

## Interactive Critiquing as a Form of Decision-Support: An Empirical Evaluation

Stephanie A. Guerlain, Philip J. Smith, Jodi Heinz Obradovich, Sally Rudmann, Patricia Strohm, Jack W. Smith, John Svrbely, and Larry Sachs, Cognitive Systems Engineering Laboratory, The Ohio State University, Columbus, Ohio

This research focused on the design of a decision support system to assist blood bankers in identifying alloantibodies in patients' blood. It was hypothesized that critiquing, a technique whereby the computer monitors human performance for errors, would be an effective role for such a decision-support system if the error monitoring was unobtrusive, and the critiquing was in response to both intermediate and final conclusions made by the user. A prototype critiquing system monitored medical technologists for: 1) errors of commission and errors of omission, 2) failure to follow a complete protocol, 3) answers inconsistent with the data collected, and 4) answers inconsistent with prior probability information. Subjects using the critiquing system had significantly better performance (completely eliminating misdiagnosis rates for 3 out of 4 test cases) than a comparable control group. Detailed analysis of the behavioral protocols provided insights into how specific design features influenced performance. Potential applications of this research include its use (after refinements) as a tool for routine use in blood banks.

- Portions of this work were completed as part of this author's dissertation at The Ohio State University. She is now at Honeywell Technology Center, Minneapolis, MN
- Requests for reprints should be sent to Phil Smith, Institute for Ergonomics, The Ohio State University, 210 Baker Systems, 1971 Neil Avenue, Columbus OH 43210-1271; Phone: 614-292-4120; Fax: 614-292-7852; Email: Phil+@osu.edu
- Now at Northside Medical Center, Youngstown, OH Running

Title | Interactive Critiquing

Key words | Critiquing, Expert Systems, Medicine, Blood Bank, Human Error

Introduction | In examining the artificial intelligence literature, one finds that there have been many attempts to build decision support systems that provide the user with a conclusion and an explanation for that conclusion. For example, in the medical domain, systems have been developed to aid with the diagnosis of diseases (e.g., MYCIN, Shortliffe, 1976 and MENINGE, François, et al., 1993), or with the management of a patient's treatment plan (e.g., ONCOCIN, Shortliffe et al., 1981). Many expert systems are knowledge-based, but some use mathematical models, such as Bayesian reasoning (Sutton, 1989). Almost all of these systems, however, follow the same model of decision support: The computer tries to solve the problem for the person and then gives its results (possibly along with an explanation) to the person for review. The user is then expected to critique that conclusion and decide whether he/she agrees with it.

Traditionally, evaluations of such systems focused on whether or not the computer system was able to generate the "gold standard" (i.e., best answer) as either the top answer or a highly rated answer on a range of cases (e.g., Bernelot Moens, 1992; Berner, et al., 1994; François, et al., 1993; Hickam et al., 1985; Nelson, Blois, Tuttle et al., 1985; Plugge, Verhey, and Jolles, 1990; Shamsolmaali, et al., 1989; Sutton, 1989; Verdager, et al., 1992; Wellwood, et al., 1992). Subsequently, however, researchers in medical informatics began to realize that focusing only on the computer's performance is a limited and unrealistic evaluation of a decision support system, if the goal is to successfully incorporate the system into actual practice (e.g., Forsythe and Buchanan, 1992; Miller and Maserie, 1990; Wyatt and Spiegelhalter, 1992). The human interface is almost always cited as a problem (e.g., Berner, Brooks, Miller et al., 1989; Harris and Owens, 1986; Miller, 1984; Shamsolmaali, et al., 1989; Shortliffe, 1990) particularly since most expert systems require that the practitioner enter data into the computer to have the information necessary to perform its reasoning. Thus, one requirement for a successful medical informatics system is to already have the necessary data on-line (Linnarson, 1993; Miller, 1984; Shortliffe, 1990).

A second usage problem is that these systems may have an incomplete knowledge base or use simplifying assumptions that make them brittle, meaning that they can fail on cases that the system was not designed to handle. This leaves the practitioner in the role of having to detect and correct any problems generated by faulty computer reasoning (Aikins, Kunz and Shortliffe, 1983; Andert, 1992; Bankowitz, et al., 1989; Bernard, 1989; Berner, et al., 1989; Gregory, 1986; Guerlain, Smith et al., 1994; Harris and Owens, 1986; Miller, 1984; Roth, et al., 1988; Sassen, et al., 1994). With this design model, the human must decide whether or not to accept the computer's diagnosis or treatment plan. A serious concern, however, is that the user in the role of critiquing the computer's answer may become overreliant, or alternatively may be lead "down a garden path," failing to adequately evaluate the computer's conclusion.

People may ignore the advice of a system, even when it is relevant, or heed the advice of a system, even when it is faulty. "Complacency" may occur when monitoring for automation failures if the automation reliability is unchanging or if the operator is responsible for more than one task (Parasuraman et al., 1993; Parasuraman et al., 1994). In addition, less than obvious failures may cause practitioners to be unduly influenced by an expert system's proposed solutions. This was demonstrated in a recent study in the domain of flight planning (Layton, Smith, et al., 1994). In this study, in a scenario where the computer's brittleness led to a poor recommendation, when the computer presented its suggestion early in the person's own problem evaluation, that person's cognitive processes were adversely biased, resulting in a 30 increase in the selection of a poor flight plan.

A second study by Guerlain, Smith et al. (1995) found similar results in a medical application. This study showed that when a problem-solving strategy is encoded into an expert system and its knowledge is applied automatically by the computer, performance can degrade significantly if the task situation is outside the computer's range of competence (the brittleness problem). Such a degradation, however, did not occur when the computer used its knowledge to critique the user who was performing the problem-solving task while the computer looked "over her shoulder." These results provided initial evidence that placing an expert system in a critiquing role may be a safer and more effective form of decision support than asking the person to critique the computer. The goal of the research reported here, then, was to examine in much greater detail the critiquing approach of decision support as a means of supporting human-computer cooperative problem solving.

**The Critiquing Model of Decision Support** | In a critiquing system, the computer system's model of expert performance is not used to try to solve the problem for the user, but instead to monitor the human's problem-solving performance for faulty reasoning, inconsistent answers given the data in the world, or violation of constraints that should be maintained when conducting a task. A critiquing system must have 1) a model of expert performance, 2) a model of the types of errors to detect, 3) a means to detect the errors, and 4) a means to notify the user of a detected problem.

**Previous critiquing studies** | The first attempt at building a large-scale critiquing system for the medical community was made by Miller (1986). He developed a prototype system, called ATTENDING, which was designed to work in anesthesiology. Based on this initial research, he also experimented with critiquing systems for hypertension, ventilator management, and pheochromocytoma workup. All of these prototypes operated in a similar manner. The user was required to enter information about the patient's status and symptoms, as well as the proposed diagnosis and treatment. The computer then critiqued the proposed solution, generating a three paragraph output summarizing its critique.

Miller saw much potential to the critiquing approach and was able to provide recommendations to other designers for developing good critiquing systems. First, Miller concluded that choosing a sufficiently constrained domain was important. ATTENDING was a system attempting to aid anesthesiologists in treating their patients, a task that takes years for people to learn and practice. Attempting to build a useful expert system in this field turned out to be too difficult due to the expanse of knowledge required. This lesson led him to switch to the more constrained hypertension domain. Second, Miller concluded that critiquing systems are most appropriate for tasks that are frequently performed, but require the practitioner to remember lots of information about the treatment procedures, risks, benefits, side effects, and costs, as these are conditions under which people are more likely to make errors if unaided, thus making the critiquing system potentially valuable.

A second critiquing system was developed by Langlotz and Shortliffe (1983), who adapted their diagnostic expert system, ONCOCIN (designed to assist with the treatment of cancer patients) to be a critiquing system rather than an autonomous expert system because they found that: "The most frequent complaint raised by physicians who used ONCOCIN is that they became annoyed with changing or 'overriding' ONCOCIN's treatment suggestion" (Langlotz and Shortliffe, 1983, p. 480). It was found that since a doctor's treatment plan might only differ slightly from the system's treatment plan (e.g., by a small difference in the prescribed dosage of a medicine), it might be better to let the physician suggest his/her treatment plan first, and then let the system decide if the difference is significant enough to mention to the doctor. In this manner, the system would be less obtrusive to the doctor. Thus, Langlotz and Shortliffe changed ONCOCIN to act as a critiquing system rather than a diagnostic expert system with the hopes of increasing user acceptance (although they did not report whether acceptance was better with this system).

A third critiquing system, called JANUS, was developed by Fischer, Lemke, and Mastaglio (1990) to aid with the design of kitchens. It is an integrated system, in that the user is already using the computer to design, and the system

uses building codes, safety standards, and functional preferences (such as having a sink next to a dishwasher) as triggering events to critique a user's design.

Another critiquing system was developed by Silverman (1992), who compared performance on two versions of a critiquing system designed to help people avoid common biases when interpreting word problems that included multiplicative probability. The first system only used debiasers, meaning that it provided criticism only after it found that the user's conclusion was incorrect. It had three levels of increasingly elaborate explanation if subjects continued to get the wrong answer. Performance was significantly improved with the critiques than without (69 correct answers for the Treatment Group after the third critique vs. 4 correct for the Control Group), but was not nearly perfect. Subsequently, a second version of the critiquing system was built that included the use of influencers, i.e., before-task explanations of probability theory that would aid in answering the upcoming problems. With the addition of these influencers, performance improved to 100 correct by the end of the third critique.

In examining these results and the performance of several other critiquing systems on the market, Silverman (1992) proposed that, to be effective, a critiquing system should have a library of functions that serve as error-identification triggers, and include the use of influencer, debiaser, and director strategies. (A director demonstrates a strategy to the user).

The final study that will be discussed was conducted in our lab (Guerlain, Smith et al., 1995; Guerlain, Smith et al., 1997). Knowledge about how to rule out alloantibodies in blood typing was encoded into a computer, and critiquing the user at the task (AIDA1) was compared to having the computer perform that subtask (AIDA2). There was no statistical difference in outcome errors for cases for which the computer's knowledge was competent. On a case for which the system's knowledge was brittle, however, misdiagnosis rates increased by 29 for users of the automated system.

Thus, the design of critiquing systems has been explored in a number of domains, but we have only been able to find two rigorous studies using objective data to evaluate actual use of such critiquing systems. Silverman's study compares alternative designs for a critiquing system, finding improved performance with certain forms of remediation when teaching students probability theory. The study by Guerlain et al. is the only source of objective data contrasting the design of decision support systems where the person critiques the computer vs. the computer critiquing the person, and looks at the processes by which actual practitioners using such a system are aided with this kind of support. Guerlain's results suggest that cooperative problem-solving is superior on brittle cases when using the critiquing model of decision support.

Overview: Evaluating a Proof-Of-Concept Critiquing System | The literature reviewed above suggests that critiquing could be a good model for the design of effective cooperative problem solving computer systems. However, although many aspects of critiquing systems have been identified as potentially effective ways to promote cooperative problem-solving, very little research has been done to test the efficacy of this claim. The focus of the research conducted here was to try to develop a proof-of-concept critiquing system, as well as to introduce a new concept: interactive critiquing. With such an interactive critic, the system monitors all of the user's data collection and interpretation activities, providing immediate feedback as soon as it sees a potential problem. To be effective, the critiquing system must therefore be designed so that it has access to data about the user's intermediate cognitive processes while he/she is solving the problem. Antibody identification, as described below, is a task that fits all the characteristics described for an effective critiquing system. It is a difficult, but routinely performed task, with the potential for users to make many intermediate and final errors in problem solving. It is also a task that contains many intermediate steps, and lends itself to critiquing because the user interface can be designed to yield information about intermediate problem-solving conclusions in an unobtrusive manner.

Antibody Identification as a Testbed | The purpose of antibody identification is to find donor blood that can safely be transfused to a patient. It has the classical characteristics of an abduction task (Josephson and Josephson, 1994), including masking and problems with noisy data. Medical technologists run a series of tests, combining test cells (red blood cells), which contain known antigens, with the patient's serum, which may contain antibodies. The test cells have been carefully typed for the presence or absence of antigens by the commercial supplier of the cells. When the test cells and the patient's serum are combined, the blood banker looks for agglutination, a visible clumping of the blood cells, which indicates that antibodies from the patient's serum have bound to some of the antigens contained in the test cells. The amount of agglutination is rated on a scale from 0 (no clumping) to 4+ (very strong; one big clump).

Initially, this process is performed with two or three different test cells that cover all of the major antibodies likely

to be formed. This is called the antibody screen test. If there is a positive reaction with any of these screening cells, then the process is repeated with many more test cells to allow the blood banker to determine which antibodies are causing the reactions. Different reagents can be added to the test cells, or the cells can be run at warm or cold temperatures, to help differentiate which antibodies are causing the observed reactions. Figure 1 shows an example data sheet, called an antigram panel, that is used to mark down reactions.

Expert Strategies | Studies by Smith et al. (1992) have shown that, like experts studied in other medical domains, expert blood bankers try to sort out which antibodies are causing the reactions by recognizing typical reaction patterns and making early hypotheses upon which to base further analyses. To minimize the chance for an incomplete or incorrect diagnosis and to protect against human error and the fallibility of the heuristic methods, the expert blood banker also tries to collect independent, converging evidence to both confirm the presence of hypothesized antibodies and to rule out all others. Thus, even though an antibody may seem very probable, it is important to rule out all other frequent, clinically significant antibodies to be sure that other antibodies are not being masked by the reactions of the first one, as well as to provide protection against slips (Norman, 1981; Norman, 1988) and the fallibility of the heuristics used (Smith, et al., 1991).

For example, when first looking at the antigram panel for the case shown in Figure 1, it looks as if anti-Fyb is a very likely candidate because the Fyb antigen is present on all cells where there is a positive reaction. (There are plus signs in the column labeled Fyb only in those rows where there is a non-zero reaction in the column labeled AHG.). However, after ruling out on this panel, four other antibodies still remain as possibilities (see Figure 2, where the labels for the antibodies that have been ruled out are blacked out). In looking at the remaining set, two other subsets could account for the positive reactions. Anti-E and anti-K together could account for the reactions because one or the other is present on all reacting cells, or anti-E, anti-K, and anti-Fyb could all be reacting together. At this point, it is necessary to run further tests that will discriminate among these three sets of answers. It turns out that, for this case, anti-E and anti-K together form the answer, not anti-Fyb as originally hypothesized. This case clearly demonstrates the importance of collecting converging evidence even though one answer may at first seem very likely.

Donor	Rh-hr										MNSs					P <sub>1</sub>	Lewis				Luth'n				Kell				Duffy				Kidd				Special Type	Test Methods					
	D	C	E	c	e	f	V	C <sup>W</sup>	M	N	S	s	P <sub>1</sub>	Le <sup>a</sup>	Le <sup>b</sup>		Le <sup>x</sup>	Lu <sup>a</sup>	Lu <sup>b</sup>	K	k	Kp <sup>a</sup>	Jk <sup>a</sup>	Fy <sup>a</sup>	Fy <sup>b</sup>	Jk <sup>b</sup>	Jk <sup>c</sup>	Xg <sup>a</sup>	IS	37°	AHG	IgG	RT	4°									
	D	C	E	c	e	f	V	C <sup>W</sup>	M	N	S	s	P <sub>1</sub>	Le <sup>a</sup>	Le <sup>b</sup>		Le <sup>x</sup>	Lu <sup>a</sup>	Lu <sup>b</sup>	K	k	Kp <sup>a</sup>	Jk <sup>a</sup>	Fy <sup>a</sup>	Fy <sup>b</sup>	Jk <sup>b</sup>	Jk <sup>c</sup>	Xg <sup>a</sup>															
1 A618	+	+	+	o	+	o	o	+	o	o	+	+	o	o	o	+	+	+	+	+	o	o	+	+	+	+	+	+	+	Dit(a+)	0	0	3+			1							
2 B459	+	+	o	o	+	o	o	+	+	o	+	+	+	o	+	o	+	+	+	+	+	o	o	+	+	o	+	+		0	0	2+			2								
3 C921	+	o	+	+	o	o	o	+	o	+	o	+	o	+	o	+	o	+	+	+	+	o	o	+	+	+	+	+		0	0	3+			3								
4 D117	+	o	o	+	+	+	+	o	+	+	o	o	+	o	+	o	+	o	+	o	+	o	o	o	+	+	+	+	Bg(a+)	0	0	0			4								
5 E305	o	+	o	+	+	+	o	o	+	o	o	+	+	o	+	o	+	o	+	o	+	o	+	+	o	+	+	+		0	0	0			5								
6 F804	o	o	+	+	+	+	o	o	+	o	+	+	+	o	+	+	+	+	+	+	+	o	o	o	+	+	o	+	Ch(a-)	0	0	3+			6								
7 G922	o	o	o	+	+	+	o	o	+	+	+	+	+	o	o	+	+	+	+	+	+	o	o	+	o	+	o	+		0	0	2+			7								
8 H523	o	o	o	+	+	+	o	o	+	+	o	+	o	+	o	+	o	+	o	+	o	+	o	+	o	+	o	+		0	0	0			8								
9 I710	+	o	o	+	+	+	+	o	+	o	o	+	+	o	o	o	+	o	+	o	+	o	+	o	+	o	o	+	Jk(b-)	0	0	0			9								
10 J386	o	o	o	+	+	+	o	o	+	+	+	+	o	+	o	o	+	o	+	o	+	o	o	o	+	+	o	+		0	0	0			10								
AutoCtrl																													0	0	0												

Figure 1. Anti-Fyb looks likely

Donor	Rh-hr										MNSs				P		Lewis		Luth'n		Kell		Duffy		Kidd		Special Type	Test Methods				
	D	C	E	e	r	y	v	C <sup>w</sup>	M	N	S	s	P <sub>1</sub>	L <sub>1</sub>	L <sub>2</sub>	L <sub>3</sub>	L <sub>4</sub>	K	k	Kp <sup>a</sup>	Jk <sup>a</sup>	Fy <sup>a</sup>	Fy <sup>b</sup>	Jk <sup>b</sup>	Jk <sup>c</sup>	IS		37°	AHG	IgG	RT	4°
1 A618	+	+	+	o	+	o	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	Dk(a+)	0	0	3+			1	
2 B439	+	+	o	o	+	o	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		0	0	2+			2	
3 C921	+	o	+	+	o	o	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		0	0	3+			3	
4 D117	+	o	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	Bg(a+)	0	0	0			4	
5 E305	o	+	o	+	+	+	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		0	0	0			5	
6 F804	o	o	+	+	+	+	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	Ch(a-)	0	0	3+			6	
7 G922	o	o	o	+	+	+	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		0	0	2+			7	
8 H523	o	o	o	+	+	+	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		0	0	0			8	
9 I710	+	o	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	Jb(b-)	0	0	0			9	
10 J386	o	o	o	+	+	+	o	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		0	0	0			10	
AutoCtrl																										0	0	0				

Case: **RJR** D C E e r y v C<sup>w</sup> M N S s P<sub>1</sub> L<sub>1</sub> L<sub>2</sub> L<sub>3</sub> L<sub>4</sub> K k Kp<sup>a</sup> Jk<sup>a</sup> Fy<sup>a</sup> Fy<sup>b</sup> Jk<sup>b</sup> Jk<sup>c</sup>

Figure 2. Anti-E and anti-K can also account for the reactions, however.

Poor problem-solving strategies | Previous studies indicate that blood bankers vary to the extent that they understand and use all the knowledge that they need to solve a case (Smith, et al., 1991). Poor performance can result from failing to use a good strategy, or from using incorrect strategies, or from making slips. In addition, some practitioners use inefficient problem-solving techniques, collecting more data than necessary. This can happen if the practitioner does not make all of the inferences possible given a test result.

The Design of the Antibody Identification Assistant (AIDA) | Based on studies of the expert strategies and erroneous/inefficient strategies found to be used in this domain, a critiquing system (AIDA) was developed. This system was developed on the Macintosh using Symantec's® Think C programming language. The system can be used as an information display tool that allows practitioners to request and interpret the various tests used for antibody identification similar to the way they normally would using paper and pencil. With all messaging turned off, subjects do not get any feedback from the computer (although it is still monitoring for errors and logging them for data analysis purposes). With messaging turned on, the system provides immediate feedback if an error is detected. Four design principles were used to guide the design of this problem-solving tool.

Design Principle 1: Use Direct Manipulation to Provide an Unobtrusive Form of Communication | The interface for AIDA is designed not only to be helpful and easy to use, but also to provide data for the computer to diagnose errors in the user's problem solving. The technologist can request test forms (via a pull-down menu) and mark hypotheses on those forms (via color-coded "pens"), so the computer is able to watch the person's problem-solving process, potentially detecting errors in the subject's procedure. Thus, no extra work is required on the user's part to feed information to the computer. Practitioners just work as they naturally would and, because of the interface design, the data on the user's problem-solving activities is rich enough for the computer to detect problems and provide immediate, context-sensitive feedback. Figure 3 shows the tests available, and Figure 4 shows the interactive user interface.

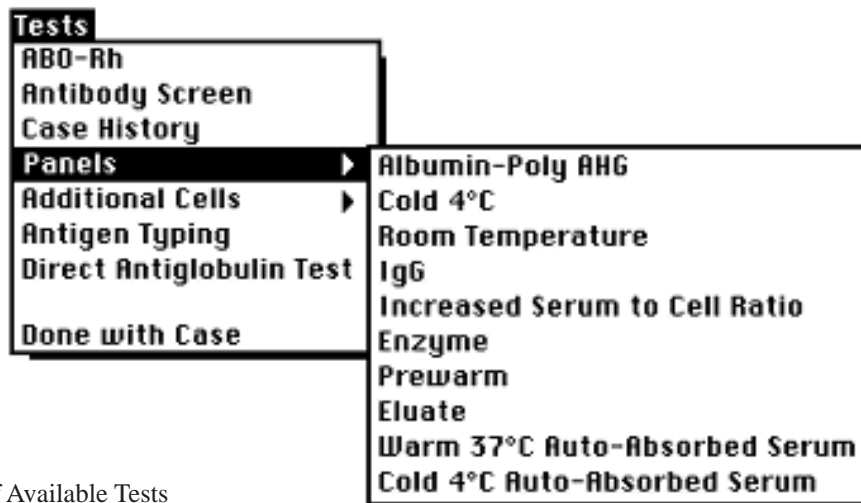


Figure 3. Menu of Available Tests

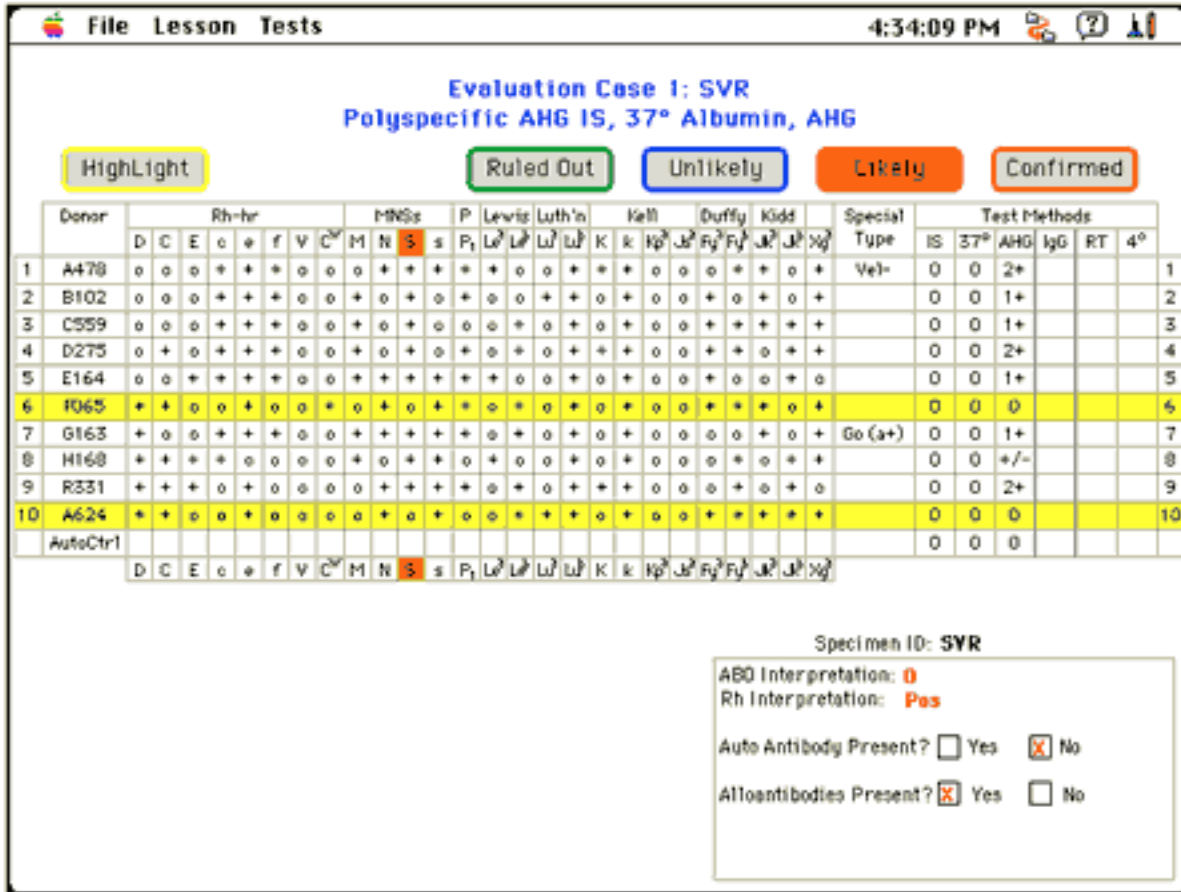


Figure 4. Sample Screen

Design Principle 2: Provide Perceptual and Memory Aids to Encourage the User to Use the System's Direct Manipulation Capabilities | The color coding functions used for highlighting rows, columns or individual cells on a form, and for marking intermediate conclusions (such as ruling out an antibody) make it easier for the technologist to complete his/her task, by reducing perceptual and memory demands. Memory load is reduced because the intermediate conclusions are propagated to new panels when the user asks for another data display. This functionality thus encourages the user to use the color coding, which in turn informs the computer about the user's thought processes.

Design Principle 3: Base the Critiquing Strategy on an Error Model for the Domain | The collection of critics was designed around a broad strategy of collecting converging evidence before completing a case. To help ensure use of this strategy, AIDA monitors for both errors of commission and errors of omission as well as ensuring that practitioners follow a good protocol, (using independent, converging evidence to rule out all non-confirmed antibodies) and ensuring that the final solution set makes sense (i.e., there are no unexplained reactions, and that the pattern of reactions is consistent with the types of antibodies confirmed) and ensuring that the final solution set does not violate constraints inherent in the domain (e.g., anti-E alone in an Rh negative patient is very rare).

Design Principle 4: Abstractly Represent the Computer's Knowledge for the User to Establish a Common Frame of Reference | A written checklist was designed to provide an explicit, high-level representation of the computer's goal hierarchy. The design of the system is such that users can apply additional strategies without interference from the computer, and can override a critique from the computer, but the checklist makes it clear what steps the computer expects the person to have done before completing a case. The computer also allows the user flexibility in deciding what order to use in completing the subgoals listed on the checklist (i.e., the computer does not monitor for the ordering of the steps listed in the checklist except when that ordering is critical to successful problem-solving).

## Experimental Procedure

**Subjects** | Two subject pools were used to test AIDA. The first was a group of four "experts" (certified Specialist in Blood Bank (SBBs)) who were tested with the system as a pilot group. These subjects came from three different hospitals. (The objective of this preliminary study was to make sure AIDA did not interfere with the performances of skilled practitioners.)

Subsequently, thirty-two blood bankers from seven different hospitals were tested. All of these technologists were identified by their supervisors as "actually performing the task of antibody identification as part of their job, but who would benefit from additional experience and training". Their years of experience ranged from 1 to 35 years (with a mean of 10 years), and they constituted 35 of the total number of practitioners at these seven hospitals.

**Experimental Design** | All of the subjects were tested on the same six cases. Half of the subjects were randomly assigned to the Control Group and the other half were randomly assigned to the Treatment Group. Figure 5 shows the experimental design for this study. All subjects (in both the Treatment and Control Group) solved the first and second case without any aid from the computer. Thus, both groups were using the control version of the system (with all critiquing messages turned off). The first case was used to give both groups the same initial training on how to use the system interface. Subjects were shown how to use the pull-down menus to select test results and how to interpret the test results on each screen. During this initial training, no knowledge specific to blood banking was discussed. Furthermore, subjects were not asked to solve the first case, but just used it to practice selecting and marking individual test results.

File Lesson Cases 4:18:35 PM

**Lesson 1: ABO and Rh Typing**  
**Practice Case 1a: PJJ**  
**ABO and Rh Grouping Results**

SPECIMEN I.D.	FORWARD TYPE			REVERSE TYPE		Rh TYPING									
	Anti-A	Anti-B	Anti-A,B	A <sub>1</sub> Cells	B Cells	Anti-D				Rh Control					
						IS	37°	AHG	CC	IS	37°	AHG	CC		
PJJ	4+	0	4+	0	3+	3+					0				

ABO Interpretation:  Rh Interpretation:

**Incorrect. Group B individuals have the B antigen but no A antigen on their red cells. You would therefore expect a reaction with Anti-B but not with Anti-A. Also, Group B individuals have Anti-A but no Anti-B in their serum, producing a reaction with the A1 cells, but not the B cells, for the reverse typing.**

**Thus, these data are inconsistent with your answer.**

Figure 5. Sample ABO/Rh Error Message

The purpose of second case (herein referred to as the "Pre-Test Case") was to get a benchmark on the practitioners' current performance strategies against which to compare the Treatment Group's strategies when using the experimental system. For example, we could determine by actions taken whether subjects performed rule-out, performed antigen typing, and hypothesized the presence of more than one antibody, (i.e., by highlighting the differential patterns separately) . The Treatment Group was then trained on the use of the checklist and shown how the systems monitored for errors (see Figure 6).

The Pre-Test Case and first Post-Test Case were two matched cases, and it was randomly determined at run-time which of the two cases a particular subject solved as a Pre-Test Case and which as a Post-Test Case. With this design, a within-subjects comparison could be made for the Treatment Group. After solving the first Post-Test Case, both groups solved three more cases, with the Treatment Group continuing to use the critiquing and the checklist and the Control Group continuing to solve the cases on their own. Performance on these cases could be examined for between-subjects differences.

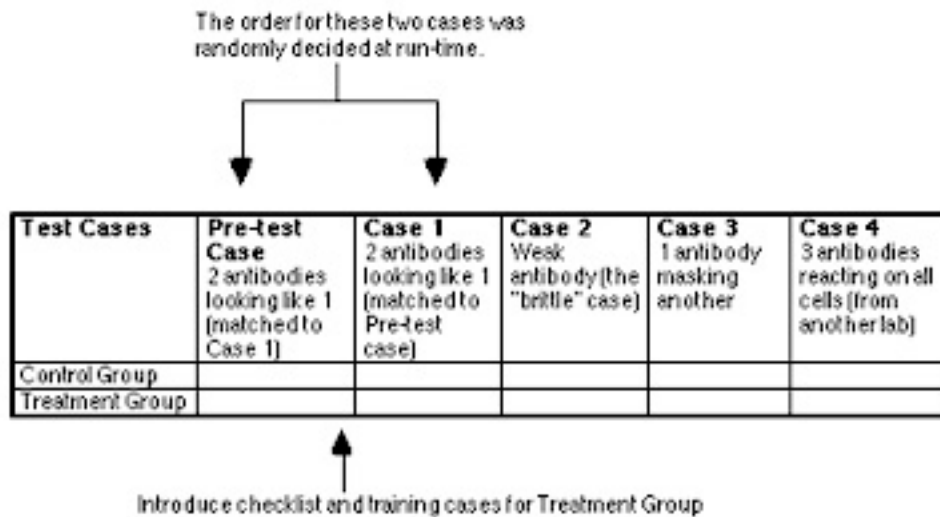


Figure 6. Experimental Design

Cases used to test AIDA | The AIDA system has a set of test cases built into it that were either designed by an expert blood banker or were taken from real patient data to ensure validity. The cases that were used for testing were carefully selected to have certain characteristics (weak antibodies, masked antibodies, etc.) that make them examples of different kinds of difficult cases. Predictions were made as to how a practitioner's performance would change depending on the case characteristics, the practitioner's strategy, and the version of the system the practitioner was using. As previously described, the Pre-Test Case and first Post-Test Case had the characteristic that the original testing panel seems to indicate that only one antibody is present, but in actuality, two different antibodies are together accounting for the reactions. The second Post-Test Case was a weak antibody case, for which the computer's knowledge was not fully competent. Case 3 was a masking case (where one antibody masks the presence of another). Case 4 was sent to us by a blood bank lab that had no specific knowledge of our work. It turned out to be a difficult three antibody case.

Data Collection | Errors for both the Control Group and Treatment Group were logged as subjects were working on the test cases. The system automatically coded the data, counting types of errors and misdiagnosis rates.

Data Analysis | Three types of analyses were made: 1) An analysis of outcome performance, as measured by misdiagnosis rates; 2) A behavioral protocol analysis to examine subjects' strategies, behaviors, and process errors. (The specific errors that the critiquing system recorded were aggregated into five groups for analysis: ruling out a hypothesis incorrectly, failing to rule out a hypothesis when it was appropriate to do so, failing to collect converging evidence, selecting an answer that was unlikely to produce the observed pattern of data, and selecting an answer that was highly unlikely based on prior probabilities alone); and 3) A questionnaire to get subjective reactions to using the critiquing system. (For brevity, results of the questionnaire can be found in Guerlain, Smith, et al., 1995.) A number of statistical tests were run to measure differences in misdiagnosis rates. McNemar's Chi Square test was

used to test the hypothesis that subjects in the Treatment Group improved in performance from the Pre-Test Case to the matched Post-Test Case. Fisher's exact test was used to test the hypothesis that the Treatment Group had better performance on each of the Post-Test Cases than the Control Group. Finally, a test for difference between the two groups was conducted using a log-linear analysis that takes into account the performances of both the Control and Treatment Groups on the Pre-Test Case.

Results and Discussion | As a basis for comparing performance with the critiquing system, we first looked at unaided performance on this task. Multiple process errors were made by unaided subjects. Table 1 shows the number of subjects (Treatment and Control groups combined) who made a particular kind of error at least once on the Pre-Test Case. (Four subjects' data were not counted in this and further analyses because it was discovered during data analysis that, due to a bug in the program, the reactions for a particular test panel that these subjects used were incorrect. The other subjects never accessed this test panel).

Error Type	Percentage of Subjects (out of 29) Who Committed that Error at Least Once on the Pre-Test Case
1. Ruling out Hypotheses Incorrectly	69%
2. Failing to Rule Out When Appropriate	48%
3. Failure to Collect Converging Evidence	90%
4. Data Implausible Given Answer	38%
5. Answer Implausible Given Prior Probabilities	38%

Table 1. Process Errors Made by Treatment and Control Group Subjects on the Pre-Test Case

Gross Performance Measures | The following sections give the results for the overall misdiagnosis rates and a comparison of the mistakes and slips made by the two groups.

Expert Subjects | The group of four experts was tested prior to the group of thirty-two less-skilled practitioners, primarily as a check to evaluate AIDA for usability and to make sure AIDA did not create difficulties or induce new errors for skilled technologists. Briefly, two of these four technologists were tested as the Control Group and two as the Treatment Group. All four subjects got all of the cases (Pre-Test and Post-Test) correct on the first try. Thus, there is no evidence from this data to suggest that the system interfered with expert problem-solving performance. No further analyses were made regarding the expert subjects' performance with the system.

Less Skilled Subjects | Of the thirty-two less skilled blood bankers tested in the actual evaluation study, sixteen were randomly assigned to the Control Group and sixteen to the Treatment Group. In analyzing the data, it was discovered that on one of the two matched cases, the reactions to one set of test results that was requested by four of the subjects were incorrect. Three subjects encountered this as their Pre-Test case, while one encountered it as Post-Test Case 1. Thus, the data from three subjects was discarded from the analyses that follow dealing with the Pre-Test Case.

As would be expected, the results showed that there was no significant difference in performances on the Pre-Test Case for the Control and Treatment Groups (using Fisher's exact Test, see Table 2). The misdiagnosis rates were eliminated for the Treatment Group from 4/15 wrong on the Pre-Test Case to 0/16 wrong on the matched Post-Test Case 1, although this difference failed to reach significance (using McNemar's Chi Square for dependent samples,  $p > .05$ ). The Control Group also failed to show a significant improvement in performance from the Pre-Test Case to Case 1, as would be expected.

Test Cases	Pre-test Case	Case 1	
	2 antibodies looking like 1 (randomly chosen from one of two matched cases, the other of which was Case 1)	2 antibodies looking like 1 (randomly chosen from one of two matched cases, the other of which was the Pre-Test Case)	
Control Group	6/14 wrong	5/15 wrong	NS
Treatment Group	4/15 wrong	0/16 wrong	

Table 2. Pre-test/Post-test comparison of misdiagnosis rates

The results showed marked differences in performance across the two groups (see Table 3). On Cases 1, 3, and 4, all subjects in the Treatment Group solved the cases correctly, while 5/15 of the subjects in the Control Group misdiagnosed Case 1, 6/16 misdiagnosed Case 3, and 10/16 misdiagnosed Case 4. Using Fisher's exact test, each of these differences is significantly different ( $p < 0.05$ ). For the case that the system was not designed to completely handle (Case 2), 8/16 subjects in the Control Group misdiagnosed the case compared to 3/16 in the Critiquing Group. This improvement in performance is not significant ( $p = 0.072$ ). Thus, with the design of a critiquing system and checklist, we were able to eliminate misdiagnoses on cases for which the system was designed (Cases 1 and 3) and on a case for which the system was not explicitly designed but for which the system's knowledge was appropriate (Case 4). Finally, misdiagnosis rates were sizably reduced on a case for which the system's knowledge was not fully competent (Case 2), but this reduction was not statistically significant.

Test Cases	Case 1	Case 2	Case 3	Case 4
	2 antibodies looking like 1 (randomly chosen from one of two matched cases, the other of which the system was not	weak antibody (for which the system was not designed to	1 antibody masking another	3 antibodies reacting on all cells (a case for which adequately handle) was the Pre-Test Case) explicitly designed, sent by another blood bank lab)
Control Group	5/15 (33.3) wrong	8/16 (50.0) wrong	6/16 (37.5) wrong	10/16 (62.5) wrong
Critiquing Group	0/16 (0.0) wrong	3/16 (18.75) wrong	0/16 (0.0) wrong	0/16 (0.0) wrong
Significance	$p < 0.05$	$p = 0.072$	$p < 0.01$	$p < 0.001$

Table 3. Post-Test Case results.

Besides individual comparisons using Fisher's exact test, a log-linear analysis was run to take into account the difference in performance between the Control and Treatment Groups on the Pre-Test Case. Both Treatment and Control Groups were subdivided into whether or not the Pre-Test Case was correct. This analysis gave very similar results on individual cases and gave a combined significance level (Weiner, 1971) of  $p < 0.000005$  favoring performance for the Treatment group.

Tables 4 and 5 give a subject by subject breakdown of misdiagnoses per case. These tables show whether a subject got a case right (represented by a 1) or wrong (represented by a 0) or, in the case of the Treatment Group on the Post-Test Cases, got some feedback from the computer regarding the plausibility of the answer. In this case, the table may show a series of answers (such as 0-0-1), the last one indicating the correctness of the final answer given by the subject.

Subject	Pre-Test Case	Case 1	Case 2	Case 3	Case 4
T1	1	1-1-1-1	1	0-1	0-1
T2	0	1	0-0	1	1
T3	1	1	0-1	1	0-1
T4	1	1	1	1	0-0-0-0-1
T5	0	1	1-1	1	1
T6	1	1	0-0-0-0-0-0-1	0-0-0-1	1
T7	1	1	0-1	1	1
T8	1	1	0-1-1	1	1
T9	invalid data	1	0-0	1	1
T10	1	1	0-0-0	1	1
T11	1	1	1	1	1
T12	1	1	0-1	1	0-1
T13	0	1	1	0-1	0-1
T14	1	1	1	1	1
T15	0	1	1	1	1
T16	1	1	1	1	1

Table 4. Correctness of Answers, Treatment Group  
(0 = wrong, 1 = right. A series of numbers indicates the subject marked a new answer more than once, in response)

Subject to critiques given by the computer).	Pre-Test Case	Case 1	Case 2	Case 3	Case 4
C1	0	0	0	0	0
C2	0	1	0	1	0
C3	1	1	1	1	1
C4	0	1	1	0	0
C5	1	invalid data	1	1	0
C6	1	1	1	1	1
C7	0	0	1	1	0
C8	1	1	1	1	1
C9	invalid data	0	0	0	0
C10	1	1	1	1	0
C11	1	1	0	0	1
C12	0	0	0	0	0
C13	0	0	0	0	0
C14	1	1	0	1	1
C15	1	1	1	1	1
C16	invalid data	1	0	1	0

Table 5. Correctness of Answers, Control Group.  
(0 = wrong, 1 = right).

Errors Made While Solving Cases | Table 6 shows the number of errors of each type for the Control and Treatment Groups. The two groups are significantly different for Error Type 3 ( $p < 0.01$ ). The primary significance of this table is the indication that the critics responsible for detecting errors during the solution of a case (Error Types 1 and 2), as well as those that fired after an answer was indicated (Error Types 3-5), were all quite active (and likely contributed to the observed reduction in error rates by the Treatment Group).

ERROR TYPE	GROUP	POST-TEST CASES			
		Case 1	Case 2	Case 3	Case 4
1. Rule out Hypothesis Incorrectly	Control	40.6	37.5	40.6	40.6
	Treatmt	28.1	34.4	18.8	28.1
2. Failure to Rule Out When Appropriate	Control	37.5	15.6	34.4	34.4
	Treatmt	34.4	21.9	9.4	25.0
3. Failure to Collect Converging Evidence	Control	37.5	34.4	31.3	34.4
	Treatmt	3.1	9.4	9.4	12.5
4. Data Implausible Given Answer	Control	15.6	21.9	3.1	21.9
	Treatmt	0.0	15.6	6.3	15.6
5. Answer Implausible Given Prior Probabilities	Control	21.9	15.6	12.5	40.6
	Treatmt	3.1	25.0	9.4	12.5

Table 6. Number of Subjects Committing Each Type of Error at Least Once Per Case (in percent) on Post-Test Cases

Proactive Training vs. Reactive Feedback (Critiquing) | One interesting question to ask is to what extent is the improvement in performance of the Treatment Group due to the initial training and use of the checklist (proactive training) and to what extent is it due to the presence of the critiquing system monitoring their performance (reactive feedback)? Clearly, a large improvement is seen from the Pre-Test Case to the matched Post-Test Case in terms of outcome errors (which were eliminated) and process errors. This indicates that the proactive training with the checklist and critiquing was immediately helpful to subjects and helped to significantly improve their performance.

It is also interesting to note, however, that subjects in the Critiquing Group did not always get a case right immediately (see Table 4). Even though they eventually got the right answer on almost all of the Post-Test Cases, this was not without assistance from the computer. For example, in 18 instances, subjects indicated that they were done with a case and the computer detected one or more errors (see Table 4). Fourteen of these instances included errors that were concerned with errors of omission in their procedure (an incomplete protocol), 17 included an inconsistency with the answer marked given the reactions, and 16 with an implausible answer given prior probabilities, for a total of 47 errors detected. On 16 out of the 18 instances, the subjects' answers were wrong and of those 16, 13 subjects subsequently changed their answer to the correct one because they were prompted by the critiques to re-examine the case (remember that the computer does NOT know any of the answers and is merely checking for particular kinds of process and intermediate inference errors). Thus, we have evidence that the presence of the critiquing system contributed significantly to the improvement in overall performance.

Detailed Analyses | Besides summary statistics, more detailed analyses of behavior for the group of less-skilled subjects were conducted from the behavioral protocol logs that were automatically generated by the computer, to determine if important behaviors with the system, whether good or bad, could be identified.

One Control Group subject, for example, made many process errors that lead to an incorrect solution on all of the cases. On the Pre-test Case, this subject ruled out by using a strategy that will fail in multiple antibody cases (ruling out using reacting cells), and which caused her to rule out both of the right answers (anti-c and anti-K). She also marked anti-S as the answer, even though it accounted for most, but not all, of the reactions exhibited. Thus, she failed to make sure there were no unexplained positive reactions. Furthermore, she made other procedural errors such as ruling out heterozygously, ruling out antibodies using results from test procedures that usually inhibit those reactions, and failing to do antigen typing for the antibody marked as the answer.

Thus, we have evidence consistent with previous studies that practicing medical technologists make a significant number of process errors and outcome errors when solving antibody identification cases.

Example Subject Interactions with the Critiquing System | As a comparison to unaided performance, this section gives the reader an idea of how the critiquing system interacted with a sample subject, detecting errors in performance and steering the subject toward a successful solution path. This subject got the Pre-Test Case wrong, but then got the rest of the cases right with the aid of the critiquing system. On two of those Post-Test Cases, she initially had an incorrect solution set, but changed her answer in response to the critiques she received.

On the Pre-Test Case, this subject correctly reviewed initial data about the patient, such as the ABO/Rh and the Case History. After seeing that the initial Antibody Screen results were positive, she selected a full panel for interpreting test results. There, she ruled out using homozygous, non-reacting cells (a good strategy) and selected additional cells for further analysis. At this point, she confirmed anti-Fyb and continued on to the next case. The correct answer for the case was anti-E plus anti-K. Anti-Fyb accounted for most of the reactions, but it did not fit the pattern of dosage (strength of reaction depending on the strength of the antigen) and did not account for two of the reacting cells, (one on the initial Antibody Screen test and one on the Additional Cells panel). This subject's erroneous conclusion stemmed from following an incomplete protocol. She did not try to rule out the alternative antibodies, and did not run an Antigen Typing test as independent evidence leading toward her answer. Furthermore, her answer did not account for two of the reactions seen, nor the strength of reactions on the reacting cells. On the matched Post-Test Case, however, this subject correctly followed a complete protocol, being sure to rule out all remaining antibodies besides the ones marked as Confirmed, and successfully solved the case.

On Post-Test Case 2, the weak antibody case, the system alerted the subject to the fact that since some of the reactions were weak, rule-out might not be an appropriate strategy and suggested first enhancing the reactions. The subject heeded this warning and enhanced the reactions before proceeding with rule-out. When looking at the reactions on the Additional Cells, this subject tried to run another test, but the system warned her that she could have ruled out more antibodies on that panel. Because of this message, she continued to rule out on the Additional Cells panel that she was looking at, and was able to finish ruling out all remaining antibodies besides anti-D. She

consequently confirmed anti-D. Thus, the system aided her by suggesting that she enhance reactions before ruling out and, once at such an enhanced phase of testing, checking to be sure that she ruled out all of the antibodies possible. In this way, the system helped her to avoid running extra tests which were not necessary.

On Post-Test Case 3, this subject solved the case to the point where anti-Fya and anti-E were the only remaining antibodies. At this point, she confirmed anti-Fya (which accounts for all of the reactions) and marked anti-E likely. The system reminded her that she had not ruled out all antibodies besides anti-Fya and warned her that anti-E was confounded with anti-Fya. This message prompted her to run the cells at Enzymes (a technique that will destroy Duffy antibodies, including Fya, and enhance Rh antibodies, including anti-E). Thus, she was able to expose the presence of anti-E and correctly add anti-E to her answer set.

On Post-Test Case 4, this subject proceeded to the point where all antibodies but anti-Jkb, anti-c, and anti-E were ruled out. Again, she marked one of them as Confirmed (anti-Jkb), but did not confirm or rule-out either of the other two remaining antibodies. The system warned her that 1) she had not ruled out all remaining antibodies, 2) the confirmed antibody did not account for many of the reactions exhibited, 3) it is rare to see anti-Jkb as the only antibody, and 4) antibodies tend to form in a certain order, and that anti-c and anti-E would be more likely to form before anti-Jkb. Thus, the system used knowledge about frequency of occurrence, as well as data specific to that case, to warn the subject that her answer was implausible. In addition, the system made the general remark that her protocol was incomplete (i.e., that she had not ruled out all remaining antibodies). In response to these messages, the subject further examined the case and included anti-E and anti-c in her answer set, getting the case right.

Analysis of the Weak D Case | It is worth examining the Weak D case (Case 2) in more detail, because this case is one where the knowledge in the system is not fully competent for solving the case. Thus, there is the potential for "brittleness" in the computer's reasoning. The question to consider is whether the critiquing system is still helpful to subjects solving such a case and if so, what are the mechanisms that are contributing to this improvement? One measure of performance is the misdiagnosis rate for the two groups. The Control Group had a 50 misdiagnosis rate as compared to the Treatment Group, who had a 18.75 misdiagnosis rate ( $p = 0.072$ ). This difference is not significant, but suggests a trend towards improved performance with the critiquing system. Thus, one may ask what aspects of the critiquing system could be contributing to this possible improved performance.

Three design features in particular had the potential to aid subjects. The first was the application of some "meta-knowledge" such that the critiquing system is aware that its rule-out strategy is fallible in the case of weak reactions. The system's "solution" in this case is to warn the user that ruling out when there are weak reactions is dangerous, and the system suggests trying to enhance the reactions first. The second possibly helpful design feature is the use of prior probability information when examining the plausibility of an answer. In particular, one common misdiagnosis on this case (based on the case characteristics and previous testing of this case on practitioners) is anti-E, since anti-E accounts for the weak reactions on the initially displayed test reactions. As part of its check for the plausibility of answers, the system "knows" that anti-E is a rare finding when the patient is Rh negative and anti-D has not also been confirmed (as is the case here). Thus, the system displays the following message: "Anti-E as the only Rh antibody is uncommon in an Rh negative person. Normally anti-D would form first. It would be better to double check and ensure that anti-D is not present by doing an enzyme panel, by increasing the serum:cell ratio or by using some other enhancement technique. (In addition, if this patient is a pregnant woman, check to see if she has been administered RhIG.)" Such a message may prompt the subject who marks this answer to re-consider the answer to the case. Other "prior probability" messages could potentially be instantiated if the person marks an answer other than anti-E. Furthermore, an answer besides anti-E will not account for all of the exhibited reactions and would thus cause the system to warn the user that the answer given does not account for all of the data seen on the case.

Figures 7 and 8 show the paths taken when solving the Weak D Case for the Control and Treatment Groups respectively. (One subject's protocol data was lost due to a computer error and thus it is not clear how he arrived at the correct answer and his path is shown as a question mark in Figure 9). In comparing these two figures, we see that a comparable group of subjects in both groups successfully solved the case "on their own" by either waiting to rule out until the reactions were enhanced (5 subjects in each group) or by enhancing the reactions after having ruled out D and subsequently confirming D (3 subjects in the Control Group and 2 subjects in the Treatment Group), such that 8 subjects in each group initially solved the case and 8 misdiagnosed it. However, the Treatment Group had the benefit of the critiquing messages that check the plausibility of a solution. Five of the subjects receiving such a message subsequently changed their answer to correctly include anti-D as part of their answer. In conclusion, it is hard to establish whether the "warning" at the beginning of the case aided subjects at all, but the end-of-case error checking (that checks the plausibility of an answer) was clearly beneficial. The benefit of these end-checkers is evident on other cases as well, causing subjects to correct their answers in three instances on Case 3 and in five

instances on Case 4 (see Table 4).

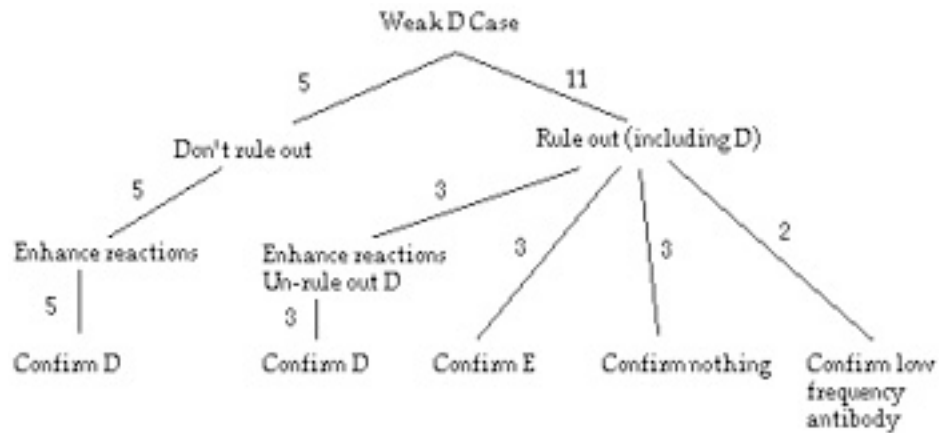


Figure 7. Paths taken to solve the Weak D Case, Control Group

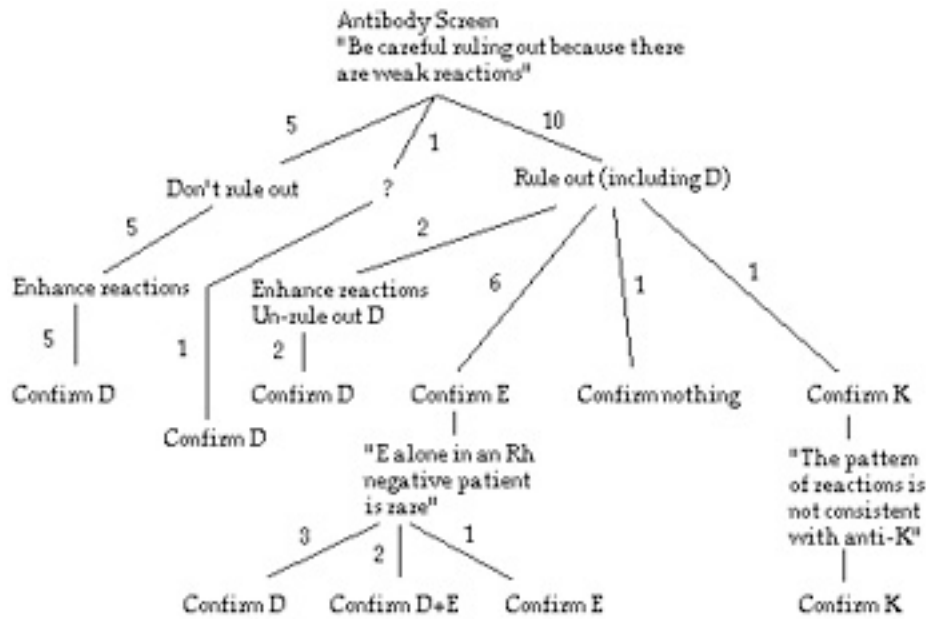


Figure 8. Paths taken to solve the Weak D Case, Treatment Group

Conclusion | This study focused on how to design a particular type of cooperative decision support system, a critiquing system. If such a design is effective, the computer should be able to detect and correct errors without inducing new errors. Specifically, the form of cooperative support studied here was to develop a representation (in the form of a written checklist) to provide guidance in the form of a high level goal structure, and to have the computer act in a critiquing role, monitoring the human's problem-solving process for potentially faulty reasoning steps.

The study presented here provides strong evidence supporting the effectiveness of critiquing as a form of decision support. On cases where the computer's reasoning was fully competent, misdiagnosis rates were completely eliminated for subjects using the critiquing/checklist system, whereas subjects with no decision support were misdiagnosing

cases 33 to 63 of the time. Furthermore, even on the case where the system was left less than fully competent, the Treatment Group correctly solved the case 32 more often than the Control Group.

Thus, critiquing, although not explored to date by very many researchers as a form of decision support, seems to be a viable solution for greatly improving performance on certain kinds of tasks, including the important, real world medical diagnosis task of antibody identification. Clearly, this is a task that medical technologists find difficult, since many of them are getting moderately difficult, yet realistic, patient cases wrong when unassisted. A well-designed critiquing/checklist system has proven to be a method for virtually eliminating the errors that it was designed to catch, and for aiding on cases for which its knowledge is incomplete.

A systems approach was taken in the design of this decision support system. The critiquing model of interaction was employed so that the human practitioners would not be adversely biased by the computer on cases where it exhibited brittle performance. There was evidence that the critiquing system aided subjects by catching errors and helping users to recover from these errors, employing five different types of error checking mechanisms, (checking for errors of commission, checking for errors of omission, checking for an incomplete protocol, checking that the data was consistent with the answer, and checking that the answer was plausible given prior probabilities). The use of a checklist was beneficial in quickly training subjects on the high-level goal structure implicit in the computer's knowledge base, and served as a reminder to subjects of the steps necessary to successfully solve a case. Finally, the success of the system's interaction with the user relied on its unobtrusive interface that allowed subjects to naturally solve antibody identification cases as they normally would using paper and pencil, while providing the computer with a rich set of data regarding the characteristics of the case and the user's reasoning about that case without requiring the practitioner to enter information that was outside of normal task requirements.

#### References |

- Aikins, J., Kunz, J., and Shortliffe, E. (1983). PUFF: An expert system for interpretation of pulmonary function data. *Computers and Biomedical Research*, 16, 199-208.
- Andert, E. (1992). Integrated knowledge-based system design and validation for solving problems in uncertain environments. *International Journal of Man-Machine Studies*, 36, 357-373.
- Bankowitz, R., McNeil, M., Challinor, S., Parker, R., Kapoor, W., and Miller, R. (1989). A computer-assisted medical diagnostic consultation service. Implementation and prospective evaluation of a prototype. *Annals of Internal Medicine*, 110, 824-832.
- Bernard, J. A. (1989). Applications of artificial intelligence to reactor and plant control. *Nuclear Engineering and Design*, 113, 219-227.
- Berner, E., Brooks, C., Miller, R., Masarie, F., and Jackson, J. (1989). Evaluation issues in the development of expert systems in medicine. *Evaluation and the Health Professions*, 12(3), 270-281.
- Berner, E., Webster, G., Shugerman, A., Jackson, J., Algina, J., Baker, A., Ball, E., Cobbs, C., Dennis, V., Frenkel, E., Hudson, L., Mancall, E., Rackley, C., and Taunton, O. (1994). Performance of four computer-based diagnostic systems. *The New England Journal of Medicine*, 330(25), 1792-1296.
- Bernelot Moens, H. J. (1992). Validation of the AI/RHEUM knowledge base with data from consecutive rheumatological outpatients. *Methods of Information in Medicine*, 31, 175-181.
- Fischer, G., Lemke, A., and Mastaglio, T. (1990). Using critics to empower users. In *CHI '90 Human Factors in Computing Systems Conference Proceedings* (pp. 337-347). New York: Association for Computing Machinery.
- Forsythe, D. and Buchanan, B. (1992). Broadening our approach to evaluating medical information systems. In *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care* (pp. 3-7). Baltimore, MD: IEEE Computer Society.
- François, P., Robert, C., Astruc, J., Begue, P., Borderon, J., Floret, D., Lagardere, B., Mallet, E., Pautard, J., and Demongeot, J. (1993). Comparative study of human expertise and an expert system: Application to the diagnosis of child's meningitis. *Computers and Biomedical Research*, 26, 383-392.

Gregory, D. (1986). Delimiting expert systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-16(6), 834-843.

Guerlain, S., Smith, P. J., Gross, S. M., Miller, T. E., Smith, J. W., Svirbely, J. R., Rudmann, S., and Strohm, P. (1994). Critiquing vs. partial automation: How the role of the computer affects human-computer cooperative problem solving. In M. Mouloua and R. Parasuraman (Eds.), *Human performance in automated systems: current research and trends* (pp. 73-80). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Guerlain, S., Smith, P. J., Obradovich, J., Smith, J. W., Rudmann, S., and Strohm, P. (1995). The Antibody Identification Assistant (AIDA), an example of a cooperative computer support system. *Proceedings of the 1995 IEEE International Conference On Systems, Man and Cybernetics*, (pp. 1909-1914).

Guerlain, S., Smith, P. J., Obradovich, J., Heintz, Rudmann, S., Strohm, P., Smith, J. W., Svirbely, J. R., and Sachs, L. (1997). Critiquing as a form of decision-support: an empirical study. CSEL Technical Report #1996-14, The Ohio State University, Columbus OH.

Harris, S. D., and Owens, J. M. (1986). Some critical factors that limit the effectiveness of machine intelligence technology in military systems applications. *Journal of Computer-Based Instruction*, 13(2), 30-34.

Hickam, D., Shortliffe, E., Bischoff, M., Scott, A., and Jacobs, C. (1985). The treatment advice of a computer-based cancer chemotherapy protocol advisor. *Annals of Internal Medicine*, 103(6 pt 1), 928-936.

Josephson, J. and Josephson, S. (1994). *Abductive inference. Computation, philosophy, technology*, Cambridge University Press.

Langlotz, C. P., and Shortliffe, E. H. (1983). Adapting a consultation system to critique user plans. *International Journal of Man-Machine Studies*, 19, 479-496.

Layton, C., Smith, P. J., and McCoy, E. (1994). Design of a cooperative problem-solving system for enroute flight planning: An empirical evaluation. *Human Factors*, 36(1), 94-119.

Linnarsson, R. (1993). Decision support for drug prescription integrated with computer-based patient records in primary care. *Medical Informatics*, 18(2), 131-142. Miller, P. (1986). *Expert critiquing systems: practice-based medical consultation by computer*. New York: Springer-Verlag.

Miller, R., and Maserie, F. (1990). The demise of the "Greek oracle" model for medical diagnostic systems. *Methods of Information in Medicine*, 29, 1-2.

Miller, R. A. (1984). INTERNIST-/CADUCEUS: Problems facing expert consultant programs. *Meth. Inform. Med.*, 23, 9-14.

Nelson, S. J., Blois, M. S., Tuttle, M. S., Erlbaum, M., Harrison, P., Kim, H., Winkelmann, B., and Yamashita, D. (1985). Evaluating RECONSIDER, a computer program for diagnostic prompting. *Journal of Medical Systems*, 9(5/6), 379-388.

Norman, D. (1981). Categorization of actions slips. *Psychological Review*, 88, 1-15.

Norman, D. (1988). *The psychology of everyday things*. New York: Basic Books.

Parasuraman, R., Molloy, R., and Singh, I. (1993). Performance consequences of automation-induced "complacency". *International Journal of Aviation Psychology*, 3(1), 1-23.

Parasuraman, R., Mouloua, M., and Molloy, R. (1994). Monitoring automation failures in human-machine systems. In M. Mouloua and R. Parasuraman (Eds.), *Human performance in automated systems: current research and trends* (pp. 45-49). Hillsdale, NJ: Lawrence Erlbaum Associates.

Plugge, L., Verhey, F., and Jolles, J. (1990). A desktop expert system for the differential diagnosis of dementia. *International Journal of Technology Assessment in Health Care*, 6, 147-156.

Roth, E., Bennett, K., and Woods, D. (1988). Human interaction with an "intelligent" machine, *Cognitive Engineering in Complex Worlds*, (pp. 23-69), London: Academic Press.

Sassen, A., Buiël, E., and Hoegge, J. (1994). A laboratory evaluation of a human operator support system. *International Journal of Human-Computer Studies*, 40, 895-931.

Shamsolmaali, A., Collinson, P., Gray, T., Carson, E., and Cramp, D. (1989). Implementation and evaluation of a knowledge-based system for the interpretation of laboratory data. In *AIME '89*, (pp. 167-176).

Shortliffe, E. H. (1976). *Computer-based medical consultations: MYCIN*. New York: Elsevier.

Shortliffe, E. (1990). Clinical decision-support systems. In E. Shortliffe and L. Perreault (Eds.), *Medical informatics. Computer applications in health care* (pp. 466-500). New York: Addison-Wesley.

Shortliffe, E. H., Scott, A. C., Bischoff, M., Campbell, A. B., van Melle, W. and Jacobs, C. (1981). ONCOCIN: An expert system for oncology protocol management. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence, Vancouver, British Columbia*, pp. 815-822.

Silverman, B. G. (1992). Building a better critic. Recent empirical results. *IEEE Expert*, April, 18-25. Smith, P. J., Galdes, D., Fraser, J., Miller, T., Smith, J. W., Svirbely, J. R., Blazina, J., Kennedy, M., Rudmann, S., and Thomas, D. L. (1991). Coping with the complexities of multiple-solution problems: A case study. *International Journal of Man-Machine Studies*, 35, 429-453.

Smith, P. J., Miller, T., Gross, S., Guerlain, S., Smith, J., Svirbely, J., Rudmann, S., and Strohm, P. (1992). The transfusion medicine tutor: A case study in the design of an intelligent tutoring system. In *Proceedings of the 1992 Annual Meetings of the IEEE Society of Systems, Man, and Cybernetics*, (pp. 515-520).

Sutton, G. C. (1989). How accurate is computer-aided diagnosis? *The Lancet* (October 14, 1989), 905-908.

Verdaguer, A., Patak, A., Sancho, J., Sierra, C., and Sanz, F. (1992). Validation of the medical expert system PNEUMON-IA. *Computers and Biomedical Research*, 25, 511-526.

Weiner, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). NY: McGraw-Hill.

Wellwood, J., Johannessen, S., and Spiegelhalter, D. J. (1992). How does computer-aided diagnosis improve the management of acute abdominal pain? *Annals of the Royal College of Surgeons of England*, 74, 40-46.

Wyatt, J., Spiegelhalter, D. (1992). Field trials of medical decision-aids: potential problems and solutions. *Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care*. 3-7.