Hyperproof : abstraction, visual preference and multimodality

Hyperproof: abstraction, visual preference and multimodality

Jon Oberlander Keith Stenning Richard Cox

1 Introduction

We have been carrying out an evaluation of the effects of teaching logic with Barwise and Etchemendy's (1994) Hyperproof, a program for teaching first-order logic (see Figure 1). Inspired by a situation-theoretic approach to heterogeneous reasoning, it uses multimodal (graphical and sentential) methods, allowing users to transfer information to and fro, between modalities (see Figure 2). One of our major findings has been that individual differences between students have a significant effect on students' responses to Hyperproof: their prior cognitive style influences both the overall effectiveness of the teaching regime, and the actual proof structures that students produce under exam conditions (cf. Monaghan 1995; Stenning, Cox and Oberlander 1995; Oberlander, Cox and Stenning 1996; Oberlander et al. 1996). In the course of this larger study, we have built up a substantial corpus of 'hyperproofs'. We believe that this corpus can provide a detailed insight into various questions concerning the paths which students follow in their pursuit of proof goals.

In particular, Barwise and Etchemendy designed Hyperproof to support *heterogeneous* reasoning, in which information from differing modalities—sentential and graphical—is combined, or transferred from one modality to another. It is obviously, therefore, a multimodal system containing a visual sub-system. But given that one group of students benefits particularly from being taught with Hyperproof, we can ask: do they do well because it is a *visual* logical system, or do they do well because it is *multimodal*?

To address this question, we first frame some hypotheses concerning the relation between the individual differences in teaching outcome which we found, and the structures to be found in students' proofs; the rest of the paper then focusses on the second of these hypotheses. As background, we outline the relevant aspects of the design of the main study, indicating how it distinguishes two styles of student. We then describe (i) the way 'proofograms' are used to track the way students deal with abstractions; and (ii) the application of bigram and trigram analyses of rule use patterns in the data corpus, demonstrating that the differing styles of student end up producing multimodal proofs of distinctive types.

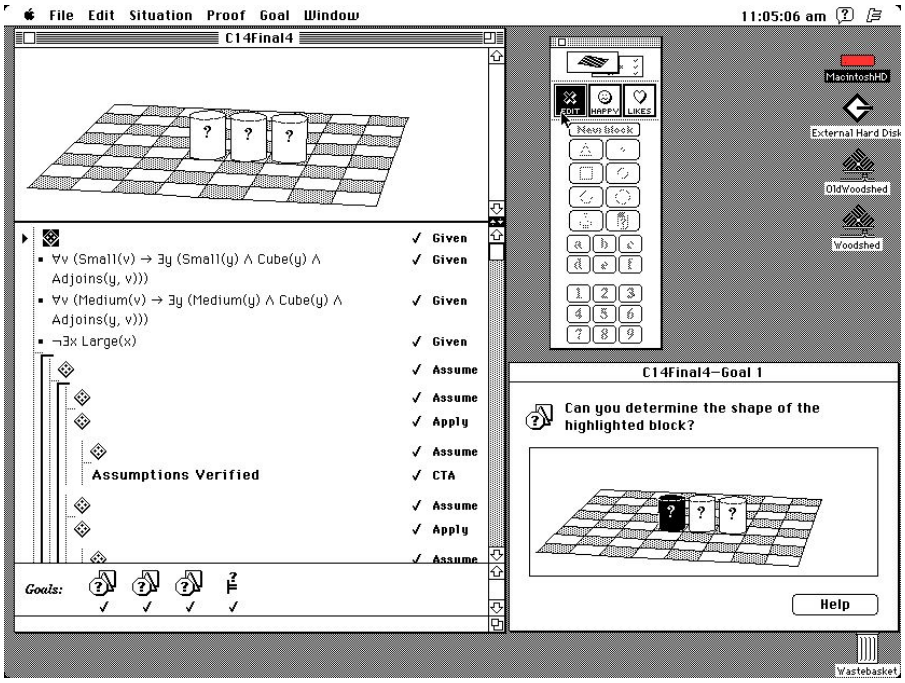

Figure 1: The Hyperproof Interface. The main window (top left) is divided into an upper graphical pane, and a lower calculus pane. The tool palette is floating next to the main window, and other windows can pop up to reveal a set of goals which have been posed.

Apply Extracts information from a set of sentential premises; expresses it graphically

Assume Introduces a new assumption into a proof, either graphically or sententially

Observe Extracts information from the situation; expresses it sententially

Inspect Extracts common information from a set of cases; expresses it sententially

Merge Extracts common information from a set of cases; expresses it graphically

Close Declares that a sentence is inconsistent with either another sentence, or the current graphical situation

CTA (Check truth of assumptions) Declares that all sentential and graphical assumptions are true in the current situation

Exhaust Declares that a part of a proof exhausts all the relevant cases

Figure 2: A set of relevant Hyperproof rules.

2 Hypotheses

The observation that graphical systems require certain classes of information to be specified goes back at least to Bishop Berkeley. Elsewhere, we have termed this property 'specificity', and argued that it is useful because inference with specific representations can be very simple (Stenning and Oberlander 1991, 1995). We have also urged that actual graphical systems do allow abstractions to be expressed, and it is this that endows them with a usable level of expressive power. Thus, Hyperproof maintains a set of abstraction conventions for objects' spatial or visual attributes. As well as concrete depictions of objects, there are 'graphical abstraction symbols', which leave attributes under-specified: the *cylinder*, for instance, depicts objects of unknown size (see Figure 1). A key step, then, in mastering an actual graphical system is to learn which abstractions can be expressed, and how.

As we describe below, our pre-tests independently allowed us to divide subjects into two cognitive style groups, on the basis of their performance on a certain type of problem item. Loosely, one group is 'good with diagrams', and the other less so. The good diagrammers turned out to benefit more from Hyperproof-based teaching than the others. Our belief is that those who benefit most from Hyperproof do so because they are better able to manipulate the graphical abstractions it offers. Call this view the *abstraction ability hypothesis*. Elsewhere, we have provided evidence in support of it (Oberlander et al. 1996).

That evidence also bears on the question in hand. Whether Hyperproof's virtue lies in its visual nature, or in its multimodality depends upon whether abstraction ability is supported by Hyperproof's visual representations—or by some other aspect of the system. One hypothesis is that the good diagrammers are simply those subjects who have a preference for the visual modality. Call this view the *visual preference hypothesis*. Another explanation would be that good diagrammers are those who are adept at translating between modalities. Call this view the *transmodal hypothesis*.

In what follows, we aim to show that the balance of evidence favours the transmodal hypothesis.

3 Distinguishing cognitive styles

In the full study, two groups of subjects were compared; one ($n = 22$ at course end) attended a one-quarter duration course taught using the multimodal Hyperproof. A comparison group ($n = 13$ at course end) were taught for the same period, but in the traditional syntactic manner supplemented with exercises using a graphics-disabled version of Hyperproof.

Subjects were administered two kinds of pre- and post-course paper and pencil test of reasoning. The first of these is most relevant to the current discussion. It tested 'analytical reasoning' ability, with two kinds of item derived from the GRE scale of that name (Duran, Powers and Swinton 1987). One subscale consists of verbal reasoning/argument analysis. The other subscale consists of

items often best solved by constructing an external representation of some kind (such as a table or a diagram). We label these subscales as 'indeterminate' and 'determinate', respectively. Scores on the latter subscale were used to classify subjects within both Hyperproof and Syntactic groups into DetHi and DetLo sub-groups. The score reflects subjects' facility for solving a type of item that often is best solved using an external representation; DetHi scored well on these items; DetLo less well. For the moment, we may consider DetHi subjects to be more 'diagrammatic', and DetLo to be less so.

4 Abstraction ability results

Both the Hyperproof and Syntactic groups contained DetHi and DetLo sub-groups. All subjects sat post-course, computer-based exams, although the questions differed for the two groups, since the Syntactic group had not been taught to use Hyperproof's systems of graphical rules. Student-computer interactions were dynamically logged, permitting a full, step-by-step, reconstruction of the process of the subject's reasoning, as well as capturing the final proof produced.

Here, we discuss only the final proofs produced by the 22 Hyperproof subjects, all of whom completed the exams. The four questions that these students were set contained two types of item: determinate and indeterminate. Here, determinate problems were taken to be those whose problem statement did not utilise Hyperproof's abstraction conventions. That is: determinate problems contained only concrete depictions of objects in their initially given graphical situation, whereas indeterminate problems—such as that in Figure 1—could contain graphical abstraction symbols in the initial situation.

4.1 Proofograms

What evidence is there for the abstraction ability hypothesis? Among the Hyperproof students, do the two sub-groups—DetHi and DetLo—use graphical abstraction symbols in characteristically different ways?

We can score each step of each proof on the basis of number of concrete situations compatible with the graphical depiction; one possible scoring method is described in Oberlander, Cox and Stenning (1996). A low score always indicates more abstraction; a higher score indicates more concreteness.

We can explore the way concreteness varies through the course of a proof by graphing it against the hierarchical structure of the proof. We call such graphs 'proofograms'. Figures 3 and 5 show how subjects C2 and C14 tackle an indeterminate exam question; Figures 4 and 6 give their proofograms. The visual differences between proofograms are quite striking: one group is 'spikey'—as in Figure 4; and the other is 'layered'—as in Figure 6. The differences are particularly pronounced on indeterminate questions, and Q-sort tests indicate that these questions reliably elicit layered proofs from DetHi subjects, and spikey proofs from DetLo (cf. Oberlander et al. 1996). The basic message appears to be that there is a 'staging phenomenon': DetHi introduce concreteness *by*

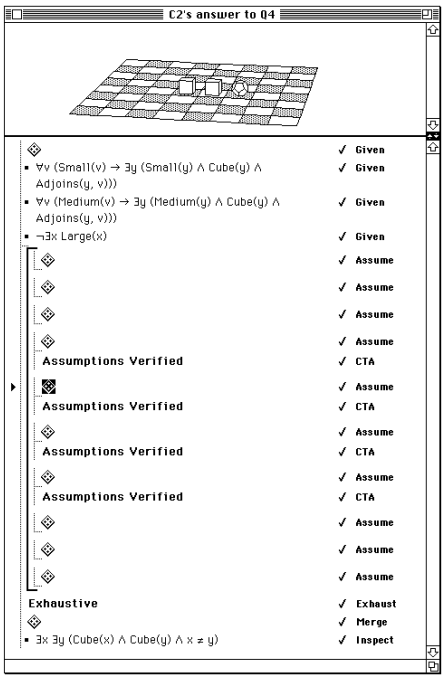

Figure 3: Submitted proof for a DetLo subject (C2) attempting an indeterminate question (Q4). The situation on view is from the 9th step of the proof.

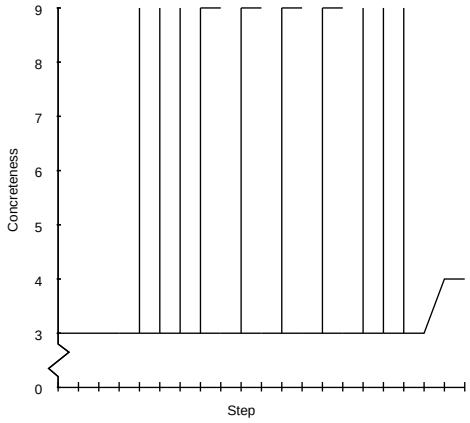

Figure 4: Proofogram for C2 attempting Q4. Proof steps are plotted on the x-axis; the concreteness of the current graphical situation is computed for each step of the proof, and is plotted on the y-axis. Horizontal lines indicate dependency structure; vertical lines indicate uses of Assume; sloping lines indicate uses of Apply or Merge. C2's proofogram is 'spikey', indicating a series of independent, concrete cases.

stages, whereas DetLo introduce it more immediately. In terms of proof structure, DetHi tend to produce structured sets of cases, with superordinate cases involving graphical abstraction; DetLo tend to produce sets of cases without such overt superordinate structure. This staging phenomenon supports the abstraction ability hypothesis: the two groups are certainly using abstractions in different ways.

4.2 Corpus analysis

Of Hyperproof's rules, only Assume, Apply and Merge increase concreteness. We therefore examined the kind of patterns in which they occur through proof-corpus analysis. The proofogram results already indicate that DetHi and DetLo differ in the way they handle concreteness. Since Assume is by far the most frequent means of adding concreteness, the corpus analysis distinguishes between uses of the rule which introduce totally concrete graphical situations, and those which leave some abstractness in the graphic. The term Fullassume denotes the former type of use, and assume denotes the latter.

Using techniques developed originally for the analysis of linguistic corpora, we have carried out bigram and trigram analyses of rule use, utilising Dunning's (1993) 'Log–Likelihood Test', which can be applied to relatively small corpora. The test is designed to "highlight particular A's and B's that are highly associated in text" (p.71). Ranking the bigrams according to this test provides a good *profile* of the individual's, or the group's, rule use in the corpus. We can then

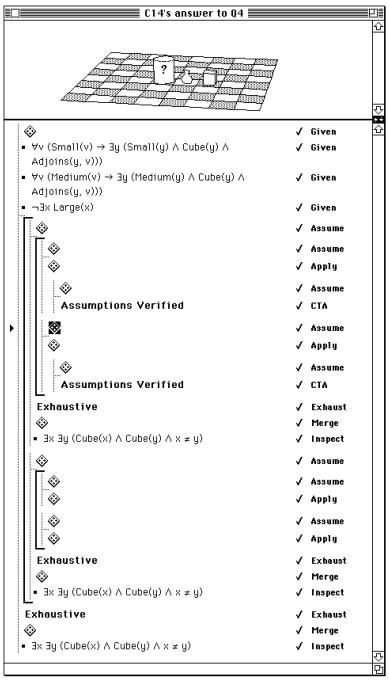

Figure 5: Submitted proof for a DetHi subject (C14) attempting an indeterminate question (Q4). The situation on view is from the 9th step of the proof.

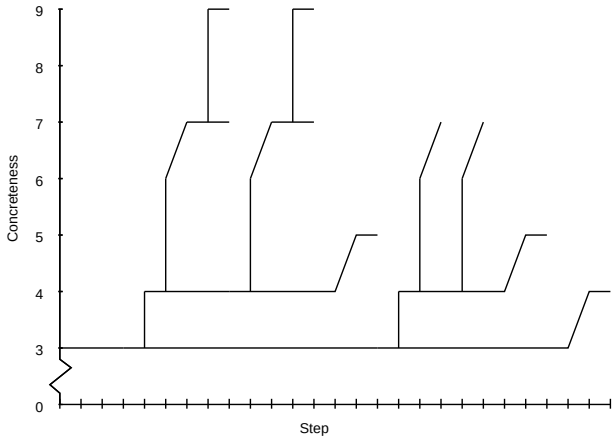

Figure 6: Proofogram for C14 attempting Q4. C14's proofogram is 'layered', indicating parallel sub-case structures with abstract superordinate cases.

compare the profiles for the sub-groups on the two question types, assessing the significance of a given bigram by using the χ^2 test on the log-likelihood value.

On indeterminate questions, we find that the bigrams assume Apply, Merge Inspect, CTA Observe, assume Close, Given assume, and assume Fullassume are significant in DetHi proofs, but not in DetLo ones. Conversely, only the bigram Inspect Merge is significant in DetLo proofs, but not in DetHi ones. The profiles are weakly but significantly correlated ($r = 0.167^{*}$).[1] When taking into account only those bigrams that are significantly associated in the profiles, the correlation is higher, but not significant ($r = 0.315, ns$).

On determinate questions, the bigrams assume Apply, CTA Observe and Close Fullassume are significant in DetHi proofs, but not in DetLo ones. Conversely, as with the indeterminate questions, the only bigram significant in DetLo proofs, but not in DetHi ones, is Inspect Merge. Here, the two subject group's profiles are significantly correlated ($r = 0.537^{**}$). The correlation between significantly-associated bigrams is even stronger and still highly significant ($r = 0.918^{**}$).

This finding accords with the proofograms' indication that it is indeterminate questions which best discriminate the two subject groups. Recall that these are the questions in which the initial graphical situation is abstract, so that all concreteness must be introduced explicitly by the subjects.

The proofogram and corpus analyses therefore support the abstraction ability hypothesis. On questions where the subject must construct the concrete graphic, it seems that DetHi subjects exhibit staging behaviour, and build their graphics incrementally, whereas DetLo subjects are prone to construct their concrete graphics in one go. The abstraction ability hypothesis seems plausible, since the 'stagers' are exactly those whom our main study showed benefit most from teaching with Hyperproof (Stenning, Cox and Oberlander, 1995).

5 Visual preference results

But why do the subject groups diverge under these circumstances? As we have suggested, one tempting hypothesis comes from identifying our DetHi—DetLo distinction with the traditional visualiser—verbaliser distinction. If it's a matter of visual preference, then the diagrammatically capable DetHi subjects are just the visualisers, and therefore, they prefer to use the graphical modality when it is available. The diagrammatically less capable DetLo are the verbalisers, and hence prefer the sentential modality—or at least, they do not show a strong preference for the graphical.

The alternative transmodal hypothesis is that DetHi subjects are better at multimodal reasoning, mixing sentential and graphical information. On this account, DetLo might be perfectly happy in the graphical modality, so long as they do not have to translate information back and forth between the graphical and the sentential.

[1] Correlations reported here are non-parametric (Spearman's ρ). Significance at the $p < .05$ level is denoted by *; significance at the $p < .001$ level by **.

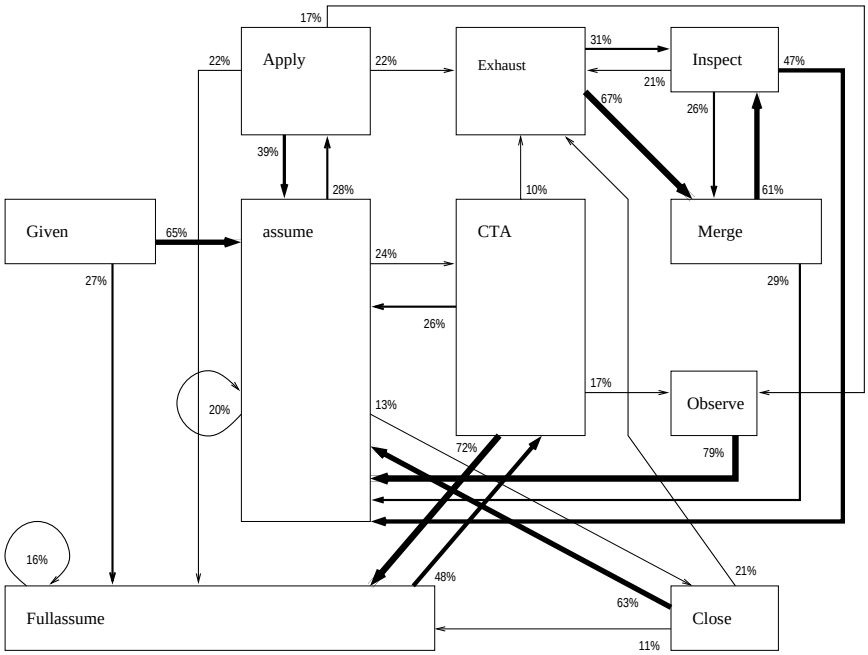

Figure 7: Transition network for DetHi behaviour on indeterminate questions. Nodes represent rules, and their areas represent the frequency at which that rule was invoked. Links represent the probability of transition from one rule to another; transitions at 10% probability and below are not shown.

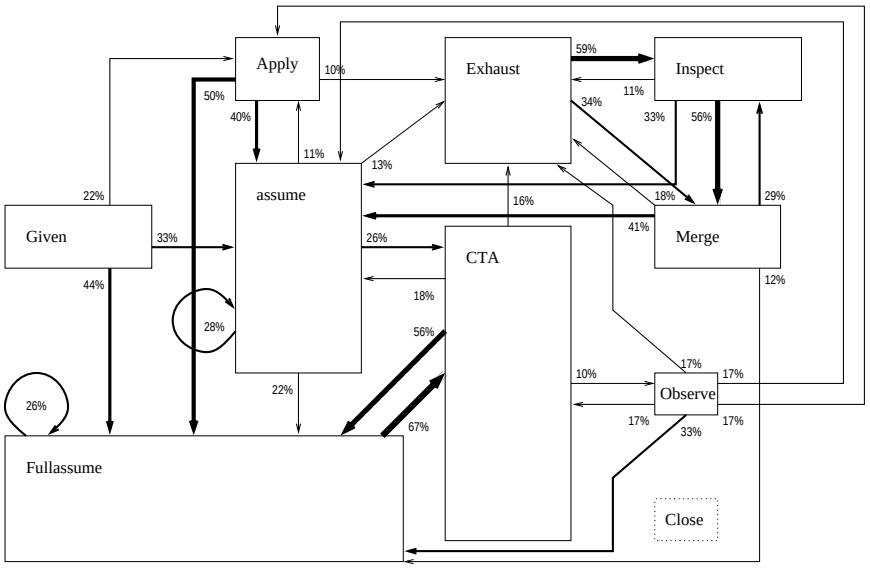

Figure 8: Transition network for DetLo behaviour on indeterminate questions. Note that Close is not visited at all.

One way of testing these competing views is to look at the overall networks of bigram transitions, for the two subject types. The transition networks in Figures 7 and 8, represent DetHi and DetLo behaviour on indeterminate questions. In the networks, the area of a node represents the frequency with which a rule is used, while the thickness of links represents the probability of taking that exit arc, given that one is in the state denoted by the node.

First, consider the left-hand parts of the networks. Any proof must start from a Given step; now, it is clear that there are several ways in which the use of assume and Fullassume varies between the DetHi and DetLo groups. First, DetHi make more use of assume than DetLo, while the latter around twice as much use of Fullassume than the former. DetLo subjects' favouring of Fullassume over assume certainly confirms that they are not 'stagers'; but in a sense, it also suggests that it is *they* who exhibit a preference for the graphical modality, moving straight into it and working entirely within it, rather than gradually transferring information into it from the sentential pane.

Notice also that around two-thirds of DetHi transitions from Given are to assume, and the rest go to Fullassume. By contrast, just one-third of DetLo transitions from Given go to assume; some 22% go to Apply, and 44% go straight to Fullassume. So, as well as favouring Fullassume over assume, DetLo subjects also often commence proof contruction by the use of Apply. This also helps to reduce subsequent interaction between the modalities, with case construction being performed only within the graphical window. Looking at Apply in more detail, it is apparent that DetHi are more likely to use it after assume (it accounts

for 28% of their transitions out of assume, as opposed to just 11% amongst DetLo). And while both groups are as likely to use assume after Apply, DetLo are more than twice as likely as DetHi to go from Apply direct to Fullassume.

The assume Apply pattern confirms that DetHi subjects tend to add information graphical window pane gradually, either by assumption, or by transfer from the sentential pane (via Apply).

Secondly, consider the top right portions of the networks. From Exhaust, DetLo are most likely to move to Inspect, and from there to Merge. By contrast, DetHi are most likely to move to Merge, and from there to Inspect. Both Inspect and Merge find common information from the set of cases declared exhaustive by Exhaust. The difference is that Inspect provides this information sententially, and Merge does it graphically. It seems from the networks that DetHi find the graphical before the sentential, while DetLo find the sentential first, and only then carry out the graphical operation.

Taking these two parts of the network together, it should be clear by now that this is not a simple matter of DetHi preferring the visual modality, not least since DetLo move to that modality more directly. Instead, the difference seems to be that the DetHi group *operate over* the graphical situations, frequently using a graphic as input, or guide, to further stages of proof construction. The DetLo, on the other hand, seem just to *output* graphics, without subsequently using them.

Finally, consider the bottom right hand corners of the networks. There is one very striking fact: DetLo subjects *never* use the Close rule on these indeterminate questions. Close is used in proofs of inconsistency: students use it show that some assumptions contradict given information. And this means that while DetLo make considerable use of proofs of consistency—evidenced by the high frequency of use of CTA—they never proceed by showing that certain cases can be explicitly ruled out as inconsistent with existing premises and assumptions. This aversion to proof by contradiction is intriguing, because it may ultimately be related to findings concerning people's ability to verify or falsify general propositions (as in the four card selection problem, discussed, for instance, in Wason 1977).

For the time being, however, it suffices that DetHi subjects do *not* show a simple preference for the visual–graphical modality. Rather, what distinguishes them is their greater tendency to *translate* between graphical and sentential modalities in *both* directions. Suppose we call people who prefer the visual modality 'artists', and people who like to switch back and forth between representation systems 'translators'. Then it seems that abstraction ability—and hence success with Hyperproof—lies not with the artists, but with the translators.

Acknowledgements

The support of the Economic and Social Research Council for HCRC is gratefully acknowledged. The work was supported by UK Joint Councils Initiative in

Cognitive Science and HCI, through grant G9018050 (Signal); and by NATO Collaborative research grant 910954 (Cognitive Evaluation of Hyperproof). The first author is supported by an EPSRC Advanced Fellowship. The paper is based on that published as Oberlander et al. (1996), but includes additional material eventually to be submitted for journal publication. Special thanks to Dave Barker-Plummer, Chris Brew, Tom Burke, John Etchemendy, Mark Greaves, Padraic Monaghan, and Richard Tobin.

References

Barwise, J. and Etchemendy, J. (1994). *Hyperproof*. Stanford: CSLI Publications.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* **19**, 61–74.

Duran, R., Powers, D. and Swinton, S. (1987). Construct Validity of the GRE Analytical Test: A Resource Document. ETS Research Report 87–11. Princeton, NJ: Educational Testing Service.

Monaghan, P. (1995). A corpus-based analysis of individual differences in proof-style. MSc Thesis, Centre for Cognitive Science, University of Edinburgh.

Oberlander, J., Cox, R. and Stenning, K. (1996). Proof styles in multimodal reasoning. In Seligman, J. and Westerståhl, D. (Eds.) *Language, Logic and Computation: Volume 1*, pp403–414. Stanford: CSLI Publications.

Oberlander, J., Cox, R., Mongahan, P., Stenning, K. and Tobin, R. (1996). Individual differences in proof structures following multimodal logic teaching. In *Proceedings of the 18th Annual Meeting of the Cognitive Science Society*, pp201–206, La Jolla, Ca., July 1996.

Stenning, K., Cox, R. and Oberlander, J. (1995). Contrasting the cognitive effects of graphical and sentential logic teaching: reasoning, representation and individual differences. *Language and Cognitive Processes*, **10**, 333–354.

Stenning, K. and Oberlander, J. (1991). Reasoning with Words, Pictures and Calculi: computation versus justification. In Barwise, J., Gawron, J. M., Plotkin, G. and Tutiya, S. (Eds.) *Situation Theory and Its Applications*, Volume 2, pp607–621. Chicago: Chicago University Press.

Stenning, K. and Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, **19**, 97–140.

Wason, P. C. (1977). Self-contradictions. In Johnson-Laird, P. N. and Wason, P. C. (Eds.) *Thinking: Readings in Cognitive Science*, pp114–128. Cambridge: Cambridge University Press.