Memory Indexing of Sequential Symptom Processing in Diagnostic Reasoning

Georg Jahn (georg.jahn@uni-greifswald.de)
University of Greifswald, Department of Psychology
Franz-Mehring-Str. 47, D-17487 Greifswald, Germany

Janina Braatz (janina.braatz@uni-greifswald.de)
University of Greifswald, Department of Psychology
Franz-Mehring-Str. 47, D-17487 Greifswald, Germany

Abstract

Explaining symptoms by the most likely cause is a process during which hypotheses are activated and updated in memory. By letting participants learn about causes and symptoms in a spatial array, we could apply eye tracking during diagnostic reasoning to trace the activation level of hypotheses across a sequence of symptoms. Fixation proportions on former locations of possible causes reflected the causal strength of initial symptoms, a bias towards focal hypotheses, and the final diagnosis. Looking-at-nothing revealing memory activation consistent with process models of diagnostic reasoning was stable even after one week.

Keywords: Diagnostic reasoning, Probabilistic inference, Eye tracking, Order effects, Spatial index

Introduction

The goal of diagnostic reasoning is to determine the most likely cause of observed symptoms. In routine cases, medical diagnosis may proceed as simple pattern classification. Often, however, symptoms are ambiguous and the clinician has to consider multiple alternative diagnoses. Then, medical diagnosis is a case of hypothesis generation and hypothesis testing (Lange, Thomas, & Davelaar, 2012; Thomas, Dougherty, Sprenger, & Harbison, 2008) as it occurs in science, criminal investigation, or searching for faults in technical systems. Diagnostic reasoning with limited information search, for example, when clinical cases are presented as case histories, requires information integration based on knowledge and multiple probabilistic cues. In the present study, we used the cover story of an accident in a chemical plant, in which workers were affected by one of four chemicals (Mehlhorn, Taatgen, Lebiere, & Krems, 2011). Participants had to decide, which chemical had caused a worker's symptoms. By laying out chemicals and symptoms in a spatial array in a learning phase, we could use eye tracking for process tracing of memory-based diagnostic reasoning to study the updating of diagnostic hypotheses.

In sequential diagnostic reasoning, the first symptom triggers a limited number of candidate hypotheses, which frame the processing of subsequent symptoms. Equally supported alternative hypotheses may be missed or rated less likely than the focal hypothesis. Similar primacy order effects have been documented for judicial decision making and social judgment, for example.

The focal hypothesis or the set of focal hypotheses is held in working memory. If symptoms have to be retained in working memory as well, capacity limits increase in importance. Cued recall of candidate hypotheses and sequential symptom processing to update the focal hypotheses' degree of support are cognitive processes, which elude observation and are altered if ratings are elicited during reasoning. If external representations of symptoms or knowledge about causes were permanently ready for inspection, patterns of information search could be recorded via behavior records (e.g., Mouselab) or eye tracking. Here, we demonstrate a similar process tracing method (Renkewitz & Jahn, in press) that is suitable for investigating purely memory-based hypothesis generation and symptom processing. It builds on the tendency to direct the gaze to locations where information was presented before when one attempts to retrieve it from memory or reactivates it in working memory (the "looking-at-nothing" phenomenon; Richardson & Spivey, 2000).

Our participants learned about the four chemicals and the symptoms that they could cause in a spatial array. During diagnostic reasoning, symptoms were presented auditorily and the participants' eye movements on the emptied spatial array were tracked. Our goal was to trace the activation, updating, and revision of hypotheses. In particular, we were interested to see whether the activation of initial hypotheses reflected the strength of support by the first symptom, whether focal hypotheses had an advantage over equally supported alternative hypotheses, and whether fixation proportions corresponded to the final diagnosis in trials with ambiguous symptom sequences. Extending previous findings, we demonstrate the stability of looking-at-nothing over an interval of one week.

Experiment

Method

Participants. Thirty-six students of the University of Greifswald (28 female, 8 male) with a mean age of 22.1 years (*SD* = 2.4) completed the first session. 32 of them returned for the second session 7.3 days later on average (*SD* = 1; range 6 to 10 days).

Materials. To prepare for the diagnostic reasoning task, participants learned about four chemicals and possible symptoms. There were six symptom classes each containing two symptoms that are listed in Table 1. The four chemicals and the symptom classes that each could cause were presented in a 2x2 arrangement as shown in Figure 1. The square in the bottom right quadrant measured 9.1° by 9.1° of visual angle. Symptoms from the symptom class in the top rectangle were “almost always” caused by the respective chemical, those in the middle and bottom rectangles were “occasionally” caused by the respective chemical.

Table 1: Symptom classes and symptoms. The original materials were in German.

Symptom Class	Symptom	Symptom
Eyes	Eyelid swelling	Lacrimation
Respiration	Cough	Difficult breathing
Skin	Acid burn	Rash
Neurological	Paralysis	Speech disorder
Circulatory Pr.	Sweating	Swoon
Pain	Twinge	Sting

As can be seen in Figure 1, each symptom class appeared with two chemicals. For example, “Eyes” symptoms were almost always caused by the top left chemical, but only occasionally by the top right chemical. Such symptoms are denoted “Ab” (frequent for A, occasional for B) or “Ba” (frequent for B, occasional for A) in the following. Furthermore, each chemical shared an occasional symptom class (Circulatory Problems, Pain) with a chemical in the diagonally opposite quadrant. Symptoms from these classes are denoted “ac” in the following.

A single trial in the diagnostic task consisted of a sequence of four symptoms, for example: Eyelid swelling, Cough, Swoon, and Difficult Breathing (Ab_Ba_ac_Ba). Note that in this example, the third symptom (a circulatory problem) disambiguates the symptoms heard up to then and leaves only “A” (the top left chemical in this example) as the final diagnosis.

Ten different item types were constructed that are listed in Table 2. The point in the sequence at which a symptom in combination with foregoing symptoms determined the final diagnosis did vary across item types. In item type 10 the symptom pattern remained ambiguous. The column denoted “Specific symptoms” shows which item types are equivalent regarding the evidence provided by the symptoms irrespective of symptom order.

The symptom orders in Table 2 were used with each of the chemicals in the A-role. This was possible because the chemicals’ symptom patterns were symmetric. Furthermore, all possible assignments of symptoms to item types were constructed with the restriction that no single symptom was repeated in a symptom sequence.

Procedure. The experiment consisted of two sessions. In the first session, the participants acquired the knowledge to be used in diagnostic reasoning and then completed two phases of diagnostic reasoning trials. In the first half of diagnostic reasoning trials, the 2x2 arrangement of geometric forms was the same as during learning. In the second half, the arrangement was changed. The bottom pair became the top pair and the top pair became the bottom pair. In the second session, which took place 6 to 10 days later (*M* = 7.1), the participants returned for diagnostic reasoning trials, in which the original arrangement was presented again. Eye movements were recorded during diagnostic reasoning only.

Knowledge acquisition. Participants were instructed that their task would be to determine the cause of a patient’s symptoms. They were told that the patients are workers in a chemical plant, which processes four chemicals. Each patient was affected by exactly one of those chemicals. They should determine, which chemical most likely had caused a patient’s symptoms. Next, they studied Table 1 and worked through test trials until the set of twelve symptoms was once assigned to symptom classes without errors.

Then, participants were told that each chemical could cause three of the six symptom classes and the frequency with which a chemical causes a symptom class would vary. The symptom class shown in line one would be caused “almost always” and the symptom classes shown in lines two and three would be caused “occasionally”. Next, the chemicals with symptom classes were presented as shown in Figure 1. They could be studied until participants felt ready to be tested.

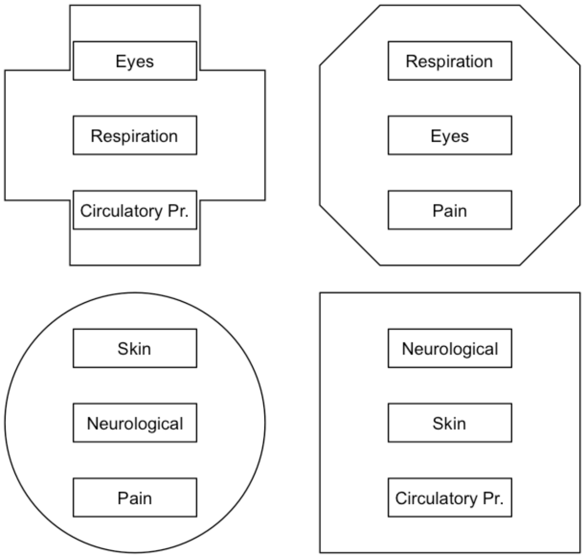

Figure 1: The four chemicals as they were presented in the learning phase. During diagnostic reasoning, the rectangular frames containing the symptom categories were empty.

In each test trial of the learning procedure, the emptied spatial array of geometric forms was shown and a symptom was presented acoustically followed by either “almost always” or “occasionally”. Participants responded by indicating the chemical that causes this symptom with this frequency. The response was given with adjacent keys on a standard keyboard (u, i, j, and k), whose arrangement approximately matched the 2x2 arrangement on the screen. Feedback was provided acoustically with a mellow or an unpleasant tone. After positive feedback, the next trial started automatically. After negative feedback, the filled arrangement was presented until participants hit the space bar to proceed to the next test trial. Testing continued until the set of 20 different testing items presented in random order was once answered without errors. Learning was completed within 19 min on average (*SD* = 7).

Before diagnostic reasoning in the second session, participants had the opportunity to refresh their knowledge by inspecting the patterns of symptom classes for the four chemicals. These were printed on separate cards within rectangular frames but without the surrounding geometrical forms. There was also one card showing the symptom classes and the single symptoms they contained.

Diagnostic reasoning. Each diagnostic reasoning trial started with a fixation cross in the center of the screen for 1000 ms followed by a screen showing the emptied spatial array and the acoustic presentation of the symptom sequence with delays of 3000 ms between symptoms that each lasted 1000 ms. After the fourth symptom, participants indicated their diagnosis with one of the four keys already used during learning. After the response, a confidence rating was collected, which is not reported further in this paper.

In the first session, each participant, worked twice through the 40 possible combinations of chemicals with item types: once viewing the original arrangement and once viewing the flipped arrangement. Participants returning for the second session, worked through the 40 possible combinations again viewing the original arrangement. In addition, there were four training trials in each session.

The order of the 40 trials in each diagnostic reasoning section was pseudo-random and balanced across participants. For each trial, the actual sequence of symptoms was drawn randomly from the 8 or 4 possible sequences for this combination of item type and chemical in the A-role.
The diagnostic reasoning sections in the first session took approximately 75 min in total. Between sections participants took a rest for about 5 min. The second session took approximately 30 min.

Apparatus. The experimental stimuli were presented on a 19" LCD-monitor at a resolution of 800 x 600 pixels, the symptom sequences were presented through headphones. During the diagnostic reasoning phases, eye movements were monitored by a desk-mounted SMI RED eye tracker (Sensomotoric Instruments, Teltow, Germany) with a sampling rate of 60 Hz and an accuracy of approximately 0.5 degrees of visual angle. The eye tracker was calibrated before each diagnostic reasoning phase. Participants sat at a distance of approximately 60 cm to the monitor. Head movements were restrained with a chin rest.

Table 2: The ten item types, the specific symptoms that they contain, the symptom orders, and the chemicals that remain as diagnostic hypotheses after each symptom.

Item type	Specific symptoms	Order	After 1st	After 2nd	After 3rd	After 4th
1	BB	Ba_ac_Ba_ac	B,(A)	A	A	A
2	ABB	Ba_ac_Ba_Ab	B,(A)	A	A	A
3	ABB	ac_Ba_Ba_Ab	A,C	A	A	A
4	AB	ac_Ba_ac_Ab	A,C	A	A	A
5	AA	ac_ac_Ab_Ab	A,C	A,C	A	A
6	AA	Ab_Ab_ac_ac	A,(B)	A,(B)	A	A
7	ABB	Ab_Ba_ac_Ba	A,(B)	A,B	A	A
8	ABB	Ab_Ba_Ba_ac	A,(B)	A,B	A,B	A
9	ABB	Ba_Ba_Ab_ac	B,(A)	B,(A)	A,B	A
10	AABB	Ab_Ab_Ba_Ba	A,(B)	A,(B)	A,B	A,B

Results

In all non-ambiguous item types the single correct diagnosis is denoted "A" (see Table 2). In the ambiguous item type AABB_10, both "A" and "B" were correct diagnoses consistent with the pattern of symptoms. In every trial, the following spatial relations held between chemicals in the A-, B-, C-, and D-roles: the B-chemical was located horizontally next to the A-chemical, the C-chemical was diagonally opposite to the A-chemical, and the D-chemical was below or above the A-chemical (see Figure 1). When the layout on the screen was flipped for the second half of the first session, eight participants immediately noticed the flipped layout, eleven participants noticed the change at some point during the 40 trials, and the remaining seventeen apparently did not notice the change at all. Because of this variability, we focus on diagnostic reasoning sections with original layouts.

Accuracy. The mean proportion of A-diagnoses for each item type is shown in Figure 2. The item types are ordered by the combination of specific symptoms that they contain and numbered as in Table 2. Overall, accuracy was high. For the ambiguous item type AABB_10, the mean proportion of A- or B-diagnoses was .99 and .98 in the first and second session, respectively (both *SE*s .01).

The five ABB item types did not differ significantly in accuracy. The two AA item types were similar in accuracy as well. Thus, we computed mean accuracy for ABB and AA for a comparison with AB and BB item types in a repeated-measures ANOVA including session and item type (AA, AB, ABB, and BB). Accuracy was higher in the second session, $F(1, 31) = 5.49$, $MSE = 0.011$, $p = .03$, the main effect of item type was significant, $F(3, 93) = 6.43$, $MSE = 0.015$, $p = .002$, and there was no significant interaction, $F < 0.9$. In both sessions, accuracy for AA was similar to AB, higher than ABB (Cohen's d = 0.64 and 0.77

in sessions 1 and 2, respectively), but only slightly higher than BB (*d* = 0.34 and 0.27).

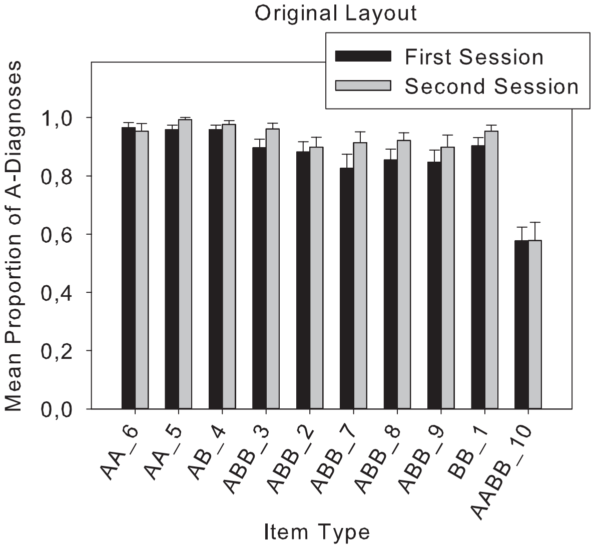

Figure 2. Mean proportions of A-diagnoses; for AABB items B-diagnoses were correct as well

Response times. Response times were measured from the onset of the fourth symptom. Median response times for A-responses were computed after trimming outliers 3 *SD* above the individual session means (2.2 % of all A-responses). As shown in Figure 3, median response time was prolonged for the ambiguous item type AABB_10 in both sessions. In the first session, median response time was also prolonged for ABB items, in which the final diagnosis was determined late with the third symptom (ABB_8 and ABB_9), and for the BB item. These three item types differed significantly from all other non-ambiguous item types with *d*s varying between 0.31 and 1.06. In the second session, these differences between the nine non-ambiguous item types were attenuated, *F*(8,232) = 2.02, *MSE* = 851180, *p* = .045.

Fixation proportions on quadrants. Gaze data were analyzed for trials with correct responses, in which the original layout had been presented. Trials with more than 40% missing gaze data were discarded (2.4 % in the first session, 4.0 % in the second session). We focus on aggregated gaze data to examine the distribution of gaze allocation between screen quadrants representing diagnostic hypotheses in response to each symptom.

The quadrants of the screen were defined as areas of interest and each trial was divided in four intervals defined by the onsets of the four symptoms and the response after the fourth symptom. For each interval in a trial, the proportions of total fixation time that fell upon the four quadrants were computed and coded as A, B, C, or D according to a quadrant's chemical's role in the respective trial. Means of these fixation proportions were computed separately by session and item type for each participant. For the ambiguous AABB item type, mean fixation proportions were computed separately for trials with A- and B-responses.

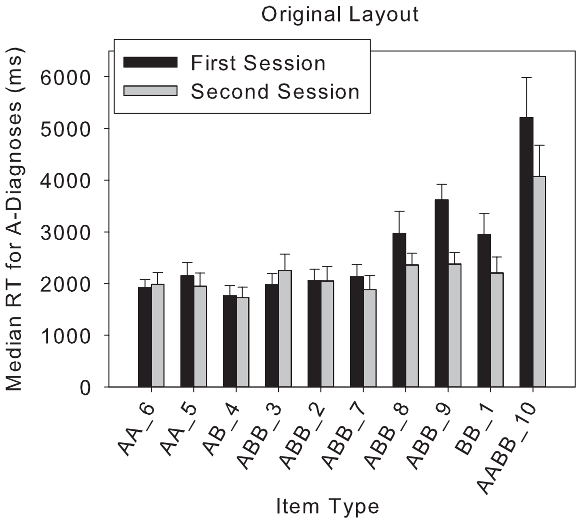

Figure 3. Median response times for A-diagnoses

Plots of mean fixation proportions across the four intervals of a symptom sequence are shown for three exemplary non-ambigous item types in Figure 4. The three non-ambiguous item types are BB_1 starting with a Ba-symptom, ABB_3 starting with an ac-symptom, and ABB_8 starting with an Ab-symptom. In addition, plots for item type AABB_10 separated for trials with A-responses and trials with B-responses are shown.

Apparent from a brief inspection, the item types induced very different fixation patterns, which were replicated for each item type with only small deviations in the second session one week later. In the interval from the onset of the first symptom to the onset of the second symptom, mean fixation proportions reflected how much the first symptom supported the individual hypotheses. With Ba as the first symptom, the largest proportion (nearly 40%) of fixations in the first interval was directed to the B-quadrant, a smaller proportion (about 20%) to the A-quadrant, and only about 10 % to the C- and D-quadrants, respectively. With Ab as the first symptom, the analogous pattern was observed except for the ambiguous AABB items that were answered with B finally. With ac as the first symptom, both A- and C-quadrants were fixated for a similar proportion (nearly 30%) and longer than the B- and D-quadrants.

As soon as an ac-symptom had occurred with either an Ab- or a Ba-symptom, the diagnosis A was determined. In the following intervals, fixation proportions for all other hypotheses dropped sharply and remained low (third and fourth intervals for BB_1 and ABB_3). Thus, later symptoms triggered only fixations to the A-quadrant, but not to quadrants with which they were associated as well and which were fixated in the first interval for the respective symptom. For example, the B-quadrant received a large proportion after Ba as the first symptom but only a small proportion after Ba as the third symptom in BB_1. Similarly, the C-quadrant received hardly any fixations after ac as a later symptom in BB_1 and ABB_8 compared to after ac as the first symptom in ABB_3.

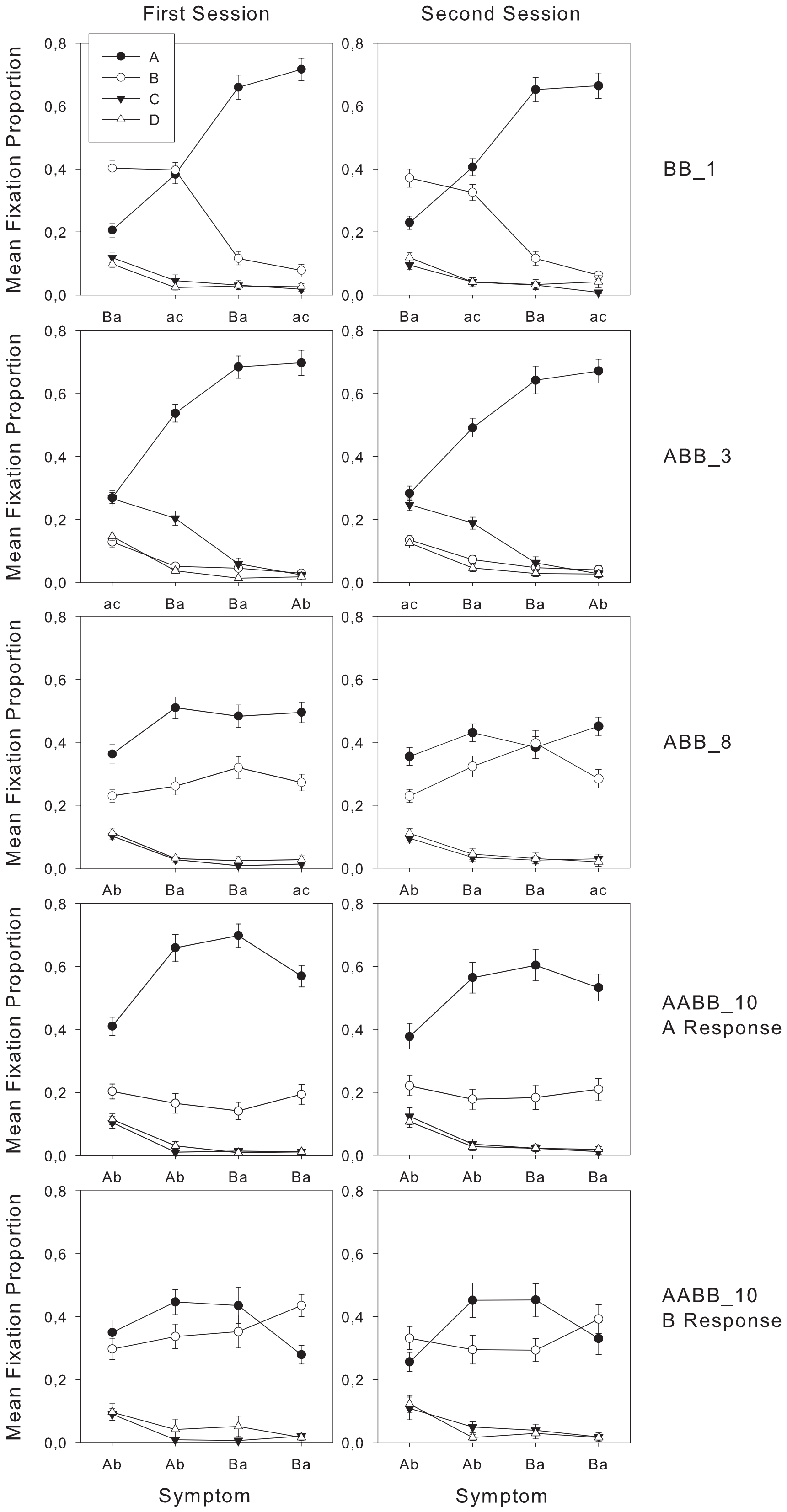

Figure 4: Mean fixation proportions on A-, B-, C, and D-quadrants in each interval of a symptom sequence. Fixation proportions for the ambiguous item type AABB_10 are shown separately for trials answered with A and those answered with B. Error bars show the standard error of the mean.

When the diagnosis was determined late with an ac-symptom as the fourth symptom (ABB_8) or not at all (AABB_10), the alternative hypothesis B received considerable fixation proportions even in the fourth interval. In AABB_10, fixation proportions were higher for B in the trials that were finally answered with B than in those finally answered with A.

Furthermore, ABB_8, in which the diagnosis was determined late, showed a strong influence of the first symptom. The A-hypothesis that was supported more strongly by the first symptom and remained consistent with the following symptoms drew a larger fixation proportion than the alternative B-hypothesis in the second interval although the alternative was equally supported. For ABB_8 in the first session, the A quadrant received a larger proportion even in the third interval although B was more strongly supported at this time. For the ambiguous item type AABB_10, in which support was equal for A and B, fixation proportions were also influenced by the first symptom but finally rose for B above A when responded with B.

Overall, fixation proportions were similar in both sessions with a tendency towards fixation proportions better reflecting support for alternative hypotheses in the second session.

Discussion

Process models of diagnostic reasoning postulate selective and changing activation of hypotheses in working memory during sequential symptom processing (Mehlhorn et al., 2011). To observe correlates of these memory dynamics, we have assigned the hypotheses to spatial locations and applied eye tracking for process tracing. Fixation proportions were influenced by memory activation because the possible causes (candidate hypotheses) had been spatially indexed during learning. Thus, process tracing was possible despite purely memory-based diagnostic reasoning. Extending previous successful applications of memory indexing (Renkewitz & Jahn, in press) and previous studies of looking-at-nothing (Richardson & Spivey, 2000), we found surprisingly stable patterns of fixation proportions one week after learning. Associations with locations and geometric figures in long-term memory strongly influenced gaze behavior similar to knowledge triggering fixations in the visual world paradigm.

The first symptom triggered fixations to its possible causes that are set up as focal hypotheses in working memory according to process models. The participants had to remember the symptom class that the first symptom belonged to and which chemicals could cause this symptom class. The result that fixation proportions were higher for more strongly supported hypotheses suggests that relative status as a focal hypothesis and the according activation level in working memory directed fixations. Hence, not just retrieval from long-term memory, but also rehearsal in working memory seems to trigger looking-at-nothing.

Excluded hypotheses did receive hardly any fixations in the subsequent intervals. Thus, fixations were not involuntarily directed towards any location associated with presented symptoms. Instead, symptoms supported the remaining focal hypothesis, and its location was fixated.

The unique information to be gained by memory indexing is clearly shown, for example, in the symptom sequence that starts with a symptom supporting strongly a hypothesis that is not the correct final diagnosis (BB_1). The time course of fixation proportions reveals the change of the initial focal hypothesis that, of course, left no trace in the final response. And for the ambiguous item, for which the final response varied, memory indexing reveals that the finally chosen hypothesis is reflected in the relative weighting of focal hypotheses right from the beginning of the ambiguous symptom sequence.

Possibly, gaze was not only a correlate of memory activation, but also actively used as a deictic pointer to support or relieve working memory. In this study, gaze as deictic pointer was particularly useful because the spatial array of hypotheses matched the arrangement of response keys. Consequently, fixation proportions may reflect both memory activation and intended memory retention. Nonetheless, memory indexing revealed the current status of hypotheses in diagnostic reasoning, which proves this method as a valuable tool for informing and testing process models of information integration in reasoning and decision making.

Acknowledgments

This research was supported by German Research Foundation (DFG) Grant JA 1761/7-1. We thank Agnes Scholz and Markus Krüger for helpful comments.

References

Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012). Data acquisition dynamics and hypothesis generation. In N. Rußwinkel, U. Drewitz, & H. van Rijn (Eds.), *Proceedings of the 11th International Conference on Cognitive Modeling* (pp. 31-36), Berlin: Universitaetsverlag der TU Berlin.

Mehlhorn, K., Taatgen, N. A., Lebiere, C., & Krems, J. F. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1391-1411.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*(1), 155-185.

Renkewitz, F. & Jahn, G. (in press). Memory Indexing: A novel method for tracing memory processes in complex cognitive tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Richardson, D. C. & Spivey, M. J. (2000). Representation, space and hollywood squares: Looking at things that aren't there anymore. *Cognition, 76,* 269–295.