Evaluation of an Intelligent Tutoring System in Pathology: Effects of External Representation on Performance Gains, Metacognition, and Acceptance

REBECCA S. CROWLEY, MD, MS, ELIZABETH LEGOWSKI, OLGA MEDVEDEVA, MS, MS, EUGENE TSEYTLIN, MS, ELLEN ROH, MD, DRAZEN JUKIC, MD, PhD

Abstract

Objective: Determine effects of computer-based tutoring on diagnostic performance gains, metacognition, and acceptance using two different problem representations. Describe impact of tutoring on spectrum of diagnostic skills required for task performance. Identify key features of student-tutor interaction contributing to learning gains.

Design: Prospective, between-subjects study, controlled for participant level of training. Resident physicians in two academic pathology programs spent four hours using one of two interfaces which differed mainly in external problem representation. The case-focused representation provided an open-learning environment in which students were free to explore evidence-hypothesis relationships within a case, but could not visualize the entire diagnostic space. The knowledge-focused representation provided an interactive representation of the entire diagnostic space, which more tightly constrained student actions.

Measurements: Metrics included results of pretest, post-test and retention-test for multiple choice and case diagnosis tests, ratios of performance to student reported certainty, results of participant survey, learning curves, and interaction behaviors during tutoring.

Results: Students had highly significant learning gains after one tutoring session. Learning was retained at one week. There were no differences between the two interfaces in learning gains on post-test or retention test. Only students in the knowledge-focused interface exhibited significant metacognitive gains from pretest to post-test and pretest to retention test. Students rated the knowledge-focused interface significantly higher than the case-focused interface.

Conclusions: Cognitive tutoring is associated with improved diagnostic performance in a complex medical domain. The effect is retained at one-week post-training. Knowledge-focused external problem representation shows an advantage over case-focused representation for metacognitive effects and user acceptance.

Introduction

Training of health professionals grows more and more difficult for a number of reasons. Federal legislation impacts residency work time and limits opportunities for residents and students to participate in patient care. Increased workloads for health professionals reduce the time remaining for teaching the next generation of practitioners, and the development of the distributed health care enterprise weakens the traditional apprenticeship relationship as mentors and mentees may travel to many different hospitals during a single rotation.

Computer based education can counter some of these trends, but generally fails to provide a structured, coached,
training environment which is the strength of apprenticeship during residency training. In contrast, intelligent medical training systems\textsuperscript{1,2} could be an important computer-based supplement to apprenticeship, because they provide opportunities for case-based problem solving combined with feedback and guidance, but in an environment where patients cannot be harmed. In other domains, Intelligent Tutoring Systems (ITS) have been highly successful.\textsuperscript{3–6} But there have been no evaluations of medical tutoring systems to determine whether these systems increase performance on complex medical tasks. We have developed a general tutoring system in Pathology which adapts the ITS paradigm. In this study, we report on an evaluation of the system to determine the effect of tutoring and problem representation on performance, metacognition, and acceptance.

**Background**

**Intelligent Tutoring Systems**

Intelligent Tutoring Systems (ITS) are adaptive, instructional systems that attempt to emulate the well known benefits of one-on-one human tutoring.\textsuperscript{7} 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. Typically, these systems incorporate an explicit encoding of domain knowledge and pedagogic expertise, which makes it possible to individualize instruction without necessarily anticipating specific interactions.\textsuperscript{8,9}

Cognitive Tutoring Systems—a subtype of ITS—incorporate domain-specific production rules that are based on a cognitive theory of skill acquisition, such as ACT-R.\textsuperscript{10} Intermediate cognitive steps are first identified using empirical methods such as cognitive task analysis.\textsuperscript{11} Most cognitive tutoring systems have been developed for domains that are highly procedural and do not require substantial declarative knowledge bases. Examples include mathematics and science instruction,\textsuperscript{3,4} flight simulator training,\textsuperscript{12} and training in the workplace.\textsuperscript{5} The instructional “gold standard” is considered to be one-on-one human tutoring, which is associated with a 2-sigma effect over classroom learning.\textsuperscript{13} Cognitive tutoring systems have been shown to bring students 1-sigma or more above standard instructional methods.\textsuperscript{3–6} In contrast, meta-analyses of many traditional computer-assisted instruction systems have exhibited only a 0.3–0.5 sigma effect.\textsuperscript{14,15}

**Medical Tutoring Systems**

Prior to the current work, a small number of medical ITS have been developed.\textsuperscript{16–23} However, to our knowledge, none of these systems have been tested to determine if they improve performance on diagnostic tasks.

Most medical ITS take a pedagogic approach that can be characterized along a spectrum, from purely knowledge-based to purely case-based. GUIDON, an Intelligent Tutoring System that incorporated MYCIN’s rule set, serves as the prototypical knowledge-based medical ITS.\textsuperscript{24,25} GUIDON made an important early contribution to research in ITS. GUIDON taught medical students to reason about infectious meningitis and bacteremia, and to identify the most likely causative organism given a patient’s history, physical examination, and laboratory results. Interaction used a mixed-initiative method of dialogue—either the student or the system could be in control of the discussion. Ultimately the MYCIN rule set proved inadequate for tutoring, and led the authors to develop NEOMYCIN.\textsuperscript{26} The diagnostic Meta-Strategy of NEOMYCIN was an important contribution to research in cognitive modeling of medical problem solving.

In contrast, MR Tutor\textsuperscript{20} provides an example of a case-based tutoring system which trains students using a focus on case similarity across instances. The system uses statistical indexes of image similarity to develop training systems in radiology. Students learn by example from a library of radiologic images. The tutor exploits differences in measurements of typicality or similarity to train clinicians to recognize the full breadth of presentations of a given entity.

**Effects of Problem Representation on Problem-solving**

ITS interfaces provide an external representation for students to express the intermediate reasoning steps in problem-solving.\textsuperscript{9} This external representation may also benefit students by replacing or supplementing less functional internal representations (mental models). Thus an important question in developing an ITS in a new domain is: What kinds of problem representation in the interface will best support learning in this domain? Research in cognitive psychology and medical problem solving offers guidance.

Cognitive psychologists have studied the effect of external representations on problem-solving and found that graphical representations are associated with enhanced learning in a number of domains including logic\textsuperscript{27} and hypothetico-deductive reasoning.\textsuperscript{28} Well-designed diagrammatic representations may be beneficial for a number of reasons including extension of working memory, support for perceptual inference, and reduction of search.\textsuperscript{29} Interestingly, some researchers have found interactions between use of external representations and individual characteristics of the learner. For example, students with higher spatial abilities and students who were less expert tended to benefit disproportionately.\textsuperscript{30} Although external visual representations can be categorized by their placement along a series of scales\textsuperscript{31,32} (see first column of Table 1), there is little work to suggest how such specific characteristics of a representation contribute to the learning process.

<table>
<thead>
<tr>
<th>Characteristics of External Representations</th>
<th>Knowledge-Focused</th>
<th>Case-Focused</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial vs Nonspatial</td>
<td>Spatial</td>
<td>Spatial</td>
</tr>
<tr>
<td>Concrete vs Abstract</td>
<td>Abstract</td>
<td>Abstract</td>
</tr>
<tr>
<td>Continuous vs Discrete</td>
<td>Discrete</td>
<td>Discrete</td>
</tr>
<tr>
<td>Numeric vs Non-numeric</td>
<td>Non-numeric</td>
<td>Non-numeric</td>
</tr>
<tr>
<td>Static vs Dynamic</td>
<td>Dynamic</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Causal vs Non-causal</td>
<td>Causal</td>
<td>Causal</td>
</tr>
<tr>
<td>Temporal vs Non-temporal</td>
<td>Temporal</td>
<td>Non-temporal</td>
</tr>
<tr>
<td>High vs Low Information Quantity</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Whole vs Part emphasis</td>
<td>Whole</td>
<td>Part</td>
</tr>
<tr>
<td>Heirarchical vs Non-heirarchical</td>
<td>Heirarchical</td>
<td>Non-heirarchical</td>
</tr>
</tbody>
</table>
Mental Representations for Medical Problem-solving

Work on medical problem-solving over the last 30 years has identified a number of potential mental representations that may guide medical problem-solving. Schmidt et al.\textsuperscript{33} have proposed that development of clinical expertise is accompanied by a transition from basic mechanistic models of disease, to illness scripts (schemas)\textsuperscript{34} to exemplars derived from experience. The proposal suggests that case-based representations predominate as expertise grows. In contrast, Mandin et al.\textsuperscript{35,36} favor a more knowledge-based representation corresponding to a disease classification tree beginning with features and ending with specific diagnoses. Some authors have suggested that multiple mental representations may be maintained and utilized by experts, especially when the default representation is insufficient for problem-solving.\textsuperscript{37}

In designing an appropriate external representation, the choice to emphasize the instance or “case” vs. the entire “knowledge-base” is a specific kind of part versus whole depiction (Table 1) with special significance in medical problem-solving and medical ITS. To our knowledge, there has been no other investigation of case-based versus knowledge-based problem representations as they relate to skill acquisition, metacognition, and student experience in an ITS.

Research Questions

1. Does cognitive tutoring improve diagnostic performance in a complex, medical domain?
2. If cognitive tutoring improves diagnostic performance, is the effect retained past the training period?
3. Does cognitive tutoring improve the metacognitive ability of students to correctly gauge when they know and don’t know the correct diagnosis?
4. Does problem representation (case-focused vs. knowledge-focused) affect learning gains or metacognitive gains made by students using the tutoring system?
5. Do students differ in their acceptance of tutoring systems that utilize case-focused versus knowledge-focused problem representations?

Methods

Study Design

The study used a prospective, between-subjects design, controlled for participant level of training by random assignment of same-level pairs. The entire study was performed in our laboratory, and consisted of two sessions (Figure 1). The first session was a full day, including (in order): (1) pretest, (2) interface training period, (3) 4.5 hour working period, (4) post-test, and (5) survey. Participants returned one week later to take a retention test. All participants received the same assessments.

During the working period, time on task was held constant. Subjects worked through a sequence of 20 dermatopathology cases of Subepidermal Vesicular Dermatitis (Table 2). The sequence was identical for all participants (Figure 1). For subjects who completed the entire set of cases before the end of the working period, the sequence started over again and they re-solved problems until the entire working period had elapsed.

Materials

1. Cases

Cases were obtained from the slide archives of five university-based pathology departments with active dermatopathology services. We selected twenty eight cases of Subepidermal Vesicular Dermatitis—representing 13 different visual patterns (combinations of evidence) and 21 different diagnostic entities. Each of these cases was used during the tutoring session and eight cases were used in learning assessments (Table 2). In many cases, only a differential diagnosis could be reached on the basis of the evidence. Whenever multiple instances of the same pattern were required for both assessment and working period (Figure 1), we randomly assigned individual cases for these purposes. Additionally, we selected four cases that were not Subepidermal Vesicular Dermatitis as control cases for use in learning assessments. Diagnosis was confirmed by a second pathologist on all cases.

2. System

SlideTutor is an intelligent tutoring system in Dermatopathology. It is one instantiation of our general VCT framework for tutoring of visual classification problem solving.\textsuperscript{38} SlideTutor and the VCT framework are Cogni-
tive Tutors—utilizing a model of skill development based on our previous research on expertise in microscopic diagnosis. The computational methods and implementation of the system have been previously described. Briefly, the system uses a client-server architecture, implemented in Java and Jess—a Java based production rule system. Cases to be solved by the student are based on virtual slides—gigabyte size images created by scanning glass slides at high resolution using a robotic microscope. Virtual slides are then annotated to produce a frame-based representation of evidence and locations in Protégé. Evidence–hypothesis relationships (domain knowledge), information about the task (task knowledge), and teaching strategies (pedagogic knowledge) are also created and maintained in Protégé. Jess rules use these frame-based representations to create a dynamic solution graph (DSG). The graph represents the current problem-state and advances with the student. The tutoring system evaluates all student actions against the current state of the DSG. Correct student actions propagate changes within the graph to produce the next problem-state, specific to that case and student. Incorrect actions are matched to a set of general errors and result in context-specific remediation that includes visual and textual explanations. Requests for help by the student produce increasingly specific hints by the tutoring system. Hints are also context-specific because they are based on the computed best-next-step which may change as the DSG advances. Hints do not advance the DSG. All student actions and tutor responses are captured and stored in an Oracle database for further analysis (see Tutoring Process Measures below).

3. Interfaces

We developed two different interfaces for this study. Figure 2 depicts the two interfaces for the identical problem state, contrasting the problem representations.

The case-focused interface (Figure 2A) presents problems using the case as the central focus. When students identify features that are present or absent in the case, they are displayed as square boxes. Students also specify descriptive qualities of these features (e.g., location, quantity) which appear within the feature boxes. When hypotheses are asserted, they are displayed as unconnected nodes. Users can test relationships between features and hypotheses by drawing support and refute links between evidence and hypotheses. Hypotheses may be moved into the Diagnostics area of the palette when a diagnosis can be made (dependent on the state of the DSG and the student model). Only the features present in the actual case are represented, but any valid hypothesis can be added and tested. At the end of each case, the diagram shows the relationships present in a single case. These diagrams will be different for each case. Therefore, while relationships are seen among features and hypotheses within an individual case, students must construct relationships across cases themselves.

In contrast, the knowledge-focused interface (Figure 2B) presents problems within the context of the entire knowl-

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Diagnosis or Differential Diagnosis</th>
<th>Pattern Appeared in All Tests</th>
<th>Case Appeared in Retention Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Lichen sclerosus et atrophicus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Blister above scar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Erythema multiforme, Mucha-Habermann disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>Erythema multiforme, Mucha-Habermann disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Bullous pemphigoid, herpes gestationis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Epidermolysis bullosa acquired, dermatitis herpetiformis, dermatitis herpetiformis-like drug eruption, linear IgA dermatosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Septic vasculitis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>Dermatitis herpetiformis, dermatitis herpetiformis-like drug eruption, linear IgA dermatosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Spider bite</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>Epidermolysis bullosa acquired, bullous pemphigoid, herpes gestationis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Porphyria cutanea tarda, variegate porphyria, erythropoietic protoporphyrria, hepatoerythropoietic protoporphyrria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>Epidermolysis bullosa dermolytic, Epidermolysis bullosa junctional, epidermolysis bullosa acquired, Bart's syndrome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Blister above scar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Septic vasculitis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Epidermolysis bullosa acquired, dermatitis herpetiformis, dermatitis herpetiformis-like drug eruption, linear IgA dermatosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>Dermatitis herpetiformis, dermatitis herpetiformis-like drug eruption, linear IgA dermatosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Bullous pemphigoid, herpes gestationis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Lichen sclerosus et atrophicus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Porphyria cutanea tarda, variegate porphyria, erythropoietic protoporphyrria, hepatoerythropoietic protoporphyrria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Dermatitis herpetiformis, dermatitis herpetiformis-like drug eruption, linear IgA dermatosis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
edge space. The problem representation is algorithmic—the diagnostic tree unfolds as the student works through the problem. Features are displayed as square boxes in the diagnostic path. Hypotheses appear at the end of the algorithm as rounded boxes. Features and hypotheses in the correct path are shown in yellow to distinguish them from the rest of the algorithm. When students complete any level of the algorithm by identifying and qualifying all features at that level, the rest of that level opens up so that students can see the other possible problem states across all cases. Steps that should have been completed earlier are displayed as boxes containing a question-mark icon (“?”). When the hypothesis fits with the current evidence it is shown connected to the current path. When the hypothesis does not fit with the current evidence, it is shown connected to other paths with the content of the associated features and attributes hidden as boxes containing the question mark icon (“?”) until students specifically request the identity of the feature or quality. A pointer is always present to provide a cue to the best-next-step. By the conclusion of problem solving the entire diagnostic tree is available for exploration. The knowledge-focused interface therefore expresses relationships between features and hypotheses both within and across cases. At the end of each case, the diagram shows the same algorithm for all cases, but highlights the evidence and diagnosis relevant to the current case.

**Participants**
The study was approved by the University of Pittsburgh Institutional Review Board prior to subject recruitment (IRB Protocol #020348). Twenty-one pathology residents from two academic pathology residency programs were recruited by e-mail. Respondents with practice experience prior to entering a US residency were excluded from the study. Subjects were assigned to one of two interface conditions, dependent on post-graduate year, in order to control for mean level of training across both conditions. Subjects were paid for participating.

**Participant Survey**
All participants completed a three-section survey at the conclusion of the first day. Section 1 contained 25 statements related to use of the tutoring system, including enjoyment, future-use, ease of use, specific system features, comparison to alternative methods of learning, self-assessment of learning, and trust in content. Responses were rated on a 4-point scale of agreement. Question polarities were varied. Sections 2 and 3 reproduced standardized instruments for measuring computer use and computer knowledge.

**Learning Assessments**
Pretest, post-test, and retention-test were computer-based. All assessments were developed by the research group in close collaboration with an attending dermatopathologist (DJ). Each test consisted of two sections:

1. **Case Diagnosis Test.** Participants inspected eight unknown virtual slides, and entered (1) a diagnosis or differential diagnosis, and (2) a justification for their answer. Cases used in the assessments had one of three possible relationships to those used in the tutoring session (Figure 1):
   a. **Untutored Patterns.** The tested pattern was not seen during the tutoring session.
Table 3 • Pretest, Post-test and Retention Test Scores

<table>
<thead>
<tr>
<th>Condition</th>
<th>Multiple Choice</th>
<th>Tutored Patterns</th>
<th>Untutored Patterns</th>
<th>Multiple Choice</th>
<th>Tutored Patterns</th>
<th>Untutored Patterns</th>
<th>Multiple Choice</th>
<th>Tutored Patterns</th>
<th>Untutored Patterns</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined</td>
<td>53 ± 13</td>
<td>12 ± 9</td>
<td>12 ± 8</td>
<td>77 ± 10</td>
<td>50 ± 23</td>
<td>10 ± 7</td>
<td>74 ± 13</td>
<td>48 ± 21</td>
<td>34 ± 17</td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td>53 ± 13</td>
<td>12 ± 8</td>
<td>12 ± 9</td>
<td>76 ± 11</td>
<td>51 ± 25</td>
<td>8 ± 6</td>
<td>75 ± 14</td>
<td>45 ± 22</td>
<td>28 ± 19</td>
<td></td>
</tr>
<tr>
<td>KF</td>
<td>53 ± 13</td>
<td>13 ± 10</td>
<td>12 ± 7</td>
<td>79 ± 11</td>
<td>50 ± 21</td>
<td>13 ± 7</td>
<td>74 ± 14</td>
<td>51 ± 20</td>
<td>41 ± 13</td>
<td></td>
</tr>
</tbody>
</table>

MANOVA, repeated contrast, $F=78.0$, $p<0.001$; MANOVA, repeated contrast, $F=3.52$, $p=0.08$; MANOVA, repeated contrast, $F=64.01$, $p<0.001$; MANOVA, repeated contrast, $F=0.25$, $p<0.62$; MANOVA, repeated contrast, $F=0.89$, $p=0.36$

b. Tutored Patterns. The tested case was not seen during the tutoring session, but the pattern (combination of evidence) represented by the tested case was seen during the tutoring session in a different case.

c. Tutored Cases. The tested case was seen during the tutoring session.

2. Multiple Choice Test. Participants completed a 51-item multiple choice test, which evaluated ability to locate features, identify features, qualify and quantify features, relate evidence to hypotheses, and select features distinguishing between hypotheses. The test covered the entire domain of subepidermal vesicular dermatitides, including material not covered in the tutoring session.

The pretest and post-test were identical. They contained the same multiple choice test and eight unknown virtual slides, including four untutored patterns and four tutored patterns. The retention test was different than the pretest and posttest. It contained a modified version of the multiple-choice test, in which questions were re-worded and re-ordered. The case diagnosis section of the retention test included four tutored patterns and four tutored cases, none of which overlapped with the pretest or post-test. No feedback was provided on test performance.

Scoring of Learning Assessments

Multiple choice questions were scored as correct or incorrect, based on answers provided by an attending dermatopathologist.

For each case in the case diagnosis test, we determined separate scores for the diagnosis and justification components. For the diagnosis component, we added 5 points for the first correct diagnosis, added 3 points for each additional correct diagnosis (when the case required a differential diagnosis), and subtracted 1 point for each incorrect diagnosis. No points were assigned for blank answers. For the justification component, we added 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.

Metacognitive Measures

After each case in all case diagnosis tests, participants scored their certainty about whether the diagnosis was correct, using a four point Likert scale.

Tutoring Process Measures

For all tutoring sessions we collected and stored detailed interaction data to the level of mouse clicks, and menu selections, but also at the level of correct actions, errors, and hint requests. Methods for data collection have been previously described. Data were used to construct learning curves and examine process differences between representation conditions and across other study variables.

Analysis

Performance on assessments was analyzed using repeated measures ANOVA. Main effects and interactions were analyzed for test, interface condition, and level of training, including repeated contrasts. For performance-certainty correlations and survey results, we used student’s t-test to compare between conditions. All analyses were performed in SPSS.

Results

Task Metrics

All students completed the entire 4 hour session. There were no significant differences between conditions for total number of cases completed during the working period. Eighteen of twenty-one students saw all twenty cases during the tutoring sessions.

Learning Outcomes

In both conditions, performance improved significantly at post-test, for both multiple-choice and case diagnosis tests for tutored pattern (Table 3), and learning gains were entirely retained at retention test.

Mean scores on the multiple-choice test increased from 52.8% on pretest to 77.0% on post-test (MANOVA, effect of test, $F=78.0$, $p<0.001$). On the case diagnosis test mean scores increased for tutored patterns, from 11.7% on pretest to 50.2% on post-test (MANOVA, effect of test, $F=64.0$, $p<0.001$). There was no improvement for untutored patterns. Learning gains were entirely retained at retention test one week later, with no significant differences when compared to post-test.

There were no significant differences between interfaces for learning gains, at post-test or retention test. One of the academic programs scored higher on all tests than the other program; however, the size of learning gains was the same across programs. The same effect was true of participants who had a previous dermatopathology rotation—these participants scored higher on each of the assessments, but the size of their learning gains did not differ from those who did not have a previous dermatopathology rotation. Learning gains also did not correlate with level of post-graduate training, pretest performance, computer experience, or computer knowledge.
### Metacognitive Measures

Certainty-performance correlations measure students’ ability to assess their own knowledge and abilities. Both conditions had significant positive certainty-performance correlations on post-test and retention test. Additionally, there was a significant positive correlation between change in case diagnosis score and change in certainty from pretest to post-test ($r = 0.46$, $p = 0.04$), indicating that as performance increased, so did participants’ certainty in their diagnosis.

Regression analysis was performed, computing the slope between certainty and performance for each test for each condition. The closer the slope is to 1, the more certainty and performance are correlated. This analysis revealed an advantage for the knowledge-focused condition: certainty-performance slope significantly improved from pretest to post-test ($p = 0.05$), and pretest to retention test ($p = 0.01$). This effect was seen only in the knowledge-focused condition.

### Participant Survey

Survey results showed higher user ratings for the knowledge-focused interface (Figure 3). Overall attitude scores were significantly higher for the knowledge-focused interface ($t = 2.66$, $p = 0.02$). In particular, participants using the knowledge-focused interface felt they were more likely to use it in the future ($p = 0.001$), and felt it was more enjoyable ($p = 0.05$). There were no correlations between attitude score and learning outcomes for either condition.

### Process Measures

Tutor interaction data were analyzed in fine-grained steps. Analysis focused on the level of correct actions, errors, and hint requests. We examined where users requested hints, what kinds of errors were common, and how hint usage and errors changed over time. Hint and error rates were computed as percent out of the total number of actions. Hint and error rate slopes were also computed as an indicator of how much hints and errors dropped over the course of the tutoring session.

Hint usage stayed fairly levels over the course of the tutoring session in both conditions (Figure 4). Errors dropped during tutoring, with a high number of errors occurring while working through the first case, then declining rapidly. For both conditions, there was a significant negative correlation
between hint and error rate—the more users relied on hints, the fewer errors they made.

Error and hint rate slopes did correlate with test performance for the case-focused condition. There was a significant negative correlation between slope and learning gains from pretest to post-test on the rationale score in the case diagnosis section, indicating that the more hint usage dropped over time (the more negative the hint/error rate slope), the more users learned. Interestingly, we saw no such correlations between slope and learning gains for the knowledge-focused condition.

**Discussion**

**Significance**

This is the first extensive evaluation of cognitive tutoring in a medical domain. Large learning gains were measured with multiple metrics including direct testing of diagnostic accuracy. Learning gains from pretest to post-test cannot be explained by differences in test difficulty, because pretest and post-test were identical. The absence of any improvement for untutored pattern cases indicates that learning gains were not related to a re-testing phenomenon.

Learning gains persisted unchanged at one week post-intervention. The tutored patterns encountered in the retention test were entirely novel cases that the students had never seen before, strengthening our conclusion that tutoring resulted in deep and sustained learning of the diagnostic patterns.

Students using case-focused and knowledge-focused problem representations achieved the same learning gains for tutored patterns, and neither group showed significant learning of untutored patterns. Students in both conditions showed improvements in their ability to correlate self-assessment to actual diagnostic performance on the diagnostic case tests. But only students in the knowledge-focused condition had a significant increase in the slope of this correlation from pretest to post-test. Thus, measures of metacognitive gains point to an advantage for the knowledge-focused interface, which requires further study. In other research by our group, we have seen absolutely no increase in metacognitive skills when students used a natural language interface that offered no visual affordance of the problem-space (unpublished data).

Why might the knowledge-focused graphical interface contribute more towards developing an appropriate “feeling-of-knowing”? From the standpoint of current theory on external representations, the knowledge-focused and case-focused interface we created differ along four important dimensions (Table 1). The knowledge-focused interface provides a more holistic, global view of the problem because it repeatedly depicts the entire knowledge base. Students may benefit from the retrieval-structure aspects of the knowledge-focused diagram, because only one retrieval structure must be traversed. Another effect of the knowledge-focused representation is that it may encourage students to see the effect of subtle differences that specific feature sets have on diagnosis, leading to improved self-assessment because students can visualize diagnostic “near-misses.” The more temporal and hierarchical representation of problem-solving in the knowledge-focused interface may further enhance this effect. On the other hand, the case-focused interface has the advantage of presenting less information for any given case (Table 1), thus reducing the cognitive load. Results of our study suggest that the effect of the unified, temporal, and hierarchical knowledge-focused interface outweigh the increased information density associated with it for this group of users.

Student ratings pointed to a second advantage of the knowledge-focused interface. Survey results indicate that students rate the knowledge-focused interface higher, feeling that it is more enjoyable to use and they are more likely to use it in the future.

Results from detailed process measures show differences between interfaces for correlations between hints, errors, and performance. The apparent absence of typical ITS process correlations for the knowledge-focused interface may relate to the more transparent knowledge-base provided by the knowledge-focused interface. Students were able to observe and learn on their own from the more holistic representation, and therefore relied less on hints. The findings have implications for student modeling as we further develop this system.

**Limitations of the Present Study**

Limitations to our work include:

1. The sample size (N=21) is relatively small, and the subject pool drew from only two residency programs. Therefore the findings may not be representative of larger, more distributed populations.

2. This study did not test the cognitive tutoring paradigm against standard practice. Although we can conclude that cognitive tutoring produces strong and sustained effects on learning gains, we cannot make any statement regarding the relative benefit of cognitive tutoring in medical domains against other methods of training.

**Implications for Future Work**

This work highlights the potential for knowledge-based cognitive tutoring in complex medical domains. Our findings also suggest that knowledge-focused external representations may provide a metacognitive advantage and may be more easily accepted by students. Further work is needed to explore the effects of external representation on metacognitive skills.

The study is the first to demonstrate a strong and sustained effect of cognitive tutoring on diagnostic performance in a medical task. Further research is needed to reproduce and verify these results and to compare effects of cognitive tutoring against existing and alternative training methods.

**Conclusions**

This study demonstrates that use of a cognitive tutoring system is associated with improved diagnostic performance in a complex medical domain. The effect is retained at one-week post-training. Both case-focused and knowledge-focused external representations are associated with equivalent learning gains. Cognitive tutoring improves metacognitive ability of students to gauge when they know and do not know the correct diagnosis, but the size of the effect may relate to the type of external representation used in tutoring. Students differed in their acceptance of the tutoring systems, with the knowledge-focused interface rated significantly higher over-
References

24. CROWLEY ET AL., Evaluation of an ITS in Pathology