# The Influence of Spreading Activation on Memory Retrieval in Sequential Diagnostic Reasoning

**Udo Böhm (udo.boehm@s2006.tu-chemnitz.de)**
**Katja Mehlhorn (katja.mehlhorn@phil.tu-chemnitz.de)**
Chemnitz University of Technology, Department of Psychology,
Wilhelm-Raabe-Str. 43, 09107 Chemnitz, Germany

**Abstract**

A crucial aspect of diagnostic reasoning is the integration of sequentially incoming information into a consistent mental representation. While research stresses the importance of working memory in such a task, it is not clear how the information represented in working memory can guide the retrieval of associated information from long-term memory. Factors that might influence this retrieval are the amount of information currently in the focus of attention (Lovett, Daily & Reder, 2000) and the time since the information first became available (Wang, Johnson & Zhang, 2006). By comparing the results of different ACT-R models to human data from a sequential diagnostic reasoning task, we show that these factors do not necessarily influence the retrieval. Our findings rather suggest that in a task where information has to be actively maintained in working memory, each piece of this information has the same potential to activate associated knowledge from long-term memory, independent from the amount of information and the time since it entered working memory.

**Keywords:** information integration; diagnostic reasoning; spreading activation; working memory; ACT-R

## Introduction

Generating and evaluating explanations for data extracted from the environment is a key component of many everyday tasks like medical or technical diagnosis and social attribution. This kind of reasoning is often called diagnostic or abductive reasoning (Josephson & Josephson, 1994; Johnson & Krems, 2001). For example, in medical diagnostic reasoning a physician needs to find the best explanation for the set of symptoms displayed by a patient. In such a task, information (e.g., the patients' symptoms) often becomes available step by step. The reasoner needs to integrate this information into a consistent mental representation that is updated every time a new piece of information becomes available. To find an explanation for the observed information, associated information (e.g., potential explanations for a set of symptoms) needs to be retrieved from memory.

Working memory has been proposed to play a crucial role in such a task. It is needed to keep track of the subsequently gathered information (Baumann, 2001) and it might hold possible explanations for this information retrieved from the reasoners long-term memory (Baumann, 2001; Thomas, Dougherty, Sprenger & Harbison, 2008). However, it is not clear how the information is represented in working memory over the course of the task and how that influences the retrieval of associated information. The goal of this paper is to develop a better understanding of how information in working memory guides the retrieval of associated knowledge from long-term memory in a sequential diagnostic task.

To achieve this, we implement different assumptions about the retrieval in ACT-R models and compare the model data to human data from a diagnostic reasoning experiment (Bauman, Mehlhorn & Bocklisch, 2007). Before we turn to describing the models, results and the related theories in detail, we want to point out that abduction in general and diagnosis in particular are complex tasks. In this paper we focus on memory retrieval, as it is a key aspect of these tasks. However, one should keep in mind that the models are a simplification of the task, as they ignore more deliberate processes (as e.g. described by. Johnson & Krems, 2001)

## Theories

Human memory might be understood as a set of elements, each of which is assigned a specific activation value. In this conception, a subset of the elements being activated above some specific threshold constitutes working memory (e.g., Just and Carpenter, 1992). In diagnostic reasoning, observations (e.g., the symptoms presented by a patient) and their possible explanations (e.g., diseases causing these symptoms) are held in memory. Given such a knowledge structure, observations can serve as a cue for the retrieval of associated knowledge. That is, information in the focus of attention (e.g., the symptoms presented by a patient) initiates a spreading activation process that activates associated information in long-term memory. Although this assumption has been made by various researchers (e.g., Arocha & Patel, 1995; Bauman et al., 2007; Thomas et al., 2008), the nature of this spreading activation process is not yet fully understood.

It has been argued that the total amount of activation that can be spread from working memory is limited and will be equally divided among the elements that spread activation (Lovett et al., 2000). Thus, the amount of activation spread by each single piece of information will depend on the amount of information that is currently held in the focus of attention. It has also been argued that information in working memory is subject to decay (e.g., Wang et al., 2006). That means that the activation spread from a specific piece of information in working memory to associated knowledge in long-term memory depends on the time since the information became available.

As noted above, in diagnostic reasoning the reasoner needs to find an explanation for information observed from the environment. As new information often only becomes available over time, the amount of information in working memory (i.e. the number of observations that need to be explained) as well as the 'age' of information in working memory (i.e. the time since the observation was made) varies. Therefore, sequential diagnostic reasoning is a field especially suited to test assumptions about information representation in working memory and its effect on retrieval from long-term memory.

To test the different possibilities, we designed different cognitive models using ACT-R. In its current implementation, ACT-R's declarative memory system consists of chunks (facts like *Influenza can cause cough and fever*) that represent declarative knowledge. Access to these memory elements depends on their activation (Anderson, 2007; Lovett et al., 2000). For each chunk, this activation is computed as the sum of its base-level activation and the associative activation from the current context (i.e. spreading activation). The base-level reflects the chunk's previous usefulness in terms of the number of times it was used and the time that has elapsed since. The associative activation reflects a chunk's usefulness in the current context and is computed as the product of the activation spread to it from some specific source (see below) and the strength with which it is related to that source (Anderson, 2007).

The source that provides the activation to be spread is information about the current problem or task. This information is represented in one of ACT-R's modules, the imaginal module. This module holds a mental representation of the problem currently in the focus of attention (Anderson, 2007). In a sequential diagnostic reasoning task, it is assumed that the imaginal module thus holds the information about all the data gathered so far. This information can then spread activation to associated knowledge held in declarative memory. To test the nature of the representation of information in working memory we implemented different modes of this spreading activation process in four ACT-R models.

The first model addresses the question if the amount of information in the focus of attention should influence spreading activation. To test this, we used the standard implementation of spreading activation in ACT-R. In this implementation, the total amount of spreading activation is assumed to be equally divided among the information stored in the source chunk (Lovett et al., 2000). Thus, the activation spread by each single piece of information depends on the amount of information in the focus of attention. The more slots the source chunk contains, the less activation can be spread by each single slot.

The second and the third model address the question whether information in working memory is subject to decay. In the second model, we use an equation for decaying activation proposed in a constraint satisfaction model (UECHO) by Wang et al. (2006). It assumes spreading activation to decay in curvilinear, negatively accelerated manner. Thus, information in working memory increasingly loses its impact over time. To test if decay needs to be negatively accelerated as proposed by Wang et al., or if a more simple assumption of decay would be sufficient, we implemented a third model using a linear decay function. In this model, information in working memory loses its ability to spread activation linearly over time.

For being able to better access the explanatory power of the above models, we implemented a fourth model that serves as a control model. This model is most parsimonious, as it assumes a constant amount of activation spread by each piece of information in working memory. Thus, in this model, spreading activation neither depends on the amount of information held in working memory, nor on the time since the information became available.

## Experiment

Human data was obtained in an experiment using an artificial diagnosis task (see also Baumann et al., 2007). Participants were told to imagine they are a doctor in a chemical plant and had to diagnose which chemical their patient had been in contact with. Therefore, they learned a knowledge structure consisting of nine different chemicals grouped into three categories. Chemicals were named with single letters and each chemical caused three to four symptoms (Table 1). Each symptom could be associated with two, three or six chemicals. Participants acquired this knowledge in an extensive training session, where they had to solve various tasks until reaching proficient performance.

In two subsequent experimental sessions, participants then worked on 340 diagnostic reasoning trials. In each of these trials, symptoms belonging to a chemical were presented sequentially on the screen. At the end of each trial, participants were asked for their diagnosis (see Figure 1 for a sample trial). As each symptom had several possible causes, only the combination of symptoms in a trial allowed for unambiguously identifying the correct diagnosis. With the number of observed symptoms, the number of plausible diagnoses could be narrowed down, leaving the correct diagnosis (consistent to all symptoms) at the end of the trial.

To track the activation of different explanations during the course of this reasoning task, a probe reaction task was used. After one of the symptoms in each trial, a single letter was shown. This could either be the name of one of the chemicals or not. Once the letter was presented on the screen, participants were to indicate as fast as possible whether it was a chemical or not. The idea of this probe reaction task is based on the idea of lexical decision tasks (e.g., Meyer & Schvaneveldt, 1971) according to which participants should respond faster to a probe that is activated higher in memory than to a probe of low activation. Using this measure, it was possible to monitor the activation of explanations over the course of the

sequential reasoning task with as little impact on the task itself as possible.

Table 1. Summary of the material participants had to learn (original material in German).

| Group | Chemical | Symptoms |
|---|---|---|
| Landin | B | cough, short breath, headache, eye inflammation |
| | T | cough, short breath, headache, itching |
| | W | cough, eye inflammation, itching |
| Amid | Q | skin irritation, redness, headache, eye inflammation |
| | M | skin irritation, redness, headache, itching |
| | G | skin irritation, eye inflammation, itching |
| Fenton | K | diarrhea, vomiting, headache, eye inflammation |
| | H | diarrhea, vomiting, headache, itching |
| | P | diarrhea, eye inflammation, itching |

Three different types of explanations were tracked in the experiment. First, the probed explanations could be an element of the current explanation (that is they were consistent to all symptoms observed so far). These probes are termed *'relevant'*. Second, the probed explanation could be an explanation that was never considered during the current trial. These probes were termed *'irrelevant'*. Third, the probed explanation could have been considered relevant until some evidence inconsistent with that explanation forced participants to reject it. These probes are called *'rejected'*. Rejected probes additionally varied with respect to the time since their rejection. They could be probed directly after rejection (*just rejected*); one symptom after rejection (*rejected 1 symptom* ago); or two symptoms after rejection (*rejected 2 symptoms ago*).

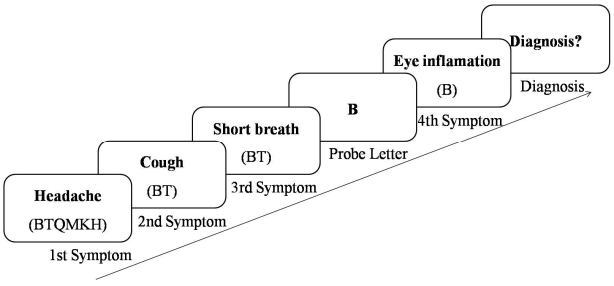


Figure 1. Sample trial from Baumann et al. (2007) with B as a relevant probe. (Letters in parentheses represent relevant explanations).

## Models

Four ACT-R models using different implementations of spreading activation from working memory were designed to test the assumptions presented above. Because ACT-R's memory system is dependent on patterns of retrieval time, the temporal order of events was modeled as closely as possible to the actual experiment. Thus, the models went through the same trials as human participants. After one symptom of each trial a probe was presented and the models had to indicate whether it was a chemical or not by typing 'Y' for Yes and 'N' for No respectively. At the end of each trial the models typed the letter representing the diagnosis. This was accomplished using ACT-R's perceptual and motor modules that allow for modeling time to process visual stimuli and performing key strokes.

As participants had received extensive training on the task prior to the experiment, the base levels of the chunks representing symptoms and probes or diagnosis respectively were all set to the same high level.

To implement the integration of the sequentially presented symptoms we assumed one chunk to be placed in the imaginal module at the beginning of each trial. Over the course of the trial, the slots of this chunk were successively filled with the symptoms seen thus far. As noted above, we assumed the imaginal module to be the source of spreading activation, thus, only information stored in this module could spread activation to associated concepts in declarative memory.

To solve the probe task, the model had to retrieve the explanation-chunk representing the probe letter. Due to spreading activation from the observed symptoms stored in the imaginal module, explanations associated to these symptoms received more spreading activation and could therefore be retrieved faster. Thus, as in human participants, the time to respond to a probe could be used as a measure for the activation of explanations in memory. As soon as the model was asked for the final diagnosis, it attempted to retrieve an explanation-chunk from memory. As the explanation most consistent to all observed symptoms obtained the highest spreading activation, this explanation was the one most likely to be retrieved.

To model the different assumptions concerning the nature of activation processes in working memory, we varied the implementation of spreading activation from the imaginal module between the different models as described in the following.

**Model 1.** In the first model, the amount of activation spread by each symptom depends on the number of symptoms observed so far. The imaginal module (holding the observed symptoms) can spread a certain amount of maximum activation that is equally divided among the symptoms:

$$W_j = W/n \quad (1)$$

with $W_j$ being the spreading activation associated with the $j^{th}$ symptom, W being the total amount of activation for the module and n being the number of symptoms hold by the module. This is the standard solution implemented in ACT-R. Thus, after the first symptom is presented, there is only one chunk in the imaginal module that can spread activation and thus, has a full spreading activation (set to 1). The more symptoms placed in the module over the course of the trial, the less activation is spread by each of these symptoms.

**Model 2.** For the second model, we implemented a function that assigned pre-specified amounts of activation

to be spread to the slots of the source chunk. The values associated with the slots were computed using a formula proposed for the decay of information in a constraint satisfaction model (Wang et al., 2006; see also Mehlhorn & Jahn, 2009) that assumed a non-linear negatively accelerated decay:

$$W_j = W_{j-1}(1-d)^{\sqrt{t}} \quad (2)$$

where $W_j$ is the spreading activation associated with the $j^{th}$ symptom, d denotes a decay parameter that was set to 0.4 and t denotes the time that has elapsed since the trial started. Thus, the most recent symptom always spreads a full amount of activation (set to 1). Over the course of the task, symptoms spread less activation the longer they are kept in the imaginal module.

**Model 3.** The third model also used a function assigning pre-specified decaying amounts of activation. However, in this model we implemented a linear instead of a negatively accelerated decay. To make sure that not the total amount of the decay, but only the slope of the decay function would influence the outcome, we used equal values as in Model 2 for the most recent and the oldest symptom:

$$W_j = W_1-(j-1)((W_1-W_4)/3) \quad (3)$$

with $W_j$ again being the spreading activation associated with the $j^{th}$ symptom and $W_1$ and $W_4$ being the spreading activation values for the most recent and the latest symptom as computed by formula (2). Thus, in this model the activation spread by symptoms decays away in a linear manner over time.

**Model 4.** A constant amount of activation associated with each slot of the source chunk was implemented in the fourth model. Thus, $W_j$ was set to a fixed value of 0.16 that had shown to provide a good fit to the human data. Thus, in this model, activation spread by a piece of information in the imaginal module neither depends on the amount of information in the module, nor on the time since the information first entered the module.

## Results

All four models were compared to the results produced by human participants on four dependent measures, namely the accuracy that was reached in the diagnosis and the probe task and the average reaction times for correct responses in these two tasks.

**Diagnosis Task.** Table 2 shows the mean accuracies and the mean reaction times for the diagnosis task. All models were able to solve the diagnosis task reaching very good to perfect performance. Inspecting the reaction times for correct diagnoses reveals that all models produced about the same reaction times as human participants.

Table 2. Mean accuracies and mean reaction times by models and human participants in the diagnosis task.

| | Mean accuracy (%) | Mean RT (ms) - correct diagnoses |
|---|---|---|
| Participants | 96.1 | 608.09 |
| Model 1 | 100 | 606.87 |
| Model 2 | 98.4 | 571.21 |
| Model 3 | 99.1 | 555.13 |
| Model 4 | 100 | 658.31 |

**Probe Task.** To analyze the accuracy of the probe task, for human data as well as for the models' data, only trials with correct final diagnoses were used. To analyze reaction times to probes, trials on which either the diagnosis or the probe response was wrong, were excluded. This was done because for human participants it remains unclear what caused the wrong diagnosis or the wrong probe response. For example, a participant might have missed a symptom and thus reached a wrong conclusion, implying that the activation measured in the probe task is not the activation of the target letter but rather that of another, possibly irrelevant one.

Human participants responded correctly to the probes in 93.1% of the trials whereas all models reached 100% accuracy. Reaction times for the different probe types are illustrated in Figure 2. For all probe types, Model 4 fits the human data best. The other models deviated more from the human data, which is not only evident in overall faster reaction times, but also in the less well fitting patterns. The different fits are reflected by the $R^2$ between participants' data and the modeling results as well as the RMSSDs; being $R^2$ = .35 and RMSSD = 2.75 for Model 1, $R^2$ = .37 and RMSSD = 3.00 for Model 2, and $R^2$ = .44 and RMSSD = 3.58 for Model 3, whereas Model 4 reached a $R^2$ of .80 and a RMSSD of .85.

As can be seen in Figure 2, for *relevant* probes, participants' reaction times decreased the closer the probe was presented to the end of the trial. Model 4 produced a pattern close to the participants' data. In all other models, reaction times were too fast at the beginning of the trials and did not change substantially during the trials, indicating that the earlier symptoms were overweighed. It is notable that none of the models fit the positive acceleration (that is, a sudden drop in reaction times from symptom 3 to symptom 4) of the participants' data.

For *irrelevant* probes, participants' response times decreased slowly over the trial. The models' reaction times decreased as well, but except for Model 4, this decrease was much faster than for the participants. Again, the slopes of the curves differed between all models and the participant data. Participants reacted increasingly faster towards the end of the trial, while the models' reactions decreased asymptotically toward some value.

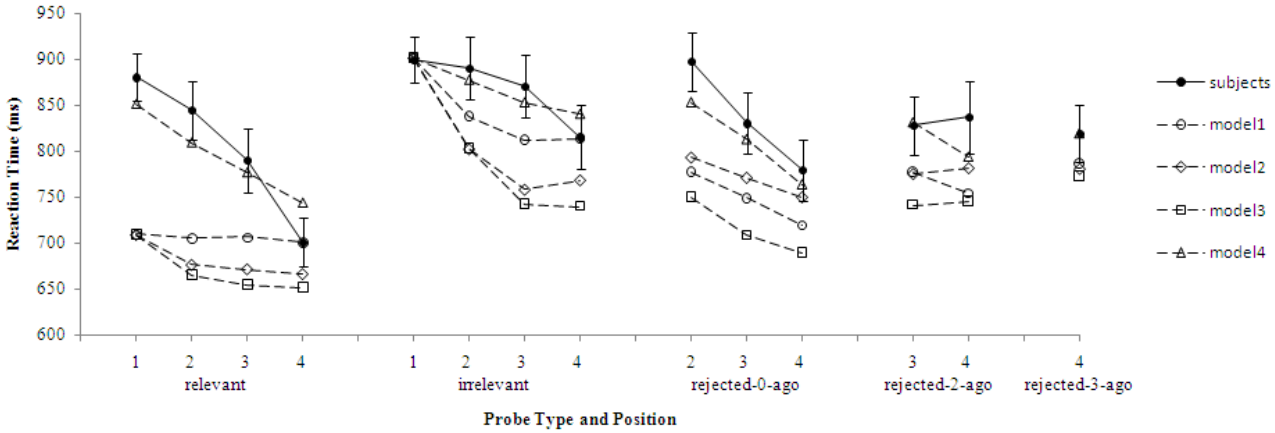


Figure 2. Models' and participants' reaction times for different probe types at different probe positions.

For the *just rejected* probes, all models' patterns matched the pattern provided by humans fairly well. However, the models' responses were too fast. Again, the fourth model's reaction times lie closest to the human data. For the probes rejected one symptom ago, all models' patterns again roughly matched the ones provided by human participants; except that Model 4 showed a slight decrease instead of an increase in the reaction times. Despite the difference in the slope, Model 4 again produced reaction times closest to the human data. Also for probes rejected two symptoms ago, the best fit to human data was provided by Model 4.

## Discussion

In this paper, we explored the influence of the implementation of different spreading activation processes in working memory during a diagnosis task. In the task, sequentially presented information needed to be integrated to find an explanation most consistent to all pieces of information. We compared the data provided by four ACT-R models that utilized different patterns of spreading activation to human data on several dependent measures. The analysis of diagnostic performance and the probe accuracy was important to show that all models were able to solve the task. However, the most interesting dependent variable is the probe reaction time. It not only provides a measure for how strongly different types of explanations are activated by the observed symptoms, but also how this activation changes over time.

As the results show, neither the standard implementation of ACT-R (Model 1), assuming the amount of spreading activation in the focus of attention to depend on the amount of information held in the source chunk nor models assuming the spreading activation of information in working memory to decay away with time (Models 2 and 3), could account for the patterns found in human data. Varying the pattern of the decay function from a negatively accelerated decay in Model 2 to a linear decay in Model 3 also had no substantial effect on the model fit. Concluding, none of the models assigning varying activation-values to the information held in working memory were able to fit the data.

Contrary to these models, our fourth model provided a pattern very close to the one provided by human participants. This model assumed the amount of spreading activation associated with each piece of information in working memory to be constant. Before discussing possible implications of this finding, we would like to address several potentially critical aspects of our approach.

One could argue that the bad fit of the Models 1, 2 and 3 might only be due to the high base levels assigned to the diagnosis chunks, thus causing the reaction times to be too short. To rule out this possibility, we also implemented the three models with lower base-levels. However, this did not improve the models' fit, because it did not affect the pattern of the response times, but only the absolute level.

Another possible source of criticism might be the different amount of total spreading activation that was used for the different models. That is, for example in Model 2, the sum of all activations assigned to the different slots of the chunk in the imaginal module was 1.56, whereas the total spreading activation in Model 1 was 1. To rule out possible criticism related to this point, we also implemented all four models in a way such that the total spreading activation was constant across the models. This, again, did not change much about the general data pattern.

## Conclusions

Our results have several interesting implications. First, they question the implementation of spreading activation currently used in ACT-R. Second, they question the assumption of decay in working memory as proposed in some constraint satisfaction models. Why could those theoretical assumptions not be confirmed by our data? Does the amount of information in working memory really have no impact on how much activation can be

spread by each piece of information? And is information in working memory really not subject to decay?

We would answer both questions with no. The results do neither implicate that there is no overall limit to the amount of activation spread from working memory nor that there is no decay. In our task, participants had to maintain a relatively small amount of information in working memory (up to four symptoms). This lies within the general range of working memory capacity (cf. Cowan, 2000). Thus, our results do not question that the total amount of activation spread from the focus of attention is related to working memory capacity (e.g., Lovett et al., 2000). Rather, this spreading activation might be assigned to the information in the focus of attention in a different way. That is, until the total capacity of working memory is reached, each piece of information seems to spread the same amount of activation.

Moreover, only information that is not currently held in the focus of attention might be subject to processes of decay. That means, as soon as some piece of information becomes irrelevant to the current task or as the amount of information in the focus exceeds its limited capacity, this information might decay away. However, in our task, the information neither became irrelevant nor did it exceed the capacity of working memory during the whole reasoning process. When new symptoms are observed, the reasoner needs to integrate them with earlier symptoms to find an explanation consistent to all symptoms. Therefore, the older symptoms need to be actively maintained, and thus they do not decay.

An interesting question for further research would be to take a closer look at what happens when the amount of information to be actively maintained during the task exceeds working memory capacity. As several authors suggest, in such cases the least activated information would be dropped from working memory (e.g., Thomas et al., 2008; Chuderski, Stettner & Orzechowski, 2006). Thus, this information should no longer be able to spread activation to associated information in long-term memory but instead it should become subject to decay.

Concluding, our results shed some light on the representation of information in working memory during a sequential diagnostic reasoning task. They suggest that in such a task, each piece of this information has the same potential to activate associated knowledge from working memory. It will be an interesting question for further research to determine in how far this finding can be generalized from diagnostic reasoning to other tasks that require information to be actively maintained in working memory.

## Acknowledgements

We thank Jelmer Borst, Niels Taatgen, Christian Lebiere and four anonymous reviewers for their helpful comments on the models and on an earlier version of this paper.

## References

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press.

Arocha, J. F., & Patel, V. L. (1995). Construction-integration theory and clinical reasoning. In C. A. Weaver, III, S. Mannes & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch.* (pp. 359-381). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Baumann, M. (2001). *Die Funktion des Arbeitsgedächtnisses beim abduktiven Schließen: Experimente zur Verfügbarkeit der mentalen Repräsentation erklärter und nicht erklärter Beobachtungen*. Doctoral dissertation, Chemnitz University of Technology, Germany.

Baumann, M.., Mehlhorn, K., & Bocklisch, F. (2007). The activation of hypotheses during abductive reasoning. *Proceedings of the 29th Annual Cognitive Science Society* (pp. 803-808).

Chuderski, A., Stettner, Z., & Orzechowski, J. (2006). Modeling individual differences in working memory search task. *Proceedings of the Seventh International Conference on Cognitive Modeling* (pp. 74-79).

Cowan, N. (2000). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Johnson, T. R. & Krems, J. F. (2001) Use of current explanation in multicausal abductive reasoning. *Cognitive Science, 25*, 903-939.

Josephson, J. & Josephson, S. G. (1994). *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Just, M A. & Carpenter, M. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review, 99*, 122-149.

Lovett, M. C., Daily, L. Z. & Reder, L. M. (2000). A source activation theory of working memory: cross-task predictions of performance in ACT-R. *Cognitive Systems Research, 1*, 99-118.

Mehlhorn, K. & Jahn, G. (2009). Modeling sequential information integration with parallel constraint satisfaction. To appear in: *Proceedings of the 31st Annual Cognitive Science Society*.

Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227-234.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M. & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*, 155-185.

Wang, H., Johnson, T. R., & Zhang, J. (2006). The order effect in human abductive reasoning: an empirical and computational study. *Journal of Experimental & Theoretical Artificial Intelligence. 18(2)*, 215–247