In Proceedings of the International Conference of the Learning Sciences (1998),
(A.S. Bruckman, M. Guzdial, J.L. Kolodner, & A. Ram, Eds.). Charlottesville, VA: AACE.

Using *Convince Me* to Assess Medical Reasoning Skills (and Vice Versa)

Jeanne Weidner*

Michael Ranney*

Alan Steinbach**

*Graduate School of Education, E-mails: jweidner@socrates.berkeley.edu, ranney@soe.berkeley.edu
** Joint Medical Program, E-mail: alanburr@socrates.berkeley.edu
University of California at Berkeley, 94720, USA

Abstract: We investigated medical students' reasoning about a neurophysiology problem under two elicitation methods. In one group, subjects verbalized about a problematic patient case during a think-aloud protocol (TAP) session. The second group used the *Convince Me* (*CM*) software system to reason about the case. All subjects solved the problem and appropriately recognized irrelevant evidence. Although *CM* subjects took longer to finish, their arguments seem more sophisticated, exhibiting trends toward (1) more alternative solution hypotheses, (2) more new, auxiliary, evidence to support their hypotheses, (3) more contradictory evidence and relationships, and (4) more relationships per proposition. *Convince Me*'s Model's Fit feature further indicated their arguments to be reasonably coherent. TAP subjects, in contrast, cited more total evidence, were more likely to draw a neurophysiological diagram, and rated their method more highly than did *CM* subjects. Findings are discussed regarding (a) past results, (b) usability issues, and (c) improving the match between methodology and instructional milieu.

1. Introduction

If a physician cannot apply his/her medical mastery in clinical settings, such a knowledge base would have little utility. This is one reason why evaluating medical reasoning skills holds great interest for medical educators [Arocha & Patel 1995]; [Rosenberg & Sackett 1996]; [Elstein et. al. 1985]. Still, we lack precise, direct, objective, measures of medical reasoning. One *indirect* measure is the Objective Structured Clinical Examination (OSCE) [Tervo et. al. 1997]. However, while an effective way to evaluate medical students in simulated clinical settings, OSCE is usually costly in time and finances [Poenaru, Morales, Richards & O'Connor 1997]. Of course, OSCE only *approximates* one's ability–and offers no opportunity to improve clinical or reasoning skills [Mavis, Henry, Ogle & Hoppe 1996].

A more common approach to assessing such skills is through more informal question-and-answer sessions during clinical rounds. This method is analogous to the use of think-aloud protocols (TAP), in which subjects vocalize their thoughts as they solve a problem [Ericsson & Simon 1984]. Clinical questioning is fraught with difficulty, especially due to the evaluators' inherent subjectivity. Naturally, it requires clinicians to make many inferences about students' reasoning skills, and penalizes those with poor communication skills. Further, when the problem solver uses complex processes to transform information into oral forms, cognitive load problems may result, leading to fewer verbalizations. In addition, anxieties due to the extemporaneous, public, nature of the process may interfere with one's reasoning, yielding underestimates of the student's skills. Such assessments are often eventually done with rating forms, which have serious limitations [Stenchever et. al. 1979]; [Tonesk & Buchanan 1987]; [Metheny 1991].

Given the drawbacks of oral accounts for evaluating reasoning skills, we sought to investigate an alternate method–one designed, in part, to compensate for the shortcomings of the verbalization process and/or to provide additional support for its utility. There has been renewed interest in how computers can support the assessment of medical reasoning [Reisman 1996]; [Mackel et. al. 1995]; [Murphy et. al. 1996]; [Kleinmuntz & Elstein 1987]. We chose to investigate a computational system, called *Convince Me (CM)*, that usually incorporates a brief reasoning curriculum.

Developed by Schank, Ranney, Hoadley [Schank, Ranney & Hoadley 1995] and others as a "reasoner's workbench," the *Convince Me* program has users generate propositions, categorize them as hypotheses and evidence, determine which explain or contradict which others, and assign a rating of believability (and for evidence, reliability) to each. The structure of the information entered is used by the underlying simulation (ECHO) to both evaluate the user's arguments and give feedback about the coherence of the user's reasoning. ECHO is a connectionist model of coherence, based on the Theory of Explanatory Coherence (TEC) [Ranney & Thagard 1988]; [Thagard 1989], which consists of principles that establish coherence among hypotheses and evidence. For example, *Convince Me* assumes that the credibility of a proposition, *ceteris paribus*, increases with 1) its reliability (if evidence), 2) the simplicity with which it is explained, 3) its coherence with plausible propositions, and 4) its contradiction with implausible propositions. *Convince Me* analyzes such information with ECHO to evaluate the user's argument and assign "activations," the computer model's relative acceptance/rejection value for each statement. The correlation between ECHO's activations and a user's elicited believability ratings is calculated. The student can receive feedback about this correlation, called a "Model's Fit," that is an estimate of the coherence of the user's reasoning with the avowed beliefs and their embedding structure. The software also provides diagrammatic and other representations of the propositions' interrelationships that enable better visualization of the structure of one's argument [Fig. 1].

Convince Me helps students develop more coherent arguments that seem to include a more appropriate amount of the relevant information, when compared to paper-and-pencil arguments [Schank 1995]. We believe it to be the only system that explicitly supports students' reasoning processes while a computational model offers them plausibility feedback based on their arguments' coherence. *Convince Me* has been used as a tool to enhance critical reasoning skills in several domains [Ranney, Schank & Diehl 1995], but it was not, heretofore, applied significantly to the field of medical reasoning.

Although typing in arguments is a time-intensive process, there are several likely advantages to using *Convince Me* as a medical reasoning tool. It provides quantitative and qualitative data, enhancing method triangulation and convergence–especially when combined with verbal reports and problem-solving ability measures. Since users identify each statement as evidence or hypothesis, the system also improves epistemic categorizations and their uses Ranney, Schank, Hoadley & Neff 1996]. It utilizes a replicable model that reduces the subjectivity inherent in verbal assessments. It also gives students desirably immediate feedback, and thus may further enhance problem-solving skills. The Model's Fit feedback addresses the coherence of a subject's reasoning process and encourages the user to reflect on and improve his/her ability to identify and contrast relevant information. *Convince Me* also serves as a multi modal external memory device, promoting the organization of knowledge in readily accessible, structured, formats. The feedback and topological information help reduce a student's cognitive load during problem solving.

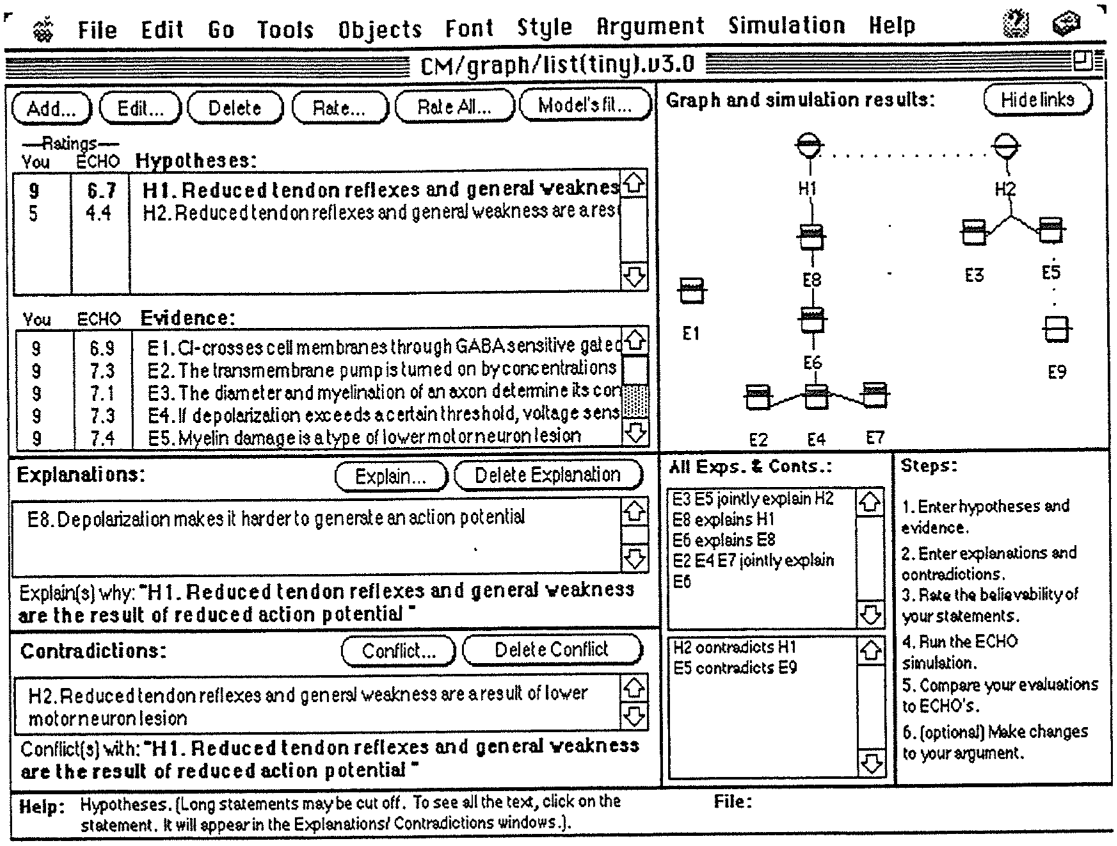
The program has utilities as a pedagogic, remedial, and/or evaluative tool, as students can either build arguments *de novo*, or edit, evaluate, and rearrange propositions entered by an instructor. Its built-in logging functions time-stamp and record each action, helping to evaluate the user's reasoning process. If *Convince Me* continues to be an effective tool for eliciting and evaluating subjects' thinking processes, it may be an efficient replacement for, or a valuable supplement to, think-aloud protocols. As a result, we conducted an exploratory study to determine if *CM* can provide valuable measures of medical reasoning and problem-solving skills.

2. Experimental Method

Rationale. We compared *Convince Me* to think-aloud protocols, as elicitation methods, in the solving of a neurophysiology problem. We sought to observe (a) differences in subjects' reasoning processes across the two methods, (b) which method better supports the assessment of reasoning skills, and (c) how well relevant and irrelevant information are distinguished during problem solving. As alluded to above, our hypothesis was that *Convince Me* would provide equally or more informative measures of a subject's problem-solving ability than would think-aloud protocols, due to its Model's Fit feedback, structured approach to argument development, and its requirement that subjects identify each proposition as hypothesis or evidence (and rate its believability).

Subjects. All twelve of the dozen first-year medical students, eight women and four men, from the University of California's (Berkeley/San Francisco) Joint Medical Program participated in this study. The study was conducted as part of the midterm evaluation for a Human Neurobiology course.

Design. Subjects read a brief patient case in neurobiology, and were to determine the physiological mechanisms accounting for the ailments. Six students were randomly assigned to use *Convince Me* to develop hypotheses about the patient's problem and link them to evidence (e.g., from the provided case). The other six reasoned about the problem during a think-aloud protocol session, which was audio taped, transcribed and coded. Both groups then



—Ratings—

You	ECHO	Hypotheses:
9	6.7	**H1. Reduced tendon reflexes and general weaknes**
5	4.4	H2. Reduced tendon reflexes and general weakness are a res

You	ECHO	Evidence:
9	6.9	E1. Cl-crosses cell membranes through GABA sensitive gate
9	7.3	E2. The transmembrane pump is turned on by concentrations
9	7.1	E3. The diameter and myelination of an axon determine its co
9	7.3	E4. If depolarization exceeds a certain threshold, voltage sens
9	7.4	E5. Myelin damage is a type of lower motor neuron lesion

Explanations:

E8. Depolarization makes it harder to generate an action potential

Explain(s) why: "**H1. Reduced tendon reflexes and general weakness are the result of reduced action potential** "

Contradictions:

H2. Reduced tendon reflexes and general weakness are a result of lower motor neuron lesion

Conflict(s) with: "**H1. Reduced tendon reflexes and general weakness are the result of reduced action potential** "

All Exps. & Conts.:

E3 E5 jointly explain H2
E8 explains H1
E6 explains E8
E2 E4 E7 jointly explain E6

H2 contradicts H1
E5 contradicts E9

Steps:

1. Enter hypotheses and evidence.
2. Enter explanations and contradictions.
3. Rate the believability of your statements.
4. Run the ECHO simulation.
5. Compare your evaluations to ECHO's.
6. (optional) Make changes to your argument.

Help: Hypotheses. (Long statements may be cut off. To see all the text, click on the statement. It will appear in the Explanations/ Contradictions windows.).

File:

Figure 1: Sample *Convince Me* argument. Propositions E1-E4 were entered by the investigators, with E1 and E3 considered irrelevant to the solution. This subject used E3 to support the alternative hypothesis.

completed a survey that elicited their views on the effectiveness of their particular method in supporting their reasoning processes [Tab. 1].

Procedure. Each subject received the same problem and accompanying review material. To control for differences in memory and knowledge, the material included both relevant and irrelevant information regarding the problem. The *Convince Me* group further received a diskette of the program and its two pages of instructions. The *CM* program pre-included four propositions (tentatively entered as evidence by the investigators)–two deemed relevant to the solution and two deemed irrelevant [Fig. 1]. Think-aloud subjects were given instructions for the TAP procedure and a simple practice problem in math prior their medical problem solving. TAP transcripts were coded by one of the authors to identify hypotheses, evidence and their relationships in the arguments. Four months later (after the summer), eleven of the twelve subjects responded to a follow-up survey, designed to explore some initial findings in more depth.

3. Results

No differences between the two groups regarding age, gender, or MCAT scores approached significance. The TAP sextet had a marginally higher self-reported mean GPA, although both groups' GPA's were quite high. An unsurprising difference was that the *Convince Me* (*CM*) subjects took (approximately three times) longer to construct their arguments than the TAP subjects (58.5 vs.19.2 min; $p < 0.001$; see [Tab. 1]). Although expected (due to the cognitive overhead of understanding and using the *CM* system), and seemingly unrelated to the arguments themselves, the effect may have influenced other variables, specifically those related to the software evaluation.

Several significant effects and trends indicate differences between the two conditions' argument structures. Statistically significant differences were fewer than we had anticipated, partly due both to the small subject population and the challenges inherent in TAP analysis.

#	Statement	TAP*	CM*	*F*	*p*
1	The think aloud process (program)** was useful in helping me to think about this case	4.33	2.83	8.27	0.017
2	Reasoning about this case could have been more (just as) easily done on paper.	2.83	4.50	8.06	0.018
3	The case was unusually complex or difficult.	1.33	1.67	1.25	0.289
4	The protocol (program) was easy to follow.	4.83	2.00	49.83	0.000
5	The instructions (program interface) were (was) confusing.	1.17	3.50	15.81	0.003
6	Thinking out loud (the program) would be useful in reasoning about other basic science cases.	4.33	3.00	7.27	0.022
7	I used the "show links" feature to view the relationships between my statements.	[CM group only]	3.50	[CM group only]	
8	The "show links" feature was helpful in visualizing the relationships between my statements.		2.67		
9	Thinking out loud (the program) would be useful to help develop my reasoning skills .	4.67	2.33	18.84	0.002
10	I was able to discover areas of weakness in my argument by thinking out loud (using the program).	4.67	2.67	18.00	0.002
11	I drew diagrams on paper to help me solve the problem.	4.33	1.67	21.23	0.001
12	Thinking out loud (the program) would be helpful in reasoning about clinical diagnoses	4.83	2.33	18.44	0.002

*1 = Strongly Disagree, 2 = Disagree, 3=Unsure, 4 = Agree, and 5 = Strongly Agree; **Parentheses show wording for *Convince Me* subjects

Table 1: Program/Protocol Initial Survey

3.1 Argument Structure Measures Tend to Favor *Convince Me*

There were no significant differences between the groups regarding total numbers of hypotheses, evidence, or links generated by the subjects. TAP subjects cited more evidence from the problem statement *(*5.7 vs. 2.7; *p< 0.05;* see [Tab. 2]), but the transcript analyses show that they did not usually relate those statements in their arguments. In most cases, their references to problem evidence were simply verbal restatements of given information. Nearly half of the TAP propositions were not connected to the argument structure, as indicated by a rate of only .56 relations/proposition, whereas *CM* subjects used almost as many links as nodes (a .87 to 1 ratio).

Four of the six *CM* subjects offered an alternative hypothesis for the patient's primary problem (of muscle weakness), yet only two TAP subjects did so–even after specific prompting at the protocol's end. TAP subjects usually just reiterated their first model and either cited additional supporting evidence or explored reasons for the patient's secondary problem (of hypokalemia). In contrast, *CM* subjects used a similar number of propositions to argue for their alternative hypotheses, and rated these competing hypotheses with reasonably similar, albeit lower, believability ratings.

This differential approach to alternatives stands in stark contrast to the uniformity of opinions during the follow-up survey. In it, all subjects agreed (with 73% strongly agreeing) that generating alternative hypotheses is important when considering a diagnosis of a patient problem, while 82% agreed that being able to write down one's thoughts facilitates this process. Another stark contrast was that *CM* subjects cited ten-fold as many contradictions in their argument as TAP students (3.7 vs. 0.3; *p*=*.11*; see [Tab. 2]), yet all subjects agreed on the follow-up survey that considering contradictory evidence is important–and only 27% agreed that it is difficult to generate such evidence.

3.2 Relevance, Evidential Support, and Diagram Use in Argumentation

No subject in either group used irrelevant information to support his/her favored hypothesis about the patient's primary problem. Most of the irrelevant information that the investigator included as evidence in the *CM* file was deleted early in the arguments that they developed. Two *CM* subjects retained one of the irrelevant propositions to explain alternative hypotheses. Hence, both groups generally recognized irrelevant evidence as such.

CM subjects were three times more likely to generate new auxiliary evidence (3.7 vs. 1.2; *p* = *0.06*), usually to support alternative hypotheses. Although non-significant, a trend indicated that *CM* subjects numerically generated 50% more links among their propositions (8.84 vs. 5.83). TAP subjects cited more relevant evidence in constructing their arguments *(6.5* vs. 3.2; *p<0.005)*, but this proportionally reflects the aforementioned finding that they also cited more evidence, in general, from the problem case. It may also reflect the possibility that *CM* subjects are often more sophisticatedly parsimonious in constructing their arguments [Schank 1995].

Five of the six TAP subjects drew physical diagrams–versus none of the *CM* sextet *(p=0.001)*. In the follow-up survey, all students agreed that they reason better if a problem can be visualized on paper. Of the *CM* subjects, three did not realize that drawing a diagram was permitted, one drew a diagram but did not turn it in, and two agreed that typing in propositions allowed them time to visualize the mechanism without having to draw it. *CM's* argument representation may thus serve as an external memory device, obviating the need to draw (e.g., cellular) diagrams; still, in follow-ups, subjects generally could not recall enough about that feature to comment on its utility.

3.3 Argument Coherence

The mean "Model's Fit" for the *CM* group–the mean of the correlations of ECHO's activations with each user's believability ratings–was 0.5. This value is called "moderately related" for subjects, and indicates that, on average, *CM* subjects constructed reasonably coherent arguments vis-a-vis their beliefs.

3.4 Program/Protocol Survey Results Favor TAP

Nine of the ten common questions about the methods showed statistically significant differences favoring the TAP method, see [Tab. 1]. The one that did not concerned the problem case's difficulty; on average, neither group felt that the case was unusually complex or difficult. TAP subjects were more positive about their method than *CM* subjects were about theirs. This finding may be partially related to the problem's simplicity, as students using *Convince Me* in a previous pilot study found the features of the program more useful as the complexity of their argument increased [Diehl 1995]. In our study, all *CM* subjects either disagreed that the program was easy to use or were unsure about this, which may have influenced their ratings of the program's utility for reasoning about cases or for developing reasoning skills. In contrast to previous *CM* studies that provided subjects with detailed verbal training on its use, a pilot study indicated it may be unnecessary for subjects of this academic level. In retrospect, though, more instruction seems warranted, as one subject made multi-sentence (e.g., seventy-five word) entries for several propositions, and one subject requested clarification of the instructions. Also, it is possible that (inferred) experimenter demand may have differentially influenced the outcome of the survey results, since all TAP subjects opted to take the survey in the investigator's presence. This lack of anonymity may account for parts of the favorable ratings credited to the TAP. As is not uncommon, though, students in both groups overwhelming (91%) agreed with the follow-up survey statement that they preferred interacting with another human being compared to interacting with a computer. Finally, as *CM* subjects ran *Convince Me* from a diskette (for data-security), the much slower speed may have yielded a negative-halo effect regarding its perceived usability and utility.

Criteria	TAP	CM	*F*	*p* =
Time spent to complete (minutes)	19.17	58.50	24.71	0.001
Hypotheses: Total	1.33	1.67	1.25	0.290
Proposed Models	1.00	1.00	0.00	1.000
Alternative Models	0.33	0.67	1.25	0.290
Evidence: Total	10.50	10.18	0.08	0.781
Relevant to Primary Problem	6.50	3.17	14.92	0.003
Relevant to Secondary Problem	2.83	3.34	0.92	0.360
Auxiliary (Newly Generated)	1.17	3.67	4.45	0.061
[From Problem Case]	[5.67]	[2.67]	[7.79]	[0.019]
Links: Total	5.83	8.84	1.34	0.274
Independent Explanation	3.50	2.67	0.38	0.551
Joint Explanation	2.00	2.50	0.65	0.438
Contradiction	0.33	3.67	3.01	0.113

Table 2: Comparison of Arguments

4. Discussion

This study yielded both promise and limitations regarding applications of *Convince Me* to medical reasoning. Compared to TAP, there were several indications of its relative strengths in terms of argument structure. However, *CM* subjects were less enamored of their too briefly-trained method than were TAP subjects–at least as configured in this experiment. An important factor in this latter finding is the time *CM* subjects took to complete the exercise. As stated above, *CM* subjects worked on the problem about three times as long as did the TAP subjects, partly due to diskette drive inefficiency, typing in propositions, and possibly re-reading material. However, the follow-up survey showed that all *CM* students agreed that they would have rated *Convince Me's* utility more highly if it ran faster and were easier to use, issues that are being addressed in the updated, Web-based, version of the program currently being completed. Even so, *Convince Me* may require speech-recognition capabilities to obviate time/usability concerns (e.g., compared to the transparency of good word-processing software). The follow-up survey also revealed *CM* subjects to be more favorably inclined to use the program to reason about clinical problems that involve deciding among several harder-to-discriminate diagnoses, as opposed to using the program to reason about physiological mechanisms (as in this study). So, some of the reasons for the observed differences may be remedied by using more difficult problems and an enhanced program and method (although TAP will likely retain some temporal advantage in the near-term technological milieu).

Several interesting trends warrant further study. As *CM* subjects appeared a bit more likely to propose new/auxiliary evidence and generate more alternative hypotheses, using *Convince Me* seems to encourage a user to explore more alternatives when reasoning about a problem. In this same vein, *CM* subjects' arguments seem more likely to include competitive and contradictory information. These findings may reflect how *Convince Me* explicitly provides users with the opportunity to consider contradictory evidence when constructing an argument, something most people rarely do well on their own. Such trends and effects are consistent with past findings that *Convince Me* fosters more elaborate and coherent reasoning through the use of its structured approach [Schank 1995]. Its interface and structure encourages one to consider alternate hypotheses and contradictory evidence, and thus may have contributed to the likelihood of *CM* subjects generating more propositions to support these alternatives.

Recall that TAP subjects were more likely to draw physiological diagrams in solving the problem. This supports the idea that the topographical argument representation provided by *Convince Me* may obviate the need for a second external representation to assist in reasoning about mechanistic problems, as explicitly indicated by two *CM* subjects. This is further supported by *CM* subjects' lack of enthusiasm about a proposed feature (in the follow-up survey) that would allow one to draw a physical diagram. Another informational advantage, the ability to evaluate the coherence of an argument by means of the Model's Fit, is a *Convince Me* feature that is of notable potential benefit in both the learning and assessment of reasoning skills [Schank 1995]. No true counterpart is available for an on-line judging of the relative coherence of the TAP subjects' arguments.

In order to more fully and/or convincingly elucidate some of the potential differences between the two methods, future studies should use more difficult problems and include more irrelevant (and potentially misleading) information. Also, since our subjects did not find this problem difficult, the irrelevant information did not represent attractive foils. Further, as in all empirical studies, larger subject samples may of course reveal more differences between the elicitation methods (although we used all subjects available in the population). For these reasons as well, more definitive studies are required. In addition, more training on the concepts underlying TEC (e.g., on propositions, the nature of evidence and hypotheses, etc.) is warranted. Clearly a structural approach to reasoning takes more than a mean of 58.5 minutes to master, even if talking aloud about a problem takes only a third as much time. (Medical students are, of course, much more used to talking aloud than using a *CM*-like system).

In closing, possible *Convince Me* applications in the medical curriculum include its use as a complement to both group and individual problem solving processes. It may also assist students with sub-optimal verbal skills, or augment/replace traditional evaluation methods in Introduction to Clinical Medicine courses (which involve more subjectivity than do *CM*'s Model's Fit). Further, it can help specifically evaluate a student's recognition of irrelevant and/or contradictory data. *Convince Me* might also better integrate ethical aspects of clinical problems –those that can appear even less structured than ones involving only clinical data. These factors warrant further studies of the uses of *Convince Me* as both an outcome and process measurement system in evaluating medical reasoning skills. Finally, more attention to the system's ease of use (e.g., with speech recognition) and training demands (e.g., on basic argumentation constructs) may well allow its uniquely promising features to be more fully expressed.

5. References

[Arocha & Patel 1995] Arocha, J. F., & Patel, V. L. (1995). Novice diagnostic reasoning in medicine: Accounting for evidence. *Journal of the Learning Sciences, 4*, 355-384.

[Diehl 1995] Diehl, C.L. (1995). Pragmatic and conceptual attributes of representational tools influence students' reasoning strategies. Presented at the annual meeting of the American Psychological Society, New York, NY.

[Elstein et. al. 1985] Elstein, A. S., Dawson-Saunders, B., & Belzer, L.J. (1985). Instruction in medical decision making. A report of two surveys. *Medical Decision Making, 5*, 229-33.

[Ericsson & Simon 1984] Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge MA: MIT Press.

[Kleinmuntz & Elstein 1987] Kleinmuntz, B., & Elstein, A.S. (1987). Computer modeling of clinical judgment. *Critical Reviews in Medical Informatics, 1*, 209-228.

[Mackel et. al. 1995] Mackel, J., Farris, H., Mittman, B., Wilkes, M., & Kanouse, D. (1995). A Windows-based tool for the study of clinical decision-making. *Medinfo, 8*, 1687.

[Mavis, Henry, Ogle & Hoppe 1996] Mavis, B.E., Henry, R.C., Ogle, K.S., & Hoppe, R.B. (1996). The emperor's new clothes: the OSCE reassessed. *Academic Medicine, 71*, 447-453.

[Metheny 1991] Metheny, W.P. (1991). Limitations of physician ratings in the assessment of student clinical performance in an obstetrics and gynecology clerkship. *Obstetrics and Gynecology, 78*, 136-141.

[Murphy et. al. 1996] Murphy, G., Friedman, C.P., Elstein, A.S., Wolf, F.M., Miller, T., & Miller, J. (1996). The influence of a decision support system on the differential diagnosis of medical practitioners at three levels of training. *Proceedings of the AMIA Annual Fall Symposium* (pp. 219-223).

[Poenaru, Morales, Richards & O'Connor 1997] Poenaru, D., Morales, D., Richards, A., & O'Connor, H.M. (1997). Running an objective structured clinical examination on a shoestring budget. *American Journal of Surgery, 173*, 538-541.

[Ranney Schank & Diehl 1995] Ranney, M., Schank, P., & Diehl, C. (1995). Competence versus performance in critical reasoning: Reducing the gap by using Convince Me. *Psychology Teaching Review, 4*, 151-164.

[Ranney Schank, Hoadley & Neff 1996] Ranney, M., Schank, P., Hoadley, C., & Neff, J. (1996). "I know one when I see one": How (much) do hypotheses differ from evidence? In R. Fidel, B.H. Kwasnik, C. Beghtol, & P. Smith (Eds.), *Advances in classification research* (Vol. 5) (ASIS Monograph Series; pp. 141-158). Medford, NJ: Learned Information.

[Ranney & Thagard 1988] Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. In V.L. Patel & G.J. Groen (Eds.), *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 426-432). Hillsdale, NJ: Lawrence Erlbaum Associates.

[Reisman 1996] Reisman, Y. (1996). Computer-based clinical decision aids. A review of methods and assessment of systems. *Medical Informatics, 10*, 179-197.

[Rosenberg & Sackett 1996] Rosenberg, W. M., & Sackett, D.L. (1996). On the need for evidence-based medicine. Therapie, 51, 212-217.

[Schank 1995] Schank, P. (1995). *Computational tools for modeling and aiding reasoning: Assessing and applying the Theory of Explanatory Coherence.* Doctoral dissertation, University of California, Berkeley. (University Microfilms No. 9621352)

[Schank Ranney & Hoadley 1995] Schank, P., Ranney, M., & Hoadley, C. (1995). Convince Me (Computer Program and Manual). In J. Jungck, V. Vaughn, J. Calley, N. Peterson, P. Soderberg, and J. Stewart (Eds). *The BioQUEST Library.* College Park, MD: Academic Software Development Group, University of Maryland.

[Stenchever et. al. 1979] Stenchever, M.A., O'Toole, B., & Irby, D. (1979). Evaluating student performance in an obstetrics and gynecology clerkship. *American Journal of Obstetrics and Gynecology, 134*, 235-237.

[Tervo et. al. 1997] Tervo, R. C., Dimitrievich, E., Trujillo, A.L., Whittle, K., Redinius, P., & Wellman, L. (1997). The Objective Structured Clinical Examination (OSCE) in the clinical clerkship: an overview. *South Dakota Journal of Medicine, 50*, 153-156.

[Thagard 1989] Thagard, P. (1989). Explanatory coherence. *Behavioral & Brain Sciences, 12*, 435-502.

[Tonesk & Buchanan 1987] Tonesk, X., & Buchanan, R.G. (1987). An AAMC pilot study by 10 medical schools of clinical evaluation of students. *Journal of Medical Education, 62*, 707-718.

Acknowledgments

The authors thank the following for their valuable comments in the preparation of this manuscript: Patricia Schank, Michelle Million, and members of the Reasoning Research Group of the University of California, Berkeley.