Accounting for Discovery in a Cognitive Architecture

Ron Sun (rsun@cecs.missouri.edu)
Xi Zhang (xzf73@mizzou.edu)
Department of CECS, University of Missouri, Columbia, MO 65211, USA

Abstract

This work explores the interpretation and the computational modeling of discovery tasks—tasks where sudden insights result from cumulative information. Such tasks are useful in further understanding human everyday reasoning, beyond rule-based reasoning prevalent in cognitive science. We interpret the situation as involving mainly implicit memory with successive accumulation of information. We then implement this analysis in a cognitive architecture CLARION, which has succeeded in capturing a variety of human learning data in prior simulations. The simulation within this architecture accurately captures the human data. This work demonstrates the significant role played by implicit memory in human everyday reasoning. Furthermore, it demonstrates how such a reasoning process falls out of the existing cognitive architecture CLARION.

Introduction

What is human everyday reasoning like? Is it different from formal models developed by logicians and psychologists? What is the role of intuition in such reasoning? In this paper, we attempt to describe one particular aspect of human everyday reasoning—intuition and insight—in computational terms. We instantiate our analysis in the form of a computational model implemented in a cognitive architecture.

Specifically, this work explores the interpretation and the computational modeling of discovery tasks—tasks where sudden insights result from cumulative information (Bowers et al 1990, Schooler et al 1993, Metcalfe 1986). Such tasks are important for understanding finer details of human everyday reasoning (e.g., implicit processes), beyond explicit rule-based reasoning prevalent in cognitive science. As shown in Bowers et al (1990), subjects could respond discriminatingly to coherence that they could not explicitly identify, and that this implicit recognition of coherence guided subjects gradually towards an explicit representation of a hunch. While such hunches might surface quite suddenly into consciousness, the underlying (implicit) cognitive processes were rather continuous. Schooler et al (1993) presented related findings. We interpret the situation as involving mainly implicit memory with successive accumulation of information. We then implement this analysis in an existing cognitive architecture—CLARION (Sun 2002).

A little background is in order here. Sun (1991) first proposed a theory of human everyday reasoning based on a combination of rule-based reasoning and similarity-based reasoning, implemented with a mixture of localist and distributed connectionist models. This theory was further developed and elaborated in Sun (1995). The theory was backed up by psychological evidence in the form of verbal protocols such as those in Collins and Michalski (1989). Later on, a new cognitive architecture, CLARION, was developed (Sun 1999, Sun et al 2001), and this model of reasoning was incorporated into the architecture. The new cognitive architecture has succeeded in capturing a variety of human learning data in prior simulations (Sun et al 2001, Sun 1999, 2002).

As will be detailed later, the simulation of discovery tasks within this architecture accurately captures the human data as well. This simulation indicates the significant role played by implicit memory in human everyday reasoning. Furthermore, it demonstrates how such a reasoning process naturally falls out of the cognitive architecture CLARION.

In the remainder of this paper, we first describe the experiments of Bowers et al (1990). We then present our interpretation of the experimental results. We turn then to describe the generic cognitive architecture, CLARION, used in capturing the human data. Next, the particular setup of the architecture for capturing this set of human experiments is described. We then describe the results of simulating the experiments of Bowers et al (1990) using CLARION. Finally, some general discussions complete the paper.

The Discovery Task

Let us examine some human data that illustrates intuition and insight in human reasoning (Tversky and Kahneman 1983, Sun 1991, Sloman 1998). We will look into the experimental data from Bowers et al (1990), focusing on their third experiment, which was concerned with gradual "warming up" to a hunch and later to a conviction.

Although intuition has been defined as "the immediate apprehension of an object by the mind without the intervention of any reasoning process" (Oxford English Dictionary), or "immediate knowledge, as in perception or consciousness, distinguished from mediate knowledge as in reasoning" (Webster's Dictionary), we instead view intuition as a kind of reasoning. Reasoning encompasses both explicit processes (explicit rules and logics) on the one hand, and implicit processes (intuition and insight) on the other (Sun 1991, 1995). In fact, intuition and insight are important components of human reasoning. They supplement and guide explicit reasoning. The latter has been amply documented in cognitive science and AI (see, e.g., Collins and Michalski 1989, Davis 1990, Yang and Johnson-Laird 2001), but not yet the former. More studies of human everyday reasoning involving intuition and insight are needed.

In the discovery task of Bowers et al (1990), during each trial, a set of 15 clue words (that is, an "item") were presented to subjects, one word at a time. Each clue word was a response to a stimulus word in the Kent-Rosanoff word association test (Kent and Rosanoff 1910). The first 12 clue words occurred 5 or less times out of 1000 as a response to the stimulus word, and they were randomly assigned to position 1-12. The last three clue words occurred more than 5 times and were randomly assigned to position 13-15. Subjects' task was to identify the word that was associated with each of the 15 clue words.

The clue words from each set were presented one at a time. Subjects were required to generate a word to which each of the clue words was associated after the presentation of each clue word. They were given 10-15 seconds after the presentation of each clue word. If subjects viewed a generated word as a potential solution, they would check-mark it (indicating a "hunch"). When they were convinced that the word was a solution, they would mark it with an X (indicating a "conviction").

As reported in Bowers et al (1990), in a sample of 100 subjects, subjects arrived at a hunch on about the 10th clue word on average (M=10.12, SD=4.55). The average number of clue words it took for subjects to go from a hunch to a conviction was about 1.79 (SD=0.96).

The Interpretation of the Data

As suggested by Bowers et al (1990), subjects could respond discriminatingly to coherence that they could not explicitly identify, and that this implicit recognition of coherence guided subjects gradually towards an explicit representation of a hunch. Subjects "warmed up" to the solution in an incremental and gradual manner. That is, whereas hunches or convictions might surface quite suddenly into consciousness, implicit cognitive processes were rather continuous. An implicit representation gradually gained strength. When the level of activation reached a certain degree, crossing a certain threshold, the implicit representation triggered an explicit one. (On the other hand, an explicit representation might also be activated as a result of related explicit knowledge, e.g., explicit rules, although usually there were relatively few such rules.) A hunch was indicated, as a result of an emerged and sufficiently activated explicit representation. Even after that point, its strength might continue to grow, and thus eventually, subjects might indicate a "conviction"—an even stronger explicit representation.

Mechanistically (i.e., computationally), we can easily imagine that people are frequently "trained", deliberately or incidentally, with word associations, for example, desk-chair, pen-paper, and so on, in everyday life. Such training forms associations between pairs of words with varying degrees of strengths, based in part on frequencies of co-occurrences. Association formation happens mainly in the implicit memory (specifically, in the part of the implicit memory that is not concerned with procedural, or action-centered, information), because of the (mainly) incidental nature of the training scenarios. At the test setting of the experiment of Bowers et al (1990), clue words are presented one at a time. At the presentation of each clue word, all associated words are activated to certain extents (some strongly while others weakly). After the presentation of each clue word, more activations are accrued. Gradually, the activations of some words in the implicit memory become stronger and stronger. As a result of the implicit memory (as well as explicit rules), explicit representations are activated in the explicit memory (specifically, the explicit memory that is concerned with the general non-action-centered knowledge about the world). Eventually, a threshold (the threshold for hunches) is crossed, and thus a hunch is found. Furthermore, when more clue words are presented, more activations are accrued to the explicit representations. A second threshold (the threshold for convictions) is crossed, and thus a conviction is declared.

Of course, the processes described above are all under the direction (control) of some executive functions. Since such executive control is of the usual variety, we will not analyze it in any more detail here.

Below we will attempt to implement the above analysis in a computational model, which serves to verify and validate this analysis. As mentioned earlier, the simulation is conducted within an existing cognitive architecture CLARION.

The CLARION Model

CLARION is an integrative model with a dual representational structure (Sun et al 2001, Sun 2002). It consists of two levels: the top level captures *explicit*

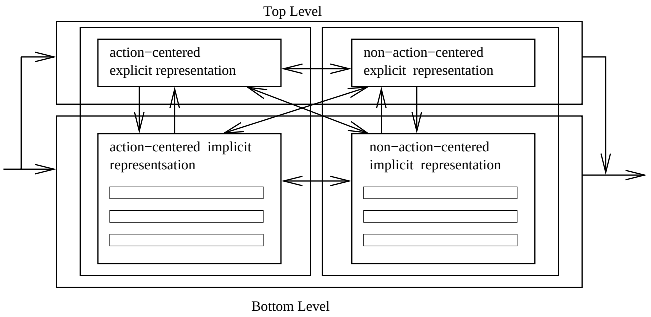

Figure 1: The CLARION architecture.

processes and the bottom level *implicit* processes. See Figure 1.

First, the inaccessible nature of implicit knowledge is suitably captured by subsymbolic distributed representations provided by a backpropagation network. This is because representational units in a distributed representation are capable of accomplishing tasks but are subsymbolic and generally not individually meaningful (see Sun 1995). This characteristic of distributed representation accords well with the (direct) inaccessibility of implicit knowledge.

In contrast, explicit knowledge may be captured in computational modeling by a symbolic or localist representation (Sun 2002), in which each unit is easily interpretable and has a clear conceptual meaning. This characteristic captures the property of explicit knowledge being (directly) accessible and manipulable (Smolensky 1988, Sun 1995).

This radical difference in the representations of the two types of knowledge leads to a two-level model whereby each level using one kind of representation captures one corresponding type of process, either implicit or explicit. The model may select to use one level or the other, or both, based on current circumstances (e.g., experimental conditions; see Sun 2002 for details).

At each level of the model, there may be multiple modules, both action-centered modules and non-action-centered modules (Schacter 1990). [1] We will refer to these two sets of modules as the *action-centered subsystem* (or the ACS) and the *non-action-centered subsystem* (or the NACS), respectively. There are also other components, such as working memory, goal structures, and so on.

In this work, we will focus on the NACS, due to the declarative (non-action-centered) nature of the discovery task. This subsystem, as stated earlier, consists of (1) a top level, which is made up of a set of explicit associative rules, and (2) a bottom level, which is made up of implicit associative memories (Sun 2002).

At the top level of the NACS, the essential elements are *chunks*, each of which is specified by a chunk identifier and a set of dimension-value pairs (i.e., attribute-value pairs) that describes an entity or an object. These dimensional values are represented in the bottom level, and thus chunks are linked to the bottom level. Links between chunks, termed *associative rules*, encodes explicit associations between pairs of chunks. These rules may be learned from externally provided information or information from the bottom level (from implicit memory). The top level of the NACS is termed the general knowledge store (the GKS).

At the bottom level of the NACS, associative memory networks (or AMNs) encode implicit associations. Specifically, among many other possibilities, a backpropagation network can be used for this memory. Associations are formed by mapping an input to an output. For example, the network may learn to hetero-associate a pattern with another. In this case, a pattern is presented as the input, and another (different) pattern to be associated with it is presented as the desired output. Through repeated training, the network may be able to output the desired pattern when a corresponding input pattern is presented. This is useful for, for example, establishing connections between concepts and facts (such as linking climate information of a region to its agricultural products, or linking a person to his professional accomplishment).

When an output of the AMN is strong enough, the chunk in the GKS corresponding to the output from the AMN will be activated to the same extent. At the same time, when an implicit association from the AMN is strong enough, an associative rule corresponding to the retrieved association from the AMN will be established in the GKS. [2]

Note that all of the operations of the non-action-centered subsystem are under the control of the action-centered subsystem, which makes action decisions each step of the way. The top level of the ACS consists of a set of explicit action rules, either externally given or extracted from the bottom level (from implicit action-centered knowledge), while the bottom level consists of implicit action decision networks (trained with reinforcement learning algorithms). For details regarding the ACS and its parameters, see Sun et al (2001) and Sun (2002).

[1] The reason for having both action-centered and non-action-centered modules (at each level) is because, as it should be obvious, action-centered knowledge (roughly, procedural knowledge) is not necessarily inaccessible (directly), and non-action-centered knowledge (roughly, declarative knowledge) is not necessarily accessible (directly). Although it was argued by some that all procedural knowledge is inaccessible directly and all declarative knowledge is directly accessible, such a clean mapping of the two dichotomies is untenable in our view (Sun 2002).

[2] The outcomes from the AMN and the GKS can be combined in various ways, for example, through a *max* function.

CLARION has been successful in simulating a variety of cognitive tasks. The simulated tasks include serial reaction time tasks, artificial grammar learning tasks, process control tasks, alphabetical arithmetic tasks, Minefield Navigation, and Tower of Hanoi (Sun 2002). We are now in a good position to extend the effort to the capturing of a wide range of human reasoning and memory processes, through simulating reasoning and memory task data. This paper is but one aspect of this effort.

Simulation Setup

During the training of the model, pairs of words were presented to the AMN (the bottom level of the NACS). The input to the AMN included three components: the current clue word, the working memory content, and the current goal. The input nodes of the AMN corresponded to dimension-value representations of words and goals. The output nodes of the AMN corresponded to dimension-value representations of words. [3] Each of the first 12 words on a list in the stimulus material was used for training, paired with the target word. That is, each of the 12 words was presented as the input to the AMN and the target word was the desired output for the AMN. [4] Each of these words was used about 4% of the times, for a total of about 48% of the training time. These associations were under-trained (thus the AMN did not perform well at the end of the training), capturing the weak, implicit associations between these pairs of words. Each of the last 3 words on a list in the stimulus material was also used for training, again paired with the target word. Each of these words was used for training 17% of the times, for a total of 51% of the training time. A total of 10 lists of words (10 training "items") was used.

Each word was represented as a chunk in the GKS (the top level of the NACS), due to the presentation during training of these words as input and output. Chunk encoding was such that each chunk had a number of dimensions and each dimension had a number of possible values. These values were represented in the AMN, and thus chunks were linked to the AMN. However, due to the relatively infrequent presentation of association pairs, there was the encoding of relatively few explicit associative rules in the GKS. (The invocation of most explicit associative rules would be below the minimum invocation frequency, and thus they would be deleted. Therefore it would be unlikely that many rules would be established in the GKS.)

During the test, clue words were presented one at a time. After the presentation of each clue word, it would be stored into the working memory. So, in effect, the partial sequence of words (seen thus far) was presented at each step. Thus, the activation of the target word became stronger and stronger in activation as a result of the "accumulating" input.

Due to accumulating evidence, a tentative winner (a hunch) first emerged, and then a final winner (a conviction) was generated. In the GKS, a retrieved chunk i was considered a hunch (a tentative winner), if $\forall j, S_i^c > S_j^c$ and $S_i^c > threshold_1$, where S_i^c indicated the strength of chunk i. A chunk i was considered a conviction (a final winner), if $\forall j, S_i^c > S_j^c$ and $S_i^c > threshold_2$. Of course, $threshold_1$ was lower than $threshold_2$, leading to partial certainty of a generated solution word.

The following action rules (among many other rules) were implemented in the ACS for this specific retrieval process:

> If goal= solving-association-task, and $\forall j, S_i^c > S_j^c$ and $S_i^c > threshold_1$, then retrieve chunk i and indicate *hunch*.
>
> If goal= solving-association-task, and $\forall j, S_i^c > S_j^c$ and $S_i^c > threshold_2$, then retrieve chunk i and indicate *conviction*.

These rules were "fixed" rules for this task acquired presumably from a priori knowledge and task instructions given to subjects prior to experiments. When both rules were applicable, a random selection was made. In these rules, the input chunks were not involved in comparison: They were specifically excluded. The goal (which was involved in the rules) was set in the goal structure, when the task instructions were given before the test began.

Simulation Results

Recall that the human data of this task indicated that (1) the average number of the clue words at which a hunch was arrived at was 10.12, with SD = 4.55; (2) the average number of the clue words it took for subjects to go from a hunch to a conviction was 1.79, with SD = 0.96. Matching the human data closely, our simulation result indicated that (1) the average number of clue words at which a hunch was arrived at was 10.23, with SD = 3.07; (2) the average number of clue words it took for subjects to go from a hunch to a conviction was 1.72, with SD = 2.09. Clearly, the match was excellent.

To understand the simulation and the interpretation of the data as embodied in this simulation better, let us examine some details. First of all, from Figure 2, we see that there was a gradual accumulation of activation on the target word over time, due to successive additions of clue words. Note that this accumulation was the result of both the bottom level (the AMN) as well as the top level (the

[3] In fact, the input and output should be phonological and morphological features of words. However, for the sake of simplifying the simulation, we used artificially constructed features. This simplification does not affect the outcome of the simulation.

[4] This process was in fact the reverse of the word association test (Kent and Rosanoff 1910).

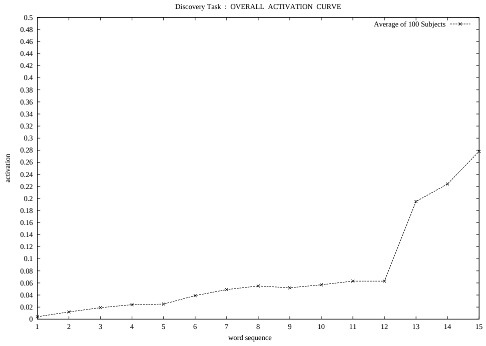

Figure 2: The accumulation of activation.

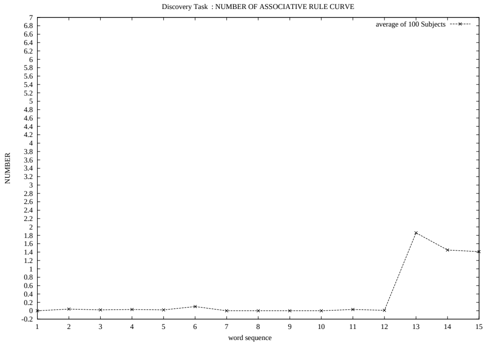

Figure 3: The number of matching rules at the top level for each clue word.

associative rules in the GKS). For example, the contribution of the top level was as shown in Figure 3, in terms of number of matching associative rules. As we can see, there was an increase of number of matching rules towards the end of the list. This led to an increasing probability of activation. However, the accumulation was also due to the bottom level, which was shown in Figure 4.

During training, the implicit association at the bottom level developed gradually. Figure 5 shows the gradual development of implicit associations at the bottom level (the AMN) over time during the course of training.

During training, explicit associative rules were also formed at the top level. A (randomly selected) sample set of rules in the GKS, extracted during training, was as follows:

Item2-Word5 ⟶ Item2-Target
Item2-Word14 ⟶ Item2-Target
Item3-Word3 ⟶ Item3-Target
Item3-Word13 ⟶ Item3-Target
Item4-Word2 ⟶ Item4-Target
Item4-Word13 ⟶ Item4-Target
Item5-Word13 ⟶ Item5-Target
Item5-Word15 ⟶ Item5-Target
Item8-Word3 ⟶ Item8-Target
Item10-Word6 ⟶ Item10-Target

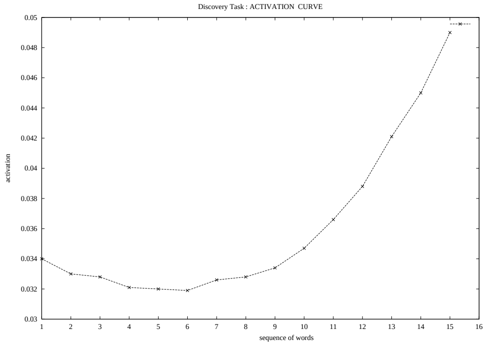

Figure 4: The accumulation of activation at the bottom level.

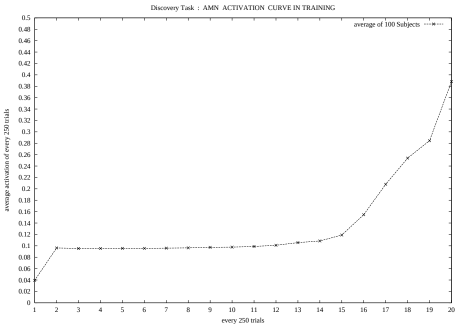

Figure 5: The training curve of the bottom level.

where "item" indicated a list of words, "word" indicated a particular word at a particular position on the list, and "target" indicated the corresponding target word. As we can see, there were, in general, more rules for words towards the end of the list than towards the beginning. This was because these words were used more frequently in training, and thus were more likely to form explicit associative rules at the top level (as well as to develop stronger implicit associations with the target word in the AMN).

Discussions

Although not a typical topic in cognitive science, intuition and insight have been documented experimentally in recently years (see, e.g., Bowers et al 1990, Schooler et al 1993, etc.). A lot of experimental data have been accumulated on that. The explanation of this phenomenon, however, is not as clear as one would like to see. Computational modeling and simulation may help in this regard. Detailed simulations help to substantiate theories concerning these experiments and help to bring out the essential cognitive processes involved.

Our simulation demonstrated the capability of Clarion in capturing this type of situation. The simulation succeeded without adding any new mechanisms or components—the simulation falls out of the existing mechanisms in CLARION, which include, in particular, implicit associative memory and explicit (non-action-centered) memory.

Our model shows that it is useful to posit the existence of these two separate memory systems: explicit versus implicit. While explicit memory encodes rules that are all-or-nothing, implicit memory allows more gradual accumulation of information. Furthermore, the simulation shows that the interaction between the two memory systems, in the sense that intuition gives rise to explicit awareness and vice versa, is important in everyday reasoning. As has been previously shown by Sun (1995), this interaction in fact underlies much of human everyday reasoning.

Compared with other existing cognitive architectures such as ACT-R (see Anderson and Lebiere 1998), CLARION embodies a different set of assumptions, which include, most notably, the separation of the two dichotomies: action-centered vs. non-action-centered knowledge, and implicit vs. explicit knowledge (Sun 2002). These alternative assumptions and structures enable CLARION to faithfully capture a variety of cognitive data (see Sun 1999, Sun et al 2001, Sun 2002 for details).

In terms of capturing human everyday reasoning in general (of which intuition is part, in our view), although logic-based models are useful, they are known to suffer from a number of shortcomings, including their restrictiveness concerning preconditions, consistency, and correctness, and their inadequacy in dealing with inexactness (see, e.g., Israel 1987, Sun 1995). In a different vein, psychological work on reasoning is relevant also. Such work mostly centers around either mental logic (Braine and O'Brien 1998) or mental models (Yang and Johnson-Laird 2001). These approaches, however, do not deal with the kind of situation embodied in this task.

Acknowledgments

This work is supported in part by Army Research Institute contract DASW01-00-K-0012. Thanks to Robert Mathews for his collaboration on the project.

References

J. Anderson and C. Lebiere, (1998). *The Atomic Components of Thought*, Lawrence Erlbaum Associates, Mahwah, NJ.

K. Bowers, G. Regehr, C. Balthazard, and Parker, (1990). Intuition in the context of discovery. *Cognitive Psychology.* 22. 72-110.

M. Braine and D. O'Brien, (eds.) (1998). *Mental Logic.* Lawrence Erlbaum Associates, Mahwah, NJ.

A. Collins and R. Michalski, (1989). The logic of plausible reasoning. *Cognitive Science*, 13(1), 1-49.

E. Davis, (1990). *Representations of Commonsense Knowledge.* Morgan Kaufman, San Mateo, CA.

D. Israel, (1987). What's wrong with non-monotonic logic? In: Ginsberg (ed.), *Readings in Nonmonotonic Reasoning*, pp.53-55, Morgan Kaufman, San Mateo, CA.

G. Kent and A. Rosanoff, (1910). A study of association in insanity. *American Journal of Insanity*, 67, 37-96.

J. Metcalfe, (1986). Dynamic metacognitive monitoring during problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12, 623-634.

D. Schacter, (1990). Toward a cognitive neuropsychology of awareness: implicit knowledge and anosagnosia. *Journal of Clinical and Experimental Neuropsychology.* 12 (1), 155-178.

J. Schooler, S. Ohlsson, and K. Brooks, (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122 (2), 166-183.

S. Sloman, (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, 35, 1-33

P. Smolensky, (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11 (1), 1-74.

R. Sun, (1991). Connectionist models of rule-based reasoning. *Proceedings of the 13th Cognitive Science Conference*, 437-442. Lawrence Erlbaum Associates, Hillsdale, NJ.

R. Sun, (1995). Robust reasoning: Integrating rule-based and similarity-based reasoning. *Artificial Intelligence.* 75, 2. 241-296.

R. Sun, (1999). Accounting for the computational basis of consciousness: A connectionist approach. *Consciousness and Cognition*, Vol.8, 529-565.

R. Sun, (2002). *Duality of the Mind.* Lawrence Erlbaum Associates, Mahwah, NJ.

R. Sun, E. Merrill, and T. Peterson, (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science.* Vol.25, No.2, 203-244.

Y. Yang and P. Johnson-Laird, (2001). Mental models and logical reasoning problems in the GRE. *Journal of Experimental Psychology: Applied*, 7 (4), 308-316.

A. Tversky and D. Kahneman, (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 439-450.