Cognitive Reflection: the 'Premature Temperature Convergence' Hypothesis

Jarbas Silva (jarbas@intuition-sciences.com)
Getulio Vargas Foundation/EBAPE, Praia de Botafogo 190 office #509
Rio de Janeiro 22250-900 Brazil

Alexandre Linhares (linhares@clubofrome.org.br)
Getulio Vargas Foundation/EBAPE, Praia de Botafogo 190 office #509
Rio de Janeiro 22250-900 Brazil

Abstract

We present a new hypothesis concerning cognitive reflection and the relationship between System 1 and System 2, corresponding roughly to intuition and reason. This hypothesis postulates a tighter integration between systems than is implied by the common framework of separate modules. If systems are tightly coupled, as we propose here, an explanation of cognitive reflection may rest in the premature convergence of an 'entropy', or 'temperature', parameter.

Keywords: Intuition, reason, cognitive models, cognitive reflection, perception, philosophy.

The cognitive reflection test

Kahneman (2003) pointed out an interesting problem in his Nobel lecture: why do people generally err in simple problems, such as "a bat and a ball cost 1.10. The bat costs 1.00 more than the ball. How much is the ball?" Frederick (2005) proposed this problem, alongside two others, as a *cognitive reflection test* (CRT)[1], a task which would measure a person's ability to suppress from responding to their first impulse and engage in more abstract cognition in order to solve a problem.

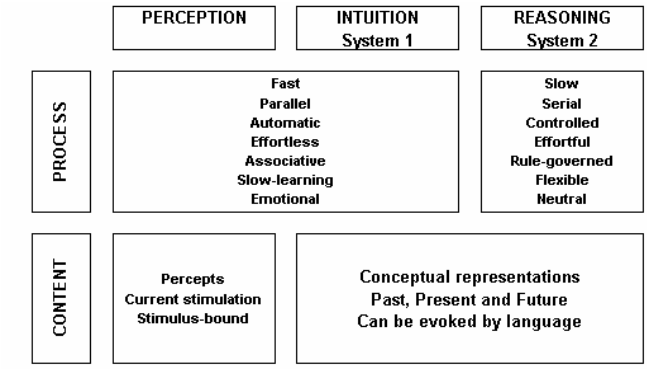

Figure 1: Process and Content in two Cognitive Systems: the "dual systems" view (Kahneman 2003).

[1] The remaining problems of the CRT are: (ii) "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?", and (iii) "In a lake, there is a patch of lilypads. Every day, the patch doubles in its size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?"

Kahneman (2003) presented the black box framework of Stanovich (1999) and Stanovich and West (2000) known as 'dual systems theory' to account for the different response types. The modularity of the presented framework implies that there is low or hardly any interaction between 'system 1' and 'system 2', and high interaction within each system. We find that position untenable, for we posit that, even in the most abstract, detached, form of thought, intuition plays a key role, by guiding reasoning processes in subtle, subconscious judgments. We therefore place the following questions concerning figure 1:

How do these systems interact? Which subsystems of an architecture could explain the different response types? In this paper we claim that the architecture of a system known as NUMBO (and, in fact, of all systems of the family to which it belongs), is able to explain the distinction in response types offered by 'system 1' and 'system 2' problems without resorting to a modular framework.

Numbo: an integrated computational architecture

Numbo is part of a family of cognitively plausible architectures which has been applied to numerous domains (see for instance French 1992; Mitchell and Hofstadter 1990; Mitchell 1993; McGraw 1995; Marshall 1999; Rehling 2001, Linhares 2005, Linhares 2007). Let us start by describing the problem with which it deals: *le compte est bon*.

The task: *le compte est bon*

The game Numble, on which NUMBO is based, is known in France as Le compte est Bon ("the total is correct") and has the objective to construct a number, called the *target* (taken from the interval between 1 and 150), from a set of 5 or less numbers, called *bricks* (taken from the the interval between 1 and 25), using only three basic operations: addition, subtraction and multiplication. The bricks can be used in only one operation in the resolution of the problem, for example:

Problem #2 Target: 87
Bricks: 8 3 9 10 7

Let us see how NUMBO is designed to solve this task in a psychologically-plausible way.

Numbo

Numbo is a parallel computational architecture which operates in multiple levels of description and consists of 5 integrated subsystems[2]. It works in a parallel fashion, with subcognitive processes we may refer to as *codelets*, or *pressing urges*. It has five principal components:

Sub-cognitive processes

The computational processes constructing the representations on short-term memory are subcognitive processes named codelets. The system perceives a great number of subtle pressures that immediately invoke subcognitive urges to handle them. These urges will eventually become impulsive processes. Some of these processes may look for particular objects, some may look for particular relations between numbers and create bonds between them, some may group numbers into chunks, or associate descriptions to chunks, etc. The collective computation of these impulsive processes, at any given time, stands for the working memory of the model. These processes can be described as impulsive for a number of reasons: first of all, they are involuntary, as there is no conscious decision required for their triggering. They are also automatic, as there is no need for conscious decisions to be taken in their internal processing; they simply know how to do their job without asking for help. They are fast, with only a few operations carried out. They accomplish direct connections between their micro-perceptions and their micro-actions. Processing is also granular and fragmented – as opposed to a linearly structured sequence of operations that cannot be interrupted. Finally, they are functional, associated with a subpattern, and operate on a subsymbolic level (but not restricted to the manipulation of internal numerical parameters—as opposed to most connectionist systems).

List of parallel priorities - Coderack

Each subcognitive process executes a local, incremental change to the emerging representation, but the philosophy of the system is that many of these pressing urges are perceived simultaneously, in parallel. So there is at any point in time a list of subcognitive urges ready to execute, fighting for the attention of the system and waiting probabilistically to fire as an impulsive process.

Any run starts with a standard initial population of bottom-up[3] codelets (with pre-set urgencies) on the list. At each time step, one codelet is probabilistically chosen to run and is removed from the current population on the Coderack (Mitchell 1993, Hofstadter and FARG 1995). The emerging representation can put other codelets in the Coderack, as well as change the urgency of the existing ones. So, the proper allocation of resources could not be programmed ahead of time, since it depends on what pressures emerge as a given situation is perceived.

STM

This is the working (short-term) memory of the model. This workspace is where the representations are constructed, with innumerable pressing urges waiting for attention and their corresponding impulsive processes swarming over the representation, independently perceiving and creating many types of subpatterns. Common examples of such subpatterns are micro-operations perceiving numbers, or the addition of two numbers, the perception that 6 is similar to 5 (in the sense of being close to it in magnitude), or that 107 is similar to 100, and so on.

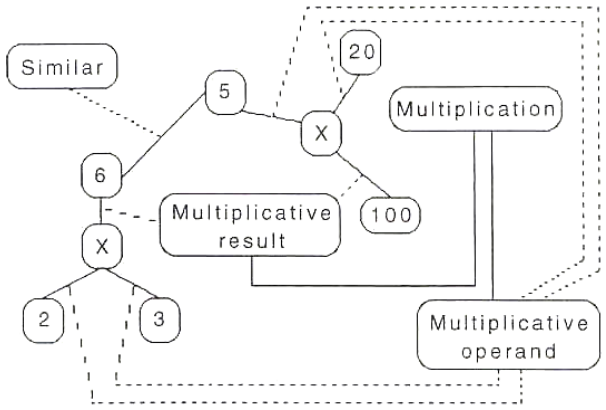

Figure 2: A fragment of Numbo's semantic network

The semantic network

The semantic associative network undergoes constant flux: The system has very limited basic knowledge; it knows the seed numbers, and some immediate relations entailed. The long-term memory of the system is embedded into a network of nodes representing concepts with links between nodes associating related concepts. This network is a crucial part for the formation of a chain reaction of conceptual

[2] Because Numbo deals with simple number problems and its description can be summarized more briefly than its other sibling-projects, it has been our preference for presenting the argument here. The reader, however, should not be misguided to conceive of it as a mere *ad hoc* program to solve such simple problems: it is, in fact, a member of a growing family of architectures to model general, abstract, human cognition in Chess, letter-string analogies, stylistic font design, Bongard problems, number sequence extrapolation, and, lately, music perception, auditory scene analysis, and discovery in Euclidean geometry (Foundalis 2006, Hofstadter and FARG 1995, Rehling 2001, Linhares 2005, Linhares 2007, Linhares and Brum 2007). All these projects embody the same fundamental architecture.

[3] Bottom-up codelets represent pressures present in all situations (the desire to make descriptions, to find relationships, to find correspondences, and so on). Top-down codelets represent pressures evoked by the situation at hand (e.g. , the desire to find similar numbers to the target 87 in the Problem #2) (Mitchell 1993, Hofstadter and FARG 1995).

activation: any specific concept, when activated, propagates activation to its related concepts, which will, in turn, launch top-down, expectation-driven, urges to look for those related concepts. This mode of computation not only enforces a context-sensitive search but also is the basis of the chain reaction of activation spreading – hence the term 'active symbols'. One of the most original features of the slipnet is the ability to "slip one concept into another", in which analogies between concepts are made (for details see Hofstadter and FARG 1995, Mitchell 1993).

Temperature

It may be clear from the above that the system does not zoom in immediately and directly into a faultless representation. The process of representation construction is gradual and tentative, with numerous impulsive processes competing with each other. At the start, the system has no expectations of the numbers to be found, so it slowly wanders through many possibilities before converging on a specific interpretation, through a process called the parallel terraced scan (Hofstadter and FARG, 1995). Embedded within it is the control parameter of temperature that measures the global amount of disorder and misunderstanding contained in the situation (Hofstadter and FARG, 1995). So, at the beginning of the process, when no relevant information has been gathered, the temperature will be high, but it will gradually decrease as intricate relationships are perceived, first concepts are activated, the abstract roles played by numbers and chunks are found, and meaning starts to emerge. Though other authors have proposed a relationship between temperature and understanding (Cagan and Kotovsky, 1997), there is still a crucial difference here (see Hofstadter 1985, 1995): unlike the simulated annealing process that has a forcedly monotonically decreasing temperature schedule, the construction of a representation for a proposed solution does not seem to get monotonically more complete as time flows. There are many instants when roadblocks are reached, when snags appear, and incompatible structures arise. At these moments, complexity (and entropy and confusion) grows, and so the temperature decrease is not monotonic. Finally, temperature does not act as a control parameter dictated by the user, that is, preset to go either down or up, but as a feedback mechanism to the system, which may reorganize itself, accepting or rejecting changes as temperature allows. As pressing urges are perceived, their corresponding impulses eventually propose changes in working memory, to construct or to destruct structures. How do these proposed changes get accepted? Temperature guides the process, very much like simulated annealing. At start it is high and the vast majority of proposed structures are built, but, as it decreases, it becomes increasingly more important for a proposed change to be compatible with the existing interpretation. And the system may thus focus on developing a particular viewpoint. Let us present two runs of the system in order to contrast its different information processing trajectories under either a system-1-type response or a system-2-type response.

Numbo's 'intuition' and 'reason': a premature temperature convergence hypothesis

By comparing possible dissimilar trajectories of "thought" due to the system 1-response types and system 2-response types, we claim that the same underlying subcognitive mechanisms account for both response types. What scientific basis supports such mechanisms? There are two principal sources: the first one is experimental psychology, and the evidence supporting mechanisms of activation-spreading, such as semantic nets. The second source of evidence comes from cognitive computational modeling, where analogous processes had already been identified, studied, and implemented in computational architectures, such that all the mechanisms that we claim to act during system 1 and system 2 problems possess a solid base in the literature, more specifically in the system known as NUMBO (Defays 1988, and personal correspondence, 2005). Let us, then, compare a system 2-type response in Numbo to a system 1-type response.

A System 2-type response

Now let us consider an execution of NUMBO in the resolution of the problem Target: 87; Bricks: 8, 3, 9, 10, 7.

It is important to point out that what is presented is a sequence of steps just for the sake of the reader's understanding, as the tasks (processes) are executed in non-deterministic parallel form, therefore, unable to follow a strictly sequential narrative.

(1) the target is read and the closest landmark, 90 = 9 x 10, is activated in the Pnet. This serves as a focus for the triggering of other activations, such as of the multiplication, subtraction and addition operations. For large targets, the system 'knows' (given the structure of its LTM) that the most likely way is though multiplication (Defays, 1995).

(2) the brick 8 is read and a codelet of syntactic comparison between the brick and the target is loaded in coderack, that, when executed, makes the system perceive a similarity between 8 and 87. Then a new codelet is loaded in the coderack, which, if executed, will increase the attractiveness of the brick.

(3) The next bricks are read in the following order: 7, 9, 10 and 3. It is important to remember that the bricks are not read from the left to the right but in a probabilistic way, which influences its behavior. Afterwards, new codelets with function of syntactic comparison between the bricks and the target will be loaded in the coderack.

(4) At random, some low urgency codelets are placed in the coderack with the function to try, for example, mathematical operations with the bricks. However, the choice of which ones is biased by their 'attractiveness', as well as for the level of activation of the mathematical operation in the Pnet.

In this way, if 8 were judged to be an attractive brick, the multiplication is activated (due to landmark 90), and, again, randomly, the next brick read was 7; block 8 x 7 = 56 is formed in the work area.

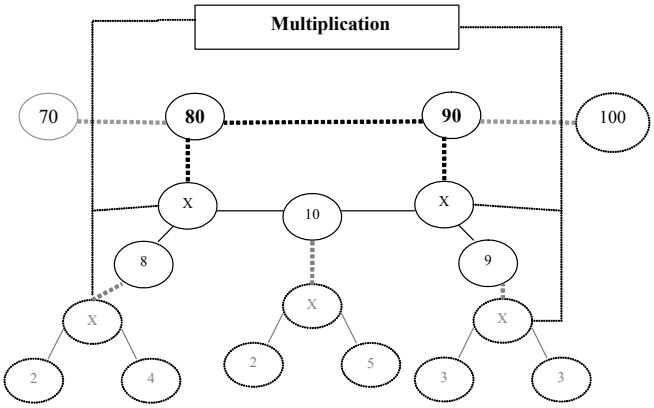

Figure 3. As target 87 is read, a great series of conceptual activations is made in the Pnet. Here we can observe some of these activations: (80=8x10), (90=9x10) - both by sheer proximity to the target. These activations are active, assuming command of the system, triggering off new processes and new activations: in the example, nodes 70 and 100 activate in smaller scale, the relation (10=5x2) also goes through a significant activation, then the system brings its focus back toward related numbers, and numbers such as 50 or 5 will receive little or no activation. The multiplication operation suffers a high activation, implying in a great probability of such operation to be promising in this scenario.

(5) The nodes activated in the Pnet trigger activations for neighboring nodes, as well as load (also via codelets) new codelets in the coderack, to verify the potential to create in the work area blocks equal to or close to the ones already activated in the Pnet. In this case, the landmark 90 = 9 x 10 is active in the Pnet, what it makes the NUMBO load codelets (of the type "seek-reasonable-facsimile") with the function to find similar bricks to 9 and 10. In this way, the system quickly perceives the existence of bricks identical to those looked (9 and 10) and loads codelets (of the type "test-if-possible-and-desirable") with the function to verify if the bricks are free and if the block to be created represents an attractive track. It is important to call attention to the fact that other alternative pathways are being explored; and that the current pathway may not be necessarily the best one. In this case, the result is the creation of the block 90 = 9 x 10 in the work area.

(6) A new codelet is then loaded in coderack with the function to compare the new block created (90) with the target. As the block and target are close, the creation of a secondary target starts to be an interesting possibility. When executed, the codelet creates the node 3 in the work area, fruit of the subtraction 90 - 87, as 'a secondary target'.

(7) Since block 56 = 8 x 7 has not been used, its attractiveness is lowered, which helps to increase the temperature of the system. With the increase of the temperature, codelets capable of breaking up that chunk are loaded in the coderack; which brings the possibility of blocks to be destroyed. The chosen blocks are those of low attractiveness and, in this case, 56 is the victim, liberating the bricks that compose it.

(8) the creation of the secondary target triggers an activation in the Pnet, as well as loads codelets of the type 'seek-reasonable-facsimile' with the function of finding an 'equal or next in size' brick. When executed, the codelet finds the brick 3 equal to the secondary target and loads codelets of the type "test-if-possible-and-desirable", in order to verify whether the brick is free and can be used. Finally, brick 3 is added to the new created block 87 = 90 - 3 and the system reaches the final answer (9 x 10) - 3.

Notice that the reading order of the bricks was important in this case, so that the system decided for some information processing trajectories, in detriment of others. Although problem #2 seems simple, NUMBO carries through some calculations and explores rival pathways until finding the definitive answer. It's important to remember that other possible answers exist, and that they might had been found through other information processing trajectories.

In contrast with this, let us examine a system 1-type response from NUMBO.

A System 1-type response

Let us see now as NUMBO solves problem # 4:: Target: 25; Bricks: 8, 5, 5, 11, 2.

(1) the target is read, amongst many other triggered activations, the node - 25 = 5 x 5 - is activated in the Pnet, which activates the operation of multiplication. Codelets of the type "seek-reasonable-facsimile" are loaded in coderack with the objective to find similar bricks to the nodes activated in the Pnet.

(2) the first brick (5) is read (randomly), which activates to a still higher level the node 5 in the Pnet. A codelet with the function to compare the newly read brick with the target is loaded, searching for associations. When executed, the codelet increases still higher the attractiveness of node 25 in the Pnet; which, in turn, triggers codelets with the function to increase the urgency of yet other codelets of the type "seek-reasonable-facsimile" already loaded in coderack.

(3) Then, as a result of codelets "seek-reasonable-facsimile", the two bricks with values equal to 5 are found and, subsequently, codelets of the type "test-if-possible-and-desirable" are loaded and will verify if the bricks are free for use and if the block created will be interesting. As the bricks are free, NUMBO creates in the work area the block 25 = 5 x 5. Temperature drops with each created structure, and, in this case, to its lowest level, as it consists in the answer to the problem.

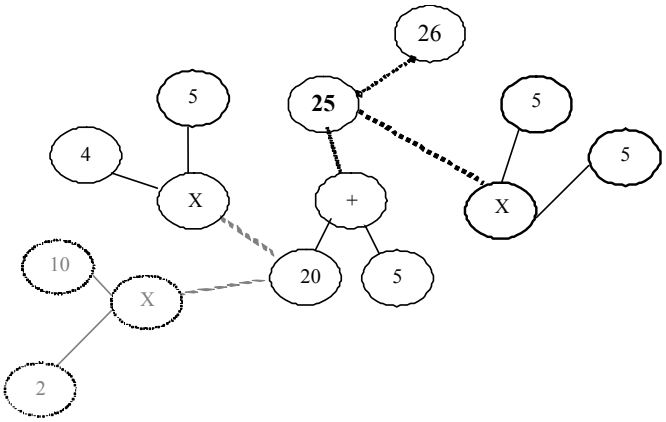

Figure 4. In the instant when target 25 is read, a great series of conceptual activations is made in the Pnet. Here we can observe some of these activations: (25=20+5), (25=5x5), (25 is close to 26), and so forth. These activations assume command of the system, trigerring new processes and activations: in the example, node 20 is active, and, in a smaller scale, relations (20=10x2) and (20=5x4), ensuring that the system immediately has its focus directed toward related numbers (and that numbers such as 41 or 3 will receive little or no activation).

The interesting point here is that NUMBO arrives at the answer to problem #4 *without executing any calculation*, only with the information stored in the Pnet. In this example of execution, brick 5 is read in first place, which facilitates the search for a solution. However, independent of any order of reading of the bricks, when the first 5 is eventually read, the activations triggered in the Pnet makes it such that the system quickly finds the solution. We claim that this immediate, no-calculations-performed, information processing is a very close approximation of intuitive human responses.

In a system 2-type response, temperature drops in a non-monotonic fashion, triggering a whole, long, chain of activation of nodes. In a system 1-type response, temperature drops rapidly and, for this reason, the solution quickly converges—whether or not to the best possible response.

It is possible that in 'system 1' problems, people rapidly activate conceptual nodes to high levels; bringing temperature down in short amounts of time. All the while, in 'system 2' problems, a much larger number of conceptual nodes are activated, the system goes through a much longer information processing, as all those active nodes strive for attention in the parallel workings of the architecture. Therefore, temperature falls rather slowly, as relevant structures take much more time to be found.

Consider again CRT problem #2: "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?" Frederick (2005) shows that this is the most difficult problem; with over 50% of his 3000+ sample finding a wrong answer. Consider, now, a variation of the problem: "If it takes 9 women 9 months to give birth to 9 babies, how long would it take 100 women to give birth to 100 babies?" (Frederick 2005).

In this second variation, of course, nobody should reply 100 months. *Yet, they are, in essence, the same problem*. So what is the differing feature of the cases? In the 'babies' case, the 'time-of-birth' conceptual node simply does not accept much variance in value (perhaps apart from premature birth cases); while people obviously conceive of a machine as working for a variable amount of time, for as long as desired. So in the 'machine' case, the enormous number of potential conceptual node activations make it much more unlikely that the relevant answer will be easily perceived. People are primed to see symmetry where there is none.

Discussion: Implications of the hypothesis

What are the implications of this hypothesis? We advocate that, in system 1-type responses, as much as in system 2-type responses, there are a great number of pressure-perceiving processes operating concurrently in multiple levels. To the extent to which these processes activate concepts in the semantic net, and structures are created in STM, the degree of "temperature" gradually drops. We claim that the basic distinction between systems 1 and 2 would be, then, the premature convergence of this measure in system 1-type responses. What would bring answers system 2 would be "the impulse not to yield to the first impulse" - the acquired ability to suppress the initial representation created after a fast drop of temperature. It's important to note that the system 1 'premature temperature convergence' is not only associated with errors and low cognitive performance, since in many situations, intuition remains the strongest tool we have to make decisions (see Klein, 1998, 2003).

The skeptical reader may ask: Why is this hypothesis of consequence? Because, as we have seen in figure 1, current theory postulates two distinct systems operating largely in separation from each other. This hypothesis postulates that there should be a tight coupling of subcognitive processes operating concurrently under both the intuition and reason systems. We thus postulate an ontological distinction between mental systems and response types; what would bring a particular response type in a particular scenario would not be the activation of system 2; it would, instead, be the premature convergence of temperature, which would lead to a smaller degree of activation of processes generally associated with system 2 responses. This explanation goes into more detail of the underlying information processing between system 1 and system 2 response types, and does not resort to a black boxes framework of human skills of intuition and reason.

Acknowledgments

The authors thank Daniel Defays, Shane Frederick, Doug Hofstadter, Harry Foundalis, Adriano Bruni, Ricardo Cardoso, and Eric Nichols. Financial support has been

generously awarded by EBAPE's propesquisa program of the Getulio Vargas Foundation.

References

Cagan, J., and Kotovsky, K., (1997). 'Simulated annealing and the generation of the objective function: a model of learning during problem solving', *Computational Intelligence*, 13, 534-581.

Defays, D. (2005). (personal communication) Source code of the NUMBO project. Université de Liège, Belgium.

Defays, D. (1998). L'Esprit en Friche: Les Foisonnements de l'intelligence artificielle. Liège : Pierre Mardaga.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*. Volume 19, Number 4, Fall 2005, pp. 24-42

French, R.M. (1992) Tabletop: an emergent stochastic computer model of analogy making. *Doctoral Dissertation*, University of Michigan, Ann Arbor.

Hofstadter, D. R. (1985). Metamagical Themas: Questing for the Essence of Mind and Pattern. New York: Basic Books.

Hofstadter, D. R.; The Fluid Analogies Research Group. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

Kahneman, D. (2003) A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, vol. 58, No.9, 697-720.

Klein, G. (1998). *Sources of Power: How people make decisions*. Cambridge, MA: MIT Press.

Klein, G. (2003). *Intuition at work: Why developing your gut instincts will make you better at what you do*. New York: Doubleday.

Linhares, A. (2005) An active symbols theory of chess intuition. *Minds and Machines*, 15, n.2, 131-181.

Linhares, A. (2007) On the nature of chess intuition: manifesto for a renaissance, *submitted for publication.*

Linhares, A., and Brum, P. (2007) Understanding our understanding of strategic scenarios: what role do chunks play? Accepted for publication, *Cognitive Science*.

Marshall, J. B. (1999). Metacat: A self-watching cognitive architecture of analogy-making and high-level perception. *Doctoral Dissertation*, University of Indiana.

Mcgraw, G. E. (1995). Letter Spirit (Part One): Emergent High-Level Perception of Letters Using Fluid Concepts. *Doctoral Dissertation*, Universidade of Indiana.

Mitchell, M., & Hofstadter, D. R. (1990). The emergence of understanding in a computer model of concepts and analogy-making. *Physica D*, 42, 322-334.

Mitchell, M. (1993). Analogy-Making as Perception: A Computer Model. Cambridge, MA: The MIT Press/Bradford Book.

Rehling, J. (2001). Letter Spirit (Part Two): Modeling Creativity in a Visual Domain. *Doctoral Dissertation,* University of Indiana.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645-665.

Stanovich, K. E. (1999). *Who Is Rational? Studies of Individual differences in reasoning*. Mahway, NJ: Erlbaum Assc.