The Influence of Experience and Context on Hypothesis Generation

In which I use behavioral experiments and an ACT-R model to investigate the respective contribution of previous experience and the current context on explanations' availability in memory.

Chapter

The Influence of Experience and Context on Hypothesis Generation

Chapter

Abstract

Recent theories of diagnostic reasoning propose that automatic memory activation processes are involved in the generation of hypotheses from memory. Two aspects have been suggested to play a role: (1) a hypothesis' past usefulness and (2) its usefulness in the current context. Based on a general theory of memory, we present two mechanisms that might explain theses aspects: (1) a hypothesis' base-level activation, reflecting its past usefulness, and (2) the spreading activation it receives from current observations, reflecting its usefulness in the current context. We conducted an experiment in which participants had to generate hypotheses, while both memory components were independently manipulated by an ostensibly unrelated secondary task. The results show an effect of both manipulations and are quantitatively predicted by an ACT-R model in which we implemented both memory mechanisms. Discrepancies between the behavioral data and the predictions of the mere memory-based model were also revealed and their potential reasons are discussed.

Introduction

A doctor trying to find the best diagnosis for his patient's symptoms, a scientist trying to understand her data, or a person trying to deduce someone else's intentions are all examples in which hypotheses need to be generated and evaluated from memory. Traditionally, cognitive psychology has focused on deliberate reasoning processes that are engaged in solving such tasks (e.g., T. R. Johnson & Krems, 2001; Klahr & Dunbar, 1988). More recently, researchers have started investigating the role of automatic memory processes for hypothesis generation (e.g., Arocha & Patel, 1995; Mehlhorn, Taatgen, Lebiere, & Krems, 2011; Thomas et al., 2008; see also Chapter 2 of this thesis).

The basic idea is that automatic memory processes can provide an adaptive subset of possible hypotheses from memory, which can serve as input to a more deliberate evaluation process (Thomas et al., 2008). Such a distinction between a memory-based and a reasoning-based component is also a central aspect of dual-process theories. They assume fast, automatic processes to provide a possible answer, which might then be justified or revised by more time consuming, deliberate reasoning (Evans, 2008).

Empirical evidence suggests that the generation of hypotheses from memory depends on two aspects. A first aspect is the hypotheses' usefulness in the past. It has been shown that from all potential hypotheses, reasoners tend to generate those that have a high a priori probability based on previous experiences (Dougherty & Hunter, 2003a; Weber et al., 1993). A second aspect is the hypotheses' usefulness in the current context. For example, Weber et al. (1993) have shown that physicians generate those diagnoses that are most likely in the light of a patient's symptoms. While both aspects have received empirical support, the underlying mechanisms, as well as their respective contribution for hypothesis generation have received relatively little attention in the literature (for an exception, see Thomas et al., 2008).

The goal of this chapter is to show the respective contribution of both aspects and to test in how far general memory mechanisms can explain the effects. We first give an overview on the memory mechanisms, before we describe an experiment in which we manipulated both aspects. Subsequently, we show how we generated quantitative predictions from a cognitive model and compare these predictions to the data from the experiment.

Memory Processes in Hypothesis Generation

A general assumption of memory theories is that "the memory system […] makes most available those memories most likely to be needed" (Anderson, 2007, p. 109). How does it do that? While theories differ on the exact proposed mechanisms and the used vocabulary, commonly, two components are assumed to determine the likelihood of an item to be needed from memory: its a priori probability based on previous experiences and its usefulness in the current context.

Previous experience. Anderson and Schooler (1991) investigated how previous experience predicts an item's likelihood to be needed from memory. Based on their

results it has been suggested (e.g., Anderson, 2007), that the inherent availability of an item in memory can be described by its base-level activation, B_i, which depends on the frequency and recency of the items past usage:

$$B_i = \ln\left(\sum_{k=1}^{n} t_k^{-d}\right) \tag{4.1}$$

where n is the number of previous encounters with item i, t_k is the time since the k^{th} encounter, and d is a decay parameter (producing the power law of forgetting). Using the example of a physician, this mechanism could for example explain why, especially during flu season, the flu will seem to be a more likely diagnosis for a patient's symptoms than throat cancer.

Current context. Various memory theories share the assumption that information in the environment can serve as a cue for the retrieval of items from memory (e.g., Anderson, 2007; Kintsch, 1998; Thomas et al., 2008). A frequently proposed mechanism underlying this cued retrieval is spreading activation between observed information and associated items in memory (e.g., Anderson, 2007; Thagard, 2000). Specifically, Anderson proposes that an item i in memory receives spreading activation, S_i, from each associated piece of information j, which is currently stored in working memory:

$$S_i = \sum_j W_j S_{ji} \tag{4.2}$$

where W_j represents a weighting of j in working memory and S_{ji} represents the associative strength between i and j. This associative strength reflects the extent to which an observed piece of information increases the likelihood of an associated item to be needed from memory. Using the physician's example again, this mechanism could explain why, in the context of symptoms that point specifically at throat cancer, this diagnosis might become available in memory.

In a previous study (Mehlhorn et al., 2011; see also Chapter 2 of this thesis), we investigated whether the current context could indeed affect the availability of hypotheses in memory as predicted by such a spreading activation account. In two experiments, participants had to generate diagnoses for sequentially presented medical symptoms, while we tracked the availability of different hypotheses in memory with a probe reaction task. As predicted, availability was found to vary over time as a function of the observed symptoms.

Respective contribution of the memory processes. While the results of Mehlhorn et al. (2011) provide evidence for the influence of the current context via spreading activation mechanisms, the respective contribution of a hypothesis' past usefulness was not investigated in that study. Anderson (2007) argued that base-level activation, B_i, and spreading activation, S_i, are independent additive components that determine the availability of an item i in memory:

$$A_i = B_i + S_i + \varepsilon \tag{4.3}$$

where A_i is an item's activation in memory and ε is a random noise component (see e.g., Anderson, 1990, for the underlying Bayesian statistics). For our physician, this could, explain why, depending on whether the base-level or the spreading-activation component are stronger, the flue or throat cancer are more strongly available in memory.

Overview of the Experiment

To investigate the memory components outlined above, we conducted an experiment in which participants had to solve a *diagnostic reasoning task,* while at the same time carrying out a secondary *choice-reaction task*. In each experimental trial, the medical symptoms of a hypothetical patient were presented one at a time on the screen. At the end of the trial, participants had to report the diagnosis that explained the patient's symptoms. During presentation of the symptoms, participants were auditorily presented with letters and had to indicate as fast as possible whether the letter was a consonant (target) or a vowel. This choice-reaction task was used to manipulate both memory components independently.

The base-level component was manipulated by the *targets* presented in the choice-reaction task. Targets were either neutral consonants or consonants that were also used to name potential diagnoses. We expected that retrieving such diagnosis-naming targets from memory would increase the base-level activation of the respective diagnoses. Consequently, performance in the diagnosis task should be reduced, as the diagnoses whose base-level activation was increased by the diagnosis-naming targets are not necessarily the correct diagnoses for the presented symptoms.

The spreading activation component was manipulated by requiring participants to *count* the targets presented in the choice-reaction task in part of the trials. The idea behind this manipulation is that counting, as well as generating hypotheses, both rely on a central working-memory resource, which can only be used for one task at a time (Borst et al., 2010; Oberauer, 2002). The to be expected working memory conflicts can result in losing part of the observed symptoms from working memory. Consequently, performance in the diagnosis task should be reduced, as the correct diagnosis receives less spreading activation from the reduced amount of symptoms in working memory.

Method

Participants

Twenty-five native German speaking undergraduate students from the University of Groningen took part in this experiment for course credit (19 female; mean age 21.2, *SD* = 1.3).

Material

Diagnostic knowledge. The knowledge participants needed to learn before solving the diagnosis task was adapted from Mehlhorn et al. (2011; see also Chapter 2 of

Table 4.1 Overview of diagnostic knowledge participants had to acquire before the experiment (original material in German).

Aggregate state and source of contamination	Chemical	Specific symptoms			Unspecific symptoms		
Gasiform, inhaled	B	Cough		Shortness of breath	Headache		Dizziness
	T	Cough	Sneezing		Headache	Fever	
	W		Sneezing	Shortness of breath		Fever	Dizziness
Crystalline, skin contact	L	Redness		Rash	Headache		Dizziness
	H	Redness	Itching		Headache	Fever	
	G		Itching	Rash		Fever	Dizziness
Liquid, drinking water	C	Diarrhea		Cramps	Headache		Dizziness
	M	Diarrhea	Vomiting		Headache	Fever	
	R		Vomiting	Cramps		Fever	Dizziness

this thesis) and consisted of nine hypothetical chemicals (all single consonants, see Table 4.1). The chemicals were grouped into three artificial categories and caused four symptoms each. To reflect the complexity of real-world diagnostic knowledge, symptoms were either specific for a category (e.g., cough) or unspecific (e.g., headache).

Audio stimuli. For the choice-reaction task we generated three sets of audio files: *chemical consonants* (the 9 chemical names), *non-chemical consonants* (the letters SKQFVDNP), and *vowels* (the letters AEIOUÄÖÜ). In the *non-chemical condition,* audio stimuli were randomly sampled from the non-chemical consonants and vowels. In the *chemical condition*, audio stimuli were randomly sampled from the chemical consonants and vowels. In this condition, we additionally varied whether the set of consonants included the correct diagnosis for the presented symptoms (*correct diagnosis primed*) or not (*correct diagnosis not primed*).

Procedure

Training. To learn the diagnostic knowledge, participants were visually presented with the four symptoms caused by one of the chemicals and had to enter a diagnosis. By receiving feedback, they learned which chemicals where associated with which symptoms. Categories were first trained separately, before the same training was repeated for the complete material. The order of categories, diagnoses, and symptoms

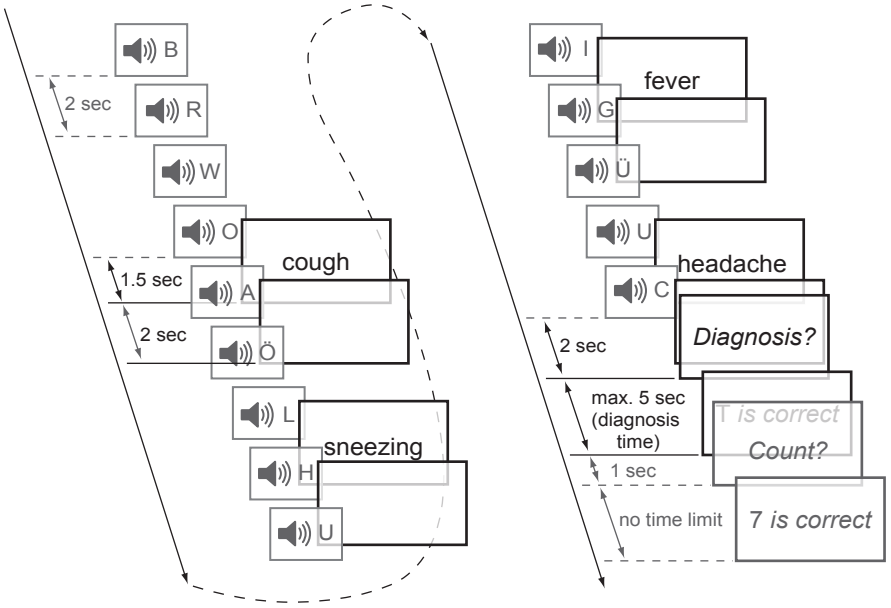

Figure 4.1 Sample trial for the chemical–count condition. In this trial, 14 audio stimuli were presented of which 7 were consonants. T was the correct diagnosis for the medical symptoms.

was randomized for each participant. After acquiring the diagnostic knowledge (performance criterion: 100%), the choice-reaction task was first practiced alone and then in combination with the diagnosis task. Participants were informed that in the experiment it was important to do both tasks as fast and accurately as possible.

Experiment. The experiment was split into 8 blocks. In each block, 9 trials from one of the four conditions (non-chemical – no count, non-chemical – count, chemical – no count; chemical – count) were presented. The conditions were assigned to the blocks in random order, with the constraint that each condition had to be presented once in the first four and once in the second four blocks.

In each trial, 12 to 14 audio stimuli were presented with a SOA of 2 s (Figure 4.1). Six to twelve of the stimuli were consonants (the exact numbers were randomly drawn from a uniform distribution for each participant for each trial). Additionally, the four symptoms of one of the chemicals were visually presented for 2 s each. The 1st symptom was presented 1.5 s after the onset of the 2nd, 3rd, or 4th audio stimulus, with the exact position depending on the total number of stimuli in the trial. Before each subsequent symptom, 3 audio stimuli were presented. The final audio stimulus was presented .5 s after onset of the 4th symptom. The presentation order of audio stimuli and symptoms was randomized for trials and participants. Each chemical occurred with equal frequency as correct diagnosis in each block. After all stimuli had been presented, participants had to enter their diagnosis within maximum 5 s and, in the count condition, to enter the number of consonants. Participants received

visual feedback for their diagnosis and count. Auditory feedback was presented for the choice reactions if the response was wrong or not given within 1.25 s.

Model

Based on a previously published model of hypothesis generation (Mehlhorn et al., 2011; see also Chapter 2 of this thesis), we implemented a cognitive model in the ACT-R architecture (Anderson, 2007). ACT-R makes precise predictions about how the memory mechanisms described above affect the probability and latency of memory retrieval. It allows for modeling the task, as solved by the participant, and thereby produces results that are directly comparable to the human data. Below we briefly describe the model (the model code, including more detailed explanations, can be downloaded from http://www.ai.rug.nl/~katja/models).[1]

The model is presented with the same tasks as the participants, that is, it has to discriminate between the auditorily perceived consonants and vowels, it has to count the consonants (in the count condition), and it has to generate a diagnosis for the visually perceived symptoms. The knowledge necessary to solve these tasks is represented in the model's long-term memory (the declarative memory).

To solve the choice-reaction task, the model tries to retrieve the perceived letter from declarative memory, assesses if the retrieved letter is a consonant or vowel, and enters its response. To count, the model keeps track of the current count in working memory.[2] The count is incremented when a retrieved letter is classified as consonant. When asked for the count, the model enters the current count. To solve the diagnosis task, the model stores the observed symptoms in working memory, from where they spread activation to associated diagnoses in declarative memory. When asked for the diagnosis, the model retrieves and enters that diagnosis from declarative memory that has the highest activation as calculated by Equation 4.3.

In the no-count condition, all observed symptoms are stored in working memory until the model is asked for a diagnosis. In the count condition, the set of symptoms that is currently stored in working memory has to be swapped out to declarative memory whenever the model needs working memory for counting, because working memory can only be used for one task at a time (see Borst et al., 2010, for empirical support). Whenever working memory is needed for the diagnosis task again, the current count is swapped out and the set of symptoms is swapped back in. Information can be lost during swapping because, due to noise, the model might erroneously retrieve older working-memory contents (e.g., an incomplete set of symptoms) from declarative memory.

[1] *To fit the model, we estimated the latency and stochasticity of memory retrievals, the base-level activation of facts in memory at the beginning of each trial, and the maximum associative strength between items. Based on earlier results (Mehlhorn et al., 2011), we assumed the associative strength of each symptom in working memory to be weighted by a constant value of W, independent of the number of symptoms (see the model code for the exact parameter values). The model was run 40 times for each participant. Results were calculated for each run and then averaged across runs.*

[2] *To model working memory we use one of the buffers of ACT-R's cognitive modules, the imaginal buffer. The imaginal buffer is commonly used to hold a mental representation of the problem currently in the focus of attention (Borst et al., 2010).*

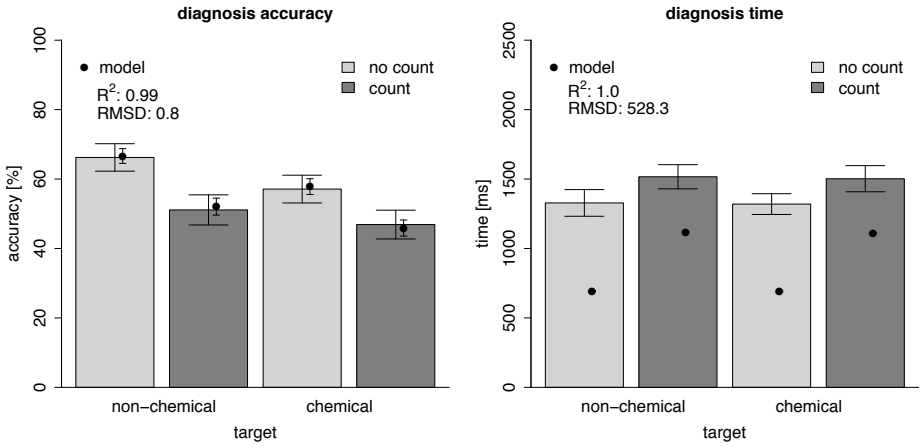

Figure 4.2 Diagnosis accuracy (left) and time (right) for the factors target (non-chemical, chemical) and counting (no count, count); $M \pm 1$ *SE* for human (bars) and model data (dots).

Results

Performance in the Diagnosis Task

Effects of target and counting. To investigate the respective impact of both manipulated memory components, we analyzed diagnosis accuracy and time for the factors target (non-chemical, chemical) and counting (no count, count). The results are shown in Figure 4.2 and Table 4.2. As correctly predicted by the model, chemical targets lead to lower diagnosis accuracy than non-chemical targets, but do not affect diagnosis times. Also the effects of counting are correctly predicted by the model: counting leads to lower diagnosis accuracies and higher diagnosis times than no counting. However, the model generally underestimates diagnosis times.

Effect of priming the correct diagnosis. To further test the effect of the base-level manipulation on diagnosis performance, we compared chemical-condition trials in which the correct diagnosis was among the presented chemical consonants (*correct diagnosis primed*) to chemical-condition trials in which this was not the case (*correct diagnosis not primed*). As shown in Table 4.3, the model predicts higher diagnosis accuracy in primed than in not-primed trials, while participants do not show this effect on diagnosis accuracy. However, participants do show an effect on diagnosis time, with primed diagnoses being faster than not-primed ones. The model correctly predicts this effect on diagnosis time, but underpredicts its magnitude.

Performance in the Choice-Reaction Task

Reaction and counting accuracy. Participants discriminated between consonants and vowels with a reasonably high accuracy (M = 88.5%, SD = 6.0), which is approximated

Table 4.2 Results of repeated-measures ANOVAs for the factors target (non-chemical, chemical) and counting (no count, count).

Dependent measure	Effects	*F*	*p*	η_p^2
Diagnosis accuracy	Main effect of target (non-chemical, chemical)	(1,24) = 11.15	**.003**	.317
	Main effect of counting (no count, count)	(1,24) = 27.47	< **.001**	.534
	Interaction of target and counting	(1,24) = 2.60	.120	.098
Diagnosis time[a]	Main effect of target (non-chemical, chemical)	(1,23) = .07	.792	.003
	Main effect of counting (no count, count)	(1,23) = 14.72	< **.001**	.390
	Interaction of target and counting	(1,23) = .01	.942	< .001

Note. *p* values <.1 are shown in bold. [a] Only correct responses. Data points that differed more than 3 *SD* from a participant's condition mean were excluded (0.1% of the diagnosis time data). Additionally, one participant had to be completely excluded from this analysis, because of a diagnosis accuracy of 0 in the non-chemical – count condition.

Table 4.3 Human and model data in the chemical-consonants condition, depending on whether the correct diagnosis was primed or not.

Dependent measure	Priming	Human *M* (*SD*)	Model *M* (*SD*)
Diagnosis accuracy [%]	correct diagnosis primed	52.1 (22.5)	58.4 (11.4)
	correct diagnosis not primed	51.7 (22.8)	45.1 (11.7)
		$t(24) = .08, p = .940$	
Diagnosis time [ms][a]	correct diagnosis primed	1339 (363)	853 (53)
	correct diagnosis not primed	1533 (551)	902 (66)
		$t(24) = -2.50, p =$ **.020**	

Note. *p* values <.1 are shown in bold. For simplicity, here we collapsed over the factor counting, which was justified by a lack of interactions with the factor. [a] Only correct responses. No diagnosis times differed more than 3 *SD* from a participant's condition mean.

well by the model (*M* = 91.3%, *SD* = 1.0). The correct count was reported in 57.1% (*SD* = 20.2) of the trials in the count condition. The model reaches a slightly lower counting accuracy (*M* = 39.7%, *SD* = 8.3).

Reaction accuracy over the trial. Due to the nature of memory activation, we also expected the diagnosis task to influence performance in the choice-reaction task. For

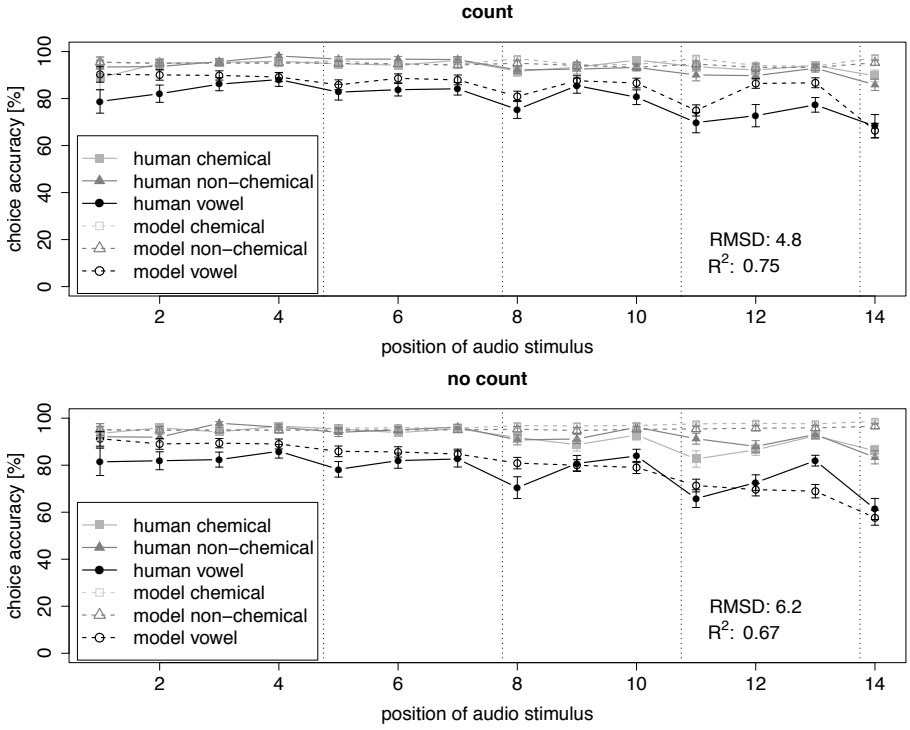

Figure 4.3 Human and model accuracy ($M \pm 1$ *SE*) for the choice-reaction task in the count (top) and no-count (bottom) condition for the different stimuli (non-chemical, chemical, vowel) at each of the 14 possible positions in the trial (positions were lined out to the last stimulus, so that each trial had a 14th stimulus, but only trials with 14 stimuli had a 1st stimulus). Vertical lines indicate the four points at which symptoms were presented.

example, if observed symptoms indeed spread activation to associated hypotheses, the availability of the respective chemical consonant letters should increase, resulting in an increased chance to retrieve consonants and thereby to more errors in reacting to vowels. To test this assumption, we analyzed the accuracy of reactions to the different types of stimuli (chemical consonant, non-chemical consonant, vowel), depending on counting, and on the stimulus' position in the trial (1-14).

As correctly predicted by the model, overall accuracy in the choice-reaction task was indeed lower for vowels than for consonants (Figure 4.3). In the count condition, the model correctly predicts a drop of response accuracy to vowels whenever a new symptom is observed. This happens because, due to the swapping of information between the diagnosis and counting task, observed symptoms spread activation to associated chemical consonants at these points. In the no-count condition, the model predicts a slightly more gradual decrease of response accuracy to vowels than found in the human data. The decrease in the model is caused by the increasing amount of spreading activation from the increasing number of symptoms in working memory.

Discussion

Empirical research has shown that, when generating hypotheses from memory, reasoners generate only a small subset of all potential hypotheses. However, this subset seems to be highly adaptive, as it contains those hypotheses that have (1) a high a priori probability based on previous experience and (2) a high usefulness in the current context. The results of our study can help to understand this adaptive selection in terms of general memory mechanisms. We presented base-level activation as a memory mechanism that is sensitive to a hypothesis' past usefulness. It predicts the availability of a hypothesis in memory to increase with the frequency and recency of its usage. We presented spreading activation as a mechanism that can regulate the influence of the current context on a hypothesis' availability in memory. It predicts the availability of a hypothesis in memory to increase with the amount of observations in working memory that are associated with this hypothesis and with the strength of their association.

While the influence of the current context via spreading activation mechanisms was already supported in an earlier study (Mehlhorn et al., 2011; see also Chapter 2 of this thesis), the respective contribution of a base-level activation mechanism that reflects a hypothesis' past usefulness had not yet been shown. To test this contribution, we manipulated both components within one experiment. Diagnosis performance showed main effects of *both* manipulations, suggesting that the components might indeed reflect two distinct aspects of memory activation. This assumption is supported by the cognitive model, which revealed both base-level activation and spreading activation to be important components for fitting the behavioral data.

The model explains the reduced performance in the chemical compared to the non-chemical condition by an increase in base-level activation of wrong diagnoses, which were presented as letters in the choice-reaction task. Finding behavioral evidence for this influence is especially interesting because, objectively, participants had to do the same choice-reaction task in both conditions: discriminate between consonants and vowels. It is additionally interesting because the letters were used in semantically different meanings in both tasks: In the choice-reaction task, they were used to discriminate between consonants and vowels, while in the diagnosis task, they were used as potential diagnoses. Nevertheless, diagnosis performance decreased, as predicted by the model, when the consonants of the choice-reaction task were names of chemicals. This demonstrates the impact of automatic memory activation on hypothesis generation.

The model explains the reduced performance in the count condition compared to the no-count condition by a decrease in spreading activation to the correct diagnosis. This decrease is caused by working-memory conflicts between the count and diagnosis task, which result in the loss of observed symptoms from working memory in the count condition. Together with the findings presented in Chapter 2, this illustrates how hypothesis generation depends on the current context that is available in working memory.

It has been proposed that automatic hypothesis-generation processes interact with deliberate hypothesis evaluation (Thomas et al., 2008). Deviations between the results of our merely memory-based model and the behavioral data suggest such an interaction also in our study. The absence of an effect of priming on human diagnosis accuracy, as well as the model's general underprediction of diagnosis times, suggest that participants did not simply enter the diagnosis made most available by memory activation as in the model. Rather, participants might have used additional time to evaluate and justify the retrieved hypotheses.

De Neys (2006) showed that the use of deliberate reasoning strategies increases with the availability of working-memory resources. Such an increased use of deliberate reasoning might explain why, in our choice-reaction data, the model which did not use any deliberate reasoning strategies fitted better in the count condition (with high working memory demands) than in the no-count condition (with lower working memory demands). However, despite participants' potential additional use of deliberate reasoning, the mere activation-based model fits the choice-reaction data quite well. This is remarkable, because this fit directly emerges from the memory activation mechanisms implemented in the model, without us adding any additional assumptions.

Are the results of our laboratory study generalizable? Real-world hypothesis generation will often be more complex and less structured than the diagnosis task participants solved in our experiment. We decided for such a simplified hypothesis-generation task, because it allowed for the experimental control necessary to test our assumptions about the subtle effects of memory activation. However, we do not expect this simplification of the task to limit the validity of our results. Higher complexity and a less well defined task structure are expected to even increase the importance of memory activation processes (Dijksterhuis & Nordgren, 2006).

Before closing, we want to stress that the main point of this chapter is not to promote one particular memory theory, but to show how taking into account the importance of automatic memory activation can help to understand hypothesis generation. While memory theories differ in the exact proposed mechanisms, many theories share the assumption that the probability of an item to be needed from memory depends on the two factors discussed in this chapter: the item's a priori probability based on previous experiences and its usefulness in the current context (e.g., Anderson, 2007; Thomas et al., 2008).