Contrasting the cognitive effects of graphical and sentential logic teaching: reasoning, representation and individual differences

Keith Stenning Richard Cox Jon Oberlander

Human Communication Research Centre
University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW

Short title Graphical and sentential logic teaching

Address for correspondence Keith Stenning, Human Communication Research Centre, University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW.

Abstract

Hyperproof is a computer program created by Barwise & Etchemendy for teaching logic using multimodal graphical and sentential representations. Elsewhere, we have proposed a theory of the cognitive impact of assigning information to different modalities. The theory predicts that the Hyperproof's devices for graphical abstraction will play a pivotal role in determining learning outcomes. Here, the claims are tested by a controlled comparison of the effects of teaching undergraduate classes using Hyperproof and a traditional syntactic teaching method. Results indicate that there is significant transfer from the logic courses to a range of verbal reasoning problems. There are also significant interactions between theoretically motivated pre-course aptitude measures and teaching method, and these interactions influence post-course reasoning performance in transfer domains. As well as being theoretically significant, the results provide support for the important practical conclusion that individual differences in aptitude should be taken into account in choosing teaching technique.

1 Introduction

Is reasoning teachable? What is the impact of learning logic upon general reasoning? What differences in cognitive effects emerge when graphical representations are used in teaching abstract subjects? These are important pedagogical questions. But the issues are also central to the development of theories of representation within cognitive science. In particular, what is the relation between logics of various kinds, and human reasoning and learning? Are heterogeneous logics, which treat information in more than the sentential modality, especially relevant to human thought? What is graphical representation good for and why? The present study seeks to explore these theoretical concerns in a highly practical setting—real undergraduate logic courses, taught using Hyperproof (HP) and conventional syntactic methods—by measuring pre- and post-course reasoning performance.

HP is a computer program for teaching first-order logic (FOL) which Barwise & Etchemendy (1994) designed and implemented on the basis of their situation theoretic approach to reasoning. HP builds on their earlier program, Tarski's World, which uses graphics to teach the syntax and semantics of FOL (Barwise & Etchemendy, 1993). To this, 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. An overview of HP is given in Section 3.

1.1 Practical perspective: logic and transfer

Teaching logic is an important practical concern. Very large numbers of undergraduates follow elementary logic courses. The prominent place of logic in the curriculum of university education was, and continues to be, justified in terms of its supposed effect on general reasoning abilities. There has been a widespread controversy amongst teachers of logic about the type of teaching which transfers to 'real-world' reasoning, with claims that 'informal logic' and 'critical thinking' courses—which eschew symbolic logic—transfer better. Quite apart from issues about what transfers best to domain independent reasoning skills, an increasing number of other disciplines require logic for specific technological purposes—notably computer science and electronic engineering. In these areas, graphical methods have been regarded with particular favour. For instance, Barwise & Etchemendy's creation of Hyperproof was motivated in part by the belief that integrating graphical and sentential representations in logic teaching can help students re-focus on the syntax/semantic relation, and thereby achieve better transfer.

Among teachers of logic and mathematics, such graphical methods remain controversial. In a recent discussion of beliefs about the nature of mathematics, Eisenberg (1992) points out that, in the mathematics community, the idea that mathematics must be communicated in a non-visual manner is "deeply rooted".

Attitudes towards graphical representations are sometimes seemingly inconsistent. Graphical representations can be assigned different status depending upon whether they are used as a conceptual aid or as a medium of communication. Eisenberg (1992) illustrates this by citing Hilbert:

> I have given a simplified proof of part (a) of Jordan's theorem. Of course, my proof is completely arithmetizable (otherwise it would be considered non-existent); but, investigating it, I never ceased thinking of the diagram (only thinking of a very twisted curve), and so do I still when remembering it.

It is therefore important to assess what impact logic courses in general, and innovative methods such as HP teaching in particular, have on subsequent reasoning abilities.

Psychologists have long been sceptical about the extent to which logical skills generalise to domain independent reasoning skills; the concern goes back at least to Thorndike (cf. Nickerson, Perkins & Smith, 1985; Nisbett, Fong, Lehman & Cheng, 1987 for discussions). Whether or not the notion of a mental logic is maintained (compare Johnson-Laird 1983 and Rips 1994), it is widely held that teaching logic influences only the algebraic symbol shuffling skills which commonly make up much of symbolic logic course exams—and for that matter, some logic courses. The argument appears to run—"Logic is a syntactic mechanism of reasoning. Humans do not reason syntactically. Therefore teaching logic will not help them reason."

This widespread belief is not empirically well-supported in the literature. There are studies showing that University courses which teach various reasoning skills do transfer to reasoning beyond course exams. For example, Nisbett et al. (1987) show that psychology courses which teach experimental method and statistical reasoning do transfer to reasoning in other domains, and that their performance in this respect exceeds that of, say, chemistry courses. Nonetheless, there has been little explicit study of the generalisation from logic teaching to general reasoning abilities. The theoretical literature on reasoning is based almost entirely on laboratory experiments within a narrow range of paradigms. Sometimes these paradigms have included 'therapy experiments' which incorporate elements of 'logic teaching', and these are sometimes seen to fail to correct reasoning fallacies. But these interventions bear little resemblance to real teaching in the logic classroom. Indeed, there is a more general problem over the relation between the laboratory tests and real-world reasoning. Studies have been made of 'real-world' reasoning on problems similar in form to laboratory paradigms, via observations of daily professional practice. Results here often indicate that practitioners are less susceptible to fallacies than decontextualised laboratory subjects; for example, in the domain of probabilistic reasoning compare Kahneman and Tversky (1972) to Koehler (1993).

For all these reasons, the psychological literature on human reasoning is much in need of empirical observations of real teaching, and the logic teaching

profession is much in need of a theoretical understanding of cognitive processes. Whilst most logic teachers are only too well aware of failures in their students' transfer of knowledge, there can scarcely be any discipline where transfer is not a problem. The transfer issue lies at the heart of the theory and practice of teaching.

1.2 Theoretical perspective: cognition and modality

Our own starting point was an account of the cognitive impact of assigning information to different modalities. The theory (described in Stenning & Oberlander (in press)) 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.

In the simplest graphical representation systems, one graphic stands for a single logical model. But most graphical representation systems observed in practical employment have various semantic devices for expressing limited abstractions. As we indicate in section 3, Hyperproof is no exception.

Our approach can be seen as both a generalisation and a refinement of Larkin & Simon (1987). While agreeing that the central cognitive advantages of graphics come from their computational tractability. they see searchability by parallel mechanisms and other properties of graphics as the major advantages. For us, weak expressiveness is the key. To see where the accounts diverge, consider the case of processing a weakly expressive sentential language equivalent to a graphical system. It will be possible to construct a processor which makes this sentential language equally tractable (and therefore computationally equivalent). The difference between the graphical and sentential systems lies in the discoverability of the limitations on expression and the necessary methods of exploiting them in inference.

Our notions of abstraction and expressiveness are also useful in distinguishing types of reasoning problem; they thus also contribute to the classification of 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 sufficient 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 abstractions. GRE-type examples are illustrated in Figure 1.

Insert Figure 1 about here

1.3 Structure of the paper

The present study is intended as an initial step towards the empirical observation of real logic teaching; it sets out to provide measurements of the impact on general reasoning, together with some analysis of the cognitive processes which determine transfer. The theoretical perspective we have just introduced helps guide this analysis: we believe that general theories of the semantics of graphics can illuminate the study of behaviour in specific problem domains. Computer-based courses offer a unique platform for assessing cognitive theories, since they permit detailed observation of students' reasoning processes.

The rest of the paper is structured as follows. Section 2 discusses a number of significant issues concerning the measurement of transfer effects. Section 3 sketches the Hyperproof interface, and Section 4 describes the method we followed in our study, including the various tests and the computer-based logging system we used. Section 5 details the results based on the analysis of the measures, and Section 6 interprets the results, discussing their broader implications.

2 Issues in Evaluating the Teaching of Reasoning

Educational computer software in general has rarely been evaluated in terms of effects upon learning outcomes, or with respect to general theories of cognitive representation and process. Reasoning in particular, as we emphasised in Section 1, is a domain in which regularities in performance across situations is a central issue. Reasoning is about generalisable inference based on problem *form*, but much of the moral of psychological investigations of reasoning is that problem form does not predict performance well—content has substantial effects. Indeed teaching logic can be seen as aimed at overcoming influences of content. The theory of reasoning is intimately involved with notions of *transfer* of performance, and so our selection of measures of pre- and post course is of more than usual relevance to our theoretical concerns. In this section we explain our choice of measures.

There have been several calls for more evaluative studies (for example, Littman & Soloway, 1988; Shute, 1990). Several computer-based FOL teaching programs have now been developed (Lancashire (1991) and Goldson, Reeves & Bornat (1993) provide a recent reviews); however, few have been evaluated in controlled comparisons. One system that was not included in either review is EPIC, a propositional (syntactic) logic tutor (Twidale, 1991). Twidale (1991) reports an evaluation of EPIC, using the users' form-filling, menu selection and text annotations as a means of detecting their rule instantiations, plans, and goals during proof development. An uncontrolled, observational study of 8 subjects revealed that EPIC provided "useful and usable" information about students' understanding. However, to the authors' knowledge, this is the only logic system about which published evaluative information is available.

In selecting outcome measures suitable to our theoretical and practical concerns it is hard to avoid 'classroom tests'. Ideally one might prefer tests of real-world reasoning that occur in the lives of our undergraduate students. Lacking the resources to identify or administer such tests, we sought a test of general reasoning ability which could be administered before and after both HP and conventional logic courses, and for which there was some evidence of real-world validity. We settled on two tests, one based on the graduate record exam (GRE) Analytical Reasoning Scale, and the other constructed to have a direct relation to HP graphics. We console ourselves that for undergraduate students, classroom tests *are* part of the real-world, and the GRE tests have an important impact on real-life outcomes.

The GRE exam is taken by US undergraduates, and, together with course grades, plays a major role in determining entry into Graduate Schools. The Analytical Reasoning Scale of the GRE was introduced into the exam in 1977, in response to the view of graduate school teaching staff that the then current exam lacked a test of abstract reasoning ability. Abstract reasoning was favoured by staff and students as a means of broadening the GRE over alternative scales measuring scientific thinking and study style (Miller & Wild, 1979).

Swinton & Powers (1983:104) write that "...the (analytical portion of the) test is intended to measure analytical reasoning abilities that, like the verbal and quantitative skills measured by the test, are assumed to develop over a relatively long period of time." It is claimed that the test does not require specialised domain knowledge and is relatively resistant to the effects of coaching. However, Swinton & Powers (1983) showed that certain types of item in the GRE analytical scale were susceptible to the effects of a "brief curriculum of special preparation". Those item types were eliminated from the analytic measure in 1981 (Emmerich, Enright, Rock & Tucker, 1991). Hence the current GRE analytical reasoning scale contains two types of item: analytical reasoning problems, which are usually constraint satisfaction puzzles for which diagrams are often useful; and logical reasoning problems, which involve argument analysis, a kind of verbal reasoning problem. Examples of items from each subscale have been provided in Figure 1.

Although the GRE is a purely verbal test, it was of some interest to us to discover that its constructors had felt it necessary to include two subscales of problems, and that these subscales were related to our prior theoretical distinction between problems suited or not suited to graphical representation.

An analysis of the content characteristics of analytical reasoning items by Chalifour & Powers (1989) revealed that the difficulty of analytical reasoning items is predicted by a number of factors. Factors that are positively correlated with difficulty include: the usefulness of drawing diagrams (the greater the usefulness, the more difficult), the number of words in the stimulus, the number of rules and the amount of information from the rules or conditions needed for a solution. The number of unvarying assignments of entities to position[1] was negatively correlated with item difficulty; that is, the more explicitly given determinate information, the easier the problem.

The GRE *verbal* and *quantitative* subscales do not add a great deal of predictive utility to the Scholastic Aptitude Test (SAT), which is taken at secondary school four or five years earlier than the GRE. In a sample of 22,923 subjects, the GRE verbal scale was observed to correlate highly with SAT verbal scores ($r = .858$) and GRE quantitative also correlates highly with SAT mathematical scores ($r = .862$) (Angoff & Johnson, 1990). Hence there was also a need for a GRE scale designed to include aspects of reasoning not measured by the SAT.

It does in fact predict graduate school performance in a wide range of disciplines separately from other measures such as domain specific undergraduate performance. Although graduate school performance might be criticised as still too academic a measure of 'real-world' reasoning, and although this psychometric evidence is necessarily correlational, it was felt that these credentials give the GRE analytical reasoning scale a certain 'street credibility' as a test of domain independent reasoning.

The GRE has an additional advantage for this study: it is a verbal test posed in English, eliciting selections of verbal answers. It presents no diagrams and there is no opportunity to present the results of reasoning diagrammatically, though there is room on the test questionnaire for the construction of external representations and therefore candidates can engage in diagrammatic and other representational activity in the course of the test. Below we analyse some of this activity as evidence of the impact of teaching on reasoning processes. Any test which used graphics in the presentation of problems or elicitation of answers might be held to favour HP-type graphical teaching strategies. As matters stand, the GRE might be seen as favouring linguistic teaching.

As a second test of reasoning ability we constructed what we call the 'Blocks-world' test based on HP graphics but posed in English so that it could be used as a pre-test with students ignorant of FOL formalism. This test is evidently more closely related to the HP curriculum than to conventional logic course problems.

[1]For example, in the office allocation example, statements of the kind "Ms Green, the senior employee, is entitled to Office 5, which has the largest window."

The test nevertheless consists of spatial reasoning problems independent of any narrow domain knowledge. It was intended to provide some insight into the students' reasoning abilities in the HP domain prior to logic teaching.

Before proceeding to describe the method we pursued in our investigations, it is perhaps worth noting what our evaluation was *not* supposed to achieve. We did not set out to find out whether multimodal teaching was 'better' than conventional syntactic approaches. Rather, we were exploring the effects of the courses on cognitive processes, using our background theory as a guide. Our methods therefore focus on the effects of teaching on different problems and different groups of students, and we concentrate on transfer effects. The reader should also bear in mind that the HP course observed in this study was the very first course taught using this system. As such it is being compared to conventional 'syntactic' teaching methods which have been developed in a rich tradition of about half a century. Verdicts would be premature, but systematic evaluation can contribute to the acceleration of future developments.

3 The Hyperproof Interface

As can be seen in Figure 2, the HP 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. In addition, control palettes are also available. The various 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. HP should be viewed as a proof-checking environment designed to support human theorem proving using heterogeneous information.

Insert Figure 2 about here

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, and grasping the employment of semantic devices to avoid combinatorial explosions. Hyperproof presents graphical views of a blocks-world, but uses various devices to overcome graphical specificities. First, a single diagram can contain symbols representing objects of unknown size and

unknown shape. When an object's size is unknown, it is represented by a cylinder, sporting a badge indicating the object's shape. When size is known, but shape is not, then the object is represented by a *paper bag* of the relevant size. When neither shape nor size is known, the object is represented by a cylinder whose badge is a question mark. Secondly, by the use of a special part of the diagram—off the chequer-board—objects with unknown location can be represented. Thirdly, disjunctions of information about an object (for instance, the fact that it is either a dodecahedron or a tetrahedron) are captured by multiple diagrams. Finally, blocks may be either presented with or without a visible label. Several of these devices are illustrated in Figure 3.

Insert Figure 3 around here

---

4 Method

4.1 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.

4.2 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 & 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—and hence syntactically oriented—instructional text (Bergman, Moor & 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. 23 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.

4.3 Pre- and post-teaching tests

The objective 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 arrangement 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 & 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. As we observed earlier, 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 (such as 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 could 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.

5 Results

We report the scores on the Blocks-world and GRE tests, followed by analysis of the representation selections made in the GRE test, and finish with analyses of the proof-styles that appear in the examination exercises of the HP students.

5.1 Blocks-world test scores

In order to examine the effects of training modality with respect to 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 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. Figures 4 and 5 show 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 group means were 4.68 (Hyperproof group) and 3.73 (syntactic group).

The main effect for DetHi/Lo was also significant ($F(1,26) = 12.81, MSe = 2.33, p < .001$). As shown in Figures 4, and 5, the DetHi subjects tend to score higher than the DetLo subjects.

Insert Figure 4 about here

Insert Figure 5 about here

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 Hyperproof subjects' scores did not differ from those of their DetLo colleagues on the pre-training test, but a post hoc test revealed that,

following the HP course, they scored significantly higher than their DetLo ounterparts ($t = 4.42, df = 15, p < .001$). Conversely, Newman Keuls post hoc tests of pre to post test changes (that is: within-subjects factor),[2] revealed that DetHi subjects in the syntactic group significantly decreased in their Blocks World performance.

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 $t = -2.29, df = 11, p = .042$. The reasons for this result are unclear, but may be due to selection biases.

5.2 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 6. 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 7). In this case HP did improve the measure more than the syntactic class, but not significantly so.

Insert Figure 6 about here

Insert Figure 7 about here

---

6 Discussion

We began this study influenced by the strong psychological folklore that led us to expect that logic courses would not transfer to more general reasoning

[2] As recommended by Winer (1971).

behaviour. We expected that Hyperproof teaching might transfer properties different from those of conventional teaching. The results show good transfer of learning on both conventional and HP courses, but strong interactions between pre-existing individual differences and methods of teaching. The results do not indicate that either course is a globally better way of teaching, though it is perhaps remarkable that the very first foray with HP should be able to stand comparison with well established methods.

Are our transfer tests adequate tests of impact on general reasoning abilities? No pencil-and-paper classroom test can provide the perfect measure of classroom learning for predicting subsequent real-world behaviour. But as we argued in the introduction, the GRE's role in graduate student selection gives it some credibility as a test of transfer. This is supported by our observations that the courses alter students' spontaneous representational constructions in the GRE test, and this is known to improve performance on the GRE, so there is at least the suggestion of one mechanism of improvement (Cox & Stenning (in preparation)). The blocks-world test has some credibility as a reasoning task from a domain commonly encountered by everyone. These verbal problems probably require more complex *reasoning* than most commonly encountered tasks. However, if they require any domain specific *knowledge*, it is of the kind that all students are bound to possess.

If the tests of transfer are reasonable, why should we find transfer of teaching when so many are of the opinion that logic teaching does not transfer? And how far can we generalise from the current finding? Part of the answer is probably that current opinion is not based on a sound empirical foundation. The laboratory studies of reasoning do not look at the effects of real logic courses taught by staff with a belief in the relevance of logic to general reasoning. By the same token, it is important to remember that the courses studied here were taught by highly skilled and dedicated teachers to classes of able, well-motivated students in a prestigious educational institution. The 'syntactic' course was taught using conventional methods, but it would be most unwise to assume that even specifying the reasoning system and the text book and the student population would standardise the results. The syntactic course was taught by a teacher who strongly believes that logic teaching is about grasping the relation between the syntax and semantics of reasoning systems. It is possible that a teacher with a more superficially syntactic approach might be able to teach both syntactic and HP courses and achieve no transfer of learning at all. Should our results have shown that HP achieved *better* transfer of learning, our results might have been interpreted as being due to a poorly taught control course. A great deal more research will be needed to find out what characteristics of courses lead to transfer. We believe this study provides a useful existence proof that logic courses of a variety of kinds *can* transfer to other reasoning.

The other unexpected message in these results was perhaps the clearest—that pre-existing individual differences between students had a substantial impact on their response to the different types of teaching. 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 an 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 conventional logic courses. (Of course they are also important in real applications of logic). 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. Post-hoc analyses of items in the blocks-world post-test suggests that the decline in reasoning performance amongst DetHi students in the syntactic course is due to difficulties with questions about logical independence. These questions are particularly susceptible to HP-style case-based model-construction strategies. It is plausible that syntactic teaching discourages these students from their natural tendency to reason in this suitable style, and thus leads to a decrement in performance.

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 effect of syntactic teaching might be expected. An examination of the GRE verbal subscale items suggests that some of the conventional curriculum on inductive reasoning and its relation to conditionals might well be added to the HP curriculum. It is perhaps surprising that syntactic teaching also improves determinate problem performance as much as HP teaching.

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 involves applying information from a set of sentential constraints to a *presented* diagram. A GRE analytical reasoning problem involves *constructing* 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 is a possible explanation of the discrepancies between the blocks-world and GRE tests. Whether graphics are constructed by subjects or merely presented to them has been shown in other settings to be a determining factor in their efficacy (Grossen & Carnine, 1990).

The present study yielded data which will eventually provide a more analytic account of the differences in mental processes between the groups of students.

Two important sources of data that we have not been able to analyse here are the 'work scratchings' which accompany students' answers on the GRE, and the proof logs from Hyperproof sessions. Of the latter, exam proofs turn out to be much richer than 'example' logs, probably because the latter were accompanied by advice which lead to rather uniform data.

The work scratchings data is reported in Cox & Stenning (in preparation). By classifying the types of representation spontaneously used by subjects in doing GRE model problem reasoning, analysis shows that kind of representation selected is an important determinant of success at reasoning. Furthermore, this ability to select 'suitable' representations is differentially affected by the different logic courses—Hyperproof teaching improved representation selection accuracy whereas syntactic teaching did not.

Preliminary analysis of exam log data is reported in Oberlander, Cox & Stenning 1994. Patterns of rule-use in performing proofs do differ between DetLo and DetHi students. Use of graphical abstraction symbols does differ between these groups on some questions, and in the direction predicted by the theory. Inspection of the proofs suggests that DetHi subjects use graphical abstraction symbols to produce more hierarchically organised proofs whereas DetLo subjects produce flat proofs which are unstructured lists of 'cases'. (See Oberlander, Cox & Stenning 1994 for an extended presentation of some examples).

Although conclusions about detailed mental processes must await further analysis of this data, the evidence presented here already indicates both that different teaching methods can induce opposite effects in different groups of students, and that the *same* teaching method administered in a strictly controlled computerised environment using the same examples, and the same advice can induce different groups of students to develop quite distinct reasoning styles.

It is particularly intriguing that the DetLo/Hi distinction is drawn on the basis of a 'purely verbal' reasoning test, which yet proves capable of predicting differences in student response to graphical and sentential teaching methods. Intuitions must give way to computational accounts. We expect further modelling of HP proof-styles 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 is compatible with the view that the cognitive style of the learner is relatively immutable, and that it is best to adapt instruction to style, rather than *vice versa*. This is the approach advocated by

Snow based on studies of Aptitude-Treatment Interactions (cf. Snow, Federico & Montague, 1980). To the authors knowledge, only one study has demonstrated that the 'visualiser—verbaliser' dimension is responsive to educational intervention (Frandsen & Holder, 1969).

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 (1993:1) that "efficient reasoning is inescapably heterogeneous (or 'hybrid') in nature". 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 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 third author is supported by an EPSRC Advanced Fellowship. Special thanks to John Etchemendy, Tom Burke and Mark Greaves.

References

Angoff, W. H. and Johnson, E. G. (1990). The differential impact of curriculum on aptitude test scores. *Journal of Educational Measurement*, **27**, 4, 291–305.

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. (1993). *The Language of First-Order Logic, Including the Macintosh Program Tarski's World 4.0*. CSLI Lecture Notes. Chicago: Chicago University Press.

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.

Chalifour, C.L. and Powers, D.E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement*, **26**, 2, 120–132.

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.

Cox, R. and Stenning, K. (in preparation). The effect of logic teaching modality on on representation selection accuracy. To be submitted to the *Japanese Journal of Cognitive Science*.

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, Princeton, NJ: Educational Testing Service.

Eisenberg, T. (1992). The development of a sense for functions. In Harel, G. and Dubinsky, E. (Eds.) *The Concept of Function*, Mathematical Association of America, MAA Notes Volume 25.

Emmerich, W., Enright, M. K., Rock, D. A. and Tucker, C. (1991). *The development, investigation and evaluation of new item types for the GRE analytical measure*, Princeton, NJ: Educational Testing Service.

Frandsen, A. N. & Holder, J. R. (1969). Spatial visualization in solving complex verbal problems. *Journal of Psychology*, **73**, 229–233.

Goldson, D., Reeves, S. and Bornat, R. (1993). A reviews of several programs for the teaching of logic. *The Computer Journal*, **36**,4, 373–386.

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

Johnson-Laird, P. N. (1983). *Mental Models*, Cambridge, MA: Cambridge University Press.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430–453.

Koehler, J. J. (1993). The base rate fallacy myth. *Psycoloquy electronic journal*, 93.4.49, November 9th. ISSN 1055-0143.

Lancashire, I. (Ed.) (1991). *The humanities computing yearbook: A comprehensive guide to software and other resources*. Oxford: Clarendon Press.

Larkin, J. H. and Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, **11**, 65–100.

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

Miller, R. and Wild, C.L. (1979). *Restructuring the Graduate Record Examinations Aptitude Test*, Princeton, NJ: Educational Testing Service.

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.

Nisbett, R. E., Fong, G. T., Lehman, D. R. and Cheng, P. W. (1987). Teaching reasoning. *Science*, **238**, 625–631.

Oberlander, J., Cox, R., and Stenning, K. (1994). Proof styles in multimodal reasoning. Presented at the International Conference on Information-oriented approaches to Language, Logic and Computation, Moraga, CA, June, 1994. To appear in Seligman, J. and Westerstahl, D. (Eds.) *Language, Logic and Computation: The 1994 Moraga Proceedings*. Stanford: CSLI Publications.

Rips, L. J. (1994). *The Psychology of Proof*, Cambridge, MA: MIT Press.

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. (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.

Swinton, S. S. and Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. *Journal of Educational Psychology*, **75**, 1, 104–115.

Twidale, M. (1991). Improving error diagnosis using intermediate representations. *Instructional Science*, **20**, 359–387.

Winer, B. (1971). *Statistical principles in experimental design.*, New York: McGraw-Hill.

Figures

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. 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.

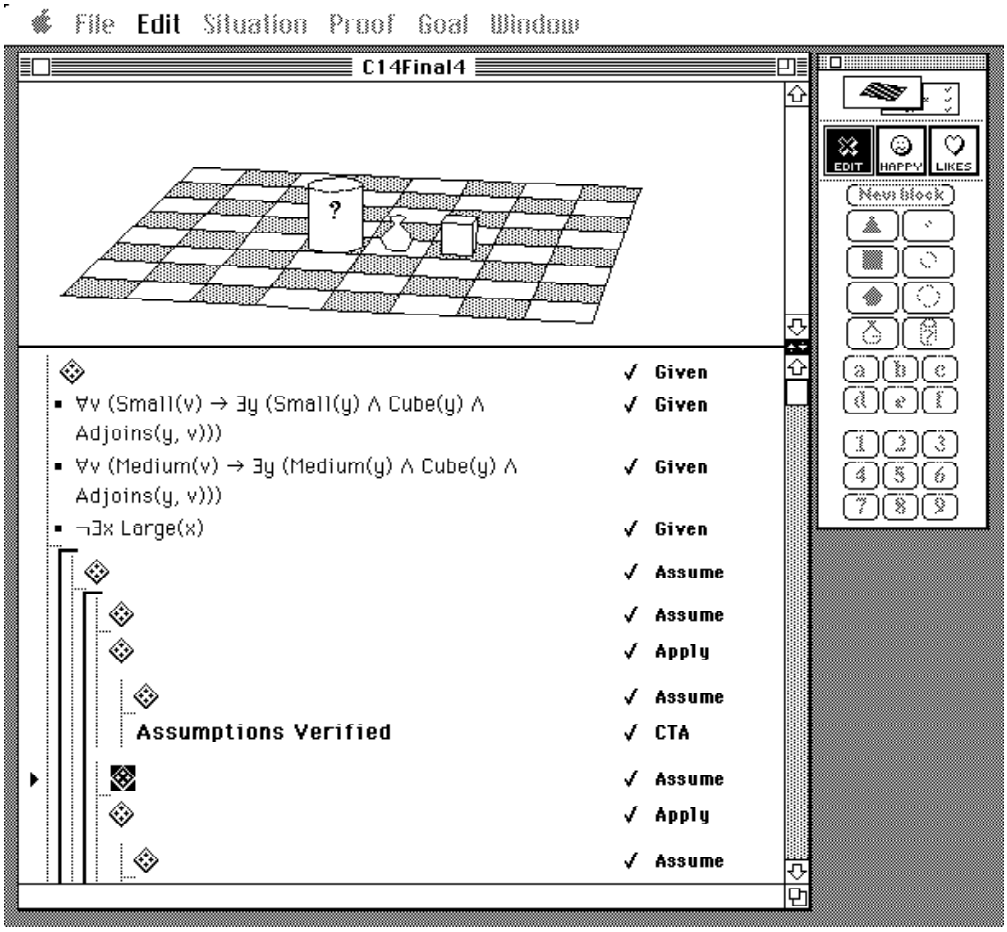

Figure 2: The Hyperproof interface. The main window panes—graphical and calculus—are supplemented by control palettes. The situation being viewed is the fifth in the course of the proof, and corresponds to the fifth diamond-shaped 'situation' icon in the body of the proof. The graphical window pane contains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

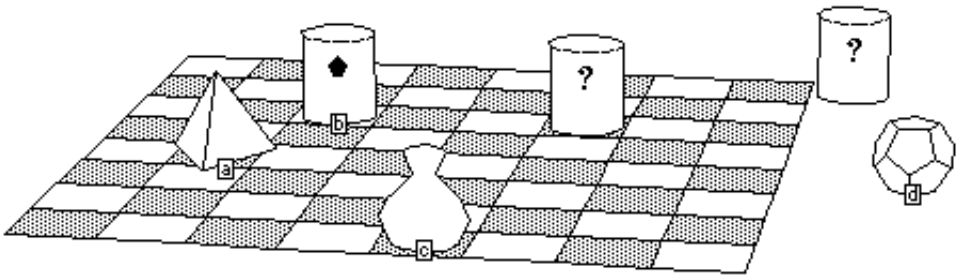

Figure 3: Abstraction in Hyperproof. The graphical situation depicted contains symbols of varying degrees and types of abstraction. The large tetrahedron labelled a is completely concrete. The cylinder with the dodecahedron badge labelled b lacks only a size attribute, and the large paper bag labelled c lacks only a shape attribute. The first unlabelled cylinder with the question mark badge lacks size, shape and label attributes, but still has a position; its twin off the chequer-board lacks even position. The neighbouring medium sized dodecahedron labelled d lacks only a position.

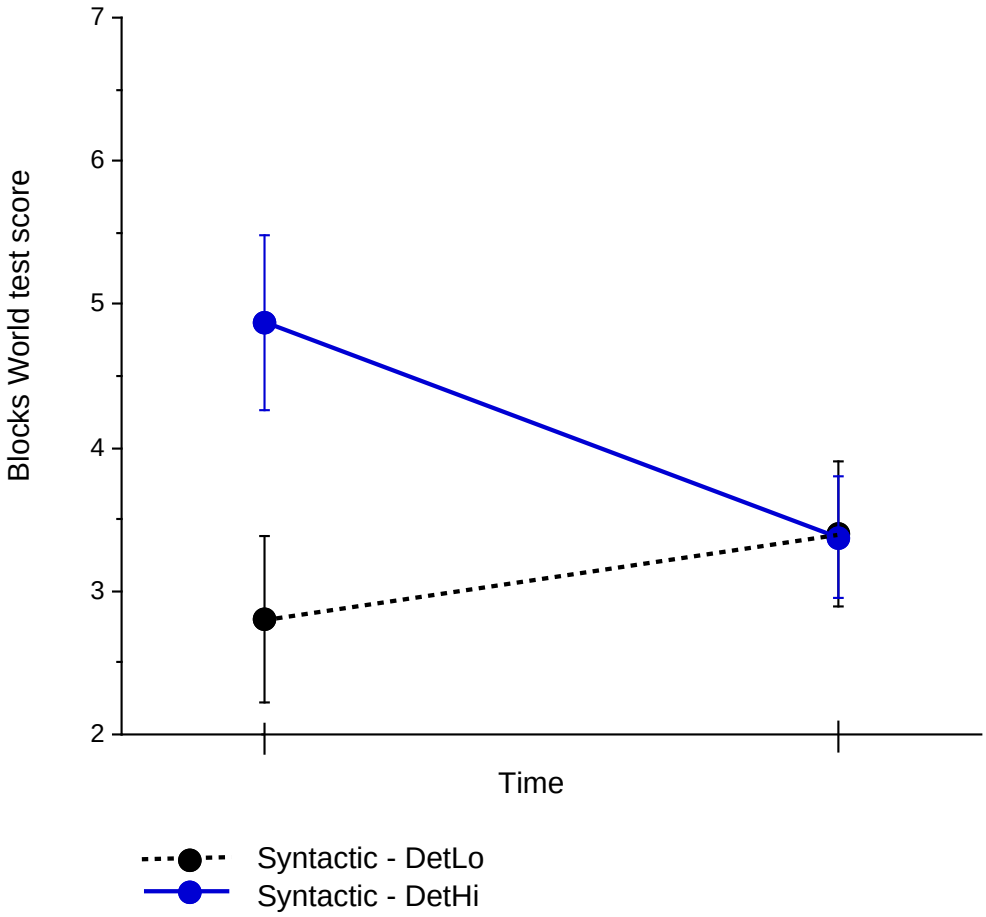

Figure 4: Mean score for Syntactic subjects on Blocks-World test as a function of subjects' performance on analytical reasoning GRE subscale (DetHi/Lo) by Time (Pre/Post)

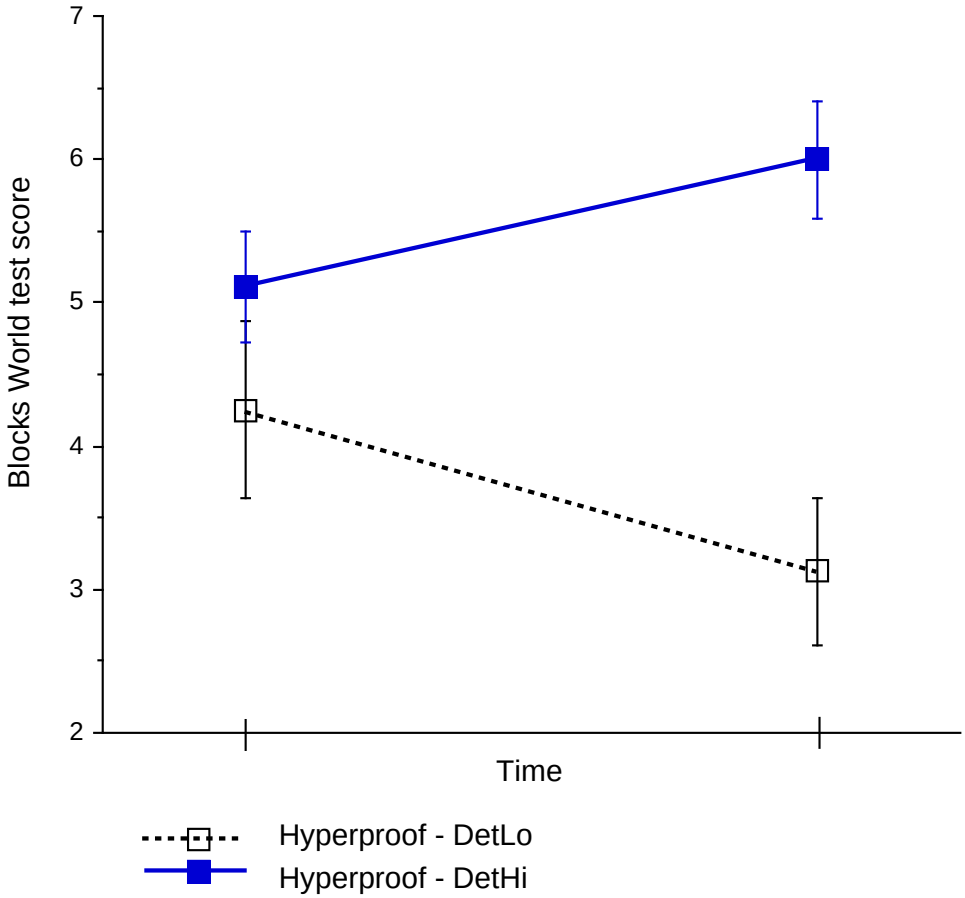

Figure 5: Mean score for Hyperproof subjects on Blocks-World test as a function of subjects' performance on analytical reasoning GRE subscale (DetHi/Lo) by Time (Pre/Post)

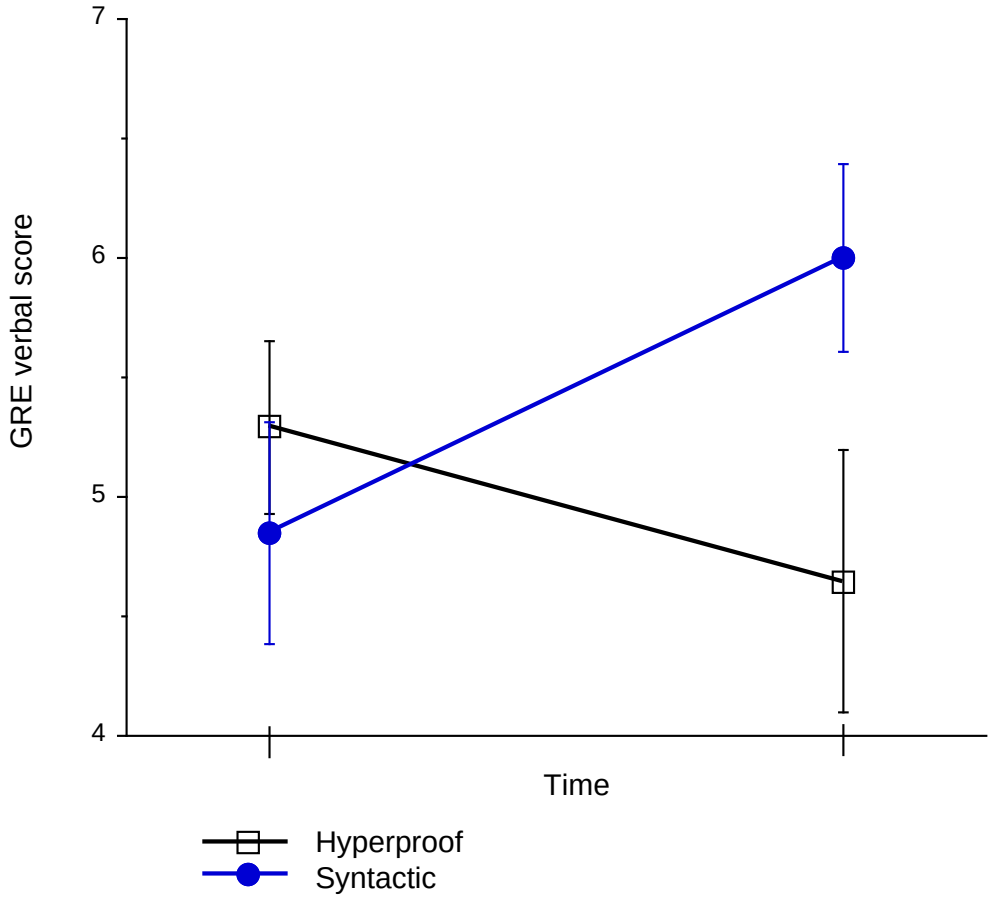

Figure 6: Hyperproof and Syntactic groups' mean scores on GRE subscales: (a) logical (argument analysis)

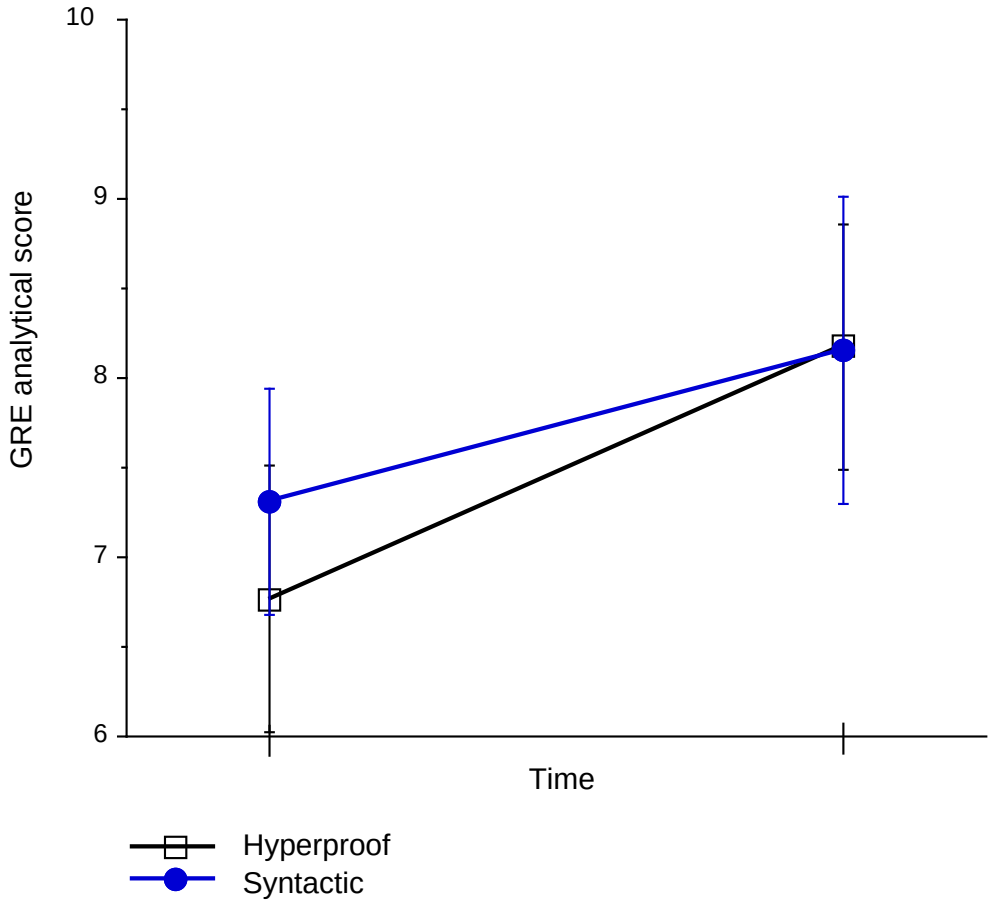

Figure 7: Hyperproof and Syntactic groups' mean scores on GRE subscales: (b) analytical

Table 1:A set of relevant Hyperproof rules.

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