Insight into Theorem Proving via Eye Movements

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

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

Abstract

We are implementing an automated theorem proving system that will in part attempt to simulate human performance on calculational logic. 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 moment-by-moment computations of the problem solver. The experiment also uncovered other interesting patterns, such as the seemingly inefficient but widely occurring tendency to attend to particular premises despite their not being used in the proof under consideration. Overall, we gained insights into real-time problem solving that directly apply to our automated system and could not have been gained merely by studying written proofs. Our findings also demonstrate that analyses of eye movements can improve our understanding of the psychology of theorem proving.

Introduction

We are interested in the psychology of logical inference and its applicability to the field of automated theorem proving, an interface with many interesting questions that have been largely unexplored by past research. 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 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 (1993), they are not specifically 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 mathematicians' written proofs and their verbal recall of the proving process.

We are investigating another natural way to join the two fields: We are designing an automated reasoning system for calculational logic that will, in part, attempt to incorporate a model of some aspects of human theorem proving behavior (Aaron & Spivey, 1998). As part of this, we are exploring the moment-by-moment computations that subserve theorem proving. To truly reflect and simulate human cognitive behavior, we must investigate not only macrocognition on the level of "What are people writing?" but also microcognitive questions such as "What are people thinking before and when they write?" or, more specifically, "Where is attention directed while people search to decide the next proof step?"

We decided to focus on calculational logic theorem proving (an extension of common calculational techniques learned by all algebra students) for several reasons, some of which we discuss later in this paper. Particularly important is the fact that calculational logic is visually-oriented, in that a constrained visual search is an essential part of every proof construction by students. Therefore, we might be able to answer some of our microcognitive questions using eyetracking methods. Although untested in the area of theorem proving, analyzing eye movements has helped experimenters in many fields of study 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; Hegarty, Mayer & Green, 1992; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995.)

We began our explorations with some primary questions about the process of theorem proving, particularly focusing on *attention during search*, i.e., the extent of attention paid to various features during visual search. Our experiments uncovered information of practical importance for our automated reasoning system, verifying expected answers — passing a baseline test that eyetracking results are informative in this context — as well as pointing to some more surprising behavior and interesting questions for future study. In this paper, we briefly describe calculational logic, present results from our eyetracking studies, and discuss possible extensions of our research.

Calculational Logic

For this brief and somewhat simplified introduction, 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 $(R \wedge (P \Rightarrow Q)) = (R \wedge (\neg P \vee Q))$.[1] This substitution process is exemplary of the well-established calculational methods used by mathematicians and computer scientists (Backhouse, 1995; Dijkstra & Scholten, 1990), permitting proof construction in a formal manner that has the feel of ordinary calculation. Gries & Schneider (1993, 1995) present it in their college-level text, along with heuristics and guidelines, as a general technique for calculational logic and discrete mathematics.

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 reference 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. 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 their 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:

$R \wedge (P \Rightarrow Q)$
$=$ <$P \Rightarrow Q \equiv \neg P \vee Q$>
$R \wedge (\neg P \vee Q)$

where the "hint" in angle brackets refers to the premise that justifies the conclusion $(R \wedge (P \Rightarrow Q)) = (R \wedge (\neg P \vee Q))$. Hints can be expressions or names that refer to them. For instance, a student could have written "Definition Of Implication" (a name for the equivalence $P \Rightarrow Q \equiv \neg P \vee Q$) as the hint in the example. 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 $\equiv$, $P \equiv P$.

Within this task domain, several microcognitive questions emerge about attention during search. Using eyetracking methods, we begin to answer some of these questions.

Method

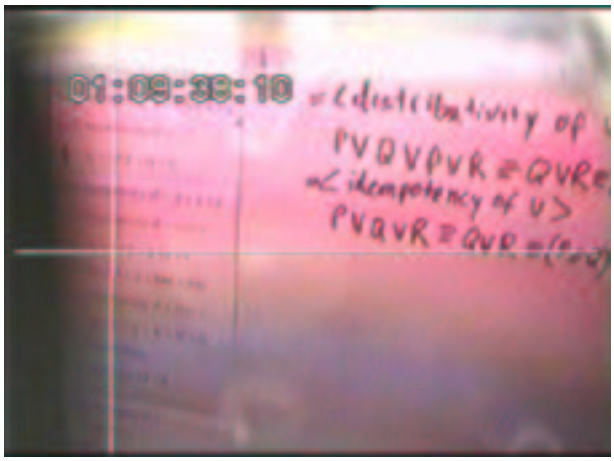
Eye movements were monitored by an ISCAN eyetracker mounted on top of a lightweight headband. The camera provided an infrared image of the left eye sampled at 60 Hz. The center of the pupil and the corneal reflection were tracked to determine the orbit of the eye relative to the head. A scene camera, 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 (see Figure 1). 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 artifact from a blink or an intermediate track during a saccade— 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. 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.

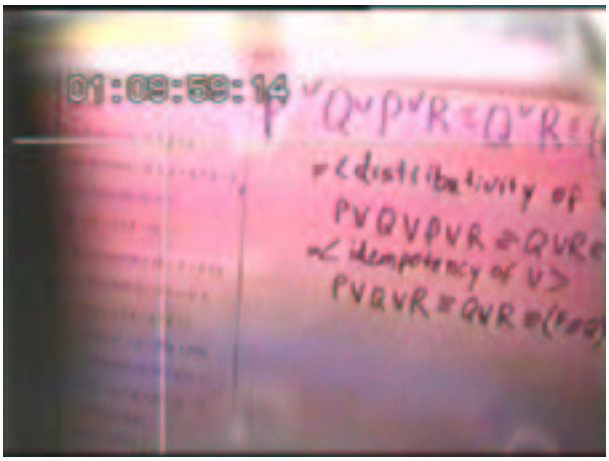
1 (a). $P \equiv P \equiv Q \equiv Q \equiv$ ***true***
(b). $P \equiv P \equiv Q \equiv Q \equiv$ ***true*** $\vee P \vee Q$
2. $P \vee Q \vee P \vee R \equiv Q \vee R \equiv (P \equiv Q) \vee (Q \vee R)$
3. $P \vee (P \equiv P \vee Q) \equiv P \vee Q \equiv P$
4. $P \vee (P \equiv Q \vee Q \equiv P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

Before wearing 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 and

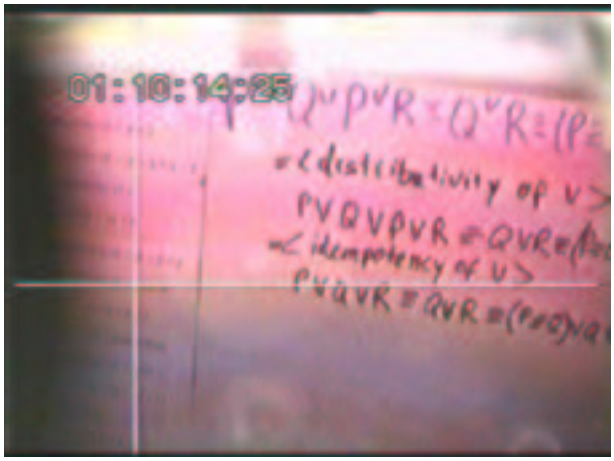
[1] The symbols $\neg$, $\vee$, $\wedge$, and $\Rightarrow$ stand for negation, disjunction, conjunction, and implication, respectively. We also use the symbol $\equiv$ in this paper to stand for *equivalence*, which is equality on truth-valued expressions except that it is read *associatively* and not *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.



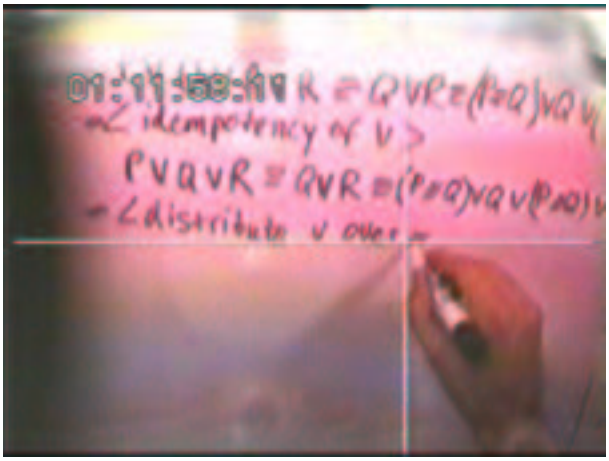
(a) Participant's gaze on Distributivity of ∨ over ≡.



(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 1. 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 (showing hours, minutes, seconds, and frames).

definitions of terminology. When subjects were ready, they sat in front of a whiteboard with a premise list on the left (see Figure 1), and the eyetracking gear was calibrated for them. They were then read brief instructions: They would be given five statements (one at a time) to prove, all the necessary premises were on the list, and they should let the experimenter know when they were done with a proof.

The premise list given to them (Figure 2) contained 18 theorems taken from the reference premise list from Gries & Schneider (1993), presented and labeled exactly as written in the text. The given 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 preceding groups. There was nothing to lead subjects to divide the premises into groups; the spacing between theorems was uniform and no verbal cues were given. The list was 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, and did not answer questions while the experiment was in progress. This is consistent with standard practice on exams about this material, in an attempt to recreate natural conditions as much as possible. The experiments were videotaped and later analyzed for data collection. No audio was considered.

(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 2. List of premises available to subjects.

Experiment 1

Clearly, different problems result in different solutions on the level of macrorepresentations (e.g., full proofs). We are also interested in microrepresentations (e.g., syntactic features considered while constructing a proof), however, 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 first 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 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%) fixated 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%) fixated one or more of the disjunction theorems during their initial search. In group B, all seven subjects (100%) fixated one or more disjunction theorems during their initial search.

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

Discussion

The added operator significantly altered subjects' attention during initial search, empirically answering a fundamental question: We cannot simply assume solvers will weigh all features the same independent of the particular problem; our model must take the features of the current problem into account each time. On different problems, students do not *pay attention* to the same features and then *use* them differently, resulting in the different proof. Instead, students actually attend to different features. This is the kind of distinction that can only be made on the microcognitive level.

The observation that these differences in attention did not lead to different choices for a first proof step also confirms an important assumption behind our research: This eyetracking method does indeed yield results that are not achievable merely by examining written proofs. 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

In this method of proof by rewriting, 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 this experiment, we examined whether the presence or absence of an operator had an observable effect on search-space constraints when constructing the next step of a proof. We considered instances where a subject performed a rewrite step that resulted in the 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. We analyzed all 15 subjects on all their work to see if this operator elimination affected the range of premises that subjects fixated when considering the proof step following the operator elimination. We considered only cases where we could make the necessary determinations with high confidence; due to poor eyetracking calibrations and similar obstacles, we excluded some proofs from consideration.

Results

In the proofs, the only operators eliminated in rewrites were ∧ and ∨, 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 ∧ was eliminated 8 times and operator ∨ was eliminated 12 times.

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

Elimination of ∧:	≡	∨	∧
before	25%	32%	43%
after	35%	65%	0%

Elimination of ∨:	≡	∨	∧
before	8%	84%	8%
after	79%	21%	0%

In the 8 proofs in which ∧ was eliminated, subjects fixated conjunction theorems 43% of the time immediately before the elimination of ∧ and not at all after the elimination of ∧. In the 12 proofs in which ∨ was eliminated, subjects fixated disjunction theorems 84% of the time before the elimination of ∨ and 21% of the time afterwards (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 experimentally support an important answer to a primary question: Natural changes in problem structure, both as different initial problems and as refinements during a proof in progress, are indeed accompanied by changes in attention 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 attention during search is not static over the course of a calculational proof, and the changes that occur are related in expected ways to the problems on which they are working.

Other Results

In addition to the interrelated results presented above, we observed other patterns in search behavior. We present one such result here; Aaron & Spivey (1998) contains a somewhat extended account of our observations.

Fixations On Unused Premises

In their coursework, students become aware that some premises are cited frequently and others extremely rarely in proofs. One might expect them to attend less to the rarely cited premises. Across all 15 subjects, however, there were many fixations on disjunction premises (the range on which our eyetracking calibration was most reliable) that were infrequently cited in proofs (see Table 2).

We would not be surprised to see subjects fixating each premise for perhaps as many as 50% of the proofs. Even given their familiarity with the premise list before the experiment, we would expect subjects to fixate every premise at least once and very likely twice in the span of the experiment. 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.

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

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

As further background, it should be noted that Symmetry of ∨, Associativity of ∨, and Distributivity of ∨ over ∨ are cited infrequently during calculational logic coursework. 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 ∨ over ∨ is also usually uncited: With the expression P ∨ (Q ∨ R), people generally (implicitly) use Associativity to get P ∨ Q ∨ R; with (P ∨ Q) ∨ (P ∨ R), people generally (implicitly) use Associativity and Symmetry and (explicitly use) Idempotency of ∨ to reduce the expression to P ∨ Q ∨ R, entirely bypassing Distributivity. So we are not surprised to note that none of these three theorems was widely used.

We did find the amount of attention paid to these theorems noteworthy, given their general inutility, suggesting the importance of pattern matching on the microcognition of calculational proving to an unexpected extent. The frequently-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 a strong feature match with one of the most useful and frequently used premises, Distributivity of ∨ over ≡ (3.27), a match extending even to the names of the theorems (which were present on the premise list). Hence, beyond even the expected emphasis that calculational logic places on it, pattern matching appears to play a surprisingly important role in attention during search for subjects at this level of expertise.

General Discussion

Some of our experimental 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 ever be cited at all. The answers to such previously unexplored questions are all of value to us as support for our attempt to incorporate empirically verified facts about human cognitive processing in an automated reasoning system. In a larger sense, however, our results also provide more evidence that eye movements are closely related to moment-by-moment cognitive processes. Indeed, our eyetracking studies yielded insight into the microcognition of theorem proving that studies of written output alone could not provide, paving the way for new applications of this experimental procedure.

For instance, consider the result about attention to uncited premises. Our subjects, all of whom were relative novices to calculational logic, generally paid attention to unused premises. Would experts demonstrate the same behavior? A method such as ours, capable of microcognitive investigation, permits the natural continuation from the results in this paper into questions of expert/novice distinctions. This could prove useful for our automated reasoning system —in which we might hope to model different levels of expertise— but it is also of independent interest to the greater field of cognitive science. Thus, we see our application of eyetracking methods as both answering existing questions and opening new ranges of questions in the psychology of theorem proving.

Although we expect that our methods would prove useful in other theorem proving domains, such as the Hyperproof system (Barwise & Etchemendy, 1994; Cox, Stenning & Oberlander, 1994) or other sufficiently visually-oriented systems, there were several reasons for choosing calculational logic for this initial eyetracking investigation. One reason was its relation to the tremendously common calculation learned by many students. Another reason is that it was a natural extension of independent work on understanding and developing automated tools for calculational logic (Aaron, in progress; Aaron & Allen, in progress). Some other reasons are technical, such as the added complexities in calculational logic compared to natural deduction and pure predicate logic; these reasons are not relevant to or evident in this paper. Some are due to the design of calculational logic, designed to capture actual practice in using formalism and provide a standardized set of heuristics so that the natural behavior of our experimental participants would not be jarringly unexpected to us. The very presence of that subject population at Cornell was another reason. Thus, although we do not believe that the ability of this approach to open and answer new questions is limited to the calculational domain, there were many compelling reasons to choose calculational logic for this initial microcognitive investigation. Indeed, based on preliminary explorations of our experimental paradigm, we believe that both other proof methods and deeper aspects of calculational proof are bases for fruitful extensions to our research.

Acknowledgments

We are grateful for the many helpful comments made by readers of earlier versions of this paper. We would particularly like to thank David Gries and Stuart Allen for their many contributions. The first author was supported by NSF grant GER-9454149.

References

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.

Aaron, E. & Spivey, M. (1998). *Designing a Calculational Logic Theorem Prover: Insight into Search Procedure via Eye Movements.* Technical Report TR98-1680, Computer Science Department, Cornell University.

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.

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.

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

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

Hegarty, M., Mayer, R. & Green, C. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology*, *84*, 76-84.

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.

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