Efficacious Logic Instruction: People Are Not Irremediably Poor Deductive Reasoners

Kelsey J. Rinella (rinelk@rpi.edu)
Department of Philosophy, Psychology & Cognitive Science
The Minds & Machines Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy, NY 12180 USA

Selmer Bringsjord (selmer@rpi.edu)
Department of Philosophy, Psychology & Cognitive Science
Department of Computer Science
The Minds & Machines Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy, NY 12180 USA

Yingrui Yang (yangyri@rpi.edu)
Department of Philosophy, Psychology & Cognitive Science
The Minds & Machines Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy, NY 12180 USA

Abstract

Cheng and Holyoak, and many others in psychology and cognitive science, subscribe to the view that humans have little context-independent deductive capacity, and that they can't acquire such a capacity through instruction in formal logic. This position is based, in no small part, upon C&H's well-known investigation of the efficacy of an undergraduate course in logic in improving performance on problems related to Wason's Selection Task, in which they found the benefit of such training to be minimal (Cheng, Holyoak, Nisbett, & Oliver, 1986). We believe, based on the encouraging results of a new study involving a similar pre-test/post-test design on a logic class at RPI, that the results obtained in the Cheng & Holyoak study serve to highlight problems with the way logic has historically been taught (related to techniques unavailable or impractical before the advent of heavy computer saturation in higher education), rather than to suggest that humans are unable to learn to reason. This prompted the reevaluation of conclusions based on C&H's research, requiring a new theory of meta-reasoning, Mental MetaLogic.

Introduction

The backlash against Piaget's claims (e.g., see the claims in Inhelder and Piaget, 1958) that humans naturally acquire competence in (elementary extensional) logic has "ruled the roost" in the psychology of reasoning for some time. Recently there has been some thought that perhaps the inherent irrationality of the species has been exaggerated (see Bringsjord, Noel, & Bringsjord, 1998; Evans & Over, 1996; Rips, 1994). This article is targeted specifically at the claims made by Cheng et. al. (1986) that not only are humans inherently bad at logic, but we are unable through training in formal logic to learn how to reason in abstract, context-independent fashion. One of the experiments they report, experiment 2, involves a pre-test/post-test design in which students in a logic class are tested on their understanding of how the conditional works in examples—the improvement they report is minimal. Using the same design, but a different instructional method, our results indicated a significantly greater improvement.

The three major reasons put forth in this presentation that the logic class at RPI differed from those in previous studies is that it taught disproofs, diagrammatic techniques, and, "Rigorous and general-purpose procedures for formalizing natural language logic problems in first-order logic so that they can then be solved by automated theorem provers". (For more on this last technique, see Bringsjord & Ferrucci, 2000.) Briefly, disproofs are proofs that one sentence does not follow from a set (possibly empty) of givens. Put another way, they are proofs that, given whatever premises one has, it is not possible to prove the goal, nor is it possible to prove the negation of the goal. The software used in the course, HYPERPROOF (Barwise

and Etchemendy, 1994), allows students to see how sentential information in first-order logic interacts with a toy world which acts as the domain of discourse (were it not for the existence of this world, it would not be possible to perform disproofs in a way remotely similar to that used in the RPI course) through experimentation and practice problems. Because of this, our students learn the meanings of the sentences in a much more understandable fashion while retaining the abstractness and universality of formal logic—these are the diagrammatic techniques. Finally, one of the most challenging tasks involved in solving many of the problems presented by psychologists of reasoning is finding the intended content in the words presented. The translation procedures mentioned above allow students to make fewer errors on these sorts of tasks. The virtues of these advances are discussed in some detail in Bringsjord, Noel, & Bringsjord (1998), with additional data presented in Bringsjord & Rinella (1999).

To demonstrate the abovementioned diagrammatic techniques from HYPERPROOF, which may be unfamiliar to many readers, consider figure 1:

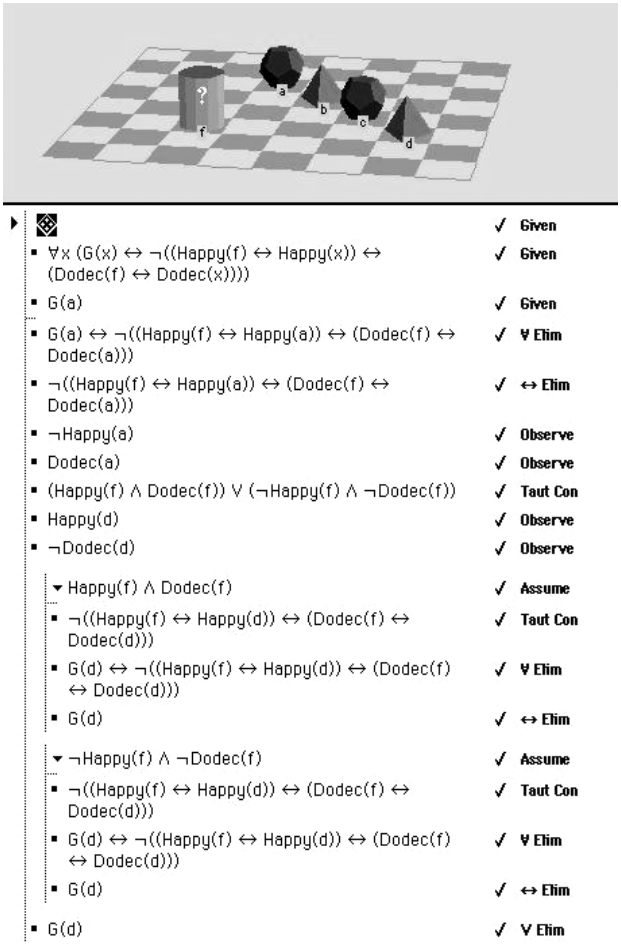


Figure 1:The THOG Problem

The HYPERPROOF window consists of two major areas: on top, there is the toy world, which shows the locations and properties of a small number of objects (in this case, five); on the bottom, sentential logic inferences proceed toward the goal. This particular example is a formalization of the THOG problem, a common problem used in the psychology of reasoning. The first given sentence claims that an object has the property G if and only if it has either the emotional state (happy or unhappy) or the shape (dodecahedron or tetrahedron) of the object f, but not both. Since we don't know f's shape (HYPERPROOF hides shape when it is not know by placing a cylindrical box over the object, which is why f appears to be a cylinder) or emotional state, we need the information in the second given, that object a has property G, to determine which of the other objects has property G. The proof proceeds by first manipulating the givens to extract the information that object a has either the emotional state or shape of f, but not both. Unlike regular sentential logic, we are then able to observe from the world the status of a, noting that it is unhappy and a dodecahedron (in some problems, it is actually necessary to use information from the sentential section to add information to the world, often allowing the user to detect an object's location, shape, or size, so this information moves up to the world as well as down to the sentences). From this, we infer that f must be either a happy dodecahedron or an unhappy object of a different shape. Finally, we show that, in either of these instances, object d also has property G, and conclude by stating that d must have G. Students using this system have the advantage that they are able to see what the sentences mean—rather than proceeding merely by manipulating the symbols of the sentences according to rules they have learned by rote, they begin to understand how different configurations of objects alter the effects of different sentences.

Method

We gave students enrolled in Rensselaer Polytechnic Institute's Introduction to Logic class in the Fall term of 1998 a pre-test including Wason's Selection Task as problem one, the THOG problem as problem two, and five other problems from previous work by psychologists of reasoning or from experience with tests of logic encountered by students in other contexts (e.g., two of the problems were straightforward adaptations of problems the Board of Regents of New York State say every New York high school students should be able to solve; Verzoni & Swan, 1995). A similar test, mathematically matched for problem type and difficulty, was given as a post-test appended to the final exam. Though there is insufficient space here to present them, the complete pre-test and post-test are available online:

http://www.rpi.edu/~faheyj2/SB/INTLOG/pre-test.f98.pdf, http://www.rpi.edu/~faheyj2/SB/INTLOG/post-test.f98.pdf. An example pair of THOG-like problems follows. In both cases, of course, students had to provide correct justifications.

> **2**
> Suppose that there are four possible kinds of objects:
>
> - an unhappy dodecahedron
> - a happy dodecahedron
> - an unhappy cube
> - a happy cube
>
> Suppose as well that I have written down on a hidden piece of paper one of the attitudes (unhappy or happy) and one of the shapes (dodecahedron or cube). Now read the following rule carefully:
>
> - An object is a GOKE if and only if it has either the attitude I have written down, or the shape I have written down, but not both.
>
> I will tell you that the unhappy dodecahedron is a GOKE. Which of the other objects, if any, is a GOKE?

The analogous problem on the post-test was the following:

> **2**
> Suppose that there are four possible kinds of objects:
>
> - an smart tetrahedron
> - a stupid tetrahedron
> - a smart cube
> - a stupid cube
>
> Suppose as well that I have written down on a hidden piece of paper one of the mental attributes (smart/stupid) and one of the shapes (tetrahedron/cube). Now read the following rule carefully:
>
> - An object is a LOKE if and only if it has neither the mental attribute I have written down, nor the shape I have written down.
>
> I will tell you that the stupid tetrahedron is a LOKE. Which of the other objects, if any, is a LOKE?

Note that a direct, unreflective transfer of reasoning brought to bear on the first of these problems to the second won't yield a solution to the second. This pair of problems (and this holds true for each pair on our pre-test/post-test combination) will not "match" at the surface level in English, nor at such a level in the propositional calculus. However, we needed pairs of problems that, at the level of proof or disproof, could be said to be very similar, formally speaking. Without such mathematical similarity we wouldn't be able to justifiably say that the problems, from the standpoint of formal deduction, are essentially the same. Figure 1 above presents a proof-theoretic solution to the first of the THOG-like problems—it is at this level of detail that difficulty must be matched without allowing the same argument form to work a second time.

Subjects who took only one of the two tests were discarded, to ensure that every participant had exposure to the entire course, leaving exactly 100 participants. After the first test, we abandoned asking the subjects whether they had seen the question before. There were two reasons for this: prior experience did not correlate with success on the questions, and the problems on the post-test were so similar in theme and difficulty that it was very likely that their experience with the pre-test would generate false positive responses. We also of course asked for justifications for their answers, hoping that out of the data we would be able to divine an appropriate scheme for categorizing the unstructured and heterogeneous responses we were likely to get.

As a preface to the first test, we gathered some biographical information, including the sex of the participants, the location of their high school, and previous logic experience. Since New York's Board of Regents has decreed that students must learn logic in their math courses, we hypothesized that attending high school in New York state would increase performance on tests of reasoning. We also hypothesized that previous experience in logic would increase scores on the pre-test, but that this effect would be reduced or eliminated by the post-test.

Results & Discussion

As expected, the averages on the pre-test were significantly lower than on the post-test, 3.89 correct compared to 5.11 correct. A paired-samples t-test reported an extremely low (t = -8.393) probability of no effect, suggesting that taking the logic class did improve students' ability to reason logically. Full results for each of the questions appear in table 1, below:

Table 1: Individual Question Results

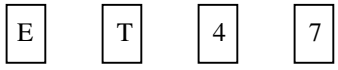
	Test 1	Test 2	t	Significance
Question 1	29	84	-9.563	0.000
Question 2	72	83	-2.076	0.040

Question 3	77	94	-3.597	0.001
Question 4	55	80	-4.639	0.000
Question 5	90	98	-2.934	0.004
Question 6	7	58	-9.768	0.000
Question 7	59	14	7.595	0.000

Though the improvement was significant at the .01 level for each of the first six questions except question two (which had a problem with a ceiling effect, but was still significant at the .05 level), there were three questions that particularly attracted our attention. Questions one and six showed extremely low initial rates of success, but great improvement—this suggests that these question types may be particularly amenable to improvement by instruction in formal logic. Question seven totally reversed our expectations—students did markedly worse on the post-test.

Individual Question Findings

The first result of some import is the comparison of Wason's Selection Task and its analogue (in each case, question one) on the post-test. These problems were chosen to test the ability of students to comprehend the use of the conditional in a context-free setting. The difficulty our subjects had on the pre-test with this problem very much agrees with the performance of Cheng & Holyoak's participants on their pre-test (1985). From this poor performance, and the lack of improvement, Cheng & Holyoak concluded that people are not good at using the conditional in a context-independent manner. On the pre-test, the problem looked like this:

> **1**
> Suppose that I have a pack of cards each of which has a letter written on one side and a number written on the other side. Suppose in addition that I claim the following rule is true:
>
> - If a card has a vowel on one side, then it has an even number on the other side.
>
> Imagine that I now show you four cards from the pack:
>
> 
>
>
> Which card or cards should you turn over in order to decide whether the rule is true or false?

The analogous problem (from Verzoni & Swan, 1995) on the post-test follows:

> **1**
> Suppose that you are doing an experiment for a biology expedition. You learn before starting on this expedition that insects can be one of two kinds, a spade fly or a bevel wasp, and that insect color is either black or green. Your task is to study insects in order to find out if a certain rule is false. The rule is:
>
> - If an insect is a spade fly, then it is black.
>
> You see an insect that is green. Which of the following would be true about the insect if it violates the rule?
>
> **a** The insect is a spade fly.
> **b** The insect is a bevel wasp
> **c** The type of insect does not matter.

Because these are the problems which are identical in underlying form to those used by Cheng et. al. in the aforementioned 1986 study, we were quite pleased to discover that our methods had induced an improvement from 29 correct responses to 84 correct responses, an extremely impressive improvement. This confirms our initial hypothesis, and allows our results to very directly be compared with previous work.

The second question which drew our attention because of the extremely poor (well below chance) performance on the pre-test. Since this was the question relating to *reductio ad absurdum,* or proof by contradiction, which is an integral part of the work our students do with HYPERPROOF during the semester. Such a proof, from the standpoint of the psychology of reasoning (which focuses on untrained reasoning), is exotic, but from the standpoint of mathematics and mathematical logic, it's thoroughly routine, and is therefore part and parcel of an introductory course of the type we offered. The full text of this question from the pre-test follows:

> **6**
> We will use lower-case Roman letters *a, b, c, …* to represent propositions. Let the symbol '¬' stand for 'it is not the case that.' Let the symbol '∨' stand for 'or.' Let the symbol '→' stand for 'if-then', so that $p \rightarrow q$ means 'if p then q.'
> Given the statements
> $\neg\neg c$
> $c \rightarrow a$
> $\neg a \vee b$
> $b \rightarrow d$
> $\neg(d \vee e)$
> which one of the following statements must also be true? (Check the correct answer.)
>
> $\neg c$

e
h
¬*a*
all of the above

Once again, of course, we gave a corresponding problem on the post-test. Alert readers will have realized that the answer to **6** is "all of the above," which of course means that *h* must be true given the quintet. The reason for this, of course, is that the quintet is inconsistent, and therefore a straightforward proof for *h* (or any other propositional variable) can be easily given.

The final question of particular interest was question seven, the results from which seemed to suggest that our course had made students worse at reasoning of this type. It involved fairly complex reasoning on statements which were presented in English, thus requiring more effort to extract meaning. Looking for an explanation, we noticed that the following sentence appeared in the pre-test version of this question (from Smullyan, 1982), "'At least one of them did [tell the truth],' replied the Dormouse, who then fell asleep for the rest of the trial." The question from the post-test, which was intended to be analogous, included the following sentence, which was supposed to play the same role, "'Well, one of them did [tell the truth],' replied Devin, who then fainted and remained unconscious for the remainder of the investigation." This difference seemed potentially problematic.

On further investigation, we noticed that many of the justifications on the second problem suggested that subjects were having problems interpreting this statement by Dr. Devin, "Well, one of them did." This can be (and was) interpreted in two common ways, as, "One and only one of them did," or as, "At least one of them did." If these are both appealing interpretations, as they seemed to be for many of the participants, there is no entirely logical way to figure out the answer. A very small number of particularly clever subjects assumed that there would be enough information given in the question to figure out the right answer, and realized that one of the interpretations, that only one of them told the truth, did not fulfill this requirement. These students then rejected this option and solved the problem. However, doing all of that, which seems to be the only way other than guessing that subjects were able to correctly answer the problem, is far more difficult than interpreting the analogous statement in the missing jam problem, which was made by the Dormouse, in response to questioning about whether the Mad Hatter and March Hare had spoken the truth: "At least one of them did." Since this is clearly much more explicit, we have considered the seventh questions on the two tests to be sufficiently different that they are no longer appropriate for comparison. Unfortunately, it was not possible to counter-balance the pre-test and post-test, because of the high degree of availability of students to each other; to make both sets of questions available in this way would have introduced an unacceptably strong confound.

Demographic Data

Without the last question on each test, averages dropped to 3.30 correct on the pre-test and 4.97 on the post-test. This improved the value of the t-statistic to t= -13.653. This indicates even more clearly that subjects did in fact improve their ability to succeed on tests of reasoning due to the instruction in logic, and that the improvement was of a fairly substantial magnitude.

Interpreting the justifications turned out to be fairly problematic. Our initial attempt was a fairly subjective rating system based on the opinions of a competent logician, but there is a potentially very important confound in this method, which is that a correct answer is much more likely to suggest to a reader that the subject knew what s/he was doing, even if this is somewhat underdetermined by the written justification. Since we are most interested in the correlation between justification quality and success rate, this rating system was unacceptable. However, the information from the justifications did turn out to be useful in checking to make sure that the questions were interpreted as we intended, and further exploration may reveal a more objective way to code this data such that it may be made more useful.

Point-biserial correlations (appropriate for categorical data of this type, rather than more common values) were calculated between sex, high school attendance in New York state/elsewhere, and previous logic experience and the two test averages, and similar Pearson correlations calculated within those two groups of factors. Nothing significant came out of sex. Surprisingly, we did not observe a significant correlation between high school state and performance on either test. Previous logic experience did correlate positively with performance on the pre-test, but not on the post-test, as expected, with Pearson correlations of .246 (significant at the 0.05 level) and -.021, respectively. This indicates that the course did make up for any disadvantage less experienced students may have had coming in, and also suggests that performance on the pre-test was actually higher than it ought to have been, because we assumed that incoming students would not have been formally trained in logic. Since only one-tenth of the subjects were so trained, and sometimes in courses that dealt only tangentially with logic, we suspect this effect was negligible.

Unsurprisingly, none of sex, high school state, and previous logic experience correlated with each other. Also unremarkable was the highly significant correlation of .465 found between the pre-test and post-test scores—subjects with higher initial ability are likely to have higher ability after the end of the course.

Conclusions

The proposition that humans are unable to learn to reason better through instruction in formal logic seems to be disconfirmed by these data. This naturally does not mean that pragmatic effects hold no power over our attempts to use what deductive competence we have developed, nor does it suggest that all tests of reasoning will show improvement following an arbitrarily-selected course in logic. However, Cheng and Holyoak's proposed pragmatic reasoning schema theory (see Cheng et. al., 1986; Cheng and Holyoak, 1985; and Holyoak and Cheng, 1995) needs revision to remain a plausible candidate explanation of human reasoning. Yang and Bringsjord (2001, 2001) have suggested an alternative theory of human and, by extension, machine reasoning, viz., Mental Metalogic (MML), which allows pragmatic reasoning schemas to continue to play a role in human cognition, but not alone. In MML, mental models and mental logic exist side-by-side with such schemas, and a higher-level choice mechanism selects the most appropriate form of reasoning for the task at hand. In this regard, it's important to note that MML draws from a part of mathematical logic hitherto untapped in cognitive science: metatheory.

Recent advances in the teaching of logic (particularly HYPERPROOF) were utilized in the course used in the study, and this may help explain the differences in the results seen by Cheng and company, and those found in our study (for a mathematical analysis of HYPERPROOF in the context of "heterogeneous" reasoning consistent with Mental MetaLogic, see Barwise & Etchemendy 1995). In addition to the technological sophistication and concomitant improvement in available techniques, our interest in matters related to the psychology of reasoning may help to explain these results. In the class at RPI, students were encouraged to think about problems from the standpoint of metatheory: to ponder the way that they might approach logic problems (e.g., from the standpoint of searching for proofs *a la* mental logic, or from the standpoint of disproofs and mental models.) We routinely presented several options and contrasted their power, and also studied the reasoning process itself. The increased introspection about the reasoning process that this may have produced in our students is another factor which distinguishes the RPI logic class from previous subjects of similar experiments.

We believe that the reason standard logic instruction has not improved performance on tests of the sort given by proponents of the pragmatic reasoning schema theory may be related to the importance of one or more of the factors we have mentioned, which are historically missing in most classes. If this is correct, contra Cheng and Holyoak, it is not the level of abstraction that keeps logic instruction from being efficacious in improving reasoning. With the right theoretical perspective (MML), and pedagogical techniques which recognize the efficacy of non-pragmatic reasoning associated with that perspective, students can easily carry out difficult context-independent deduction suggestive of that of a professional logician or mathematician.

References

Barwise, J., & Etchemendy, J. (1994). *HYPERPROOF*. Stanford, CA: CLSI.

Barwise, J. & Etchemendy, J. (1995) Heterogeneous Logic. In *Diagrammatic Reasoning*, Glasgow, J., Narayanan, N.H., and Chandrasekaran, B., eds. Cambridge, MA: MIT Press.

Bringsjord, S. & Ferrucci, D. (2000) *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, A Storytelling Machine*. Mahwah, NJ: Lawrence Erlbaum.

Bringsjord, S., & Rinella, K. Hard Data in Defense of Logical Minds. *Annual International Conference on Computing and Philosophy*. Carnegie-Mellon University, August 6, 1999.

Bringsjord, S., Noel, R., & Bringsjord, E. (1998). In Defense of Logical Minds. *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 173-178). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus Syntactic Approaches to Training Deductive Reasoning. *Cognitive Psychology*. 18, 293-328.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic versus Syntactic Approaches to Training Deductive Reasoning. *Cognitive Psychology*. 17, 391-416.

Evans, J., & Over, D. E. (1996). *Rationality and Reasoning*. Hove, East Sussex, UK: Psychology Press.

Holyoak, K. J., & Cheng, P. W. (1995). Pragmatic Reasoning About Human Voluntary Action: Evidence from Wason's Selection Task. In S. E. Newstead & J. Evans (Eds.), *Perspectives on Thinking and Reasoning*. Englewood Cliffs, NJ: Lawrence Erlbaum Associates.

Inhelder, B., & Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books.

Smuyllan, R. (1982). *Alice in Puzzleland*. New York, NY: Morrow.

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

Verzoni, K. & Swan, K. (1995) On the Nature and Development of Conditional Reasoning in Early Adolescence. *Applied Cognitive Psychology*. 9, 213-234.

Yang, Y., & Bringsjord, S. (2001). Mental Possible World Mechanism and Logical Reasoning in GRE. (under submission).

Yang, Y., & Bringsjord, S. (2001). Mental MetaLogic: A New Paradigm in Psychology of Reasoning. (under submission).