From: AAAI Technical Report SS-95-06. Compilation copyright © 1995, AAAI (www.aaai.org). All rights reserved.

Proofs as discourse: an empirical study

Jon Oberlander and **Richard Cox** and **Keith Stenning**
Human Communication Research Centre
University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW
{J.Oberlander | R.Cox | K.Stenning}@ed.ac.uk

Abstract

Computer-based logic proofs are a form of 'unnatural' language discourse, but the structure and process of proof can be observed in considerable detail, and analysis is leading to a number of general insights. We have been studying how students respond to multimodal logic teaching. First, psychological measures indicate that students' pre-existing cognitive styles have a significant impact on teaching outcome. Secondly, a large corpus of proofs has been gathered via automatic logging of proof development. Frequency analysis and cluster analysis of this corpus indicate that students' cognitive styles influence the structure of their logical discourse. Our current objective is to apply further statistical methods to the proof development logs, to derive various transition frequencies, and then construct process models which explain the differences in discourse style.

Introduction: multimodal logical discourse

Computer-based multimodal tools are giving people the freedom to express themselves in brand new ways. But what do people actually *do* when given these tools? Does everyone end up generating the same forms of multimodal discourse? Does multimodality lead to better performance than monomodal systems?

These questions arise in many areas, but are particularly important in educational applications, since multimodality is believed to be especially helpful to novices (cf. di Sessa, 1979; Schwarz & Dreyfus, 1993). Hyperproof is a program created by Barwise and Etchemendy for teaching first-order logic. It uses multimodal graphical and sentential methods, and is inspired by a situation-theoretic approach to heterogeneous reasoning (Barwise & Etchemendy, 1994). A distinctive feature of Hyperproof is its set of 'graphical' rules, which permit users to transfer information to and fro, between graphical and linguistic modes.

We have been carrying out a series of experiments on Hyperproof, to help evaluate its effects on students learning logic. Amongst other things, we have built up a substantial corpus of proofs. These 'hyperproofs' are an unusual form of discourse, for two main reasons. Firstly, they are primarily used for *self*-communication: a student arranges proof steps and rules in an external representation as an aid to their individual problem-solving activities. Secondly, hyperproofs are, of course, *multimodal* discourse: they involve both language and graphics, and are therefore in some ways more complex than text or speech.

Elsewhere, we have argued that graphical systems possess a useful property—over-specificity—whereby certain classes of information must be specified (Stenning & Oberlander 1991; in press). The property is useful because inference with such specific representations can be very simple. We have also urged that actual graphical systems—such as Hyperproof—do allow abstractions to be expressed, and it is this that endows them with a usable level of expressive power. We have recently established that there are important individual differences in the way students respond to different types of logic teaching, and that these differences are linked to the way students respond to Hyperproof's abstraction mechanisms (Cox, Stenning & Oberlander, 1994; Stenning, Cox & Oberlander, in press).

Here, we focus on the hyperproofs themselves, and show how our empirical methods are revealing subtle patterns in multimodal discourse structure. To this end, we first introduce Hyperproof, and indicate how multimodal proofs can be considered as structured discourses. We then outline our experimental method, briefly summarising the results regarding teaching outcomes. We then discuss the existing results from the proof logs, and indicate the current objects of analysis. We conclude by drawing some general morals from the study.

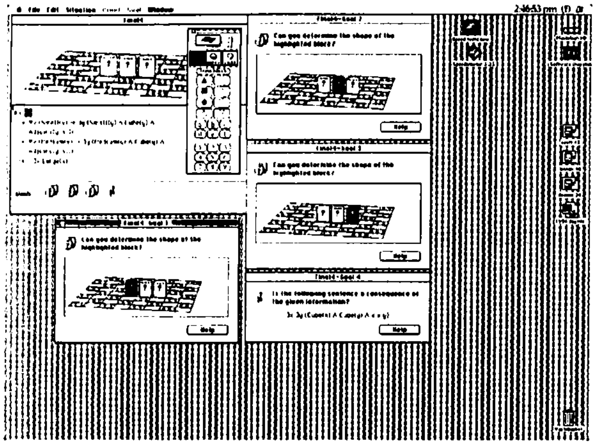

Figure 1: The Hyperproof Interface. The main window (top left) is divided into an upper graphical pane, and a lower calculus pane. The tool palette is floating on top of the main window, and the other windows reveal a set of goals which have been posed. To achieve them, a proof must be developed, by applying a set of multimodal inference rules to the graphical and calculus premises given.

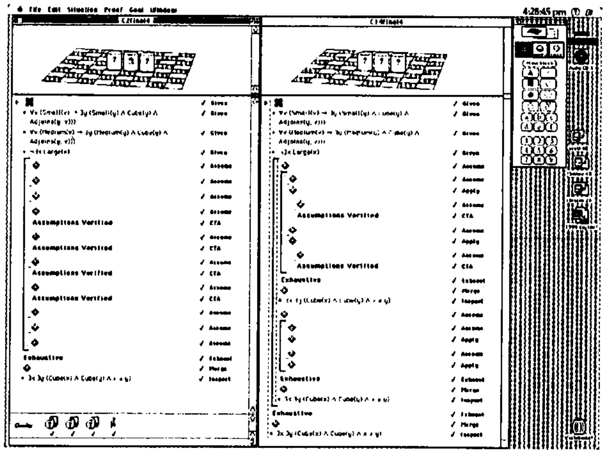

Figure 2: Two different subjects' proofs given in answer to the exam question in Figure 1. When—as here—little is fixed in the graphical situation we term a question *indeterminate* in type, and contrast it with those *determinate* questions in which all the relevant information is specified. The subjects differ in their cognitive styles; as we argue later, the structures of their proofs reflects this.

Hyperproof and discourse structure

The Hyperproof Interface

As can be seen in Figure 1, the interface contains two main window panes: one presents a diagrammatic view of a chess-board world containing geometric objects of various shapes and sizes; the other presents a list of sentences in predicate calculus; control palettes are also available. These window panes are used in the construction and editing of proofs. Several types of goals can be proved, involving the shape, size, location, identity or sentential descriptions of objects; in each case, the goal can involve determining some property of an object, or showing that a property *cannot* be determined from the given information. A number of rules are available for proof construction; some of these are traditional syntactic rules (such as $\wedge$-elimination); others are 'graphical', in the sense that they involve consulting or altering the situation depicted in the diagrammatic window. In addition, a number of rules check properties of a developing proof. Hyperproof should be viewed as a proof-checking environment designed to support human theorem proving using heterogeneous information.

Discourse structures in multimodal proofs

A proof produced using Hyperproof can be thought of as an artefact of multimodal self-communication. Consider the two differing proofs displayed in Figure 2. These are the answers produced by two students (C2 and C14) to the exam question displayed in Figure 1. Take C14's proof. Each line corresponds to a single utterance. Some utterances are linguistic (and are represented by formulae in the calculus pane); others are graphical (and are represented by a diamond icon in the calculus pane, and a particular situation in the graphical pane). Each utterance is associated with a single rule, which specifies its functional role within the proof. The rule is therefore similar to McKeown's (1985) notion of a rhetorical predicate. Some rules require explicit dependencies to be established between the current utterance and others; these are introduced by the student, and displayed by highlighting. The dependencies are akin to anaphoric links. As well as this dependency structure, there is a hierarchical structure, reflecting the grouping of common cases in the argument. This structure is similar in kind to Grosz & Sidner's (1986) linguistic structure.

Notice that C14's proof is more hierarchically structured than C2's. The differences between their proofs are representative of broader distinctions between cognitive styles, which we discuss below.

Method

Two groups of subjects were compared; one group ($n = 22$) attended a one-quarter duration course taught using the heterogeneous reasoning approach of Hyperproof. A comparison group ($n = 13$) were also taught for one quarter but in the traditional syntactic manner

Determinate problem An office manager must assign offices to six staff members. The available offices are numbered 1–6 and are arranged in a row, separated by six foot high dividers. Therefore sounds and smoke readily pass from one to others on either side. Ms Braun's work requires her to speak on the phone throughout the day. Mr White and Mr Black often talk to one another in their work and prefer to be adjacent. Ms Green, the senior employee, is entitled to Office 5, which has the largest window. Mr Parker needs silence in the adjacent offices. Mr Allen, Mr White, and Mr Parker all smoke. Ms Green is allergic to tobacco smoke and must have non-smokers adjacent. All employees maintain silence in their offices unless stated otherwise.

- The best office for Mr White is in 1, 2, 3, 4, or 6?
- The best employee to occupy the furthest office from Mr Black would be Allen, Braun, Green, Parker or White?
- The three smokers should be placed in offices 1, 2, & 3, or 1, 2 & 4, or 1, 2 & 6, or 2, 3, & 4, or 2, 3 & 6?

Indeterminate problem Excessive amounts of mercury in drinking water, associated with certain types of industrial pollution, have been shown to cause Hobson's Disease. Island R has an economy based entirely on subsistence level agriculture with no industry or pollution. The inhabitants of R have an unusually high incidence of Hobson's' Disease.
Which of the following can be validly inferred from the above statements?

i. Mercury in the drinking water is actually perfectly safe.

ii. Mercury in the drinking water must have sources other than industrial pollution; or

iii. Hobson's Disease must have causes other than mercury in the drinking water.

- (ii) only?
- (iii) only?
- (i) or (iii) but not both?
- (ii) or (iii) but not both?

Figure 3: Examples of two types of reasoning problem. Determinate problems provide premisses which determine a (nearly) unique logical model; indeterminate problems do not. The former are closely related to what the graduate record exam (GRE) analytical test calls the *analytical reasoning* subscale; the latter to the test's *logical reasoning* subscale.

supplemented with exercises using a graphics-disabled version of Hyperproof (to control for the motivational and other effects of computer-based activities). A fuller description of the method and procedure is provided in Cox, Stenning & Oberlander (1994).

All subjects were administered two kinds of pre and post-course paper and pencil test of reasoning. The first test was of 'analytical reasoning' and contained two kinds of item derived from the GRE-type of scale of that name (see for example, Duran, Swinton & Powers, 1987). We refer to this test as the 'GRE' test. The first 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. Examples items are displayed in Figure 3.

The second paper and pencil test we term 'Blocks world'. This test requires reasoning about blocks-world situations like those used in Hyperproof, but is couched in natural language rather than first order logic.

Both groups also sat post-course, computer-based Hyperproof exams. The questions differed for the two groups, however, since the syntactic group had not been taught to use Hyperproof's systems of graphical rules. The four questions set the Hyperproof group, though, contained two types of item: determinate and indeterminate. Here, determinate problems were taken to be those which did not utilise Hyperproof's abstraction conventions for objects' spatial or visual attributes. As well as concrete depictions of objects, Hyperproof allows 'graphical abstraction symbols', which leave attributes under-specified: the *cylinder* depicts objects of unknown size; the *paper bag* depicts objects of unknown shape. Figure 2 illustrates Question 4, one of the two indeterminate questions. Student-computer interactions were dynamically logged—this approach might be termed 'computer-based protocol taking'. The logs were time stamped and permitted a full, step-by-step, reconstruction of the time course of the subject's reasoning. The results reported in Section are based on analyses of those protocols.

Scores on the determinate subscale of the GRE test were used to classify subjects within both Hyperproof and syntactic groups into DetHi and DetLo subgroups. The score reflects subjects' facility for solving a type of item that often is best solved using an external representation; DetHi scored well on analytical reasoning items, like the office allocation problem in Figure 3; DetLo scored less well on such items. Loosely, we may consider DetHi subjects to be more 'diagrammatic', and DetLo to be more 'verbal'. DetHi and DetLo subjects in the Hyperproof and syntactic groups responded differently to traditionally versus heterogeneously taught courses. See, for example, Figure 4; the results are reported in full in Cox, Stenning & Oberlander (1994); Stenning, Cox & Oberlander (in press). Typically, learning style studies that have investigated the visualiser–verbaliser distinction use *psychometric* instruments as the basis for classifying subjects. For example, the paper-folding test has been used by Mayer and Sims (1994) in a recent study of learning from computer-generated animation; and by Campagnoni and Ehrlich (1989) in a study of individual differences in hypertext navigation. However, it is currently unclear how strongly internal behaviour (as measured by paper-and-pencil psychometric tests) is related to external reasoning performance. Therefore, in the current study, subjects were classified according to their *performance* on diagrammatic reasoning items.

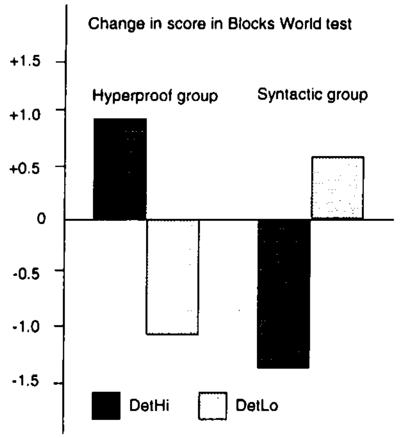

Figure 4: Changes in scores between Blocks-World pre- and post-tests. Notice that DetHi students benefit from Hyperproof, and suffer under the Syntactic regime. In contrast, DetLo students do better under the syntactic regime.

Table 1:A set of relevant Hyperproof rules.

Rule	Description
Apply	Extracts information from a set of sentential premises; expresses it graphically
Assume	Introduces a new assumption into a proof, either graphically or 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
Observe	Extracts information from the situation; expresses it sententially
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

Results: Computer-activity logs

We gathered logs of all students' course exercises, and their exam answers. The former proved rather uniform, probably because of the extensive course guidance offered. We therefore concentrate here on students' logged exam answers.

Exam problem proofs

Preliminary analyses were performed on several parameters of these examination proofs. Each proof-log was coded for score (number of proof goals validated), time (time spent on proof), number of proof steps, the proof depth (the depth of nested subproofs the subjects used in their solution), and the frequency with which each of the Hyperproof logical rules was used (rule use frequency).

Absolute values of Hyperproof and Syntactic subjects could not be directly compared because these groups answered different questions. There was a tendency for DetHi subjects to produce 'better' (that is, longer, quicker, more accurate, more nested proofs) than their DetLo counterparts within the Hyperproof group, whereas the converse was the case within the syntactic class. The difference between Hyperproof DetHi and DetLo subjects on the time parameter approached statistical significance ($t = -2.06, df = 14, p = .058$). No other comparisons were statistically reliable.

However, the trends in these results support an interpretation in terms of the interaction between cognitive style and teaching modality. Diagrammatic reasoners (DetHi) benefit more from instruction in the graphical modality than non-diagrammatic reasoners (DetLo). The opposite trend holds with syntactic instruction. This dissociation on the exam parameters is commensurate with the blocks-world test results and results on the GRE verbal reasoning test.

Within the Hyperproof students, interesting differences were also noted between performance on the determinate and indeterminate exam items. These were not differences in terms of the score, time, steps or depth parameters—the differences were in terms of rule use patterns. It is to these that we now turn.

Rule use patterns

The Hyperproof group data throws light on the nature of the individual differences observed between DetHi/Lo subjects, and on our theoretical predictions about graphical reasoning and communication. Hyperproof supports the use of both the traditional syntactic rules of FOL and special graphical rules. The most important of these are summarised in Table 1; see Barwise & Etchemendy (1994) for a full account of Hyperproof's rule system. Our Hyperproof logging recorded all uses of rules, and internal system responses to user input (called *manoeuvres* below). The system responses and feedback to the user were also recorded.

Analyses of variance (ANOVA) were conducted on the log data. ANOVA is a statistical technique designed to analyse the separate and combined effects of predictor variables upon some outcome measure or dependent variable. ANOVA overcomes the "multiple comparison" problem that would arise if multiple univariate comparisons (such as t-tests) were used instead. A readable overview of multivariate statistical techniques for use in the social sciences is provided in Chapter 1 of Harris (1975).

A two-factor ANOVA for subjects (DetHi, DetLo) and item determinacy (determinate, indeterminate) was conducted separately for each of seventeen rules/-manoeuvres/system responses, with frequency of rule use as the dependent variable. The results of these analyses revealed that all subjects used the following rules and manoeuvres significantly more frequently in developing proofs for the 2 indeterminate questions than for the 2 determinate questions:[1] Assume, Apply, CTA, create step, create subproof, move focus, cite step in support, update sentences in proof, and delete step. The Close rule was used significantly more on the *determinate* than on indeterminate questions. The remove step manoeuvre was used significantly more by DetLo subjects than by DetHi subjects.

A two-way interaction was significant in one of the analyses: the Apply rule was used more on determinate questions by DetLo subjects than by DetHi subjects. Conversely, on indeterminate questions, DetHi subjects used it more frequently than DetLo subjects.

Abstraction patterns

As well as patterns of rule-use, students' use of graphical abstraction devices are characteristics of their proof-styles. We scored each step of each proof on the basis of number of concrete situations compatible with the graphical depiction; the scoring method is described in more detail in Oberlander, Cox & Stenning (in press). Basically, we give each graphical symbol in a situation a score: for each attribute (size, shape, location, and label), a symbol scores 1 if that attribute is specified, and 0 otherwise. By totalling the scores, we can give each situation in a proof a score. This graphical concreteness score was then used to derive overall scores for use of abstraction devices for all the DetLo and DetHi subjects. A low score indicates more abstraction; a higher score indicates more concreteness.

Analysis of questions 1, 2 and 3 revealed no differences between DetHi and DetLo subjects in their graphical concreteness scores. Question 4 showed a different pattern of results. It was no more difficult than the other questions, but it contained substantially more graphical abstraction in its initial reasoning situation. Considering only the subjects who succeeded in proving the proof goals, a one-tailed t-test between DetLo and DetHi subjects' graphical concreteness scores reveals a small but reliable difference between the scores of DetLo and DetHi subjects ($t = 1.83, df = 18, p < .05$). The mean concreteness score for DetLo was $7.92, SD = 0.88$ and for DetHi it was $7.13, SD = 0.98$. The lower mean score for DetHi indicates more use of abstraction in the steps of the proof, a result that is consistent with a greater facility on the part of DetHi subjects for using the graphical abstraction conventions of Hyperproof, such as the *paper bag* and *badged cylinder* devices.

Rule clustering

The Hyperproof rule use frequency data was subjected to cluster analyses. This statistical technique creates homogeneous groups of entities; a useful introduction to the topic is provided by Aldenderfer & Blashfield (1984).

Cluster analysis is often used as an exploratory data analysis technique and it was used in this study as a method of distinguishing different *patterns* of rule use. Clustering reveals *correlations* between rule uses and suggests trends that deserve further investigation. Part of the interest lies, of course, in the fact that correlations give a different view of rule patterns from that delivered by frequency analysis.

Thus far, we have considered informally the extent to which question type and cognitive style influence rule clustering, by examining the dendrograms generated using average linkage. By way of example, Figure 5 indicates the rule usage clusters for DetHi and DetLo attempting exam question 4. A number of observations are suggested by this exercise. First, in general, for DetLo subjects, CTA clusters particularly closely with other items in their rule repertoire, whereas Exhaust plays this role among DetHi subjects. Secondly, DetLo subjects seem to have a more stable set of relationships between their rules: the only rule which clusters less well for them on indeterminate questions is Close. Thirdly, DetHi subjects appear to be more flexible. They use CTA on indeterminate questions more frequently than on determinate questions, but the rule does *not* correlate closely with the rules which cluster together well. By contrast, Apply, and Inspect do seem to cluster well, on indeterminate (but not determinate) questions. Finally, like DetLo subjects, DetHi subjects use Close less frequently on indeterminate questions, but as can be seen from Figure 5, it still clusters relatively well on question 4.

Next step: corpus-based statistics

These exploratory analyses are sufficient to show the existence, after the Hyperproof course, of differences in the discourse structures produced by the two groups of subjects identified by pre-test aptitudes. More analytic characterisation of these differences is an objective of

[1] As evidenced by significant main effect for the determinacy factor in each analysis.

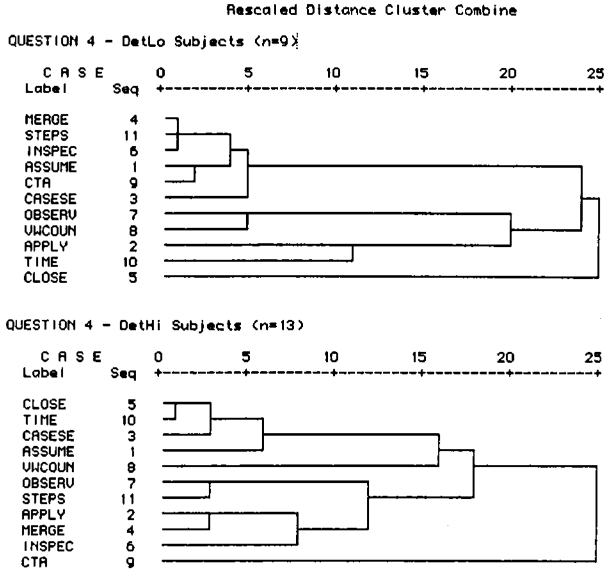

Figure 5: Average Linkage Dendrogram, representing cluster analysis of DetLo/Hi subjects' rule usage on exam question 4. Note how CTA, Close and Apply cluster with other rules: DetLo and DetHi subjects exhibit contrasting tendencies.

current research.

The existing analyses do not take adequate account of *order*, or *hierarchy*. We are therefore currently extracting all sequences of rules from the proof development logs, and imposing a finer-grained classification on the sequences, depending on how each rule influences the graphical concreteness of the relevant graphical situation. The aim is to compute bigram and trigram transition frequencies for these rules, allowing us to construct models which will generate appropriate rule sequences. We predict that the models for DetLo and DetHi subjects will diverge, at least for indeterminate exam questions.

The aim is to report results from this new phase at the spring symposium. We hope to articulate the way in which the differing cognitive styles lead to diverging semantic competences, and show how this in turn determines the characteristic discourse structures produced by the subjects.

Conclusions

Computer-based logic proofs are a form of multimodal self-communication. If each line of a hyperproof is an utterance, then the proof as a whole functions as an organised discourse, possessing hierarchical structure, inter-utterance dependencies, and rhetorical structure.

It might seem that such artefacts are merely 'unnatural' language discourse. However, we believe that there are three reasons for studying them. First, results from the GRE test indicate that the experience of being taught first-order logic generalises to other kinds of linguistic skill: from reasoning about proofs in a formal language (first-order logic) to reasoning in natural language. Logical discourse may be unnatural, but it is certainly connected to natural language discourse. Secondly, computer-based protocol taking has allowed us to observe the structure and process of a type of discourse production in very considerable detail. The study therefore represents an approach that could be replicated for more natural forms of language. Finally, our empirical analysis has demonstrated that students' pre-existing cognitive styles interact in a significant way with both teaching modality, and with the structures of logical discourse they learn to generate. 'Diagrammatic' (DetHi) and 'verbal' (DetLo) students grasp graphical abstractions in different ways, and their multimodal discourses end up with characteristically different structures. If someone can use abstractions effectively, then they generate deeper, longer and more structured proofs, opened and closed with particular rule sequences. A more general lesson may be drawn from this: individual differences could prove to be a particularly exciting area of empirical study for discourse theorists.

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: Specificity of Information in Graphics and Natural Language); and by NATO Collaborative research grant 910954 (Cognitive Evaluation of Hyperproof). The first author is supported by an EPSRC Advanced Fellowship. Special thanks to John Etchemendy, Tom Burke and Mark Greaves. Note that this paper is based on Stenning et al. (in press) and Oberlander et al. (in press), but that the verbal presentation at Stanford should focus on new results from the corpus-based statistical analysis currently underway.

References

Aldenderfer, M. S. and Blashfield, R. K. (1984). *Cluster Analysis.* London: Sage Publications.

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

Cox, R., Stenning, K. and Oberlander, J. (1994). Graphical effects in learning logic: reasoning, representation and individual differences. In *Proceedings of the 16th Annual Meeting of the Cognitive Science Society*, pp237–242, Atlanta, Georgia, August.

Campagnoni, F. R. and Ehrlich, K. (1989). Information retreival using a hypertext-based help system. *ACM Transactions on Information Systems*, **7**, 271–291.

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.

Grosz, B. and Sidner, C. L. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, **12**, 175–204.

Harris, R. J. (1975). A Primer of Multivariate Statistics. London: Academic Press.

McKeown, K. (1985). *Text Generation*. Cambridge: Cambridge University Press.

Mayer, R. E. and Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Jounral of Educational Psychology*, **86**, 389–401.

Oberlander, J., Cox, R. and Stenning, K. (in press). Proof styles in multimodal reasoning. To appear in Seligman, J. and Westerstahl, D. (Eds.) *Language, Logic and Computation: The 1994 Moraga Proceedings*. Stanford: CSLI Publications.

Schwarz, B. and Dreyfus, T. (1993) Measuring integration of information in multirepresentational software. *Interactive Learning Environments*, **3**, 177–198.

di Sessa, A. A. (1979) On 'learnable' representations of knowledge: A meaning for the computational metaphor. In Lochhead, J. and Clement, J. (Eds.) *Cognitive Process Instruction: Research on teaching thinking skills*. Philadelphia, Pennsylvania: The Franklin Institute Press.

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., Cox, R. and Oberlander, J. (in press). Contrasting the cognitive effects of graphical and sentential logic teaching: reasoning, representation and individual differences. To appear in *Language and Cognitive Processes*.

Stenning, K. and Oberlander, J. (in press). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. To appear in *Cognitive Science*. Available as Research Report HCRC/RP–20, Human Communication Research Centre, University of Edinburgh, April 1992.