day, as illustrated in Fig. 1. A description of each phase of the timeline is provided in the "Procedure" section.

Pathology cases

Cases were obtained from the University of Pittsburgh Medical Center (UPMC) slide archive and from private slide collections. Diagnoses were confirmed by a collaborating dermatopathologist prior to use in the study. A total of 53 de-identified dermatopathology cases in the domain of nodular and diffuse dermatitides (NDD) were used for all phases of the study, including pre-test, post-test, immediate feedback, and metacognitive scaffolding tutoring interfaces. We selected NDD as the domain for tutoring in this study because it was unlikely that residents would have complete knowledge of this diagnostic area. However, the residents' prior familiarity with the domain was not tested in this study. Cases used in this study were typical instances of the entities encountered by residents in the clinical setting. For each case, a knowledge engineer and an expert dermatopathologist collaborated in defining all present and absent findings and their locations on the slides (case annotation), and in developing the relationships among findings and diagnoses (knowledge-base development). The diagnoses included sets of one or more diseases that matched the histopathologic pattern. A pattern is a combination of evidence, comprised of findings and absent findings identified in a particular case, which form the basis for the hypotheses and diagnoses.

Intelligent Tutoring System: SlideTutor

The SlideTutor ITS (http://slidetutor.upmc.edu) was modified for use in this study. The computational methods and implementation of the system have been previously described

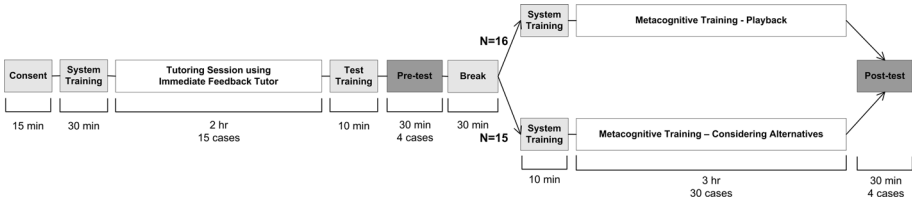

Fig. 1 Study timeline

(Crowley and Medvedeva 2006). Briefly, the system uses a client–server architecture implemented in the Java programming language, and utilizes the Jess production rule engine (http://www.jessrules.com/jess/). SlideTutor provides users with cases to be solved under supervision by the system. Cases incorporate virtual slides, which are gigabyte-sized image files created from traditional glass slides by concatenating multiple images from a high-resolution robotic microscope. Virtual slides are annotated using Protégé ontology editing environment with an in-house developed plug-in, which was an extension to Protégé, that required authors to link discrete findings with their respective locations on the slide. Protégé is a Java-based open-source platform for ontology development. An ontology describes the concepts and relationships that are important in a particular domain (e.g., taxonomies, schemas, and classifications) (see http://protege.stanford.edu). A separate Protégé-Frames expert knowledge base consists of a comprehensive set of evidence–diagnosis relationships for the entire domain of study. Jess rules utilize these static knowledge representations to create a dynamic solution graph (DSG) representing the current problem-state and all acceptable next steps, including the best-next-step (Crowley and Medvedeva 2006).

For both immediate feedback training and the two metacognitive scaffoldings, participants use a graphical user interface (Fig. 2) to examine cases and describe their reasoning. Participants pan and zoom in the virtual slide, point to findings using the mouse, and select from lists of findings and qualifiers (e.g., size and type). These named findings then appear as evidence nodes in the diagrammatic reasoning palette. Hypotheses can be asserted using a separate tree-based menu. Once asserted, they also appear as nodes in the diagrammatic reasoning palette. Support and refute links may be drawn between evidence and hypothesis nodes to indicate relationships. Finally, one or more hypotheses may be selected as the final diagnosis(es) before completing the case.

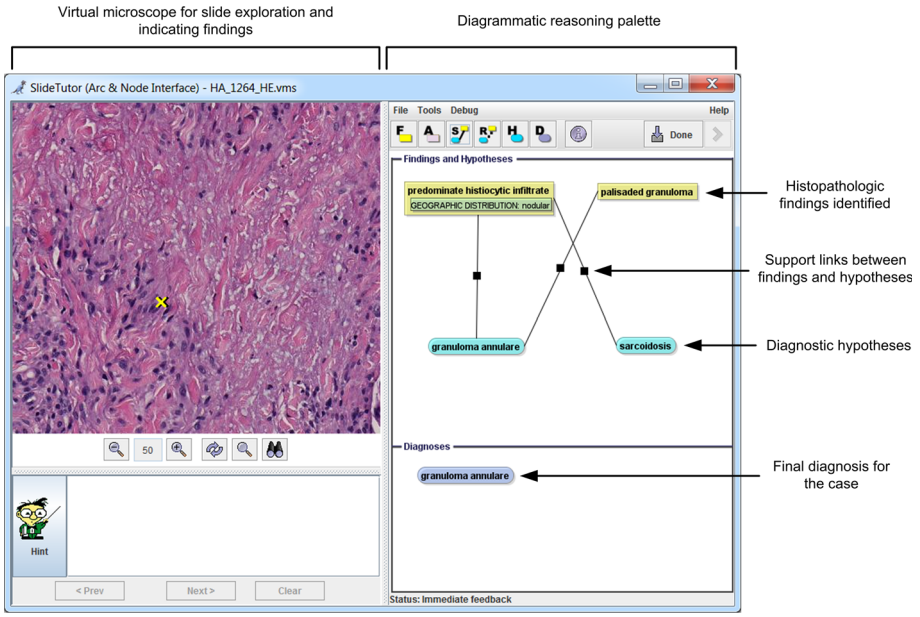

Fig. 2 SlideTutor interface

During *immediate feedback training,* the system evaluates each student action against the current state of the DSG and provides feedback on every intermediate step. Correct student actions modify the graph to produce the next problem-state for that case and student. Incorrect actions are matched against a set of specific errors and produce visual and textual explanations to help the learner identify and correct his mistakes. When the participant requests a hint, the system provides context-specific help using the computed best-next-step of the DSG, which is constantly changing as the student progresses through the case. Hints are increasingly specific, and the last hint provides the student with all declarative and procedural knowledge required to complete that step. The immediate-feedback tutoring system reproduces the general behavior of cognitive tutoring systems (Koedinger and Aleven 2007).

During *metacognitive scaffolding*, immediate feedback is removed, and feedback is provided only after completion of each case. Participants use one of two versions of the system, which differ based on the type of feedback given at the conclusion of each case. In the *Considering Alternatives* condition, participants are shown a *knowledge tree* that highlights the path taken to diagnosis, as well as the correct path for the case based on the expert model. Anything incorrect is highlighted in purple and outlined in red, and the correct path for the case is highlighted in green (Fig. 3a). We chose this metacognitive intervention since our previous study (Crowley et al. 2007) indicates that participants who use a knowledge representation that exposes the underlying structure of the decision-making process show a significant increase in metacognitive gains when compared to those who do not have this affordance. This could be due to the more holistic and global view of the problem-solving algorithm in this type of representation, and to its impact in assisting participants in seeing the effects of subtle differences that specific sets of features (i.e., patterns) have on the final diagnosis. This in turn could lead to enhanced accuracy of self-assessments because they can recognize diagnostic near-misses visually. The use of a diagnostic decision tree in our previous study (Crowley et al. 2007) offered only a passive opportunity; however, in this study, the student model has been used in addition to presenting the expert solution path, which provides learners with the opportunity to reflect on their decision-making process and compare it with the expert diagnostic decision-making path.

In the *Playback* condition, participants are shown a "*play-by-play*" of their own actions while working on the case, and are then shown a "*play-by-play*" of the optimal problem solution to get to the correct diagnosis. For the optimal solution, the tutoring system simply concatenates all best-next-steps provided by the expert model (Fig. 3b). The ability to recognize one's own errors is a complex cognitive process which affects cognitive as well as metacognitive performance. It is often difficult to notice one's own errors because the cognitive load of task performance uses the limited resources of working memory, making it difficult to simultaneously self-monitor. Thus, an alternative approach is to help participants learn to recognize their own errors by watching themselves perform the task while reflecting on it and then watching the correct solution path for the case. Viewing mistakes following each problem-solving activity could trigger metacognitive awareness in the learner which would lead to recalibration of his/her metacognitive judgments. Active monitoring and paying attention to the playback of one's problem-solving steps would lead to a change in the learner's internal standards (Winne and Hadwin 1998, 2008) toward becoming more accurate in future self-assessments of performance.

During the metacognitive scaffolding period, all participants use the *coloring book interface* (El Saadawi et al. 2010) to provide FOK judgments at the conclusion of each case (Fig. 4). The *coloring book interface* appears only after participants complete their examination of the case and click "Submit." Participants are asked to reflect on the

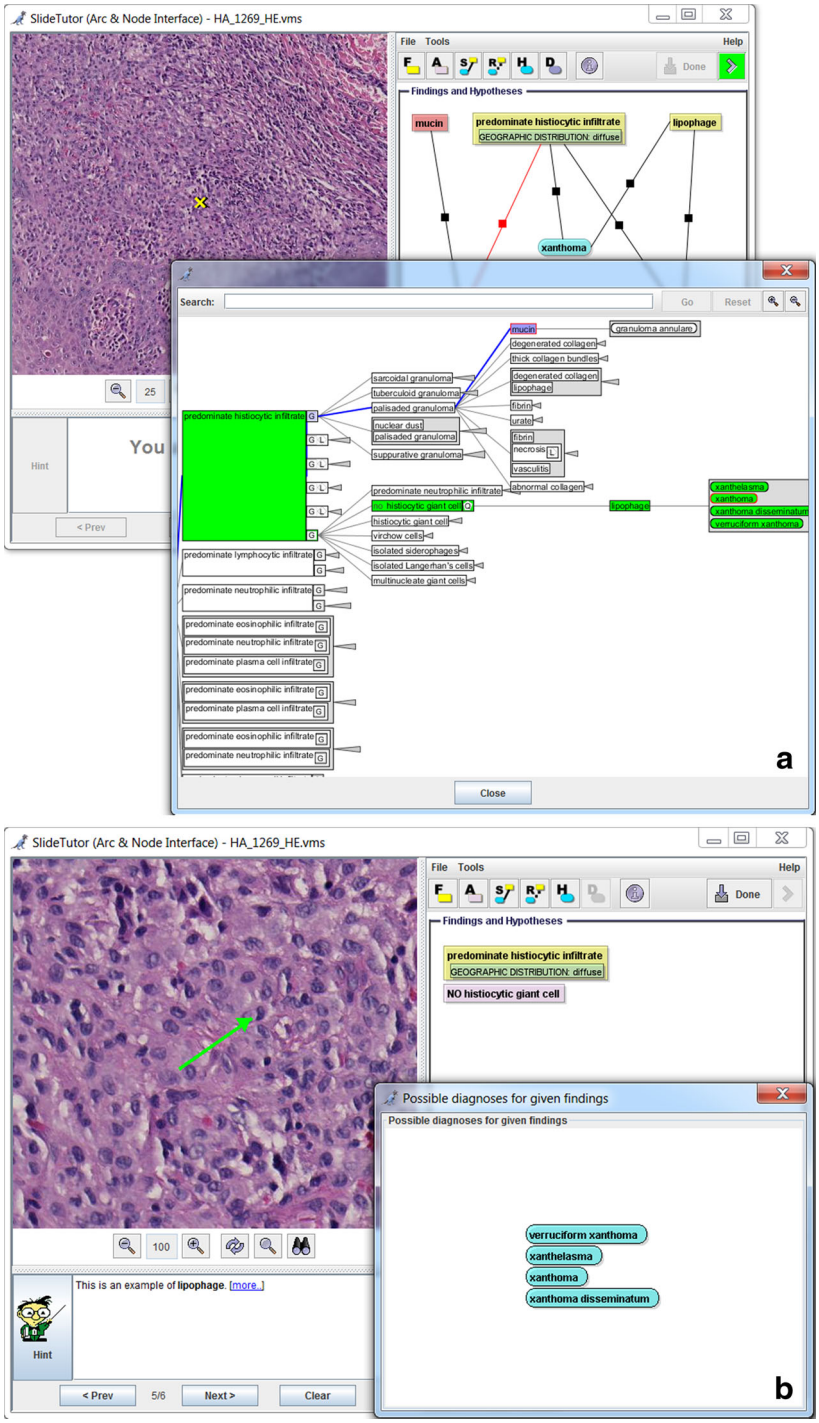

Fig. 3 Metacognitive scaffoldings. Differing metacognitive feedback for **a** Considering Alternatives and **b** Playback

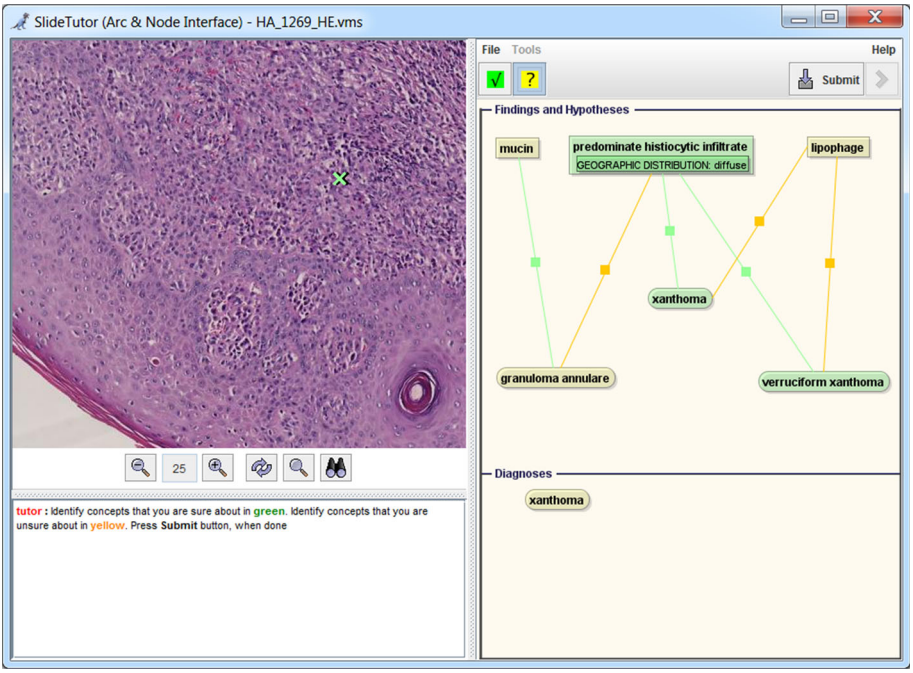

Fig. 4 'Coloring Book' interface for collecting FOK judgments. Participants were asked to right click and select sure or unsure for each finding, hypothesis, diagnosis, and link

previously asserted findings, hypotheses, links, and diagnoses for each case and to estimate their degree of confidence in the correctness of each assertion by coloring items green that they are sure about, and coloring items yellow that they are unsure about.

For all system interfaces (immediate feedback, coloring book, and metacognitive scaffoldings), every student action and tutor response is stored in an Oracle database for further analysis. Interaction data includes participant actions, errors, and hint requests, as well as detailed interface information such as mouse clicks and menu selections. All confidence judgments indicated during metacognitive scaffoldings are also saved to the database for later analysis. A detailed description of the information model, including data stored from each student session, has been previously published (Medvedeva et al. 2005).

Procedure

At the beginning of the experiment, all participants signed a consent form, and were trained to use the immediate feedback tutoring system. To limit the effect of prior knowledge on learning to use the system, participants were trained using a toy task that has identical requirements: categorizing dinosaurs based on cartoon images. Following system training, participants were presented with fifteen NDD cases, for which they received immediate feedback on all intermediate actions over a working period of 2 h. During this part of the experiment, all students used the same version of the *immediate feedback tutoring system* to gain some diagnostic skill in this complex area of pathology so that they could benefit from metacognitive training received later in the experimental session.

We then administered a 30-min pre-test that consisted of four cases that were not previously seen. The pre-test included two cases with tutored patterns (cases with the same set of

histopathologic findings and diagnoses as those seen during immediate feedback tutoring) and two cases with untutored patterns (cases with a different set of histopathologic findings and diagnoses than those seen during immediate feedback tutoring). Before solving each case in the pre-test, participants were asked how difficult they thought it would be to diagnose the current case, and they responded on a 6-point Likert-type scale where choices ranged from "very difficult" to "very easy". After solving each case in the pre-test, participants were asked to express their confidence on all findings, hypotheses, links, and diagnoses identified for the case by coloring the items they were sure about green and the items they were unsure about yellow. No cognitive or metacognitive feedback was provided during the pre-test.

Next, participants received *metacognitive training* using one of two modified versions of SlideTutor (*Playback* or *Considering Alternatives*). The scaffolding period lasted 3 h. Cases were presented to the participants, and they were free to navigate through the slide until they were ready to solve the case, which they indicated by clicking on the "Solve Case" button. At this point, participants were asked how difficult they thought it would be to diagnose the present case, and responded using a 6-point Likert-type scale ranging from "very difficult" to "very easy" (i.e., before case judgment). Participants then identified findings, hypotheses, diagnoses, and links without feedback from the tutoring system. After solving the case, they used the *Coloring Book* interface to indicate their confidence in all asserted findings, hypotheses, links, and diagnoses. They were also asked to separately rate overall confidence on their diagnostic accuracy using a 6-point Likert-type scale ranging from "not confident" to "very confident" (i.e., after coloring judgment). Participants then received either *Considering Alternatives* or *Playback* feedback for the case, depending on the experimental condition. Participants were asked how confident they would be in solving similar cases in the future, and responded on a 6-point Likert scale from "not confident" to "very confident" (i.e., after tutor feedback judgment). The timeline of prompts for the metacognitive judgments is illustrated in Fig. 5.

Following metacognitive training, we administered a 30-min post-test which consisted of four cases that were not previously seen. The post-test included two cases with tutored patterns (cases with the same set of histopathologic findings and diagnoses as those seen during immediate feedback tutoring) and two cases with untutored patterns (cases with a different set of histopathologic findings and diagnoses as those seen during immediate feedback tutoring). Participants were asked to make a metacognitive judgment before solving each case on the post-test, and after solving the case, they were asked to express their confidence in asserted findings, hypotheses, links, and diagnoses using the coloring book. No cognitive or metacognitive feedback was provided during the post-test.

Data analysis

Pre-test and post-test

In order to measure the prior knowledge of the participants and cognitive gains after interacting and solving cases with SlideTutor, for each test question in the pre-test and post-test, two scores were computed: diagnosis and rationale. For the diagnosis score, 5 points were added for the first correct diagnosis, and 3 points were added for each additional correct diagnosis (for cases with a differential diagnosis). Additionally, 1 point was subtracted for each incorrect diagnosis (with 0 being the lowest possible score given for a case). For the rationale score, 2 points were added for each correct finding, and 1.5 points were added for each finding with incorrect or missing attributes. The total points possible could vary across questions; therefore, scores were normalized to produce equal weight by

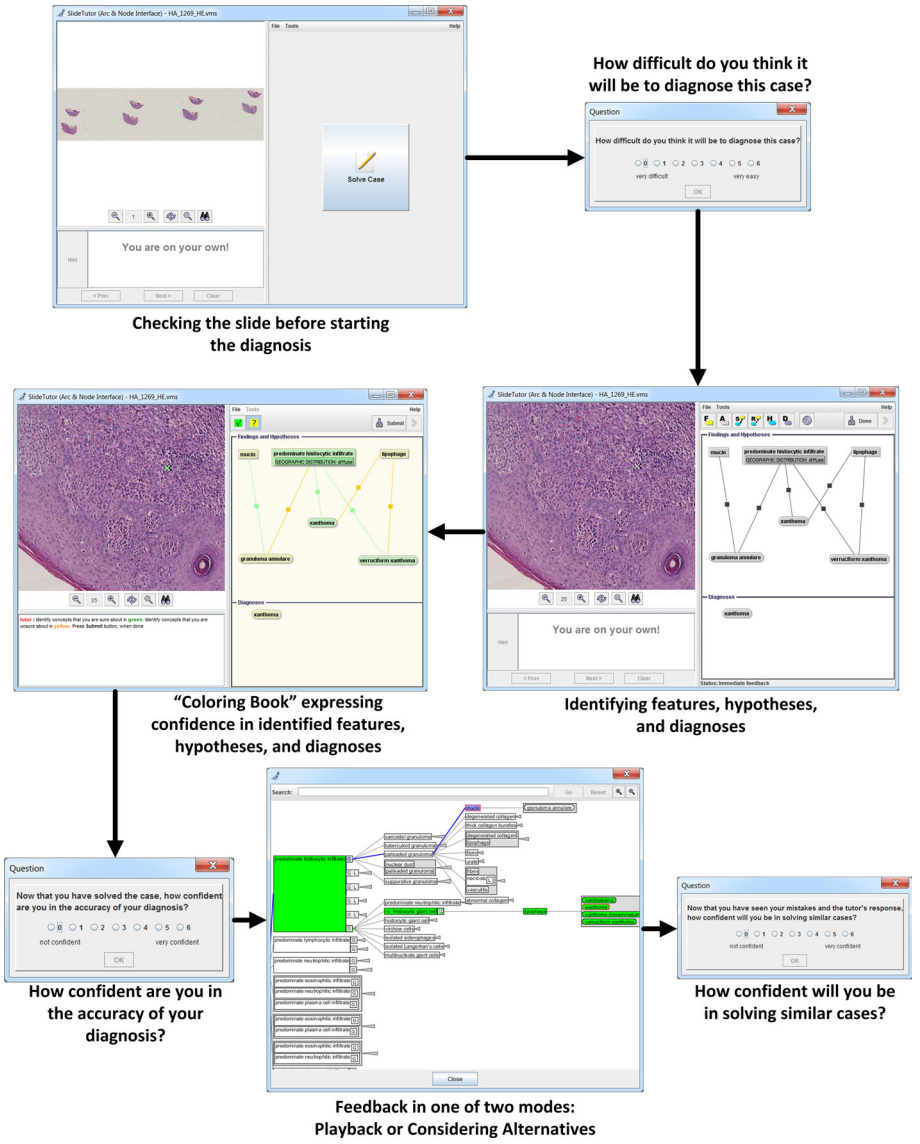

Fig. 5 Timeline of prompts for metacognitive judgments. Participants in both scaffolding groups were prompted for metacognitive judgments before and after each case and received feedback before beginning the next case

case. Overall test scores (reported in the "Results" section) are the average of the diagnosis and rationale test scores.

Diagnostic accuracy

FOK judgments were obtained from residents' assessments of their confidence in the identified features, hypotheses, and diagnoses in slides. Data from coloring support or

refute links were not included in our FOK analyses because of the variability in frequency of use by participants. In order to determine the accuracy of FOK judgments made throughout the experimental session, three measures were used: bias, discrimination, and Goodman–Kruskal Gamma correlation (G; see Schraw 2009). Measures were based on a two-by-two contingency table that compared the performance on the item (correct vs. incorrect) and the feeling of knowing ('sure' or 'unsure'). Formulas for calculating the three accuracy measures are shown in the Appendix.

We compared cognitive and metacognitive results by condition using independent samples *t* tests, repeated-measures ANOVA, and ANCOVA. All analyses were conducted using IBM SPSS Statistics v19.

Results

Cognitive learning gains

Pre-test and post-test scores for each participant were used to calculate the cognitive learning gains from the scaffolded working period. Scores for tutored patterns, untutored patterns, and overall scores were analyzed separately (Table 1). The results of independent sample *t* tests show no significant difference in the pre-test or post-test scores between the two groups, *Playback* and *Considering Alternatives* (pre-test: $t(29) = 1.266$, $p > .05$; post-test: $t(29) = 0.840$, $p > .05$). A repeated-measures ANOVA revealed a significant difference between total pre-test and post-test scores for the participants in both conditions ($F(1,29) = 14.104$, $p < .05$, partial $\eta^2 = .327$), with no significant effect of condition ($F(1,29) = 1.66$, $p > .05$). Moreover, no significant interaction effect of test and condition was found (test * condition: $F(1,29) = 0.003$, $p > .05$). A significant difference was seen between pre-test and post-test for scores on tutored patterns ($F(1,29) = 32.128$, $p < .05$, partial $\eta^2 = .526$), but not for untutored patterns ($F(1,29) = 0.202$, $p > .05$). In sum, participants in both conditions showed significant performance improvement overall and on tutored patterns. No significant differences in learning gains were observed between groups. These findings replicate results demonstrated in previous studies of our tutoring system (Crowley et al. 2007; El Saadawi et al. 2008, 2010).

In order to determine whether there is a significant difference between the post-test scores of participants in the two conditions taking into account their pre-test scores, an

Table 1 Results of *t* tests for pre-test and post-test

	t	df	Sig	Playback		Considering Alternatives	
				M	SD	*M*	SD
Pre-test							
Overall	1.266	29	0.215	29.79	10.18	23.99	15.02
Untutored patterns	1.615	29	0.117	22.58	15.52	14.93	10.07
Tutored patterns	0.468	29	0.643	37.01	21.33	33.05	25.64
Post-test							
Overall	0.840	29	0.408	45.22	20.30	38.95	21.21
Untutored patterns	0.086	29	0.932	17.03	26.79	16.22	25.27
Tutored patterns	1.382	29	0.177	73.40	21.19	61.69	25.91

ANCOVA was conducted on post-test scores with pre-test as the covariate. The results indicated no significant effect of condition for post-test scores ($F(1,30) = 0.378$, $p > .05$).

Accuracy of metacognitive judgments

The accuracy of metacognitive judgments made on identified findings, absent findings, hypotheses, and diagnoses in the *Coloring Book* interface of SlideTutor were analyzed using three measures of metacognitive accuracy: Goodman–Kruskal Gamma (G), bias, and discrimination. The analyses were conducted in order to investigate the effect of the two metacognitive scaffolds (*Playback* and *Considering Alternatives*) on the metacognitive accuracy of residents while diagnosing pathology cases.

In order to compare residents' metacognitive accuracy across the two groups, repeated-measures ANOVAs were conducted. The results indicated that there was a significant main effect of condition for overall G, G for tutored patterns, G for diagnoses, overall discrimination, and discrimination for tutored patterns (Table 2). For all of these measures, the participants in the *Considering Alternatives* condition had significantly higher scores than those in the *Playback* condition (Table 2). However, not all of these measures improved from pre-test to post-test. Moreover, a significant effect of test was found for discrimination index for hypotheses ($F(1,29) = 8.38$, $p < .05$, partial $\eta^2 = .201$). The discrimination index for hypotheses improved for both groups from pre-test to post-test (Playback condition: pre-test $M = 0.20$, post-test $M = 0.37$; Considering Alternatives: pre-test $M = 0.11$, post-test $M = 0.56$).

The results of one-way ANOVA on the three measures of metacognitive accuracy at pre-test and post-test indicated that there was a significant difference between the participants in the two conditions at pre-test in overall G ($F(1,29) = 5.554$, $p < .05$, $\eta^2 = .16$), G for diagnoses ($F(1,29) = 10.312$, $p < .05$, $\eta^2 = .262$), and overall discrimination ($F(1,29) = 5.298$, $p < .05, \eta^2 = .154$). Furthermore, G for tutored patterns at post-test was significantly higher for participants in the Considering Alternatives condition ($F(1,29) = 7.423, p < .05, \eta^2 = .203$).

Since participants in the *Considering Alternatives* condition had higher G, bias, and discrimination scores than participants in the *Playback* condition at pre-test, we conducted further statistical analyses with ANCOVA to take into account the effect of metacognitive indices at pre-test on participants' metacognitive performance at post-test. ANCOVA was conducted with G, discrimination, and bias measures as dependent variables, and pre-test as the covariate. For each of the three metacognitive accuracy measures, ANCOVAs were

Table 2 Significant repeated-measures ANOVA results for coloring accuracy measures (effect of condition)

	F	Sig	Pre-test		Post-test	
			Playback	Alternatives	Playback	Alternatives
G						
Overall	7.328	0.011	0.23	0.55	0.36	0.52
Tutored patterns	8.675	0.006	0.34	0.46	0.37	0.78
Diagnoses	7.970	0.009	0.04	0.66	0.19	0.54
Discrimination						
Overall	4.719	0.038	0.12	0.31	0.24	0.30
Tutored patterns	8.318	0.007	0.15	0.37	0.28	0.59

conducted for overall measures, tutored and untutored patterns, findings, hypotheses, and diagnoses. The results of ANCOVA indicated that there is a significant effect of condition for G and discrimination for tutored patterns, both of which are in favor of the participants in the *Considering Alternatives* condition (Table 3). This indicates that participants in the *Considering Alternatives* condition had significantly greater metacognitive gains than those in the *Playback* condition in the confidence judgments of correctness of their identified findings, hypotheses, and diagnoses in tutored patterns (as indicated by the higher G value for the *Considering Alternatives* condition). Moreover, they had greater gains in their confidence about correct items as opposed to incorrect items in tutored patterns in comparison to the participants in the *Playback* condition (as indicated by a higher positive discrimination index for the Considering Alternatives condition). Also, a significant effect of pre-test was found for G for findings and diagnoses, overall bias, and bias for tutored and untutored patterns, and bias for findings and diagnoses (Table 3). As shown by the mean bias values in Table 3, we find that participants in both conditions were minimally underconfident for tutored patterns, but neither overconfident nor underconfident overall or for untutored patterns. In contrast, participants were slightly overconfident in findings and diagnoses. ANCOVAs indicated that there was no significant effect of condition for any of the bias scores. Moreover, no significant effect of pre-test was found for discrimination scores on tutored and untutored patterns, findings, and diagnoses.

Tutor questions during the experimental session

A 2 (*Playback* and *Considering Alternatives*) × 3 (before case, after coloring, and after feedback) repeated-measures ANOVA for comparison of global assessments made before case, after coloring, and after tutor feedback (Likert-type scale) for the participants in the two conditions indicated no significant effect of condition: $F(1,58) = 0.121$, $p > .05$. In other words, the participants in the two conditions did not significantly differ in the judgments made before case, after coloring, and after tutor feedback. A significant main effect for time was found depending on when the judgments were made (before case, after case, and after feedback): $F(2,116) = 24.843$, $p < .001$. The comparison of means

Table 3 Significant ANCOVA results for coloring accuracy measures (pre-test as covariate)

	Playback		Alternatives		Effect of test		Effect of condition	
	M	SD	*M*	SD	*F*	Sig	*F*	Sig
G								
Tutored patterns	0.37	0.51	0.78	0.29			8.602	0.007
Findings	0.41	0.48	0.43	0.34	7.398	0.011		
Diagnoses	0.19	0.67	0.54	0.55	4.243	0.049		
Bias								
Overall	−0.08	0.31	−0.02	0.19	15.583	0.000		
Tutored patterns	−0.12	0.33	−0.01	0.18	7.435	0.011		
Untutored patterns	−0.06	0.35	−0.01	0.29	4.959	0.034		
Findings	0.04	0.20	0.17	0.15	9.486	0.005		
Diagnoses	0.15	0.49	0.08	0.34	11.344	0.002		
Disc								
Tutored patterns	0.28	0.42	0.59	0.45			4.462	0.044

indicated that participants in both conditions felt significantly more confident about solving similar cases after receiving feedback than before solving the case and after coloring (Table 4).

Tutor confidence questions in pre-test and post-test

In pre-test and post-test, participants responded to a global assessment of confidence question (Likert-type scale) before each of the four cases they solved. The comparison of mean confidence ratings for the two tutored and two untutored questions in the pre-test and post-test indicated that the participants in both conditions improved significantly in their confidence on tutored patterns from pre-test to post-test (Playback condition: $t(28) = -2.98$, $p < .05$; Considering Alternatives condition: $t(28) = -2.24$, $p < .05$). For participants in both conditions, confidence ratings were lower on untutored pattern cases than on tutored pattern cases on both pre-test and post-test (Table 5).

Discussion

Research in self-regulated learning, metacognition, and intelligent tutoring systems (in science, medicine, etc.) has revealed that providing students with metacognitive scaffolds and feedback on their performance improves their cognitive and metacognitive gains. Although human face-to-face tutoring leads to superior learning gains and is very effective in achieving improved learning outcomes (VanLehn 2011) in some disciplines, especially ones like medicine and pathology where the model of training novices is the apprenticeship model, the expert-novice interaction and scaffolding is restricted to training hours. ITSs can provide an environment in which residents can safely practice at any time while receiving some of the same benefits of human tutoring, such as scaffolding and feedback.

Table 4 Descriptive statistics for global confidence judgments

Condition	Judgment	*M*	SD
Playback	Before case	2.605	0.100
	After coloring	2.838	0.110
	After feedback	3.290	0.119
Alternatives	Before case	2.756	0.100
	After coloring	2.831	0.110
	After feedback	3.007	0.119

Table 5 Mean global confidence ratings at pre-test and post-test

	Pre-test (*M*)		Post-test (*M*)	
	Untutored	Tutored	Untutored	Tutored
Playback	1.68	2.10	1.88	3.10
Considering Alternatives	2.23	2.50	2.23	3.24

Although several authors have suggested that overconfidence and underconfidence may result in diagnostic error (Berner and Graber 2008; Croskerry and Norman 2008), there has been little empirical research in the medical domain evaluating approaches to reduce these biases.

In this study, we examined the effect of two types of metacognitive scaffolds, *Considering Alternatives* and *Playback*, on cognitive gains and the accuracy of metacognitive judgments in a medical ITS. Overall, the results of this study suggest that participants in both metacognitive scaffolding conditions significantly improved in cognitive skills from pre-test to post-test, as indicated by overall test scores and scores on tutored patterns in the tests. When participants solved the cases on which they had not been tutored in the immediate feedback tutoring phase, no significant change was observed in their scores from pre-test to post-test. The improved performance on tutored patterns is likely a benefit of receiving immediate cognitive feedback on the steps taken in diagnosing the cases. These findings confirm our first hypothesis and are in agreement with those of previous evaluations of this system (Crowley et al. 2007; El Saadawi et al. 2008, 2010). Importantly, we observed no differences between conditions for learning gains. Given that participants learned an equivalent amount, we then investigated the accuracy, bias, and discrimination of metacognitive judgments made during the experimental session to see whether participants also became more accurate and less biased after solving cases with the help of SlideTutor in one of the two scaffolding conditions.

Bias scores revealed slight underconfidence for participants in both conditions overall, and for tutored and untutored patterns. On the other hand, slight overconfidence was observed for participants in both conditions in findings and diagnoses identified in slides. There was no significant difference between bias scores of participants in the two conditions, or in their under- or overconfidence in the correctness of the items they identified in cases in SlideTutor. However, the slight under- or overconfidence might have been a result of the relatively short exposure to the scaffoldings during the experimental session; longer interactions with SlideTutor and the solving of more cases while receiving prompts for metacognitive judgments and feedback on accuracy might further improve metacognitive bias.

Participants in the *Considering Alternatives* condition had significantly greater gains in metacognitive accuracy scores (overall G, G for tutored patterns, G for diagnoses, overall discrimination, and discrimination for tutored patterns) than did participants in the *Playback* condition. In particular, G for tutored patterns at post-test was found to be significantly higher for participants in the *Considering Alternatives* condition. Thus, participants in this condition became more accurate in their judgments about items identified in cases that included patterns on which they had been tutored during the immediate feedback phase. By controlling for the difference of the two groups in the pre-test by ANCOVA, we found that there was still a significant effect of condition for G and discrimination for tutored patterns in favor of participants in the *Considering Alternatives* condition. In solving cases with tutored patterns, participants in the *Considering Alternatives* group were more accurate in their judgments and more confident about the correctness of items identified in slides than were participants in the *Playback* condition. This is an important finding, given that participants had a relatively short amount of time interacting with SlideTutor and solved no more than 30 cases in the *Considering Alternatives* scaffolding condition. Nevertheless, they managed to achieve significantly greater metacognitive gains than the residents in the *Playback* condition. These findings confirm our second research hypothesis. The lower metacognitive accuracy and discrimination gains for participants in the *Playback* condition might be the result of the higher cognitive load imposed by

watching one's own actions and then watching the correct (expert) actions play-by-play, since they were required to remember all steps they had taken after watching their own actions and then to compare them to the correct steps in the expert solution. It is also possible that the speed of playback has a variable effect on metacognitive gains, and was perhaps too fast in this experiment, limiting the ability of participants to sufficiently reflect on the actions taken.

In summary, the findings of this study corroborate our previous findings on the effectiveness of immediate feedback as indicated by improved cognitive and metacognitive outcomes at the post-test. We also found that both metacognitive scaffolds lead to better metacognitive accuracy and discrimination indices in the experiment, meriting further research into improvements in the two scaffolds to make them even more effective. Furthermore, we found that Considering Alternatives produces a superior effect on metacognitive gains when compared with Playback. The superior effectiveness of the Considering Alternatives scaffolding suggests a specific benefit of presenting a visual schema or solution path with colors directing attention to correct versus incorrect solution paths.

Instructional implications

In a broader sense, the findings of this study have implications for the design and development of medical ITSs that can be used train physicians and simultaneously debias them. Decision trees and displaying of learner versus expert problem-solving paths (as in the Considering Alternatives metacognitive scaffold) can be used in the design of instructional materials and development of training courses for pathology and dermatology residents, and residents in visual diagnosis area (e.g., radiology) in general, in order to assist them in making more accurate diagnoses and decrease their diagnostic bias by helping them recalibrate their metacognitive judgments.

Limitations and future directions

One of the limitations of this study was its small sample size ($N = 31$), which limits the generalizability of the findings of this study. The small sample size was partly due to the narrow subject pool of pathology and dermatology residents and fellows, and the long duration of the study (~ 8 h). Despite the small number of participants in this study, however, we found significant differences between metacognitive performance in the two conditions. The inclusion of participants from five different institutions improved the representativeness of our sample. Only one domain of dermatopathology (NDD) was used for selecting cases for this study; therefore, the results cannot be widely generalized across medical domains. Future studies can investigate the effectiveness of the scaffolding used in this study in other domains of pathology, or even in other domains. Another possible limitation of this study is the inclusion of participants from different residency years of pathology and dermatology; this could be associated with potential differences in familiarity with the domain used in this study. It would be interesting to control for prior domain familiarity, perhaps leading to a clearer distinction between the findings from the two interventions.

In future work, we would like to examine the effect of an open student model on metacognitive gains. Like the *Playback* condition that we used in this experiment, open student models provide a means for comparing performance to an expert model. However, open student models have the advantage that they aggregate student data over time, limiting the cognitive load on the student. We suspect that this scaffold may produce more

explicit engagement in metacognitive monitoring. In future work, we will test whether an open student model further increases the accuracy of metacognitive judgments.

Using open learner models (Bull 2004; Bull and McKay 2004), the cognitive performance and the degree of the accuracy, bias, and discrimination of the metacognitive judgments made by participants while solving cases can be presented to them visually (e.g., as bar graphs or skill meters) at different intervals to make them more aware of their cognitive and metacognitive performance, which might lead to improved cognitive and metacognitive gains.

Additionally, we did not collect eye-tracking data in this study. These data would provide evidence on participants' gaze patterns and allocation of attention during the *Playback* or *Considering Alternatives* phase, which would help in making more accurate inferences about the cognitive and metacognitive processes in the learners' minds. Without evidence on learners' attention allocation from eye-tracking data, we are not able to state with certainty if and how participants paid attention to the playback of their problem-solving steps or the decision-making tree in the *Considering Alternatives* phase, and how it influenced their metacognitive calibration and cognitive gains.

Acknowledgments The authors gratefully acknowledge support of this research by the National Library of Medicine through Grant 5R01LM007891. We thank Ms. Lucy Cafeo for her expert editorial assistance. This work was conducted using the Protégé resource, which is supported by Grant LM007885 from the United States National Library of Medicine. SpaceTree was provided in collaboration with the Human-Computer Interaction Lab (HCIL) at the University of Maryland, College Park.

Appendix

See Tables 6 and 7.

Table 6 FOK contingency table

		Performance: Correct	Performance: Incorrect
Feeling of Knowing	Sure	True Positive a	False Positive b
	Unsure	False Negative c	True Negative d

Table 7 FOK accuracy statistics and equations

Bias	= Confidence − judgment $= \frac{a+b}{a+b+c+d} - \frac{a+c}{a+b+c+d}$
Discrimination	= Confidence on correct items − confidence on incorrect items $= \frac{a}{a+c} - \frac{b}{b+d}$
G	= (Concordant pairs − discordant pairs)/(concordant pairs + discordant pairs) $= \frac{ad-bc}{ad+bc}$

References

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition: Implications for the design of computer-based scaffolds. *Instructional Science, 33*, 367–379.

Azevedo, R., & Lajoie, S. (1998). The cognitive basis for the design of a mammography interpretation tutor. *International Journal of Artificial Intelligence in Education, 9*, 32–44.

Azevedo, R., Moos, D., Greene, J., Winters, F., & Cromley, J. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development, 56*(1), 45–72.

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

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine, 121*(5A), S2–S23.

Bull, S. (2004). Supporting learning with open learner models. In *Proceedings of 4th Hellenic conference with international participation: Information and communication technologies in education*. Keynote. Athens, Greece.

Bull, S., & McKay, M. (2004). An open learner model for children and teachers: Inspecting knowledge level of individuals and peers. In R. Luckin, K. R. Koedinger, & J. E. Greer (Eds.), *Intelligent tutoring systems* (pp. 232–251). Berlin: Springer.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

Chi, M. T. H., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22*, 363–387.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471–534.

Clancey, W. (1987). *Knowledge-based tutoring: The GUIDON program.* Cambridge: MIT Press.

Corbett, A. T., Koedinger, K. R., & Hadley, W. H. (2002). Cognitive tutors: From the research classroom to all classrooms. In P. S. Goodman (Ed.), *Technology enhanced learning: Opportunities for change* (pp. 199–224). Mahwah, NJ: Erlbaum.

Croskerry, P., & Norman, G. (2008). Overconfidence in clinical decision making. *The American Journal of Medicine, 121*(5A), S24–S29.

Crowley, R., & Gryzbicki, D. (2006). Intelligent medical training systems (guest editorial). *Artificial Intelligence in Medicine, 38*(1), 1–4.

Crowley, R. S., Legowski, E., Medvedeva, O., Tseytlin, E., Roh, E., & Jukic, D. (2007). Evaluation of an intelligent tutoring system in pathology: Effects of external representation on performance gains, metacognition, and acceptance. *Journal of the American Medical Informatics Association, 14*(2), 182–190.

Crowley, R. S., & Medvedeva, O. (2006). An intelligent tutoring system for visual classification problem solving. *Artificial Intelligence in Medicine, 36*(1), 85–117.

Crowley, R. S., Medvedeva, O., & Jukic, D. (2003a). SlideTutor: A model-tracing intelligent tutoring system for teaching microscopic diagnosis. In J. Kay, F. Verdejo, & U. Hoppe (Eds.), *Proceedings of the 11th International Conference on Artificial Intelligence in Education*, Sydney.

Crowley, R. S., Naus, G. J., Stewart, J., & Friedman, C. P. (2003b). Development of visual diagnostic expertise in pathology—An information processing study. *Journal of American Medical Information Association, 10*(1), 39–51.

Crowley, R. S., Tseytlin, E., & Jukic, D. (2005). ReportTutor—An intelligent tutoring system that uses a natural language interface. In *Proceedings of American Medical Informatics Association Symposium*, Austin, TX (pp. 171–175).

El Saadawi, G., Azevedo, R., Castine, M., Payne, V., Medvedeva, O., Tseytlin, E., et al. (2010). Factors affecting feeling-of-knowing in a medical intelligent tutoring system: The role of immediate feedback as a metacognitive scaffold. *Advances in Health Sciences Education, 15*(1), 9–30.

El Saadawi, G. M., Tseytlin, E., Legowski, E., Jukic, D., Castine, M., Fine, J., et al. (2008). A natural language intelligent tutoring system for training pathologists: Implementation and evaluation. *Advances in Health Sciences Education: Theory and Practice, 13*, 709–722.

Feyzi-Behnagh, R., Khezri, Z., & Azevedo, R. (2011). An investigation of accuracy of metacognitive judgments during learning with an intelligent multi-agent hypermedia environment. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 96–101). Austin, TX: Cognitive Science Society.

Freidman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., et al. (2005). Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. *Journal of General Internal Medicine, 20*, 334–339.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*(268), 732–764.

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165*(13), 1493–1499.

Graesser, A. C., Bowers, C. A., Hacker, D. J., & Person, N. K. (1997). An anatomy of naturalistic tutoring. In K. Hogan & M. Presley (Eds.), *Effective scaffolding of instruction* (pp. 145–184). Cambridge, MA: Brookline Books.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kruez, R., & The Tutoring Research Group. (2000). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1*, 35–51.

Johnson, A. M., Azevedo, R., & D'Mello, S. K. (2011). The temporal and dynamic nature of self-regulatory processes during independent and externally assisted hypermedia learning. *Cognition and Instruction, 29*(4), 471–504.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1304–1316.

Katz, J. (1984). Why doctors don't disclose uncertainty. *The Hastings Center Report, 14*(1), 35–44.

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

Koedinger, K., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*(3), 239–264.

Koedinger, K., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 61–78). New York: Cambridge University Press.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147–162.

Lajoie, S. P. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from avionics and medicine. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 61–83). New York: Cambridge University Press.

Lajoie, S. P., & Azevedo, R. (2006). Teaching and learning in technology-rich environments. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 803–821). Mahwah, NJ: Erlbaum.

Lepper, M. R., Drake, M. F., & O'Donnell-Johnson, T. (1997). Scaffolding techniques of expert human tutors. In K. Hogan & M. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues* (pp. 108–144). Cambridge, MA: Brookline.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*(4), 723.

Mann, D. (1993). *The relationship between diagnostic accuracy and confidence in medical students*. Atlanta, GA: American Educational Research Association.

Maries, A., & Kumar, A. (2008). The effect of student model on learning. In I. Aedo, N. Chen Kinshuk, D. Sampson, & L. Zaitseva (Eds.), *ICALT'09, 8th IEEE International Conference on Advanced Learning Technologies* (pp. 877–881). Piscataway, NJ: IEEE Computer Society Press.

McGinnis, K. S., Lessin, S. R., Elder, D. E., DuPont Guerry, I. V., Schuchter, L., Ming, M., et al. (2002). Pathology review of cases presenting to a multidisciplinary pigmented lesion clinic. *Archives of Dermatology, 138*(5), 617–621.

Medvedeva, O., Chavan, G., & Crowley, R. S. (2005). A data collection framework for capturing ITS data based on an agent communication standard. In *Proceedings of the 20th Annual Meeting of the Association for the Advancement of Artificial Intelligence (AAAI)* (pp. 23–30). Pittsburgh, PA: AAAI Press.

Metcalf, J., & Dunlosky, J. (2008). Metamemory. In H. Roediger (Ed.), *Cognitive psychology of memory* (Vol. 2, pp. 349–362). Oxford: Elsevier.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133.

Nelson, S. B. (1996). Why study? How reasons for learning influence strategy selection. *Educational Psychology Review*, *8*(4), 335–355.

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*(2), 258–271. doi: 10.1177/0013164404273945.

Obradovich, J. H., Smith, P. J., Guerlain, S., Rudmann, S., & Smith, J. (2000). Field evaluation of an intelligent tutoring system for teaching problem-solving skills in transfusion medicine. In *Proceedings of the IEA 2000/HFES 2000 Congress. 44th Annual Meeting of the Human Factors and Ergonomics Society*, San Diego, CA. London: Taylor & Francis.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.

Rogers, E. (1995). VIA-RAD: A blackboard-based system for diagnostic radiology. *Artificial Intelligence in Medicine, 7*, 343–360.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. doi:10.1007/s11409-008-9031-3.

Sharples, M., Jeffery, N., du Boulay, B., Teather, B., Teather, D., & du Boulay, G. (2000). Structured computer-based training in the interpretation of neuroradiological images. *International Journal of Medical Informatics, 60*, 263–280.

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach (Ed.), *Adaptive technologies for training and education* (pp. 7–27). New York: Cambridge University Press.

Smith, P., Obradovich, J., Heintz, P., et al. (1998). Successful use of an expert system to teach diagnostic reasoning for antibody identification. In B. P. Goettl, H. M. Halff, C. L. Redfield, & V. J. Shute (Eds.), *Proceedings of the 4th International Conference on Intelligent Tutoring Systems, ITS '98* (pp. 54–63). San Antonio, TX: Springer.

Troxel, D. B. (2006). Medicolegal aspects of error in pathology. *Archives of Pathology and Laboratory Medicine, 130*(5), 617–619.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Voytovich, A. E., Rippey, R. M., & Suffredini, A. (1985). Premature conclusion in diagnostic reasoning. *Journal of Medical Education, 60*, 302–307.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In J. Douglas, J. D. Hacker, & A. C. Graesser (Eds.), *Self-regulated learning and academic achievement: Theoretical perspective* (pp. 153–190). Mahwah, NJ: Erlbaum.

Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Motivation in educational theory and practice* (pp. 227–304). Mahwah, NJ: Erlbaum.

Winne, P., & Hadwin, A. (2008). The weave of motivation and self-regulated learning. In D. Schunk & B. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and application* (pp. 297–314). Mahwah, NJ: Erlbaum.

Woo, C. W., Evens, M. W., Freedman, R., et al. (2006). An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning. *Artificial Intelligence in Medicine, 38*(1), 25–46.

Woolf, B. P. (2009). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Boston: Morgan Kaufmann/Elsevier.

Zimmerman, B. (2006). Development and adaptation of expertise: The role of self-regulatory processes and beliefs. In K. A. Ericsson, P. Charness, P. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 705–722). New York: Cambridge University Press.