Designing a Calculational Logic Theorem Prover: Insight into Search Procedure via Eye Movements

Eric Aaron (aaron@cs.cornell.edu)
Department of Computer Science; Cornell University
Ithaca, NY 14853

Michael Spivey (mjs41@cornell.edu)
Department of Psychology; Cornell University
Ithaca, NY 14853

Abstract

We are designing and implementing an automated theorem prover that will in part attempt to simulate human performance on calculational logic theorem proving. To support this project, we recorded and analyzed people's eye movements while they constructed calculational proofs. Our findings confirm some expected behaviors (based on strategies and principles taught to students) that previously may have seemed untestable, such as the influence of the form of the current proof step and of syntax in general on the microcognition of the problem solver. The experiment also uncovered other interesting patterns, such as the seemingly inefficient but widely occurring tendency to attend to premises that are not used in the proof under consideration. Overall, we gained insights into microcognition that could not have been gained merely by studying written proofs. We expect these insights to directly impact the theorem prover under development, but they may also find a wider audience, appealing to educators and logicians who are familiar with calculational methods and student performance on calculational proofs. Our findings also support the notion that analyses of eye movements can improve our understanding of the way people perform some theorem proving tasks.

Introduction

Simulating intelligent human behavior has long been a goal of artificial intelligence. Understanding logical inference has long been a goal of psychology. Rarely, however, have these two goals overlapped in the field of automated theorem proving. Frequently, the psychology of inference has focused on very specific tasks, such as categorical syllogisms or card selection tasks (e.g., Wason & Johnson-Laird, 1972), or on broad ranges of behavior that have led to correspondingly broad projects such as SOAR (Newell, 1990). While some important and informative models of logical inference fall between these two extremes, such as PSYCOP (Rips, 1994) and the logic-oriented portion of the Mental Models theory from Johnson-Laird (1983), they are not truly models of theorem proving. Research continues to try to unite automated theorem proving and the psychology of inference, such as the work of Melis (1994) on deriving useful information for automated theorem provers by analyzing how mathematicians prove theorems.

We are attempting to bring together these fields of research by designing a theorem prover that, in part, approximates human behavior on calculational logic[1] proving tasks (Aaron, in progress; Aaron & Allen, in progress). Calculational methods are well established, used by mathematicians and computer scientists (Backhouse, 1995; Dijkstra & Scholten, 1990) and taught to students in logic and discrete mathematics courses (Gries & Schneider, 1993, 1995a, 1995b). They provide a syntactically oriented framework for expressing proofs that is more restricted than traditional proof formats such as natural language, a framework that we hope will readily permit applications of psychological methods to theorem prover design. The experiments reported in this paper demonstrate successful applications of eyetracking methods, providing results on which we may base cognitively plausible portions of our prover.

Fundamental to our philosophy in designing this prover is our focus on *the microstructure of cognition*, or *microcognition* (see Rumelhart, McClelland, et al., 1986). Roughly, in the context of distributed representations, microcognitive structures refer to the components of a larger cognitive structure. A similar notion applies in our project: Our goal is not to approximate human behavior solely by modeling principles on the macrostructural level of proofs themselves. Instead, we would like the proofs to emerge from components such as search and attention, which are microcognitive components of the full proving process. That is, in our context, the distinction between macrocognition and microcognition is reflected in the sorts of questions we ask, questions either on the macrocognitive level of "What are people writing?" or on the microcognitive level of "What are people thinking when they write?"

For this reason, any attempt to construct a descriptively accurate model of human performance on an inference task

[1] In using the word "calculational" to describe a logic, a proving task, or a related concept in this paper, as opposed to general calculational methods, we refer to the logic and associated methodology described in the text by Gries & Schneider (1993) and related papers.

merely by studying the coarsest levels of behavior — for instance, by studying solutions to exercises that students submit — will not provide insight into many aspects of attention and search that are important for the model. While nothing short of a full mental map would be comprehensively illuminating, if the inference task is visual in some significant way, analyzing the eye movements of participants can help experimenters glean some unique insight into information processing in general and into the real-time microcognition of inference in particular. (For recent demonstrations of the wide applicability of this methodology, see Ballard, Hayhoe & Pelz, 1995; Epelboim & Suppes, 1996; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995.)

But are such methods applicable to studies of logical mathematical inference? Although many traditional proof methods do not have significant visual components, there are some exceptions. The multimodal Hyperproof system (Barwise & Etchemendy, 1994; Cox, Stenning & Oberlander, 1994; Oberlander, Cox, Monaghan, Stenning & Tobin, 1996) adds graphical representations of the logical statements involved in a given proof. The calculational predicate logic from Gries & Schneider (1993) is also visually oriented in that a constrained visual search is an essential part of nearly every proof construction, making it possible for eyetracking to aid in understanding the microcognition of people using calculational logic.

Designing search processes and constraining search spaces are critical to the success of any automated theorem prover, particularly one with strong constraints on the output proofs. By analyzing eye movements, we were able to derive answers to some primary questions about how to model the search processes of people constructing calculational proofs. In particular, we focused on microcognitive *attention during search* — the extent of attention paid to various features during visual search — and answered some of the many questions that concern it. Although the design of automated theorem provers is typically more concerned with engineering considerations than descriptive accuracy, we now have a beginning of a methodology for genuine descriptive accuracy in our prover that could not have been achieved simply by studying written proofs. In this paper, we give some background on the calculational approach and the prover-in-progress, and we present results from our eyetracking studies about search in calculational logic.

Calculational Logic

Calculational logic came about as a formalization of general calculational methods, which attempt to emphasize simple syntactic calculation as much as possible in the course of problem solving (Backhouse, 1995). In this section, we review essentials of calculational logic that are necessary for this paper; by no means is it a complete specification.

We restrict our discussion of calculational logic to the inference rules and techniques that support a particular method of proving equalities using a chain of equality-preserving rewrites. The method is based on an observation about these calculational contexts with which all participants in our studies are familiar: If E and F are expressions that have been previously proved equal, then substituting F for E in an expression does not alter the value of that expression. For example, since $(P \Rightarrow Q) = (\neg P \vee Q)$, we have $(R \wedge (P \Rightarrow Q)) = (R \wedge (\neg P \vee Q))$.[2] Justified by an inference rule called "Leibniz", this substitution process permits the construction of proofs in a formal manner that has the feel of ordinary calculation. The practice emerged from work by computer scientists on the formal development of algorithms (Backhouse, 1995; Dijkstra & Scholten, 1990), but Gries & Schneider (1993) present it, along with heuristics and guidelines for students, as a general technique for logic and discrete mathematics. They believe that it is an excellent way for students to gain comfort with formalism and proof, building on previously learned intuitions and practices from algebra and calculation. Their college-level text provides material for a discrete mathematics course built around the method.

A typical calculational proof task for a student involves a statement to be proved, a list of premises (i.e., axioms and previously proved theorems), and whatever procedural knowledge the student brings to the task. Students are generally given the standard premise list (a theorem list from the back of the text) and not required to memorize it. The theorems are listed in the order of their presentation in the text and are thus roughly grouped by concept, with simpler theorems for a concept often presented first. By using heuristics and principles elucidated in the above-mentioned text, including pattern matching with the premise list, students decide how to proceed in constructing their proofs.

By a series of rewrites, each justified by one of the premises on the list, students prove a statement by either transforming the entire expression into a previously established theorem or by transforming one side of an equality into the other side. For example, a calculational proof step that establishes the equality $(R \wedge (P \Rightarrow Q)) = (R \wedge (\neg P \vee Q))$ might have the following form:

$$
\begin{array}{ll}
 & R \wedge (P \Rightarrow Q) \\
= & \langle P \Rightarrow Q \equiv \neg P \vee Q \rangle \\
 & R \wedge (\neg P \vee Q)
\end{array}
$$

where the expression in angle brackets, called a "hint", refers to the premise of the instance of Leibniz that yields the conclusion $(R \wedge (P \Rightarrow Q)) = (R \wedge (\neg P \vee Q))$. Hints can be

[2] The symbols $\neg$, $\vee$, $\wedge$, and $\Rightarrow$ stand for negation, disjunction, conjunction, and implication, respectively. We also use the symbol $\equiv$ in this paper to stand for *equivalence*, which is equality on truth-valued expressions except that it is read *associatively* and not *conjunctionally*. The symbol = is used conjunctionally, i.e., $A = B = C$ is understood to mean $(A = B) \wedge (B = C)$; the symbol $\equiv$ is used associatively, i.e., $A \equiv B \equiv C$ is understood to mean either $(A \equiv B) \equiv C$ or $A \equiv (B \equiv C)$, both of which have the same truth value for all Boolean A, B, and C. The logical operators have the following order of precedence: $\neg$; =; $\wedge$, $\vee$; $\Rightarrow$; $\equiv$. Note that = binds more tightly than $\equiv$, so $P \Rightarrow Q = \neg P \vee Q$ and $P \Rightarrow Q \equiv \neg P \vee Q$ have different meanings. Only the second one represents the traditional relation between implication and disjunction.

logical statements or names that refer to them. For instance, a student could have written "Definition Of Implication" as the hint in the example step given, since that is the name associated with the equivalence $P \Rightarrow Q \equiv \neg P \vee Q$. (This significant simplification of the full process is sufficient for our purposes.) The following is an example of a proof constructed by a participant in our studies:

$P \vee (P \equiv Q \vee Q \equiv P \vee Q) \equiv P \vee P$
= <Idempotency of $\vee$>
$P \vee (P \equiv Q \equiv P \vee Q) \equiv P$
= <Golden Rule>
$P \vee (P \wedge Q) \equiv P$
= <Absorption, 3.43b>
$P \equiv P$ — <Reflexivity of $\equiv$>

It proves that $P \vee (P \equiv Q \vee Q \equiv P \vee Q) \equiv P \vee P$ is a theorem by demonstrating that it is equivalent to the previously proved theorem Reflexivity of $\equiv$, $P \equiv P$.

Logic or discrete mathematics courses based on Gries & Schneider (1993) are offered at Cornell University and other schools. Gries, experienced in teaching such a course at Cornell, cites numerous cases of students overcoming their fears of formalism due to the calculational approach. Some students, however, have difficulties with what they feel is unnecessary formalism in the method. There is also an interesting criticism by those who say there is too little room for traditional ("semantic") comprehension of the tools of logic in a framework that stresses such a formal ("syntactic") approach and that there is no clear way to extend the calculational framework to incorporate conventional tools. Although there is certainly a tendency for increased reliance on pattern matching and syntactic methods, semantic knowledge is represented in this framework, and some of the elegant proofs that emerge from the approach (in Gries & Schneider (1993), for instance) are the results of both semantic knowledge and syntactic manipulation.

These points aside, the fact that calculational theorem proving is a general logical inference task with an intrinsic visual component and a standardized proof report format makes it well-suited for our research. As Aaron (1996) argues, this project occupies a sort of middle ground amid other inference research in cognitive science, a feasible basis for a computational model of human performance on a general-purpose proof task in which microstructural elements of the model could be empirically tested and determined prior to implementation.

Modeling Calculational Proving

Instead of implementing a full model and then testing it, we intend to combine experimentally-derived and stipulative portions into a framework general enough to permit piecewise refinement. It is for this reason that we refer to the general prover design as a *framework* and a particular implementation of all portions of this framework as a *model*. Our model is not yet fully specified. The present studies collected information for use in designing the search and attention processes of the model.

One necessary initial decision concerned the foundation upon which we would build this mathematical and procedural framework. We chose the Nuprl proof development system (Constable, et al., 1986) for this foundation; we feel it is capable of permitting us to represent both the cognitive microstructures and macrostructures necessary for our task. We defined a language for calculational logic using the facilities in Nuprl (Aaron, in progress; Aaron & Allen, in progress), and we have committed to using the *tactic* structure (based on Gordon, Milner & Wadsworth (1979)) of Nuprl for modeling behavior on this proving task. Briefly, a tactic is a program that describes how to carry out an inference in some already established logic, an abstraction away from the primitive rules of the logic to more general inference patterns. Tactics in Nuprl often correspond to human-level inferences, making it even more feasible to represent our model in Nuprl. As an example, the Nuprl tactic "*BackThruLemma* <lemma name> THEN *Auto*" is an augmented backchaining tactic that first matches the current goal against the lemma named by <lemma name>, resulting in new subgoals corresponding to the premises of that lemma, and then applies the tactic *Auto* (which performs a variety of simple, obvious inferences) to each of those subgoals. When applied to an assertion, it would fail in cases where no match was possible with the named lemma, and when it succeeded, it would result in either a completed proof or a new proof structure with new assertions to be proved.

Since the fundamental rule set of Nuprl can be altered and there are no restrictions on the content of tactics (within the framework presented), Nuprl is a sufficiently general foundation for our purposes. Flexibility and broad applicability are intended strengths of Nuprl, and we are reasonably confident that our choice will not unduly inhibit the cognitive plausibility or explanatory capacity of the resulting framework. We believe, for instance, that Nuprl can accommodate the observations and suggestions of Melis (1994) for incorporating human-like capabilities into an automated theorem prover. We are not confined to using search processes or similar features native to Nuprl. Instead, within its framework, we can implement our own procedures based on our experimental results.

Having selected the foundation in which we will implement our model, several sorts of primary questions emerge about modeling microcognition. When someone starts to prove an exercise, at what point do they begin searching the premise list? It is clear that the symbols present in the current statement of a problem and the goal of that problem affect the form of the final constructed proof, but do they also affect microcognition in analogous ways? Do people's strategies for attention during search change as they proceed through a proof, or do people start off with certain priorities in attention during search and then keep those priorities throughout a proof? How important is pattern matching to calculational logic? Does it influence people to the point where they attend to premises that they are very unlikely to use, or will people not attend to even strong pattern matches

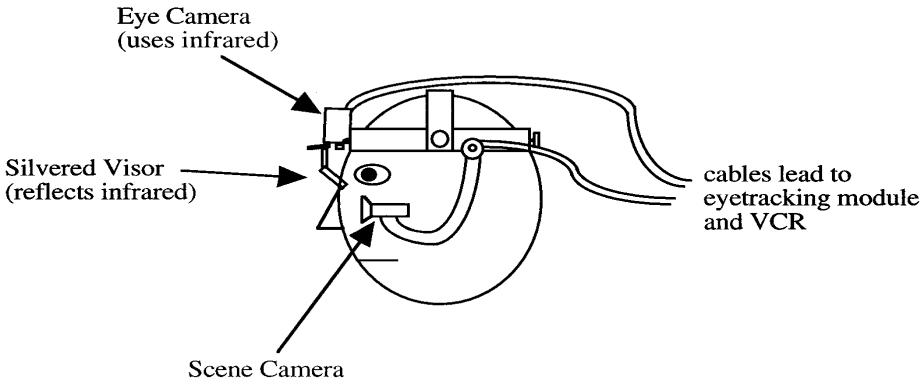


Figure 1. Diagram of the ISCAN headmounted eyetracker.

if other evidence suggests the premises matched by the pattern would not be useful? These sorts of basic questions might not be answerable on the microcognitive level except for eyetracking methods like those used in our investigations. The following sections describe the methods and experiments we used to help begin answering some of the above questions.

Method

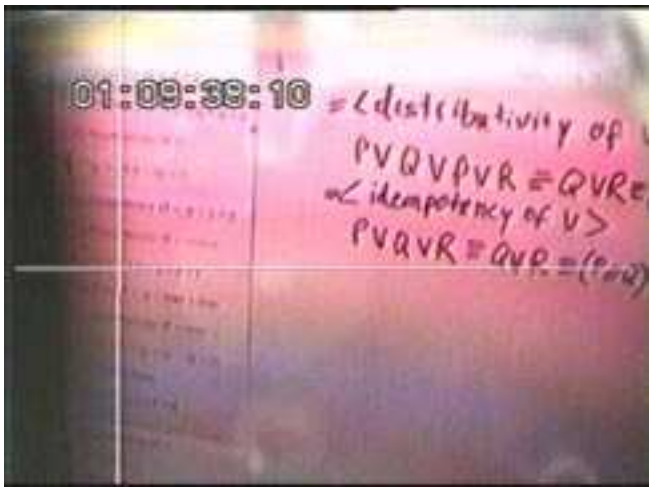
Eye movements were monitored by an ISCAN eyetracker mounted on top of a lightweight headband (Figure 1). The camera provided an infrared image of the left eye sampled at 60 Hz. The center of the pupil and the corneal reflection were tracked to determine the orbit of the eye relative to the head. A scene camera, yoked with the view of the tracked eye, provided an image of the subject's field of view. Gaze position (indicated by crosshairs) was superimposed over the scene camera image and recorded onto a Hi8 VCR with 30 Hz frame-by-frame playback (Figure 2). Accuracy of the gaze position record was about one degree of visual angle over a range of +/- 25 degrees. For purposes of determining *fixations* — instances where a participant's recorded glance on an object lasted long enough to indicate significant attention to that object and not an insignificant or random eye placement — we used a threshold of roughly 200ms, or six frames of video playback.

Participants were 15 Cornell students who had completed a course for which the Gries & Schneider (1993) book was the primary text. One was a TA; the other 14 had taken it as a standard course. Group A (n=8) was given problem 1(a) as their first problem, and group B (n=7) was given 1(b). (For some results, this distinction was not relevant, and we considered the 15 subjects as a group, undivided by this condition.) The other four problems given to participants to prove are listed here as numbers 2-5.

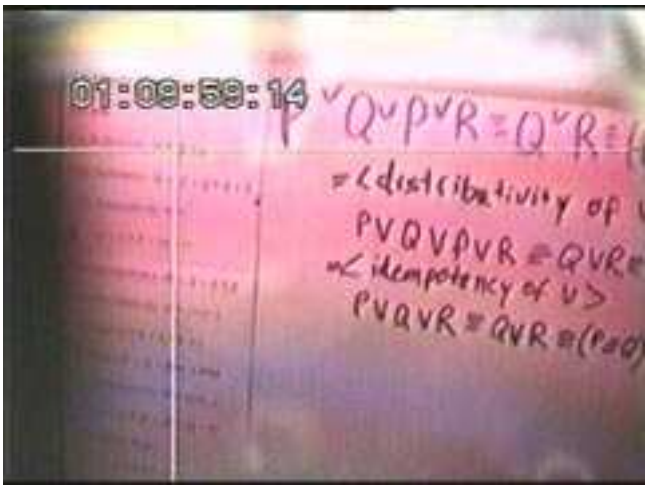
1 (a). $P \equiv P \equiv Q \equiv Q \equiv \mathit{true}$
(b). $P \equiv P \equiv Q \equiv Q \equiv \mathit{true} \vee P \vee Q$
2. $P \vee Q \vee P \vee R \equiv Q \vee R \equiv (P \equiv Q) \vee (Q \vee R)$
3. $P \vee (P \equiv P \vee Q) \equiv P \vee Q \equiv P$
4. $P \vee (P \equiv Q \vee Q \equiv P \vee Q) \equiv P \vee P$
5. $\mathit{true} \wedge Q \equiv \mathit{true} \vee Q \equiv Q$

Before encountering the eyetracking equipment, subjects had the option of working through a warm-up exercise, a proof with no significant overlap with features central to the studies. If needed, reminders were given about points of general technique, such as the order of precedence of operators, definitions of terminology, the process of instantiation of variables, and common proof formats. When subjects were ready, they were seated in front of a whiteboard with a premise list on its left (see Figure 2), and the eyetracking gear was calibrated for them. They were then read brief instructions about the task: They would be given five statements (one at a time) to prove, all the necessary premises were on the list to their left, and they should let the experimenter know when they were ready to move from one proof to the next.

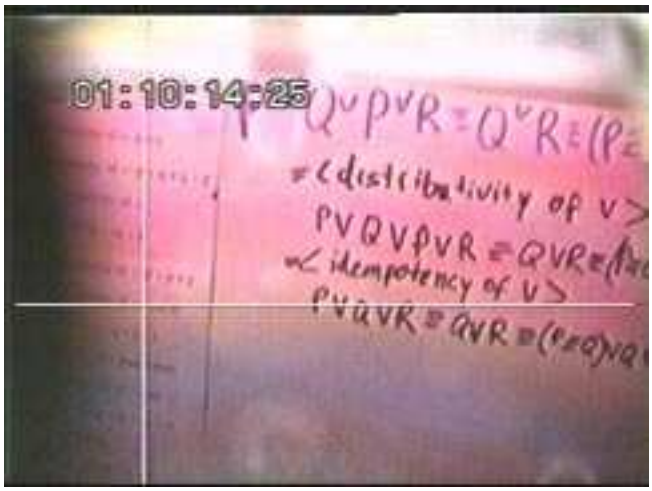
The premise list given to them (Figure 3) consisted of 18 theorems taken from the reference list of premises at the back of the text (Gries & Schneider, 1993). The premises, including the names and numbers labeling them, were presented exactly as written in the text. Above the premises at the top of the list was a brief note reminding subjects that Associativity of ≡ (an unlisted premise) could be used implicitly in the normal way and stating that the list contained all the premises available to them. The list consisted of theorems about the three logical operators that appeared in the statements to be proved, presented in the order in which they appeared in the text, effectively but not explicitly grouped into theorems about ≡ (equivalence, labeled by numbers (3.2-3.5)), ∨ (disjunction, labeled (3.24-3.31)), and ∧ (conjunction, labeled (3.35-3.50)). Theorems in a group mentioned only the operator used to classify them and operators that appeared in



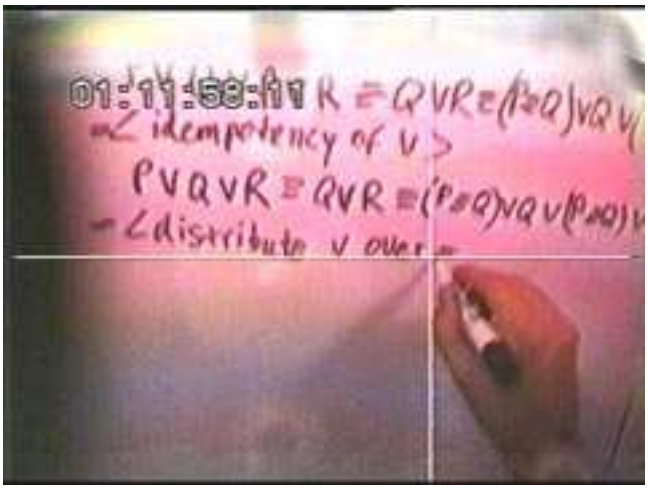
(a) Participant's gaze on Distributivity of ∨ over ≡.



(b) Participant's gaze on Reflexivity of ≡.



(c) Participant's gaze returns to Distributivity of ∨ over ≡.



(d) Having decided to use Distributivity of ∨ over ≡, the participant writes the hint in the proof step.

Figure 2. Videotaped images from the headmounted eyetracker. The participant is working on a proof he has already started, deciding what premise to use next. He looks at the one he will eventually use (Distributivity of ∨ over ≡), looks at another possibility, returns to Distributivity, and finally writes his choice. The white crosshairs indicate the particular place on the premise list (the list on the left, visible in images a, b, and c) or the workspace (the whiteboard on which he writes in image d) at which the participant is looking at the moment indicated by the timestamp.

preceding groups. There was no visual aspect of the list to lead subjects to divide them into groups that way; the spacing between theorems was uniform and no verbal cues were given. The list was chosen to be coherent, referring to only these three operators, and adequate to prove the five statements subjects were given. Although (as teachers know) it is possible to use almost any premise in an atypically creative proof, we expected that some of the theorems on the list would serve as distracters, not useful in any of the five proofs.

The experimenter presented the five statements to subjects by writing them on a whiteboard, erased the board when subjects indicated that they were ready to proceed to the next proof, and did not answer any questions while the experiment was in progress. This is consistent with standard practice on exams about this material, except for the eyetracking gear and the experimenter presenting the problems and clearing off the workspace, in an attempt to recreate classroom and homework conditions as much as possible. The experiments were videotaped and later analyzed for data collection. No audio was considered.

Experiment 1

A primary concern is the effect of the initial problem on attention during the initial search of the problem solver. While we know that different problems result in different solutions on the level of macrorepresentations (i.e., proofs),

(3.2) Symmetry of ≡: $p \equiv q \equiv q \equiv p$
(3.3) Identity of ≡: ***true*** $\equiv q \equiv q$
(3.4) ***true***
(3.5) Reflexivity of ≡: $p \equiv p$
(3.24) Symmetry of ∨: $p \vee q \equiv q \vee p$
(3.25) Associativity of ∨:
$(p \vee q) \vee r \equiv p \vee (q \vee r)$
(3.26) Idempotency of ∨: $p \vee p \equiv p$
(3.27) Distributivity of ∨ over ≡:
$p \vee (q \equiv r) \equiv p \vee q \equiv p \vee r$
(3.29) Zero of ∨: $p \vee$ ***true*** $\equiv$ ***true***
(3.31) Distributivity of ∨ over ∨:
$p \vee (q \vee r) \equiv (p \vee q) \vee (p \vee r)$
(3.35) Golden Rule:
$p \wedge q \equiv p \equiv q \equiv p \vee q$
(3.36) Symmetry of ∧:
$p \wedge q \equiv q \wedge p$
(3.38) Idempotency of ∧: $p \wedge p \equiv p$
(3.39) Identity of ∧: $p \wedge$ ***true*** $\equiv p$
(3.43a) Absorption: $p \wedge (p \vee q) \equiv p$
(3.43b) Absorption: $p \vee (p \wedge q) \equiv p$
(3.49) $p \wedge (q \equiv r) \equiv p \wedge q \equiv p \wedge r \equiv p$
(3.50) $p \wedge (q \equiv p) \equiv p \wedge q$

Figure 3. List of premises available to subjects.

we would like our model to function on the level of micro-representations, so we examined whether different initial problems resulted in different initial microcognitive behavior.

We compared groups A and B to determine whether the simple but non-trivial addition of an operator to the initial problem statement would affect the subjects' initial search of the premise list or their choice of initial step in that proof. We compared the number of fixations on Zero of ∨ (3.29), a theorem useful to group B but not group A, and disjunction-related theorems in general (3.24-3.31) by members of each group.

Results

In group A, none of the eight subjects (0%) made recorded fixations on theorem (3.29) when considering their first step in the proof. In contrast, in group B, four of the seven subjects (57%) made recorded fixations on theorem (3.29). We find similar results when considering all the disjunction theorems (3.24-3.31). In group A, three out of eight subjects (38%) made recorded fixations on one or more of the disjunction theorems during their initial search. In group B, all seven subjects (100%) made recorded fixations on one or more disjunction theorems during their initial search.

Notably, this did *not* result in a greatly increased usage of disjunction-oriented premises in the first step of the proofs of group B. Only one person in group B made an initial step that involved the added operator; the rest used an equivalence-oriented theorem, a step as viable for group A as for group B.

Discussion

The presence of the added operator significantly altered subjects' attention during initial search, answering one of the most primary questions about implementing our model: We cannot simply assume solvers will weigh all features the same independent of the particular problem; we must take the features of the current problem into account each time. While this may not seem surprising, reflecting problem-solving skills we expect the students to have, it is important to note that our observations support this answer on two different levels: They support more than the notion that different problems will result in different proofs; they also support the notion that different problems will result on some level in *differences in search strategies*, which (we presume) lead to the different proofs.

The observation that these differences in search strategies did not lead to different choices for a first proof step also serves to support an important assumption behind our research: This eyetracking paradigm can indeed yield results that are not achievable merely by examining written proofs. Subjects in group B attended to cues that they did not immediately use in constructing their proofs, so written proofs alone would not give enough information to answer the question "Do subjects attend to the disjunction operator right away, or do they postpone it?"

Experiment 2

Having investigated the effect of the initial problem on initial microcognitive behavior, a sensible next question is whether the changing form of a problem yields corresponding changes in microcognition as the solver works though a proof. The calculational method of proof by rewriting means that new symbols may emerge in the course of a proof, or symbols present at one time may be rewritten away and never reappear. This variation in the symbols present obviously leads to a variation in the premises that subjects actually *use* in their proofs. Does it also alter their behavior on the microcognitive level, changing attention during searches of the premise list in some significant way?

In this experiment, we examined whether the presence or absence of an operator in the current step of a proof had an observable effect on search-space constraints when constructing the next step of the proof. We considered instances where a subject performed a rewrite step that resulted in the complete and final elimination of an operator from the proof, such as eliminating the last ∧ in a statement, and this elimination step was not the conclusion of the proof. We then compared subjects' attention during the search immediately preceding the elimination step to their attention during the search immediately following it, analyzing the performance of all 15 subjects on all their work to see if this operator elimination affected the range of premises on which subjects made recorded fixations when considering the proof

step following the operator elimination. We considered only cases where we could make the necessary determinations with high confidence; poor eyetracking calibrations and similar obstacles resulted in the exclusion of some proofs from consideration.

Results

In the proofs, the only operators eliminated in rewrites were $\wedge$ and $\vee$, and we present the results accordingly. We found 19 proof steps in which a subject eliminated some operator by a rewrite. In one such step, two operators were eliminated simultaneously, so we consider that operator $\wedge$ was eliminated 8 times and operator $\vee$ was eliminated 12 times.

In the 8 proofs in which $\wedge$ was eliminated, subjects made recorded fixations on conjunction theorems 43% of the time immediately before the elimination of $\wedge$ and not at all after the elimination of $\wedge$. In the 12 proofs in which $\vee$ was eliminated, subjects made recorded fixations on disjunction theorems 84% of the time before the elimination of $\vee$ and 21% of the time after the elimination of $\vee$ (see Table 1).

Discussion

The elimination of an operator in a proof step had a notable effect on subjects' attention during search. This result combines with that of Experiment 1 to support an important, if expected, answer to a primary implementation question: Natural changes in problem structure, both as different initial problems and as refinements during a proof in progress, are accompanied by changes in the attention of the problem solver during search. (By a "natural change" in problem structure, we mean one that is not contrived, one that would occur, for instance, in the course of a student's doing a typical problem set.) People's search strategies are not static over the course of building a calculational proof, and the changes that occur are related in expected ways to the problems on which they are working. We find it satisfying to see such a strongly-held hypothesis empirically confirmed.

Table 1: Percentage of fixations on theorem groups before and after the elimination of an operator.

Elimination of $\wedge$:	$\equiv$	$\vee$	$\wedge$
before	25%	32%	43%
after	35%	65%	0%

Elimination of $\vee$:	$\equiv$	$\vee$	$\wedge$
before	8%	84%	8%
after	79%	21%	0%

Other Results

In addition to the interrelated results presented above, we observed other patterns in the search behavior of participants.

When Does Search Of The Premise List Begin?

It is hard to imagine an automated theorem prover starting to prove a theorem before receiving the entire problem as input. Would we need to sacrifice descriptive accuracy about people's microcognitive behavior to implement our prover that way? That is, do people begin to search for a solution based on partial information, or do they instead wait until the entire problem is presented?

Many factors come into play. In the context of this sort of problem solving, the principle "Before attempting to solve a problem, make sure you know what the problem is" is sometimes espoused; in terms of our experiments, this would translate at least partially to "Look at the whole problem before beginning to search the premise list." In different contexts, however, people regularly act in disagreement with this principle, using partial information and beginning to make inferences before they know the full inferential context. Which of those strategies applies to calculational logic? Moreover, the amount of time it takes to present a problem to a participant might also have an effect. If a problem was short and took very little time to write down, it is quite likely that subjects would not have time to look at the premises before the theorem was completely written. For longer problems or problems that took a great deal of time to present, it is quite plausible that subjects might begin searching the premise list before the theorem was completely presented.

We examined data from all 15 subjects for instances where we could determine with very high confidence whether they made any recorded fixations on the premise list after a non-trivial portion of a problem (i.e., more than just a single variable) was presented and before that entire problem was presented. We found 35 such proofs. In only 6 out of these 35 (17%) did subjects make recorded fixations on the premise list before the problem was completely presented and after sufficient information was available for a non-trivial pattern-match. Despite occasional pauses (on the order of seconds long) by the experimenter in the process of presenting their problems, subjects did not typically search the premise list until the entire problem was presented.

Fixations On Unused Premises

From our recorded fixations on theorems associated with disjunction (the range on which our eyetracking calibration was most reliable), we observed a significant number of fixations on premises that were infrequently cited in proofs (n=73) across all 15 subjects (see Table 2). Through their work, students become aware on some level that some premises are cited frequently and others extremely rarely in proofs, and one might expect them to pay less attention in visual search to the rarely cited premises. This experiment shows otherwise.

Table 2: Percentages of proofs in which disjunction theorems were fixated upon and used.

Theorem	Fixated	Used
3.24 (Symmetry of $\vee$)	71%	15%
3.25 (Associativity of $\vee$)	85%	4%
3.26 (Idempotency of $\vee$)	88%	58%
3.27 (Distrib. of $\vee$ over $\equiv$)	90%	41%
3.29 (Zero of $\vee$)	64%	32%
3.31 (Distrib. of $\vee$ over $\vee$)	62%	5%

We would not be surprised to see subjects making recorded fixations on each premise for perhaps as many as 50% of the proofs. Even given their familiarity with the premise list before the experiment begins, we would expect subjects to look at every premise at least once and very likely twice in the span of the experiment. (Some participants did not do all 5 of the proofs.) The fact that all the percentages of fixations here are higher than 60% suggests to us that participants paid more than minimal attention to all the disjunction premises, considering them all potentially useful. As further background, it should be noted that Symmetry of $\vee$, Associativity of $\vee$, and Distributivity of $\vee$ over $\vee$ are cited infrequently during a calculational logic course; teachers encourage students to use Symmetry and Associativity implicitly without citing them, and our experiment did nothing to discourage that practice, encouraging natural behavior. Distributivity of $\vee$ over $\vee$ is usually unused because when confronted with the expression $P \vee (Q \vee R)$, people generally (implicitly) use Associativity to get $P \vee Q \vee R$, and when given $(P \vee Q) \vee (P \vee R)$, people generally (implicitly) use Associativity and Symmetry and (explicitly use) Idempotency of $\vee$ to reduce the expression to $P \vee Q \vee R$, entirely bypassing Distributivity. So we are not surprised to note that none of these three theorems were widely used.

What we did find significant was the amount of attention paid to these theorems, given their general inutility, supporting the importance of pattern matching on the microcognition of calculational proving to an extent we did not expect. The highly-observed but largely uncited premises in the range (3.24-3.29) have a strong feature-match with many of the expressions that resulted when constructing the proofs in this experiment. Premises (3.24) and (3.25) are at the top of the disjunction section, with (3.24) therefore being the first premise to feature-match with the disjunction operator in a serial, top-down search and (3.25) being the first premise to match both disjunction and parentheses, two common and frequently co-occurring features, in a serial search. Premise (3.31) may seem an exception to the influence of serial search, but it has an extremely strong feature match with one of the most useful and frequently used premises, Distributivity of $\vee$ over $\equiv$ (3.27), a match extending even to the names of the theorems (which were present on the premise list). Hence, as encouraged by calculational logic, it appears that syntactic feature matching plays a very important role in shaping the search processes of subjects at this level of expertise.

General Discussion

Some of the results are satisfying, if not surprising, to people familiar with calculational logic, confirming for the first time their intuitions about students' microcognitive behavior. Other results are more unexpected, such as the extent of attention paid to premises that are not immediately used or not likely to be used at all. These may be particularly interesting to educators and others concerned with understanding how students use calculational methods in theorem proving.

Another kind of conclusion also emerged from this research: Eyetracking studies can indeed provide insight into the microcognitive processes of people performing logical inference tasks that experimenter intuition and studies of written output alone could not provide. For instance, upon hearing our results about fixations on unused premises, Gries made the conjecture that more experienced users of calculational logic would make fewer fixations on such unused premises. While we believe this is an interesting and likely true conjecture, we feel it is even more noteworthy that the groundwork now exists for expressing and resolving it. This application of eyetracking methods does more than just provide information for the design of an automated theorem prover; it seems to expand the range of psychological and educational questions that can be fruitfully posed and answered. If researchers wish to construct a descriptively accurate model of people's performance on other suitable visually-oriented proof tasks, we believe this hybrid eyetracking-before-implementation paradigm would yield similarly useful results.

At the current stage of this project, we have made no effort to search for or control for differences among subjects based on the professor who taught them. There may be systematic similarities within groups distinguished by teacher, on the levels of either eye movements or macrocognitive behavior, but our experiments were not tuned to this dimension. General data, across all classroom backgrounds, are adequate for this first investigation of a calculational logic framework.

We are convinced from our first exploration of this experimental paradigm that it offers valuable information for use in better understanding student performance and implementing a model that can simulate that performance on calculational logic. We intend to extend this research with experiments geared toward further elaborating the calculational proving framework. Based around more difficult statements to prove and with a wider range of skill levels among subjects, we expect future eyetracking data to be even more rewarding.

Acknowledgments

We would like to thank David Gries for his many contributions to this paper. We would also like to thank Stuart Allen for his comments and assistance. The first author was supported by NSF grant GER-9454149.

References

Aaron, E. (1996). *Middle ground: A place for C-proofs in cognitive science?* Presentation at the Cornell Cognitive Studies Graduate Research Forum. URL http://www.cs.cornell.edu/home/aaron/forum96.ps

Aaron, E. (in progress). An implementation of calculational logic inference.

Aaron, E. & Allen, S. F. (in progress). Justifying calculational logic by a conventional metalinguistic semantics.

Backhouse, R. (Ed.) (1995). *Information Processing Letters, Special Issue on The Calculational Method*, Vol. 53, No.3.

Ballard, D., Hayhoe, M. & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7*, 68-82.

Barwise, J. & Etchemendy, J. (1994). *Hyperproof.* CSLI Lecture Notes. Chicago: Chicago University Press.

Constable, R. L., et al. (1986). *Implementing mathematics with the Nuprl proof development system.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Cox, R., Stenning, K. & Oberlander, J. (1994). Graphical effects in learning logic: reasoning, representation and individual differences. *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (pp. 237-242). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dijkstra, E. W. & Scholten, C. S. (1990) *Predicate calculus and program semantics*. New York: Springer-Verlag.

Epelboim, J. & Suppes, P. (1996). Window on the mind? What eye movements reveal about geometrical reasoning. *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (p. 59). Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, M., Milner, A. & Wadsworth, C. (1979). Edinburgh LCF: *A mechanized logic of computation, Lecture notes in Computer Science, Vol. 78*. New York: Springer-Verlag.

Gries, D. (1981) *The science of programming*. New York: Springer-Verlag.

Gries, D. & Schneider, F. B. (1995a). A new approach to teaching discrete mathematics. *PRIMUS V, 2*, 113-138.

Gries, D. & Schneider, F. B. (1995b). Teaching math more effectively through calculational proofs. *The Mathematical Monthly, 102*, 691-697.

Gries, D. & Schneider, F. B. (1993). *A logical approach to discrete math.* New York: Springer-Verlag.

Johnson-Laird, P.N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Melis, E. (1994). How mathematicians prove theorems. *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (pp. 624-628). Hillsdale, NJ: Lawrence Erlbaum Associates.

Newell, Allen. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Oberlander, J., Cox, R., Monaghan, P., Stenning, K. & Tobin, R. (1996). Individual differences in proof structures following multimodal logic teaching. *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 201-206). Mahwah NJ: Lawrence Erlbaum Associates.

Rips, L. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., et al. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* Cambridge, MA: MIT Press.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science, 268*, 1632-1634.

Wason, P. C. & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* Cambridge, MA: Harvard University Press.