



Published in final edited form as:

*Adv Health Sci Educ Theory Pract.* 2010 March ; 15(1): 9. doi:10.1007/s10459-009-9162-6.

## Factors Affecting Feeling-of-knowing in a Medical Intelligent Tutoring System – the role of Immediate Feedback as a Metacognitive Scaffold

Gilan M. El Saadawi, MD, PhD<sup>1,2</sup>, Roger Azevedo, PhD<sup>3</sup>, Melissa Castine<sup>1</sup>, Velma Payne, MS<sup>1</sup>, Olga Medvedeva, MS<sup>1</sup>, Eugene Tseytlin, MS<sup>1</sup>, Elizabeth Legowski<sup>1</sup>, Drazen Jukic, MD, PhD<sup>1,4,5</sup>, and Rebecca S. Crowley, MD, MS<sup>1,4,6</sup>

<sup>1</sup> Department of Biomedical Informatics, University of Pittsburgh School of Medicine

<sup>2</sup> Department of Health and Community Services, University of Pittsburgh School of Nursing

<sup>3</sup> Department of Psychology, University of Memphis

<sup>4</sup> Department of Dermatology, University of Pittsburgh School of Medicine

<sup>5</sup> Department of Pathology, University of Pittsburgh School of Medicine

<sup>6</sup> Intelligent Systems Program, University of Pittsburgh School of Arts and Sciences

### Abstract

**Objective**—Previous studies in our laboratory have shown the benefits of immediate feedback on cognitive performance for pathology residents using an Intelligent Tutoring System in Pathology. In this study, we examined the effect of immediate feedback on metacognitive performance, and investigated whether other metacognitive scaffolds will support metacognitive gains when immediate feedback is faded.

**Methods**—Twenty-three (23) participants were randomized into intervention and control groups. For both groups, periods working with the ITS under varying conditions were alternated with independent computer-based assessments. On day 1, a within-subjects design was used to evaluate the effect of immediate feedback on cognitive and metacognitive performance. On day 2, a between-subjects design was used to compare the use of other metacognitive scaffolds (intervention group) against no metacognitive scaffolds (control group) on cognitive and metacognitive performance, as immediate feedback was faded. Measurements included learning gains (a measure of cognitive performance), as well as several measures of metacognitive performance, including Goodman-Kruskal Gamma correlation (G), Bias, and Discrimination. For the intervention group, we also computed metacognitive measures during tutoring sessions.

**Results**—Results showed that immediate feedback in an intelligent tutoring system had a statistically significant positive effect on learning gains, G and discrimination. Removal of immediate feedback was associated with decreasing metacognitive performance, and this decline was not prevented when students used a version of the tutoring system that provided other metacognitive scaffolds. Results obtained directly from the ITS suggest that other metacognitive scaffolds do have a positive effect on G and Discrimination, as immediate feedback is faded.

**Conclusions**—Immediate feedback had a positive effect on both metacognitive and cognitive gains in a medical tutoring system. Other metacognitive scaffolds were not sufficient to replace immediate feedback in this study. However, results obtained directly from the tutoring system are not consistent with results obtained from assessments. In order to facilitate transfer to real-world tasks, further research will be needed to determine the optimum methods for supporting metacognition as immediate feedback is faded.

### Keywords

Intelligent Tutoring Systems; Diagnostic Reasoning; Clinical Competence; Cognition; Diagnostic Errors; Education, Medical; Educational Technology; Feeling-of-knowing; Pathology; Problem Solving; Metacognition

## INTRODUCTION AND BACKGROUND

Intelligent Tutoring Systems (ITS) are adaptive, personalized instructional systems that are designed to mimic the well-known benefits of one-on-one human tutoring. ITS support “learning by doing” —as students work on computer-based problems or simulations of real-world tasks, the system offers guidance and explanations, points out errors, and organizes the curriculum to suit the needs of that individual<sup>1-3</sup>. ITS typically incorporate an explicit encoding of domain knowledge and pedagogic expertise, and maintain a dynamic model of the student’s skills and abilities<sup>4</sup>. This makes it possible to individualize instruction without necessarily anticipating specific interactions. Despite the enormous potential for medical education, there have been few ITS developed in Medicine<sup>5-10</sup>, only a small number of which have been evaluated<sup>11-13</sup>.

Cognitive tutoring systems<sup>1</sup> are a subtype of ITS that combine a cognitive model with immediate feedback. These systems can therefore help students identify and explain errors in intermediate cognitive steps. Students are interrupted and redirected when they are moving down invalid paths in the problem space, and encouraged and reinforced when they move down valid paths in the problem space. When students are unsure what to do, the model steps forward and can be used to correctly traverse the problem space while the student observes.

Previous research has demonstrated that immediate feedback is associated with efficient learning<sup>14</sup>, including within medical domains<sup>12</sup>. However, critics of “immediate feedback” have suggested that the rigid one-to-one action-feedback cycle prevents students from learning the metacognitive skills needed to evaluate their own problem-solving and sense when they are making errors<sup>15,16</sup>. In effect, the student may become overly dependent on the immediate feedback, such that they are unable to generate an internal sense of their progress towards a correct answer.

Improving the ability of clinicians to accurately sense when they are right or wrong is of particular interest in medical cognition and education. When clinicians over-estimate their performance, they may reach diagnostic closure too quickly, resulting in diagnostic errors<sup>17,18</sup>. But when they under-estimate their performance, they may inappropriately use consultative services or additional testing, increasing the risk of iatrogenic complications or delay in treatment. Learning to self-monitor and control these cognitive processes is termed self-regulated learning (SRL), and is an important part of expertise<sup>19,20,21</sup>.

Models of SRL can be used to understand the effect of metacognition on cognitive performance. For example, Winne and Hadwin’s<sup>22,23</sup> information processing model of SRL states that the learner goes through a series of recursive phases. In the first phase, the learner engages on prior knowledge and beliefs to interpret the task properties and set goals. In the second phase,

strategies and tactics are utilized to satisfy these goals. In the third phase, the learner monitors and assesses their actions in respect to their beliefs; this may result in setting new goals, modifying existing goals, or re-examining and changing tactics and strategies. As the learner works towards achieving the goal, the learners' beliefs are recalled and related to the solution path i.e. the learner monitors both his domain knowledge and learning resources<sup>24</sup>. Once a specific solution is favored, the learner starts evaluating this solution to assess its degree of correctness. However, even if the learner is not able to reach a correct solution; this process results in long-term alterations to the beliefs, motivation and strategies used in subsequent cases. Thus, it is during this third phase, that learning impacts the feeling-of-knowing (FOK) which represents the relationship between one's certainty and one's performance<sup>24-26</sup>.

Scaffolds are tools, strategies, and guides used by human and computer tutors, teachers, and animated pedagogical agents during learning to enable students to develop understandings beyond their immediate grasp 27·28. Traditionally these scaffolds were studied to evaluate their effect on the students' cognitive abilities. Recently, these same instructional scaffolds have been found to impact the third phase of self-regulated learning. Support can be in the form of pre-stocked static questions, dynamic support that is tailored to the student and context, computer-based tools that guide students in their thinking, etc. Studies have indicated that when students learn about complex topics with computer-based learning environments (CBLEs) in the absence of scaffolding, they show limited ability to regulate their learning, leading to a failure of conceptual understanding 29-31. These studies demonstrated that students who received scaffolding moved towards more sophisticated mental models and increased the frequency of use of SRL strategies, when compared to those who received no scaffolding (control group).

A number of different metacognitive scaffolding techniques have been used in computer-based learning environments, including ITS, to facilitate novices in activating, deploying, and monitoring the success of self-regulated learning. Azevedo et al. and Hadwin et al<sup>15</sup> used adaptive scaffolding based on on-going evaluation and calibration to the individual learner which may include some degree of fading. Others have attempted to engage the student in self-diagnosis with no other form of individualized support or fading<sup>32-34</sup>. Self-explanation prompts<sup>35</sup> have also been used to facilitate problem solving and help determine the adequacy of one's understanding of the topic.

Cognitive ITS depend heavily on immediate feedback as a cognitive scaffold. But ultimately, the purpose of the ITS is to facilitate transfer of performance gains to real-world tasks that do not provide such direct feedback on intermediate steps. The dilemma of removing intermediate feedback as a cognitive scaffold is that we may also inadvertently remove an important metacognitive scaffold. In this study, we explore the effect of immediate feedback as a metacognitive scaffold, and determine whether other metacognitive scaffolds can support self-regulated learning during the fading of immediate feedback. The results of this study can inform the development of medical educational systems that both improve task performance and enhance the ability of clinicians to correctly sense whether they are right or wrong.

## RESEARCH QUESTIONS

1. What is the effect of immediate feedback as a metacognitive scaffold?
2. Can other forms of metacognitive scaffolding sustain self-regulated learning during fading of immediate feedback?

## METHODS

### System Design and Architecture

SlideTutor<sup>9</sup> is an ITS in Pathology that teaches histopathologic diagnosis and reporting skills. The system is currently instantiated in two areas of Dermatopathology – inflammatory diseases and melanocytic lesions. SlideTutor is a client-server application in which students examine virtual slides using various magnifications, identify visual features, specify qualities of these features, make hypotheses and diagnoses, and write pathology reports. As a student works through a case, SlideTutor provides feedback including error analysis and confirmation of correct actions. At any time during the solution, students may request hints. Hints are context-specific to the current problem state, and provide increasingly more targeted advice based on a system generated ‘best-next-step’. In order to distinguish correct from intermediate steps, categorize errors, and provide hints, SlideTutor maintains a cognitive model of diagnosis using a Dynamic Solution Graph (DSG) and a set of ontologies that represent relationships between diagnoses and pathologic findings. SlideTutor contains a pedagogic model that describes the appropriate interventions for specific errors, and maintains a probabilistic model of student performance that is used to adapt instruction based on the student model. SlideTutor uses a variety of interfaces including a diagrammatic interface for reifying diagnostic reasoning, and a natural language interface that interprets and evaluates diagnostic reports written by students. The architecture of the system has been previously described<sup>9</sup>.

Student actions during intermediate problem-solving steps are tested against an *expert model*. The expert model consists of a domain model ontology, domain task ontology, case model, and problem solving methods (PSMs). The *domain model* consists of domain knowledge and defines the relationship between evidence, or feature sets, and disease entities. Within the model, disease entities are associated with a set of features, and feature attribute value sets such as location and quantity that further refine each feature. The *domain task* represents cognitive goals of the classification problem solving process including identifying a feature; specifying an absent feature; refining a feature by designating an attribute such as location or quantity; asserting hypotheses and diagnoses; asserting a supporting relationship connecting a feature to a hypothesis; asserting a refuting relationship between a feature and a hypothesis, and specifying that a feature distinguishes a hypothesis from a competing hypothesis. The *case model* is a representation of the name, location, and attributes of the features present in each case. The *problem solving methods* of the expert model utilize all components of the expert model to create a Dynamic Solution Graph (DSG) that models the current problem-space and valid next steps for case solution. At every point in each student’s problem-solving, the DSG maintains a context-specific ‘best-next-step’ which is used as advice if the student requests a hint.

The *instructional model* is composed of pedagogic knowledge, pedagogic task model, student model data, and problem solving methods. The primary objective of the instructional model is to respond to requests for help from the student and to intervene in the form of alerts when the student has made an error. The *pedagogic knowledge base* contains the declarative knowledge of responses to student errors and requests for help. Hints are based on a hint priority based on the state of the problem at the time the student requests help. Hints also have levels of specificity. Initially, hints offer general guidance; as the student continues to seek help, the hints become more specific and directive. The *pedagogic task* contains information related to the system’s model of how to help specific students related to the problem state and the current state of the *student model*<sup>36</sup>. The *problem solving methods* of the instructional model are used to match the student actions to the help categories in the pedagogic model in order to determine how the system should intervene. A demonstration of the basic SlideTutor system can be accessed at <http://slidetutor.upmc.edu>. For the current study, we modified SlideTutor’s interface and configured the system in two different ways. The first version (System A)

incorporated metacognitive scaffolding using three different techniques. The second version (System B) incorporated no metacognitive scaffolding, but provided other cognitive activities to control for time on task.

### Study Design and Timeline

We used a repeated measures design with both within-subjects (feedback: immediate feedback and faded feedback) and between-subjects (supplementary metacognitive scaffolds: present, absent) factors. Study participants were randomly assigned to one of two groups. The study was conducted on two consecutive days (Figure 1), but groups differed in treatment only on day 2.

On day 1, all participants were given (1) a domain review which consisted of feature identification steps to follow in order to narrow the scope of possible diagnoses within the Superficial Perivascular Dermatitis domain being studied, as well as static images of features to look for. After completing the domain review, participants (2) completed a baseline metacognitive assessment (Test 1), (3) were trained to use the baseline tutoring system, (4) used the baseline tutoring system in the *immediate feedback condition* for 2.25 hours, and (5) completed a second metacognitive assessment (Test 2).

On day 2, participants (6) were trained to use either a system with supplementary metacognitive support “MC group” (System A) or without supplementary metacognitive support (System B), (7) used system A or B in the *faded feedback condition* for 2.25 hours. When using the faded feedback tutor, feedback was gradually removed over time. The ratio of student actions performed with feedback vs. actions performed without feedback changed over the course of the session. The session began with a ratio of 9:1 (9 actions with feedback, 1 action done on their own), then progressed to 8:2, 7:3, etc until it reached 1:9 (1 action performed with feedback, 9 actions performed on their own without feedback). After the feedback condition ended, participants (8) completed the third metacognitive assessment (Test 3). During all working periods, we controlled for time on task, allowing a variable number of cases to be seen in the working period.

### System interfaces and metacognitive scaffolds

All participants used the *baseline tutoring system* during day 1. This version of the system provides feedback to participants immediately after each action in the interface. In previous studies of the baseline system, we have shown marked learning gains<sup>11,12</sup>. Participants used the baseline tutoring system first to gain baseline knowledge of the domain so that they would be able to correctly solve some problems but not others. Participants must have some baseline cognitive skill in order to benefit from metacognitive training.

Participants in the intervention group used System A on day 2. This system provided multiple supplementary metacognitive scaffolds to enhance the students’ understanding of their own cognitive strengths and weaknesses, as well as strategies to improve. These scaffolds included:

- A *coloring book interface* which requires students to revisit their findings and hypotheses at set intervals during the case and at the end of each case, and estimate their confidence in the correctness of each assertion (Figure 2). Data from these estimations was analyzed for the intervention group.
- Following the coloring book, the self-check summary (Figure 2 inset) lets students see whether their assessments were correct. The coloring book interface and self-check summary are introduced multiple times within a case. All confidence estimations were saved to the system database for later analysis.

- An *inspectable student model* that shows all individual skills within the context of the entire knowledge base tree. The tree provides two views: (1) what the student has learned (Figure 3), and (2) what the student has previously seen (not shown). Students can toggle between these views. Assessments of what the student has learned are probabilistic, predicting the performance of each skill based on previous opportunities that the user had to demonstrate each skill. Predictions are based on our modification of a common method that uses first-order Hidden Markov Models, which have shown high accuracy<sup>37</sup>. The student model is available for inspection by the student after the completion of each case.
- Pseudo-dialogue using prestocked, static questions derived from the metacognitive awareness inventory (MAI) developed by Pintrich and DeGroot<sup>38</sup>. The questions provided an opportunity for students to assess their past performance and also offered ideas for new learning strategies. Appendix 1 lists some of the questions used.

Participants in the control group used System B on day 2, which interrupted problem-solving for the same amount of time, but provided no metacognitive scaffolds. Activities that replaced the metacognitive scaffolds from System A included:

1. A *foil coloring book* where students were given instructions to color code features identified by low or high power.
2. A *foil dialogue* that asked purely cognitive questions about the findings or hypotheses for a case. Appendix 2 lists some of the questions used.

## Participants

The study was approved by the University of Pittsburgh Institutional Review Board (IRB Protocol # 0212087). Twenty-three ( $N = 23$ ) participants were recruited from University of Pittsburgh Medical Center Pathology Program and Allegheny General Hospital. All participants were pathology residents; nine of the subjects were first year residents, 7 were second year residents, 3 were third year residents, 3 were fourth year residents, and 1 was a fifth year resident. Eleven residents had a previous dermatopathology rotation, and nine subjects had previous medical training in another country. All subjects were volunteers solicited by email and received \$500 for the two day study.

## System Training

Participants were trained to use the baseline tutoring system on the first day, and to use System A or System B on the second day. Participants watched a 20 minute video demonstration of the system, and then practiced using the system while being monitored by the trainer to ensure they explored the full functionality of the system.

## Assessments

Three computer-based assessments were developed by the research group in close collaboration with an attending dermatopathologist. Each test was composed of four virtual slide cases not seen during the tutoring sessions, including two *tutored patterns* and two *untutored patterns*. *Tutored patterns* are cases with combinations of evidence that students had the opportunity to see during the tutoring session. *Untutored patterns* are cases with combinations of evidence that students did not have the opportunity to see during the tutoring session. Use of both tutored and untutored patterns provides an opportunity for the participant to assess whether they can accurately distinguish what they know and don't know.

For each case in the assessment, students were asked to provide both a diagnosis or differential diagnosis, and a justification of their conclusion in the form of individual findings supporting the diagnosis. Students also rated their certainty about each finding and hypothesis as a Boolean



value indicating whether they felt they were sure or unsure about the veracity of the finding or hypothesis.

The relationship of the assessments to the training sessions is depicted in Figure 1. Test 1 was administered before any tutoring, to determine baseline feeling-of-knowing. Test 2 was administered after using the baseline tutoring system with immediate feedback as a metacognitive scaffold. Test 3 was administered after use of system A or B in the faded feedback condition.

### Assessment Scoring

For each case (question) on the tests, we determined separate scores for the diagnosis and feature components. For the diagnosis component, we gave 5 points for the first correct diagnosis, and added 3 points for each additional correct diagnosis (when the case required a differential diagnosis). We subtracted 1 point for each incorrect diagnosis. No points were assigned for blank answers. For the feature component, we gave 2 points for each correct feature, and subtracted ½ point for each incorrect attribute (quality) of the feature. Because questions could vary for total points possible, scores were normalized to produce equal weighting by case. Overall diagnosis and feature test scores were the average of the scores from the individual questions. The combined score is an average of the diagnosis and feature scores.

### Case Selection

All cases were obtained from the University of Pittsburgh's Department of Pathology with the assistance of an Honest Broker who provided de-identification of the slides and histories. Cases that met our initial criteria for inclusion (based on pattern and typicality) were approved by our dermatopathology expert to be used by the tutor. The baseline tutoring system included 15 cases representing 5 different patterns that repeated through the session. System A and B used 25 identical cases representing 17 random patterns: 4 of these patterns were repeated from the baseline tutoring system (session 1) and 13 were newly introduced patterns. The higher case/pattern ratio used during the baseline tutoring session was designed to allow students to reach mastery on a subset of the problems during the first tutoring session.

### Data analysis

We used three measures to describe the degree to which feeling-of-knowing correlated with performance. Metrics are based on a two-by-two contingency table created by comparing the feeling-of-knowing ('sure or 'unsure') to the performance on the finding or case ("correct" or "incorrect"). All possible combinations of outcomes are shown in Figure 4A. For the coloring book data, only actions performed without any prior feedback from the system were used in our calculations. For the tests, all features and diagnoses mentioned by participants were counted as separate items and their overall count was used to calculate the three measures. Additionally, these measures were calculated for features and diagnoses separately. Since these measures are evaluating the participants FOK, i.e. how sure or unsure of the identified feature with no relevance to whether it is correct or incorrect, we did not separate tutored and untutored patterns.

Bias<sup>39</sup> is a measure of the overall degree to which confidence matches performance (Figure 4B). The first term represents the relative proportion of items judged as known (total known items divided by all items), and the second term represents relative performance on items (total correct items divided by all items). Bias scores *greater than zero* indicate over-confidence and bias scores *less than zero* indicate under-confidence, with an optimum value of zero indicating perfect matching between their confidence and performance.

Discrimination<sup>39</sup> is a measure of the ability to discriminate performance on correct and incorrect items. Scores are calculated as the difference between the proportion of correct items judged as known and the proportion of incorrect items judged as known. As the discrimination score increases from zero, it reflects improving metacognitive performance.

The Goodman-Kruskal Gamma correlation (G) is an integrative measure of overall FOK accuracy<sup>40</sup>, computed as a proportional difference of concordant and discordant pairs (Figure 4B). G correlation ranges from  $-1.0$  (perfectly negative correlation) to  $+1.0$  (perfectly positive correlation) with zero indicating no correlation. It is interpreted as a probabilistic value and not in terms of variance<sup>39,40</sup>. The optimum desired value of G is  $+1.0$ ; reflecting a perfect correlation such that when they are correct they are sure, and when they are incorrect they are unsure.

Repeated measures MANOVA was used to compare FOK metrics and learning gains obtained at multiple time points. All analyses were performed using SPSS, v. 15.

## RESULTS

### 1. Effect of Immediate Feedback

We determined differences in learning, FOK accuracy, and confidence before (test 1) and after (test 2) use of the baseline immediate-feedback tutoring system for all subjects.

**Learning gains**—We used the total test scores (all cases) for each subject to calculate the learning gains. Scores for tutored patterns and untutored patterns were analyzed separately. Both the control and MC groups show a significant feature learning gain (Figure 5) between test 1 and test 2 (one-factor MANOVA, effect of test,  $F(1,22)=40.2$ ,  $p<0.001$ ) which in turn accounts for the increase in the total score ( $F(1,22)=8.04$ ,  $p=0.01$ ). When we further divided the results into tutored and untutored patterns, we observed significant learning gains of tutored patterns (one-factor MANOVA, effect of test,  $F(1,22)=13.16$ ,  $p=0.001$ ) and no learning gains in untutored patterns (one-factor MANOVA, effect of test,  $F(1,22)=0.49$ ,  $p=0.49$ ). The findings replicate results demonstrated in previous studies of our tutoring system<sup>11,12</sup>.

**FOK accuracy**—Learning gains supported by immediate feedback are associated with an overall increase in FOK accuracy (Table 1) as measured by a significant increase in G correlation between test 1 and 2 (one-factor MANOVA, effect of test,  $F(1,22)=5.04$ ,  $p=0.04$ ) after using the immediate interface.

**Confidence level**—The discrimination index, which is a measure of the ability of the student to discriminate between correct and incorrect answers, increases significantly after immediate feedback (Table 1) ( $F(1,22)=5.92$ ,  $p=0.02$ ). However, immediate feedback did not affect the participants' overall over-confidence or under-confidence, measured by the Bias index, which shows no significant difference from test 1 to test 2 ( $F(1,22)=0.07$ ,  $p=0.79$ ).

In summary, immediate feedback is associated with significant learning gains for tutored patterns, an improvement of overall FOK accuracy, and an increase in the student's ability to distinguish correct and incorrect responses.

### 2. Effect of supplementary metacognitive scaffolds during fading of immediate feedback

We determined differences in learning, FOK accuracy, and confidence before (test 2) and after (test 3) using the delayed feedback interface, between subjects using a version of the system with metacognitive tutoring (System A) and subjects using a version of the system with no metacognitive tutoring (System B).



**Effect on learning gains**—There was no learning gain between test 2 and test 3 when using the faded interface (feature score:  $F(1,21)=0.11$ ,  $p=0.75$  and total score:  $F(1,21)=0.81$ ,  $p=0.38$ ). The same effect was observed for tutored patterns (feature score:  $F(1,21)=0.11$ ,  $p=0.74$  and total score:  $F(1,21)=0.43$ ,  $p=0.52$ ). The results show that there was no learning gain attributed to the metacognitive scaffolding during the removal of immediate feedback (Figure 6).

**Effect on FOK accuracy**—There was a trend toward a decline in G correlation (Table 2) from test 2 to test 3 (two-factor MANOVA, effect of test,  $F(1,21)=3.25$ ,  $p=0.09$ ), but there was no significant difference in this decline between the two groups (two-factor MANOVA, effect of condition,  $F(1,21)=0.73$ ,  $p=0.40$ ).

**Effect on confidence level**—Bias did not significantly change from test 2 to test 3 (two-factor MANOVA, effect of test,  $F(1,21)=0.24$ ,  $p=0.63$ ), and there was no interaction between test and condition (Table 2). Discrimination significantly declined between test 2 and test 3 (two-factor MANOVA, effect of test,  $F(1,21)=5.22$ ,  $p=0.03$ ), but there was no significant difference in this decline between the two groups (two-factor MANOVA, effect of condition,  $F(1,21)=0.60$ ,  $p=0.45$ ).

**Effect on feeling-of-knowing during the tutoring session**—Interestingly, the apparent decrement in metacognitive performance observed between test 2 and test 3 conflicts with observations made during the tutoring session itself. Students asked to indicate their degree of confidence in each feature, hypothesis and diagnosis showed a gradual increase in G correlation (Figure 7). Bias markedly improved (essentially reaching the optimum point half way through the problem set).

In summary, removal of immediate feedback in this study appeared to have important negative effects on FOK accuracy and discrimination, which were not significantly diminished by supplementary metacognitive scaffolds. FOK metrics collected during the tutoring session conflict with those obtained during interval assessments, showing a gradual increase in G correlation and marked improvement in Bias.

## DISCUSSION

This study demonstrates the positive effect of metacognitive scaffolds on student self regulation. We used four forms of metacognitive scaffolding; three adaptive (immediate feedback, coloring book, and inspectable student model) and one static scaffold (pseudo-dialogue using pre-stocked questions), and specifically tested immediate feedback against the other three.

Critics of *Immediate feedback* have suggested that this scaffold may in fact reinforce false metacognitive beliefs, making problem-solving a process of immediate single discrete steps rather than a reflective process, and that the rapid progression of students through a series of tasks may lead to over-estimation of their performance<sup>41–43</sup>. This could be a significant limitation to the use of the cognitive tutor paradigm in medical education, because the benefit of enhancing performance must be accompanied by an improvement in feeling-of-knowing. Clinical acumen encompasses accurate performance correlated to level of confidence.

Our other interventions provided a spectrum of other kinds of scaffolding. The *coloring book interface* allows direct comparison of student beliefs and performance encouraging students to reexamine errors committed due to confusion or inappropriate rule- application. The inspectable *student model* is a holistic presentation of the system's assessment of student skill across the entire domain. The use of inspectable student models, such as skill-o-meters<sup>10,44, 45</sup>, has been well-established as a cognitive scaffold. But there is no conclusive data related to

its use as a metacognitive scaffold. The final method we used was a pseudo-dialogue with *prestocked questions* - a static form of metacognitive scaffolding.

This study tested immediate feedback against the other metacognitive scaffolds because we needed to independently evaluate the role of immediate feedback, given its strong criticisms. Additionally, we sought to determine whether other metacognitive scaffolds can support SRL as immediate feedback is withdrawn. The results show that immediate feedback is associated with an improvement of overall FOK accuracy and an increase in the student's ability to distinguish correct and incorrect responses. This effect supports current theoretical models of SRL<sup>46,47</sup>. As proposed by Azevedo et al., when the discrepancy between goals and outcomes are reduced and students are making progress, a positive effect reinforces self-regulation. Feedback in an ITS can be provided after their tutoring session or throughout the entire session. Feedback given after the session only provides a state of achievement to the student with little or minimum external guidance to self regulation<sup>48</sup>. On the other hand, feedback throughout the entire session provides cues directly affecting student's self regulation<sup>49,50</sup> by enhancing the student's ability to associate these cues with actual performance or correctness<sup>51</sup>.

The dilemma of immediate feedback is that it must ultimately be removed to allow transfer to real world tasks. Therefore, effective use of ITS in medical education requires that we are able to gradually eliminate this cognitive scaffold without any adverse metacognitive effects. In this study, we were not able to remove immediate feedback without observing negative effects on FOK in the interval test assessments. FOK accuracy and discrimination values dropped in both the experimental and control groups, suggesting that the supplementary metacognitive scaffolds were insufficient to support SRL. However, during the tutoring session itself, we did observe a gradual increase in G correlation (towards positive direction) and marked improvement in Bias (towards zero value), essentially reaching the optimum point half way through the problem set. Thus, students using the metacognitive scaffolds appear to improve in their ability to judge when they are correct or incorrect - an integral part of SRL. But interestingly, this did not appear to transfer to assessments in our study. One explanation for this apparent discrepancy is that the experiment was conducted over a very short period of time relative to other demonstrations of the effects of metacognitive scaffolds<sup>22,32-34</sup>. Two days may not be sufficient to achieve both cognitive gains and metacognitive gains in such a complex domain. If this is the case, then we may have simply removed the immediate feedback too quickly for other metacognitive scaffolds to compensate.

Our first study of the metacognitive scaffolds of SlideTutor has raised many questions requiring further study. First, what explains the differences between metacognitive performance observed in independent, periodic assessments and those observed during the tutoring process itself? How does the increase in FOK accuracy and discrimination seen during tutoring supported by metacognitive scaffolds compare to what is observed when no scaffolds are provided? Our study was not designed to specifically investigate this question, because only the intervention group assessed their accuracy using the coloring book. Further studies could more directly address this question.

Second, are the effects observed during the tutoring process indicative of early metacognitive gains that do not yet transfer to assessments? If so, what factors are associated with transfer of metacognitive skills and how can the tutoring system support this transfer? Our first study suggests that timing of withdrawal may be a critical factor. Ultimately, the timing of feedback withdrawal can be considered to be yet another parameter that could be set by the student model in order to individualize the fading of the scaffolds. What student modeling methods will be needed to identify when the "time is right"? Are rate of withdrawal, cycle-length, and overall interval also important variables? How can these variables be manipulated to support metacognitive gains as we gradually move students towards more authentic task conditions

that do not provide immediate feedback? Further research in this area will help elucidate the complex mechanisms that underlie medical cognition and metacognition, and focus researchers on developing computer-based environments that are able to enhance both of these processes.

## LIMITATIONS

This study was designed to test the effects of immediate feedback and other scaffolds on metacognition. The study is limited by the relatively small sample size. It is possible that with a larger sample size, a difference between control and intervention groups could have been detected for the metacognitive scaffolding during feedback fading, although the size of this difference would most likely be quite small. Only a single domain was studied, and therefore it is not possible to widely generalize these results across medical domains. Finally, the study showed a difference in FOK during tutoring which was not seen during assessments. Because it was not possible to measure FOK in the foil coloring book, the authors cannot determine whether the change in FOK during tutoring is directly related to the use of the metacognitive scaffolds. Further research is needed to more fully understand the findings of this study.

## Acknowledgments

Work on this project was supported by a grant from the National Library of Medicine (R01 LM007891). The work was conducted using the Protege resource, which is supported by grant LM007885 from the United States National Library of Medicine. We thank Lucy Cafeo for editorial assistance.

## References

1. Anderson, JR. *Rules of the Mind*. Erlbaum, Hillsdale, NJ: Lawrence Erlbaum Associates; 1993.
2. Corbett AT, Koedinger KR, Hadley WH. *Cognitive Tutors: From the research classroom to all classrooms - Technology Enhanced Learning: Opportunities for Change*. 2001
3. VanLehn K. The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Educ* 2006;16(3):227–65.
4. Wenger, A. *Artificial intelligence and tutoring systems—computational and cognitive approaches to the communication of knowledge*. Los Altos, CA: Morgan Kaufmann Publishers Inc; 1987.
5. Azevedo R, Lajoie S. The cognitive basis for the design of a mammography interpretation tutor. *Int J Artif Intell Educ* 1998;9:32–44.
6. Clancy W. Knowledge-based tutoring: the GUIDON program. *J Computer-based Instruct* 1983;10:8–14.
7. Sharples M, Jeffery N, du Boulay B, Teather B, Teather D, du Boulay G. Structured computer-based training in the interpretation of neuroradiological images. *Int J Med Inform* 2000;60:263–80. [PubMed: 11137471]
8. Smith, P.; Obradovich, J.; Heintz, P., et al. Successful use of an expert system to teach diagnostic reasoning for antibody identification. *Proceedings of the Fourth International Conference on Intelligent Tutoring Systems*; San Antonio TX. 1998. p. 354-63.
9. Crowley RS, Medvedeva O. An intelligent tutoring system for visual classification problem solving. *Artif Intell Med* 2006 Jan;36(1):85–117. [PubMed: 16098717]
10. Maries, A.; Kumar, A. The Effect of Student Model on Learning. *Advanced Learning Technologies, 2008 ICALT '08 Eighth IEEE International Conference*; 2008. p. 877-81.
11. Crowley RS, Legowski E, Medvedeva O, Tseytlin E, Roh E, Jukic D. Evaluation of an intelligent tutoring system in pathology: effects of external representation on performance gains, metacognition, and acceptance. *J Am Med Inform Assoc* 2007 Mar–Apr;14(2):182–90. [PubMed: 17213494]
12. Saadawi GM, Tseytin E, Legowski E, Jukic D, Castine M, Crowley RS. A natural language intelligent tutoring system for training pathologists: implementation and evaluation. *Adv Health Sci Educ Theory Pract*. 2007 Oct 13; Epub ahead of print.

13. Woo CW, Evens MW, Freedman R, et al. An intelligent tutoring system that generates a natural language dialogue using dynamic multi-level planning. *Artificial Intelligence in Medicine* 2006;38(1):25–46. [PubMed: 16352427]
14. Corbett, AT.; Anderson, JR. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Seattle WA: ACM NY; 2001a. Locus of feedback control in computer-based tutoring: impact on learning rate, achievement and attitudes; p. 245-52.
15. Azevedo R, Hadwin AF. Scaffolding Self-regulated Learning and Metacognition-Implications for the Design of Computer-based Scaffolds. *Instructional Science* 2005b;33:367–79.
16. White B, Frederiksen J. A Theoretical Framework and Approach for Fostering Metacognitive Development. *Educational Psychologist* 2005;40(4):211–23.
17. Graber ML, Franklin N, Gordon R. Diagnostic Error in Internal Medicine. *Arch Intern Med* 2005;165(13):1493–99. [PubMed: 16009864]
18. Voytovich AE, Rippey RM, Suffredini A. Premature conclusions in diagnostic reasoning. *J Med Educ* 1985;60:302–07. [PubMed: 3981589]
19. Winne, PH. Self-regulated learning viewed from models of information processing. In: Douglas, J.; Hacker, JD.; Arthur, C.; Graesser, S., editors. *Self-regulated Learning and Academic Achievement: Theoretical Perspective*. Mahwah NJ: Lawrence Erlbaum Assoc; 2001. p. 153-90.
20. Azevedo R, Moos D, Greene J, Winters F, Cromley J. Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development* 2008;56(1):45–72.
21. Zimmerman, B. Development and adaptation of expertise: The role of self-regulatory processes and beliefs. In: Ericsson, KA.; Charness, P.; Feltovich, P.; Hoffman, R., editors. *The Cambridge handbook of expertise and expert performance*. 2006. p. 705-22.
22. Winne, PH.; Hadwin, AF. Studying as self-regulated learning. In: Douglas, J.; Hacker, JD.; Arthur, C.; Graesser, editors. *Metacognition in Educational Theory and Practice*. Manwah NJ: Lawrence Erlbaus Assoc; 1998. p. 277-304.
23. Winne, P.; Hadwin, A. *The weave of motivation and self-regulated learning*. NY: Taylor & Francis; 2008. p. 297-314.
24. Azevedo, R.; Witherspoon, AM. Self-regulated learning with hypermedia. In: Graesser, A.; Dunlosky, J.; Hacker, D., editors. *Handbook of metacognition in education*. Mahwah, NJ: Erlbaum; 2008. in press
25. Nelson, T.; Narens, L. Metamemory: A theoretical framework and some new findings. In: Bower, G., editor. *The Psychology of Learning and Motivation*. San Diego CA: Academic Press; 1990.
26. Metcalf, J.; Dunlosky, J. Metamemory. In: Roediger, H., editor. *Cognitive psychology of memory*. Vol. 2. Oxford: Elsevier; 2008.
27. Reiser B. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences* 2004;13(3):273–304.
28. Graesser A, McNamara D, VanLehn K. Scaffolding deep comprehension strategies through Pint&Query, AuthTutor and iSTRAT. *Educational Psychologist* 2005;40(4):225–34.
29. Azevedo R, Cromley JG, Seibert D. Does adaptive scaffolding facilitate student's ability to regulate their learning with hypermedia? *Contemp Educ Psych* 2004;29(3):344–70.
30. Green BA. Project-based learning with the world wide web: A qualitative study of resource integration. *Educ Technol Rsch & Dev* 2000a;48(1):45–66.
31. Hill JR, Hannafin MJ. Cognitive strategies and learning from the world wide web. *Educational Technology Research and Development* 1997;45(4):37–64.
32. Choi I, Land SM, Turgeon AY. Scaffolding peer-questioning strategies to facilitate metacognition during online small group discussion. *Instructional Science* 2005;33:483–511.
33. Dabbagh N, Kitsantas A. Using web-based pedagogical tools as a scaffolds for self-regulated learning. *Instructional Science* 2005;33:513–40.
34. Puntambekar S, Hubscher R. Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist* 2005;40:1–12.
35. Aleven V, Koedinger KR. An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science* 2002;26:147–81.

36. Yudelson, MV.; Medvedeva, O.; Legowski, E.; Castine, M.; Jukic, D.; Crowley, RS. Mining student learning data to develop high lever pedagogic strategy in a medical ITS. Proceedings: AAAI Educational Data Mining 21st Nat'l Conf, Educ Data Mining Workshop; Boston MA: AAAI press; 2006. p. 82-90.
37. Yudelson MV, Medvedeva OP, Crowley RS. Multifactor Approach to Student Model Evaluation in a Complex Cognitive Domain. *User Model User Adapt* 2008 Sep;18(4):315–82.
38. Pintrich PR, De Groot EV. Motivational and Self-Regulated Learning Components of Classroom. *Journal of Educational Psychology* 1990;82(1):33–40.
39. Kelemen WL, Frost PJ, Weaver CA III. Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition* 2000;28 (1):92–107.
40. Nelson T. A Comparison of Current Measures of the Accuracy of Feeling-of-Knowing Predictions. *Psychological Bulletin* 1984;95(1):109–33. [PubMed: 6544431]
41. Nikos Mattheos AN, Falk-Nilsson Eva, Attström Rolf. The interactive examination: assessing students' self-assessment ability. *Medical Education* 2004;38(4):378–89. [PubMed: 15025639]
42. Kulhavy RW, Stock WA. Feedback in written instruction: The place of response certitude. *Educational Psychology Review* 1989;1:279–308.
43. Kulhavy RW, Yekovich FR, Dyer JW. Feedback and content review in programmed instruction. *Contemporary Educational Psychology* 1979;4:91–8.
44. Loboda, TD.; Brusilovsky, P. Proceedings of the 2006 ACM symposium on Software visualization. Brighton, United Kingdom: ACM; 2006. WADEIn II: adaptive explanatory visualization for expressions evaluation.
45. Mitrovic, Antonija; Martin, Brent. Evaluating the Effects of Open Student Models on Learning. Vol. 2347/2002. Springer: Berlin/Heidelberg; 2002. p. 296-305.
46. Carver C, Scheier M. Origins and Functions of positive and negative affect: A control-process view. *Psychological review* 1990;97:19–35.
47. Azevedo, R.; Witherspoon, AM. Self-regulated use of hypermedia. In: Graesser, A.; Dunlosky, J.; Hacker, D., editors. *Handbook of metacognition in education*. Mahwah NJ: Erlbaum; in press
48. Butler D, Winne PH. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research* 1995;65(3):245–81.
49. Winne PH. Minimizing the balck box problem to enhance the validity of theories about instructional effects. *Instructional Science* 1982a;11:13–28.
50. Winne PH, Marx RW. Students' and teachers' views of thinking processes for classroom learning. *Elemental School Journal* 1982b;82:493–518.
51. Balzer W, Doherty M, O'Conner R. Effects of cognitive feedback on performance. *Psychological Bulletin* 1989;106:410–33.

## Appendix 1 Metacognitive Pseudo-dialogue

### During case

1. Having a strategy usually helps in solving any problem. What sentence would best describe your strategy at the beginning of a case?
  - a. I try to place the case first in a general category of disease.
  - b. I try to remember a similar case.
  - c. I try to look at the whole slide and then focus on finding features.
  - d. I like to hypothesize and then confirm or dispute my hypothesis.
2. At any given time in problem solving, it is usually helpful to have a strategy in place. What sentence would best describe your strategy during viewing a case?
  - a. I try to look for a pathognomonic feature that would narrow the differential diagnosis as early as possible.

- b. I have a systematic approach that would narrow the differential diagnosis in a gradual step by step fashion.
  - c. I identify all the features in the order I see them until I come up with a differential diagnosis.
  - d. I ask for help to have an idea of features I should be looking for in the case.
  - e. I try to place the case in a general schema (visual representation in my head) of concepts and diagnoses.
- 3. What is the specific purpose for the strategy you are using?
  - a. This strategy has worked in the past.
  - b. This strategy is effective in this case.
  - c. I don't know of another strategy.
  - d. I don't have any strategy.
- 4. Did you set specific goals before viewing this case? What are they?
- 5. What features are most important to identify in this case? Why are they important?
- 6. Did you ask yourself questions about the case before you began? What were these questions?
- 7. What sentence would best describe your goals at this point?
  - a. I have a hypothesis but I am not able to confirm or dispute it.
  - b. I am trying to locate all the features on the slide before making a hypothesis.
  - c. I am just looking around the slide trying to find something familiar.
  - d. I am completely lost and I need help.
- 8. Did you slow down when you encountered important features like <random\_feature>?
- 9. What features are possibly associated with the <random\_hypothesis>?
- 10. One way that expertise can be acquired is to develop a schema or representation of concepts such as visual features and their relationships to different diagnoses. Do you have a schema (visual representation) in your head that helps you integrate the information you learn after each case?
- 11. Studies show the amount you learn is related to your motivation. How motivated are you to learn something new from the tutor?
- 12. It is important to identify ones intellectual strengths and build on them to account for ones weakness. What sentence would best describe your intellectual strengths and weaknesses?
  - a. I am better at viewing the slide when I have studied and know the educational concepts for a case.
  - b. I can learn how to diagnose the case even without knowing anything about the subject.
  - c. I like to read a textbook before attempting to view cases.
  - d. I cannot learn about a case until I have been taught the educational concepts for a case



13. Locking onto a salient feature at initial presentation and failing to shift from your first impression of the case is a common heuristic error used in medical decision making. Which of the following heuristics does this describe?
- Anchoring
  - Pseudo-Diagnosticity
  - Satisficing
  - Representativeness
  - I do not know
14. A common heuristic error is the tendency to not search for other possible diagnoses once a satisfactory solution has been reached leading to premature diagnostic closure of the case. Which of the following heuristics does this describe?
- Anchoring
  - Pseudo-Diagnosticity
  - Satisficing
  - Representativeness
  - I do not know
15. Seeking features that confirm an initial diagnosis but not seeking features that support a competing diagnosis is known as which of these options?
- Anchoring
  - Pseudo-Diagnosticity
  - Satisficing
  - Representativeness
  - I do not know
16. A heuristic error commonly made through categorization of cases depending on one prototypical feature is known as which of these options?
- Anchoring
  - Pseudo-Diagnosticity
  - Satisficing
  - Representativeness
  - I do not know

### End of case

1. Are you consciously focusing your attention on important features? What are these features?
2. It is a good strategy to stop and review a case to help understand important relationships between features and diagnoses. How frequently are you reviewing the case to help understand these relationships?
3. Now that you have reached <random\_diagnosis>, what sentence would best describe what you did?

- a. I think it would have been easier to reach the diagnosis if I had asserted a hypothesis earlier.
  - b. I think I should have asked for more help.
  - c. I think I should have studied the slide more carefully and identified more features before I attempted a hypothesis.
  - d. I worked through the case efficiently and should not have done anything differently.
4. What sentence would best describe your asking for help?
  - a. I ask for help only when I need it.
  - b. I frequently ask for help to be sure of my work.
  - c. I never ask for help.
  - d. I only ask for help if everything else fails.
5. It is a good strategy to link relationships between features and diagnoses in different cases. Is a <random\_feature> related to what you have already seen in previous cases? Explain.
6. It is a good strategy to stop and reevaluate your actions to ensure that they are consistent with your goal. How did you reevaluate <random\_hypothesis>?
7. What sentence would best describe how you learn?
  - a. I have a schema (visual representation in my head) for features and their relationships with a diagnosis.
  - b. I just take each case individually.
  - c. I try to remember previous cases and link them to what is in the slide.
  - d. I try to remember what I read about the subject.
8. What sentence would best describe how correct you are in identifying the features and reaching a diagnosis?
  - a. I am always sure when I am correct and when I am incorrect.
  - b. Most of the time, I am sure when I am correct.
  - c. Most of the time, I am sure when I am incorrect.
  - d. I am never sure how correct or incorrect I am.
9. What sentence would best describe how much you learned in comparison to what the tutor expected you to learn in this case?
  - a. I think I learned most of what the tutor wanted me to learn.
  - b. I don't think I learned what the tutor wanted me to learn.
  - c. I think I got some of what the tutor wanted me to learn.
  - d. I don't think the tutor was teaching me anything new.
10. Did you ever lock onto a salient feature at initial presentation and fail to shift from your first impression of the case? This is a common heuristic error made in the medical field and is known as Anchoring.

11. Did you consider multiple hypotheses for the case? A common heuristic error known as Satisficing is the tendency to not search for other possible diagnoses once a satisfactory solution has been reached leading to premature diagnostic closure of the case.
12. When trying to reach your definitive diagnosis from the list of hypotheses, did you seek both confirming and disputing findings? A common heuristic error known as Pseudo-Diagnosticity occurs when data is sought that confirms one but not other competing hypotheses.
13. Do you try to identify multiple features to support a hypothesis? Representativeness is one of the heuristic errors commonly made through categorization of cases depending on one prototypical feature.

### Inspectable Student Model (Knowledge Explorer) Questions at End of Case

1. Using the Knowledge Explorer, can you tell if you were right or wrong about <wrong\_diagnosis\_or\_hypothesis>? If you were wrong, what error did you make?
2. Looking at the Knowledge Explorer, what things have you seen but haven't learned?
3. Click on the Self Check Summary tab. For items you were wrong or unsure about (refer to the self check column), how much knowledge does the tutor think you have about it?

## Appendix 2: Foil Dialogue

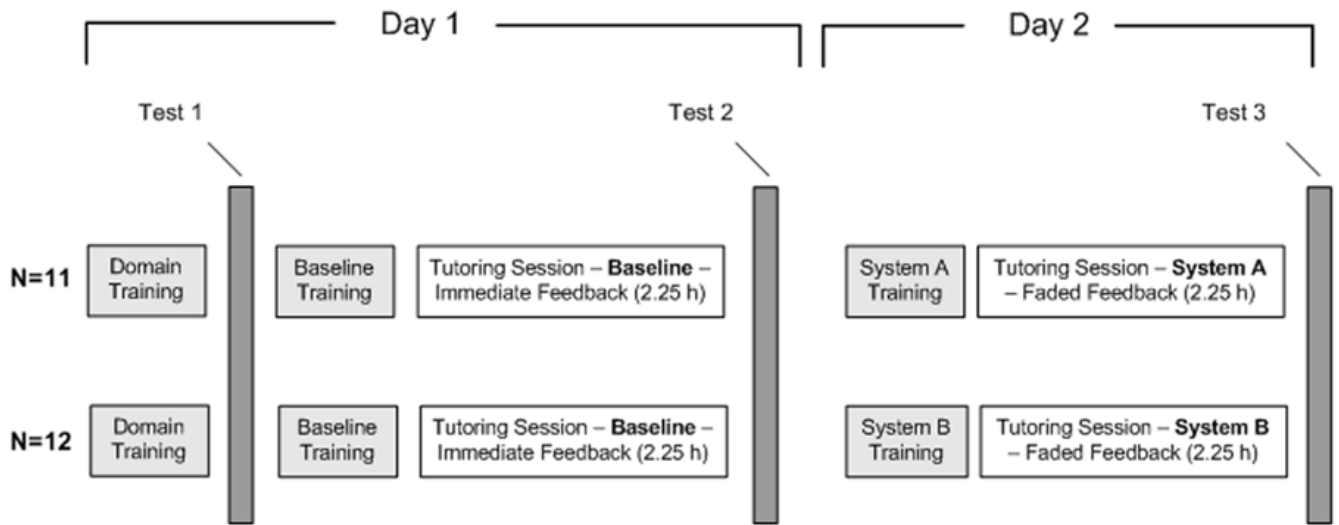
### Early in case

1. Why do you need high power for the diagnosis of this case?
2. Why do you need low/medium power for the diagnosis of this case?
3. Do you have a set of feature(s) you look for in this kind of case? If yes, please list.
4. Why did you feel *feature x* was important in coming to a diagnosis?
5. What other feature(s) can you confuse with *feature X*?
6. What is the differential diagnosis you think of when you see *feature X*?
7. What are some important features related to *hypothesis X*?
8. Does the *attribute Z* of *feature X* have other values?
9. What are the hypotheses supported by *feature X*?

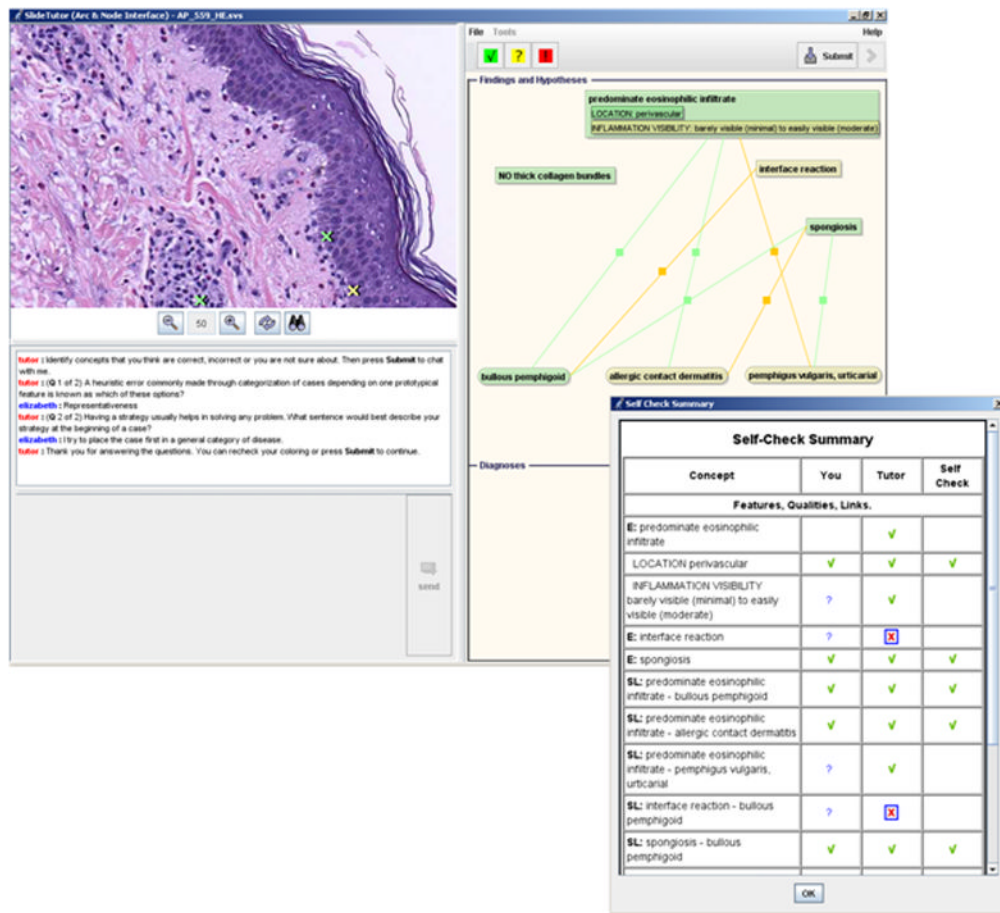
### Later in case (have suggested at least one diagnosis)

1. What feature(s) do you think is the most crucial in coming to the diagnosis of this case?
2. What are some important features that you have learned in diagnosing *diagnosis x*?
3. What do you think is the stain used in this slide? When you sign-out, would you like to have more stains done on the same specimen?
4. What are the features easily identified by using H&E?
5. What are the features easily identified by using Pas D?
6. What are the features easily identified by using colloidal iron ?

7. If you had to sign out this case, are there any additional information/tests/that you would like to have completed? If so, why?
8. Are there any additional features that you would have liked to identify in coming to the diagnosis that were not in the tutor?
9. The *diagnosis X* is supported by presence of what features?



**Figure 1.**  
Study Timeline

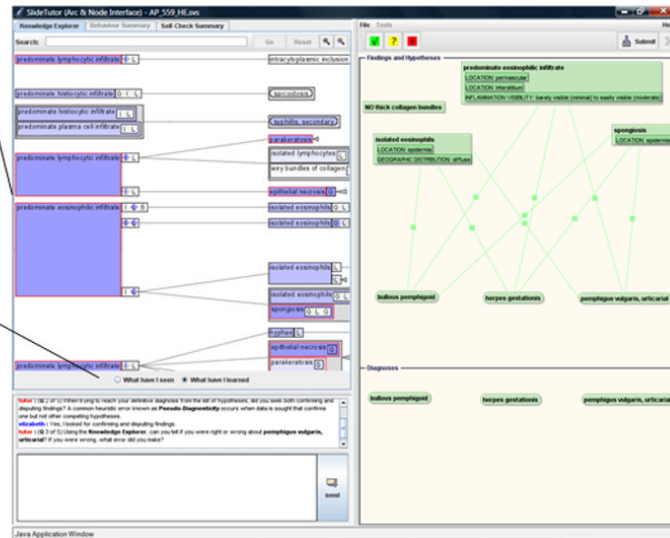


**Figure 2.** System A coloring book interface. Participants revisit their findings and are provided a self-check summary (inset)



Graph-based representation of knowledge base highlighting what the student has learned (shown here) or seen (not shown). Shade of box is relative to degree of mastery computed by Student Model. Darker shading indicates increasing mastery.

Students can toggle between "what have I learned" and "what have I seen"



**Figure 3.** System A inspectable student model. Participants see what they have mastered and what they have seen across all cases and knowledge base at the conclusion of each case.

**A**

**Performance**

		correct	incorrect
Feeling-Of-Knowing	sure	True Positive a	False Positive b
	unsure	False Negative c	True Negative d

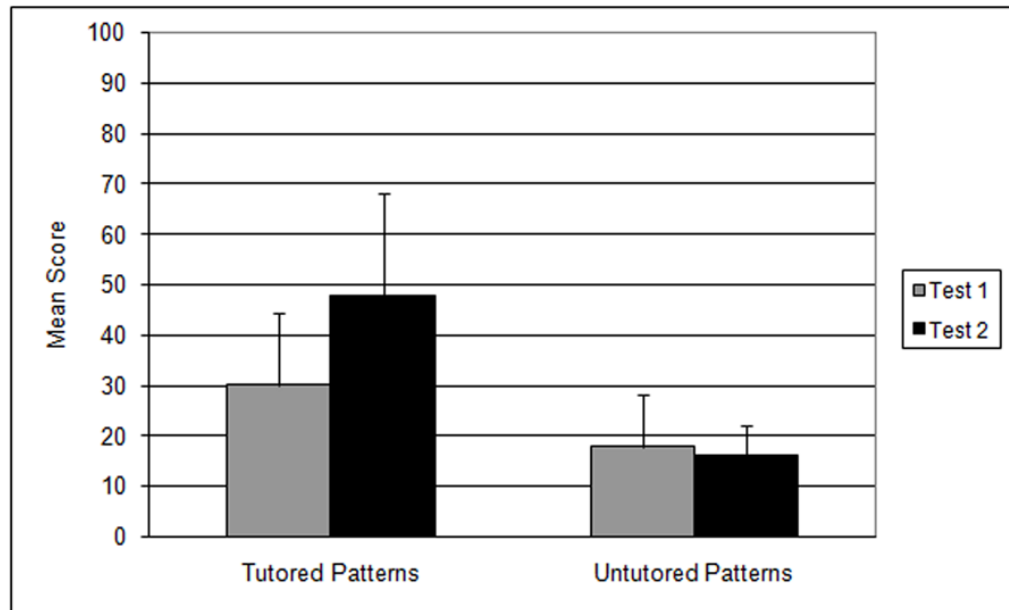
**B**

Bias = confidence - performance  
 $= (a+b)/(a+b+c+d) - (a+c)/(a+b+c+d)$

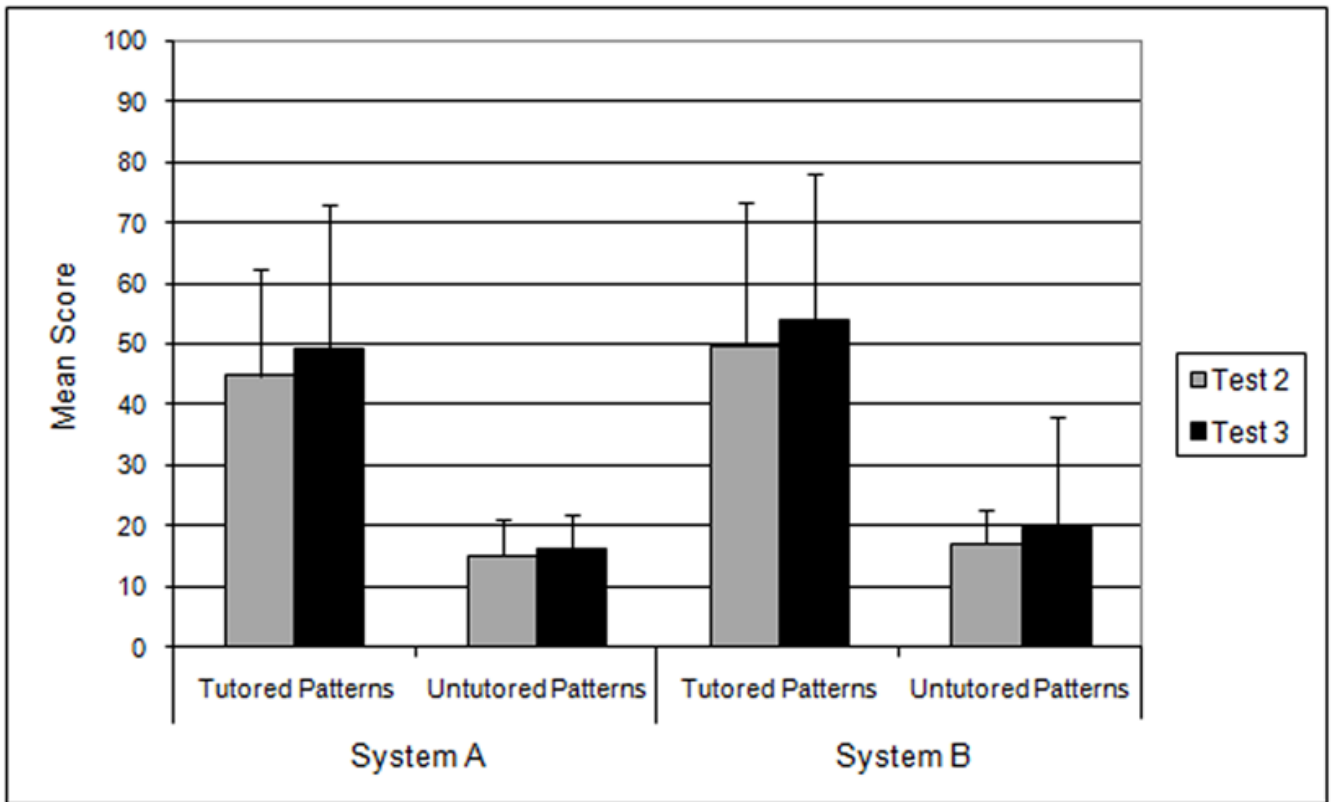
Discrimination = confidence on correct items - confidence on incorrect items  
 $= a/(a+c) - b/(b+d)$

**G** = (concordant pairs - discordant pairs)/(concordant pairs + discordant pairs)  
 $= (ad - bc) / (ad + bc)$

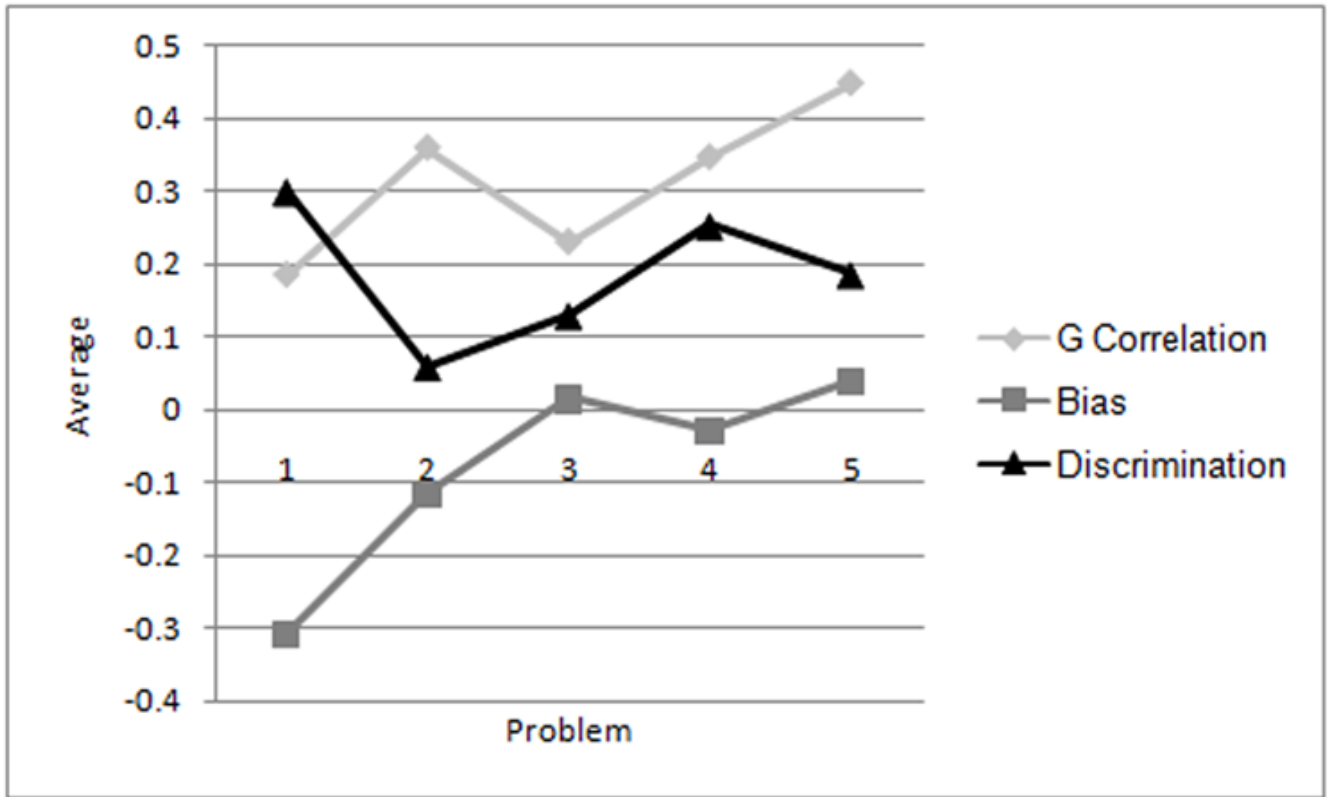
**Figure 4.**  
 Definition of FOK metrics: (A) Contingency Table; (B) Equations



**Figure 5.**  
Total score learning gains using immediate feedback



**Figure 6.**  
Total score learning gains during removal of immediate feedback



**Figure 7.**  
FOK measures obtained during tutoring session (intervention group only)

**Table 1**

Effect of immediate feedback on FOK

	Test 1	Test 2	Test 1 to 2		
			Delta	F	p-value
G Correlation	0.57	0.74	0.17	5.04	0.04
Bias	0.13	0.13	0.0	0.07	0.79
Discrimination	0.30	0.42	0.12	5.92	0.02



**Table 2**

Effect of metacognitive training on FOK

	Test 2	Test 3	Test 2 to 3				Effect of Test		Interaction	
			Delta	Effect of Condition		F	p-value	F	p-value	
				F	p-value					
<b>G Correlation</b>										
MC	0.80	0.62	-0.18	0.73	0.40	3.25	0.09	0.24	0.63	
Control	0.69	0.58	-0.11							
<b>Bias</b>										
MC	0.09	0.13	0.04	0.10	0.75	0.24	0.63	2.13	0.16	
Control	0.16	0.09	-0.07							
<b>Discrimination</b>										
MC	0.49	0.30	-0.19	0.60	0.45	5.22	0.03	2.14	0.16	
Control	0.36	0.31	-0.05							