Graphical effects in learning logic: reasoning, representation and individual differences

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

Abstract

Hyperproof is a computer program created by Barwise and Etchemendy for teaching logic using multimodal graphical and sentential methods, inspired by their theories of heterogeneous reasoning (Barwise and Etchemendy 1994). Elsewhere, we have proposed a theory of the cognitive impact of assigning information to different modalities (Stenning and Oberlander 1992). Our view is that where diagrams are advantageous, it is because they enforce the representation of information, leading to *weak* expressiveness, thereby facilitating inference. The present study tests and develops these claims by comparing the effects of teaching undergraduate logic classes using Hyperproof and a control syntactic teaching method. Results indicate that there is significant transfer from the logic courses to logical and analytical reasoning problems. There are also significant interactions between theoretically motivated pre-course aptitude measures and teaching method; the interactions influence post-course reasoning performance in transfer domains. Hyperproof boosts students previously weak on items which benefit from diagram use, whereas the syntactic course appears to degrade the same group of students' graphical strategies. As well as being theoretically interesting, these results provide support for the important practical conclusion that individual differences in aptitude should be taken into account in choosing teaching technique.

Theoretical Background

Is reasoning teachable? What is the impact of learning logic upon general reasoning? Can graphical material be used effectively in teaching abstract subjects? These are important pedagogical questions; the issues are also central to the development of theory within cognitive science. In particular, how are we to relate symbolic and heterogeneous logics to human reasoning and learning? What is graphical representation good for and why? The present study seeks to explore such theoretical concerns in a highly practical setting—real logic courses.

Hyperproof (HP) is a computer program for teaching first-order logic (FOL) which Barwise and Etchmendy (1994) designed and implemented on the basis of their situation theoretic approach to reasoning. HP builds on an earlier program Tarski's World, which uses graphics to teach the syntax and semantics of FOL; HP adds the teaching of inference by incorporating heterogeneous reasoning rules which move information back and forth between graphical representations of blocks-worlds and sentences of FOL. Computer based courses offer a unique platform for assessing cognitive theories, since they involve real learning over sustained periods, and permit detailed observation of students' reasoning processes.

Our own motivation was to extend a general theory of the cognitive impact of assigning the same information to different modalities. Graphical logic teaching methods provide a domain in which information in different modalities (sentences and diagrams) can be accurately equated and where there is a well defined task (learning and performance of reasoning). The theory (described in Stenning & Oberlander 1992, (submitted)) sees the distinctive property of graphical semantics as the enforcement of the representation of certain classes of information. This curtailment of abstraction leads to weak expressiveness (in the logical/computational sense) but tractable inference. The theory connects this property to usability through relations between: the abstractions required by tasks; the abstractions expressible in graphical systems; and the availability of knowledge of these expressiveness properties to different classes of users. The fewer superfluous abstractions a system expresses, the more useful it will be for a task, but a user must be in a position to exploit these constraints. On the other hand, if a task requires abstractions which cannot be expressed, then the system will hamper performance.

So, the cognitive theory of graphical reasoning places a special emphasis on the skills involved in exploiting the information enforcements which are inherent in graphical representations. HP, for example, uses various devices to overcome graphical specificities: (a) a single diagram can contain symbols representing objects of unknown size and unknown shape; (b) by the use of a special part of the diagram, objects with unknown location can also be represented; (c) arbitrary disjunctions of information about an object (for instance, the fact that it is either a dodecahedron or a tetrahedron) are captured by multiple diagrams. Device (a) is illustrated in Figure 2.

Our approach also provides a basis for distinguishing types of reasoning problem and so for classifying reasoning aptitudes. Some problems provide premisses which determine a unique (or nearly unique) logical model, from which numerous conclusions can be drawn. Other problems' premisses do not provide sufficent information to specify a unique model and must be approached by isolating relevant premisses and exploiting different inferential techniques. We call these problems *determinate* and *indeterminate* problems respectively. They are closely related to what the graduate record exam (GRE) analytical test calls the *analytical reasoning* and *logical reasoning* subscales respectively (Duran, Powers & Swinton, 1987). Determinate (or nearly determinate) problems lend themselves to graphical representation because representation of a single model does not require inexpressible

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 Hobsons' 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 1: Examples of two types of reasoning problem. Determinate problems provide premisses which determine a (nearly) unique logical model; indeterminate problems do not.

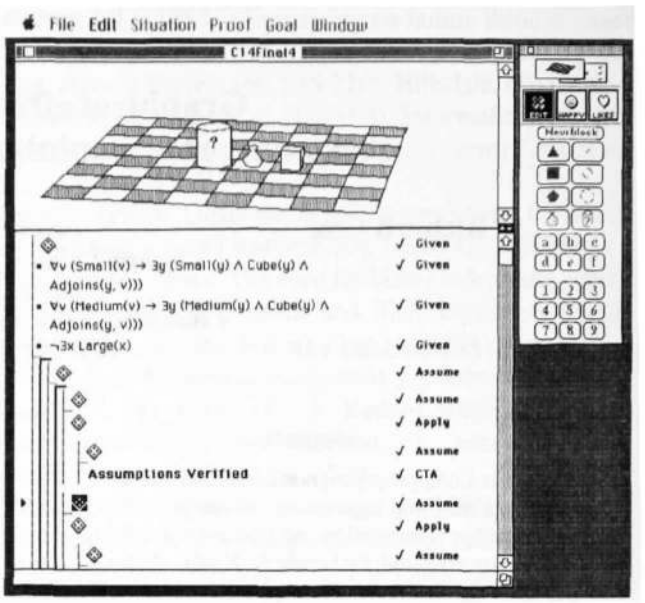

Figure 2: The Hyperproof interface. The main windows—graphical and calculus—are supplemented by control palettes. The situation being viewed is the fifth in the course of the proof, and corresponds to the fifth diamond-shaped 'situation' icon in the body of the proof. The graphical window contains three symbols of varying degrees of abstraction.

abstractions. GRE-type examples are illustrated in Figure 1.

Evaluating Teaching Outcomes

The evaluation was performed using a range of outcome measures. Educational computer software is rarely evaluated in terms of effects upon learning outcomes, or with respect to general theories of cognitive representation and process. There have been several calls for more evaluative studies (eg Littman and Soloway, 1988; Shute, 1990). Several computer-based FOL teaching programs have now been developed (Goldson, Reeves & Bornat (1992) provide a recent review); but few have been evaluated in controlled comparisons.

The outcomes we examine here are chosen to have some plausibility as tests of reasoning in domains beyond logic courses. We expect them to cast some light on the vexed question of the transferrability of reasoning skills. We used GRE-type pre- and post-tests, and 'blocks world' pre- and post-tests, as detailed in section . The former involve somewhat farther transfer than the latter, at least for the HP students. To be sure, the problems are verbally stated; nonetheless, as can be seen from Figure 1, they are different in both form and content from Hyperproof problems. The GRE analytical reasoning scale is widely recognised to be predictive of graduate school success and so may lay some claim to be one test of 'real world' reasoning skill.

Amongst teachers of logic, graphical methods remain controversial. Part of the dispute may hinge on whether teaching elementary logic is undertaken to improve general reasoning, or to prepare students for further advanced symbolic logic courses and their applications in, say, computer science. The present study should therefore be viewed as investigating the effects of the courses on cognitive processes relative to theory, rather than deciding which approach is 'best'. We focus on effects of teaching on different problems and different groups of students, and we concentrate on effects for performance on problems which are external to the courses.

Method

The Hyperproof Interface

As can be seen in Figure 2, the HP interface contains two main windows: one presents a diagrammatic view of a chessboard world containing geometric objects of various shapes and sizes; the other presents a list of sentences in predicate calculus; control palettes are also available. The main windows 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. HP should be viewed as a proof-checking environment designed to support human theorem proving using heterogeneous information.

Subjects

The subjects were 35 first-year Stanford undergraduates attending courses on introductory logic. Two groups were compared. Group 1 attended a course taught using HP (22 subjects).

A second group (13 subjects) attended a course taught syntactically. The HP class was taught in the Fall quarter of 1992 and the Syntactic class was held in the Spring quarter of 1993. While it was not possible to to randomly assign students to the two courses, the students were unaware of any differences in the courses prior to enrolment, and are drawn from the same general population of undergraduates required to take an elementary logic course.

Teaching

Hyperproof Group The course consisted of a 12 week (one quarter) course on first-order logic. The HP class were taught using HP plus HP curriculum material (Barwise and Etchemendy, 1994). The course included 72 computer-based exercises covering the use of HP graphical rules and, to a limited extent, the use of syntactic rules in the development of proofs. Eight of 30 students (27%) dropped out of the HP class before completing the course.

Syntactic Group The syntactic class was taught a course of the same duration as the HP group. The syntactic course was based around a standard, traditional (ie syntactically oriented) instructional text (Bergman, Moor and Nelson, 1990). In order to control for the motivational effects of computer use and other factors, the syntactic group also used HP. However their version had its graphics window disabled (with an empty chessboard). The 'syntactic' students used only the syntactic rules of HP and worked exclusively in the sentence window. Twenty three of their computer-based exercises were adapted from their coursebook; a further 54 exercises in the use of HP's sentential rules were taken from the HP resource book. Thus there were 77 HP-based exercises in total for this group. Nine of 22 students (41%) dropped out of the syntactic class before completing the course. The level of attrition was therefore higher than the HP group's. In both cases, the drop-out rate is attributable at least in part to the general practice of signing on for more courses than will ultimately be taken.

Pre and post teaching tests

The object was to provide tests of reasoning skill which were sufficiently independent of course content to be administrable before the course, and which would provide some test of transfer to reasoning beyond the course material. Two tests were developed which we will refer to as the blocks-world test and the GRE test. The blocks-world problems are slightly 'nearer' transfer tests (at least for the HP course) than the GRE tests, but even they are more closely related to real-world reasoning than typical intra-course exam items from a logic course. Both classes were administered the same battery of paper and pencil tests before and after the course.

The blocks world tests were based on HP graphics but couched in natural language. Their items consist of a diagram of an arrangment of blocks on a checkerboard, and some statements constraining assignments of names to blocks. Questions were about what is provable or not provable from this information, or specified modifications of it. These tests are further described in Cox and Oberlander (1993).

The second outcome test consisted of pseudo graduate record examination (GRE) analytical reasoning test items (selected from a crammer for the test), and divided into two sub-scales: logical reasoning (argument analysis) and analytical reasoning. We observe that these two subscales of the GRE analytical reasoning test align closely with our theory's distinction between determinate and indeterminate problems which are either suitable or unsuitable for graphical reasoning methods. Empirical psychometric results (eg: Duran, Powers & Swinton, 1987) have supported this distinction; it reinforces our proposal that this is an important dimension for categorising reasoning problems. It also suggests that individual differences in cognitive style may be important factors in applying the theory to empirical data.

Parallel forms of each test were used on pre- and post-course tests. Suitable tests with population norms were not available and this means that absolute comparisons between pre- and post-test scores must be interpreted with caution. However, most of the interesting comparisons are between groups of students and between subscales but within post-test scores, and so relative changes in performance are the focus. These relative changes can be assessed as long as these points are born in mind. Of the 22 Hyperproof subjects, 16 completed both the paper and pencil pre- and post-course tests, and all 22 completed the post-course computer-based exam. The 13 syntactic students completed the pencil and paper pre- and post-tests, but 11 completed the syntactic computer-based exam.

Results

Blocks world test results

In order to examine the effects of training modality upon cognitive style differences, subjects were classified as DetHi or DetLo on the basis of scores on the analytical subscale of the GRE-based test. The former scored well on analytical reasoning items; the latter scored less well. The resulting 'level of determinacy' (DetHi/Lo) factor was entered as an additional factor in the analysis of the 'blocks world' test data.

A 3 factor ANOVA (groups by DetHi/Lo by time) was performed on the 'blocks world' test results. The first factor was groups, the second (DetHi/DetLo) was nested under the first, and the third (time) was a repeated measure. Figure 3 shows the means for DetHi and DetLo scorers in the 2 groups. The ANOVA revealed that the main effect for group was significant ($F(1,26) = 6.23, MSe = 2.33, p < .02$). The main effect for DetHi/Lo was also significant ($F(1,26) = 12.81, MSe = 2.33, p < .001$). As shown in Figure 3, the DetHi subjects tend to score higher than the DetLo subjects.

The 3-way interaction (group by DetHi/Lo by time) was also significant ($F(1,26) = 9.45, MSe = 1.6, p < .01$). Thus the experience of learning logic graphically had different effects upon DetHi and DetLo scorers within the HP group. DetHi subjects' scores did not differ from those of their DetLo colleagues on the pre-training test, but following the HP course they scored significantly higher than their DetLo counterparts. Conversely, DetHi scorers in the syntactic group significantly (according to a post-hoc comparison) decreased in their blocks world test performance compared to their DetLo counterparts.

A post-hoc comparison showed that DetLo subjects in the syntactic group scored significantly lower than their DetHi counterparts on the blocks world pre-training test. The reasons

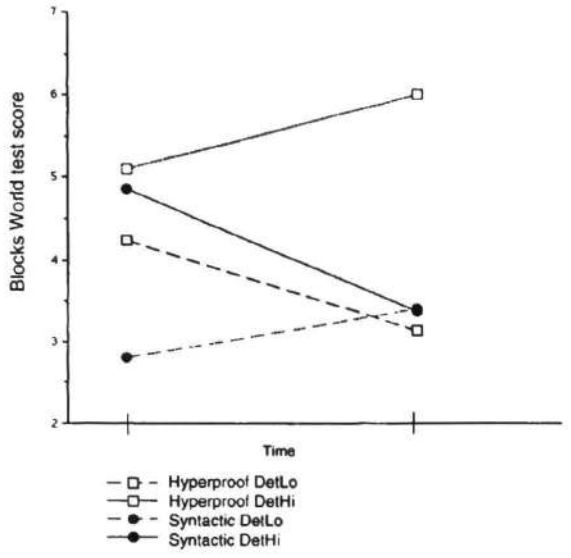

Figure 3: Mean score on Blocks World test as a function of subjects' performance on analytical reasoning GRE subscale (DetHi/Lo) by Group (HP/Syn) by Time (Pre/Post)

for this result are unclear, but may be due to selection biases.

GRE-based analytical test results

Subscale scores on the analytical reasoning test were subjected to a 2 factor (groups by time) ANOVA with repeated measures on the second factor. There were two levels of each factor.

Logical reasoning subscale scores No main effects were significant. However the group by time interaction was significant ($F(1,28) = 4.93, MSe = 2.42, p < .05$) shown in Figure 4a. On the verbal reasoning items, the syntactic group improved significantly more than the HP students. This suggests that the experience of being taught first-order logic syntactically generalises to other kinds of linguistic reasoning: from reasoning about proofs in a formal language (first-order logic) to reasoning in natural language.

Analytical reasoning subscale scores Both groups improved significantly in terms of their scores on this subscale. The main effect for time was significant ($F(1,28) = 18.78, MSe = 4.39, p < .05$). The main effect for groups and the group by time interaction were not significant (Figure 4b). In this case HP did improve the measure more than the syntactic class, but not significantly so.

Discussion

Both HP and syntactic courses were effective in teaching reasoning which transferred to other domains. This is, in itself, significant, since the existence of such transfer effects has been disputed (cf. Nickerson, Perkins & Smith, 1985, for a discussion). The results do not indicate that either course is a globally better way of teaching. The most prominent effects are interactions between student aptitudes, types of problem, and teaching methods.

For students who are able at determinate problems, a syntactic logic course actually *decreases* their scores on blocks world problems. Their colleagues on the same course who start out less able at determinate problems actually improve slightly, so the two groups are indistinguishable at course-end. Teaching students who are able at determinate problems a HP course increases their ability at blocks-world problems. Their colleagues on the same course who start out less able at determinate problems actually decline slightly in blocks world reasoning and finish indistinguishable on this measure from the syntactically taught subjects. This pattern of results supports the idea that syntactic teaching may actually interfere with 'model-construction' reasoning modes which are important outside logic courses. HP appears to enhance the performance of students who are already able at this sort of reasoning but does not yet help the students who are initially weaker. It will be important to characterise in process terms what these differences are and seek remedies.

GRE test performance presents a slightly different picture. Though both courses improve performance on both subscales of the GRE post-test, syntactic teaching increased GRE verbal subscale scores significantly more than HP teaching, whereas the two courses were not significantly different in their effect on the diagrammatic subscale scores. The greater indeterminate problem improvement of syntactic teaching might be expected, but syntactic teaching also improves determinate problem performance.

We are currently analysing students' 'work scratchings' on the GRE problems in terms of Cox and Brna's (1993) 7 categories, which include *tables*, *map/plans*, and so on. We can compare the representations used by students with those recommended by the 'crammer' from which the test item was derived (Brownstein, Weiner & Weiner Green, 1990; Educational Testing Service, 1992). These results have a pattern similar to that in the blocks-world tests. Strong determinate problem solvers (DetHi) appear to retain their effective strategies through either type of course. But weak determinate problem solvers (DetLo) respond oppositely to the two courses. The HP course moves these student's strategies significantly in the direction of the recommended methods of GRE problem solution (and the strategies of their DetHi peers). The syntactic course actually moves DetLo students significantly *away* from the recommended strategies, but without necessarily making them worse reasoners. A finer interpretation of these results awaits an analysis of the detailed methods of reasoning among these groups.

Thus blocks-world and GRE analytical reasoning subscale performance react somewhat differently to the two teaching methods. The two types of reasoning problem, though both based on determinate models, differ in several ways. The blocks world problem requires the application of information from a set of sentential constraints to a *presented* diagram. A GRE analytical reasoning problem requires the *construction* of a determinate model (often represented in a self-constructed diagram or table) from a set of sentential constraints. It would be possible to transpose such GRE problems into HP and to include them in our pre- and post-tests to assess whether this

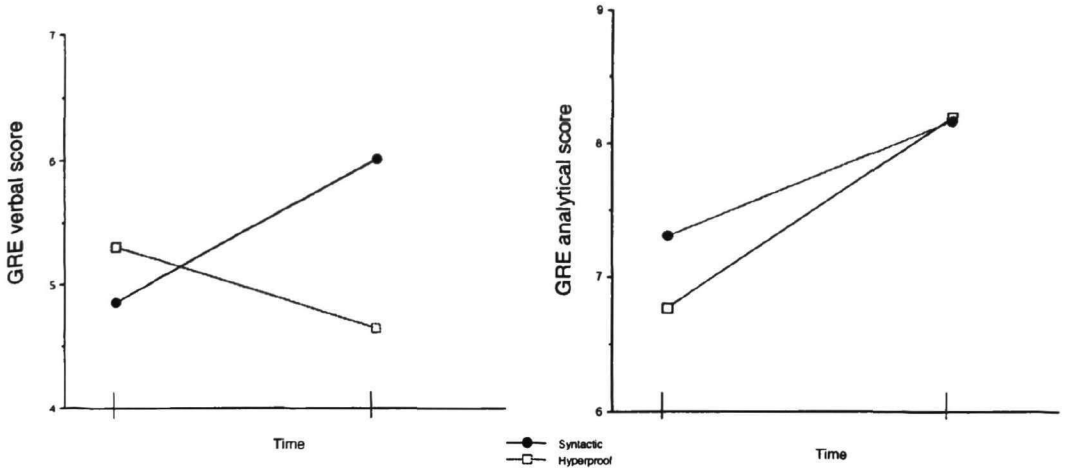

Figure 4: Hyperproof and Syntactic groups' mean scores on GRE subscales: (a) logical (argument analysis) (b) analytical

is a possible explanation of the blocks-world GRE discrepancies. Whether graphics are constructed by subjects or merely presented to them has been shown to be determining factor in their efficacy (Grossen & Carnine, 1990).

This difference in direction of movement of information may be related to some of the proof-style differences observed in HP use. DetHi and DetLo subjects seem to use graphical abstraction symbols in differing parts of the proof structure (cf. Oberlander, Cox and Stenning, 1994). DetHi subjects tend to exploit such symbols more extensively during the course of developing proofs. This appears to lead to more systematic and structured proof by cases, with deeper and more frequent use of nested structures. We are currently analysing these differences in the styles of final proofs, and we hope in due course to extend these analyses to examine the very rich HP logfile data, which reveal the time course of proof construction.

This tension in the results underlines an important potential contribution of basing a cognitive theory on *computational* properties of representations rather than on intuitively grounded concepts such as 'visual' and 'verbal' thinking. We expect further modelling of proof style to make a contribution to a cognitive characterisation of just what it is computationally to be a 'verbal' or 'visual' thinker.

These results suggest a number of possible improvements in the way that HP could be used in teaching. Perhaps the main conclusion of the present study is that those developments are almost certainly going to have to be sensitive to the pre-course aptitudes of different students. However, it is encouraging that at least some important aptitudes can be diagnosed very simply and could be built into student models within the HP environment.

The educational implications of these individual differences are far from clear. Should all students be taught to use graphical reasoning methods or should students be encouraged to follow their existing representational modality preferences? The second position suggests that instruction should be adapted to the (relatively immutable?) cognitive style of the learner. This is the approach advocated by Snow (eg Snow, Federico & Montague, 1980) based on studies of Aptitude-Treatment Interactions. It remains to be demonstrated, however, that the 'visualiser—verbaliser' dimension is unresponsive to educational intervention. Perhaps a domain-independent 'graphics curriculum' should be devised and generally taught? The authors tend towards the view that students should be encouraged to broaden their representational repertoires. We agree with Barwise (1993) that "efficient reasoning is inescapably heterogeneous (or 'hybrid') in nature" [p1]. We strongly disagree with those such as Dijsktra (1989, cited by Myers, 1990) who has described the use of graphical visualizations in teaching computer programming as "an obvious case of curriculum infantilization".

Acknowledgements

The support of the Economic and Social Research Council for HCRC is gratefully acknowledged. The work was supported by NATO Collaborative research grant 910954, and UK Joint Councils Initiative grant G9018050. The third author is supported by an EPSRC Advanced Fellowship. Special thanks to John Etchemendy and Mark Greaves.

References

Barwise, J. (1993). Heterogeneous reasoning. In G. Allwein & J. Barwise (Eds.) *Working Papers on Diagrams and Logic*. Preprint No. IULG–93–24, Indiana University Logic Group, May 1993.

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

Bergman, M., Moor, J. and Nelson, J. (1990). *The Logic Book*, New York: McGraw-Hill.

Brownstein, S. C., Weiner, M., and Weiner Green, S. (1990). How to prepare for the GRE. Barron's Educational Series, New York.

Cox, R. and Brna, P. (1993). Reasoning with external representations: Supporting the stages of selection, construction and use. In P. Brna, S. Ohlsson and H. Pain (Eds.) *Artificial Intelligence in Education, Proceedings of AI-ED93 World Conference on Artificial Intelligence in Education*, Charlottsville, VA: Association for the Advancement of Computing in Education (AACE).

Cox, R. and Oberlander, J. (1993). Graphical effects in learning logic: reasoning, representation and individual differences. In J. Oberlander (Ed.) *Semantic Issues in Graphical Representation*, ESPRIT Basic Research Action P6296, Deliverable 3.1, August.

Dijkstra, E. W. (1989). On the Cruelty of Really Teaching Computer Science. The SIGCSE Award Lecture, *CACM*, **32**, 1403–1404.

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

Educational Testing Service GRE Board, (1992). *Practicing to take the GRE General Test Number 9*. New Jersey: Warner.

Goldson, D. & Reeves, S. (1992). Using programs to teach logic to computer scientists. In D. Bateman & T. Hopkins (Eds) Proceedings of the Conference on Developments in the teaching of computer science, University of Kent, Canterbury, 167-176.

Grossen, G. & Carnine, D. (1990). Diagramming a logic strategy: Effects on difficult problem types and transfer. *Learning Disability Quarterly*, 13, 168–182.

Littman, D. and Soloway, E. (1988). Evaluating ITSs: The cognitive science perpective. In M.C. Polson and J.J. Richardson (Eds.) *Foundations of Intellident Tutoring Systems*, Hillsdale, N.J.: Lawrence Erlbaum Associates.

Myers, B. A. (1990). Taxonomies of Visual Programming and Program Visualization. *Journal of Visual Languages and Computing*, **1**, 97–123.

Nickerson, R. S., Perkins, D. N. and Smith, E. E. (1985). *The Teaching of Thinking*, Hillsdale, N.J.: Lawrence Erlbaum Associates.

Oberlander, J., Cox, R., and Stenning, K. (1994). Proof styles in multimodal reasoning. Presented at the 4th International Conference on Information-oriented approaches to Language, Logic and Computation, Moraga, Ca., June, 1994.

Shute, V.J. (1990). Golden promises of intelligent tutoring systems: Blossom or thorn? *Paper presented at the Space Operations, Applications and Research (SOAR) Symposium*, Albuquerque, N.M. June.

Snow, R.E., Federico, P-A. and Montague, W.E. (Eds.) (1980). *Aptitude, learning and instruction Volume 1: Cognitive process analyses of aptitude*, Hillsdale, N.J.: Lawrence Erlbaum Associates.

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