Order effects in diagnostic reasoning with four candidate hypotheses

Felix G. Rebitschek (felix.rebitschek@uni-greifswald.de)
University of Greifswald, Department of Psychology

Agnes Scholz (agnes.scholz@psychologie.tu-chemnitz.de)
Chemnitz University of Technology, Department of Psychology

Franziska Bocklisch (franziska.bocklisch@psychologie.tu-chemnitz.de)
Chemnitz University of Technology, Department of Psychology

Josef F. Krems (josef.krems@phil.tu-chemnitz.de)
Chemnitz University of Technology, Department of Psychology

Georg Jahn (georg.jahn@uni-greifswald.de)
University of Greifswald, Department of Psychology

Abstract

Sequentially observed symptoms in diagnostic reasoning have to be integrated to arrive at a final diagnosis. In our experiments employing quasi-medical problems, four sequentially presented symptoms were consistent with multiple diagnostic hypotheses. We tested whether symptom order creates biases in symptom evaluation. Early symptoms induced a bias towards the initial hypothesis even though an alternative hypothesis was equally supported. In two experiments, stepwise ratings were prompted to explicitly highlight alternative hypotheses. Explicit highlighting eliminated the bias towards the initial hypothesis if only two hypotheses competed, but the bias remained if more than two hypotheses were associated with symptoms in a sequence. Our results are consistent with process models of information integration that specify how early information can frame the processing of later information. Extending previous results obtained with fewer contending hypotheses, we show limits in impartially considering more than two hypotheses.

Keywords: Order Effects; Diagnostic Reasoning; Multiple Candidate Hypotheses; Construction Integration Theory

Introduction

When humans explain observations in their environment, they apply knowledge about possible causes and the effects that each cause can bring about. Explaining observed symptoms by a diagnosis that specifies the most probable cause can be difficult for symptoms which are ambiguous and thus consistent with multiple diagnoses or inconsistent and hard to subsume under a single diagnosis (Johnson & Krems, 2001). Imagine the sequential integration of symptoms in medical diagnosis in its simplest form: You, a physician, become aware of a symptom pointing towards different possible diseases your new patient might have caught. Bit by bit you take notice of a second, third and a fourth symptom. Some of them are unspecific, others strengthen your belief in a diagnosis and weaken alternatives, but none is decisive by itself. The order in which symptoms are encountered can influence the final diagnosis because the initial diagnostic hypotheses may affect how the subsequent symptoms are weighed and integrated (e.g., Chapman, Bergus & Elstein, 1996). If symptoms are observed in sequence, the initially encountered symptoms trigger diagnostic hypotheses (Mehlhorn, Taatgen, Lebiere, & Krems, 2011).

Sequential symptom processing towards the initial hypothesis demonstrates a confirmation bias (Nickerson, 1998), which would be overcome if all alternative diagnostic hypotheses could be considered in parallel. In previous studies such impartial symptom integration sometimes succeeded for two alternative diagnoses (McKenzie, 1998), but doubts have been raised whether more than two alternative diagnoses can be considered impartially in parallel (Dougherty & Hunter, 2003).

According to normative Bayesian information integration, the order of symptom presentation should not matter. Symptom patterns equally supportive of two alternative diagnoses should produce equal proportions of these diagnoses. However, already updating of a single hypothesis can be biased by the order, in which pieces of evidence are encountered (Wang, Johnson, & Zhang, 2006). Hogarth and Einhorn (1992) specified circumstances under which normative updating of a single belief is possible and no order effects should occur (e.g. stepwise simple evaluation of short and consistent sequences). Yet, models of sequential information integration including the belief adjustment model of Hogarth and Einhorn (1992) typically postulate a disproportionately large influence of early encountered information resulting in a *primacy effect*.

When multiple diagnostic hypotheses compete, such a strong influence of early information can take the form of a bias towards the diagnosis that is most strongly supported by the first symptom. The memory dynamic resulting in a confirmation bias in sequential symptom integration can be described in terms of the construction-integration theory of text comprehension by Kintsch (1998) (Baumann, Mehlhorn, & Bocklisch, 2007). After observing the first symptom, the first construction-integration cycle results in high activation of the candidate hypothesis most strongly supported by the first symptom. Subsequent construction-integration cycles start from this state. Thus, initial symptoms and preliminary hypotheses frame the processing of later symptoms.

Recently, HyGene (Thomas, Dougherty, Spenger, & Harbison, 2008), which models memory processes in hypothesis generation for a specified set of symptoms has been extended to capture effects of sequential symptom processing in detail (Lange, Thomas, & Davelaar, 2012). The activations of symptom representations compete in working memory as symptoms are sequentially encountered, however, framing of symptom processing by preliminary hypotheses is not yet implemented in HyGene.

Our main goal was to study framing of symptom processing by preliminary hypotheses and its effects on final diagnoses in diagnostic reasoning with multiple candidate hypotheses. So, the reported experiments examine sequential diagnostic reasoning with four candidate diagnoses. Differing from previous studies (Koehler, 1991), we will not set a single hypothesis, whose probability has to be rated. Instead, participants have to choose among four candidate hypotheses. We determine effects of symptom order by evaluating proportions of final diagnoses for ambiguous symptom sequences, which equally support alternative diagnoses.

Whereas framing by preliminary hypotheses should bias towards initial hypotheses, increasing the saliency of alternative diagnoses should decrease biased symptom processing. We examine both explicit and implicit highlighting of alternative diagnoses. Alternative diagnoses were explicitly highlighted by asking participants to rate the current support for each possible diagnosis after each symptom. This procedure constantly reminds participants of the competing hypotheses.

Implicit highlighting of alternative diagnoses was attempted by presenting inconsistent symptom sequences. Symptoms inconsistent with the initial hypothesis could increase the salience of diagnostic alternatives. In terms of support theory (Tversky & Koehler, 1994), symptoms inconsistent with the focal hypothesis that strongly suggest specific alternatives could unpack the complement of the focal hypothesis into specified alternative diagnoses.

Experiments

Participants were told that they should evaluate symptoms of workers in a chemical plant to determine which of four chemicals had most likely affected each worker. In all four experiments, the diagnostic reasoning tasks referred to the same knowledge about symptoms and causes, which was acquired in a learning phase. Firstly, participants learned which symptoms belonged to which symptom classes and subsequently, with which probability each of the four chemicals caused symptoms from a symptom class (see Table 1).

Each diagnostic reasoning trial consisted of four sequentially presented symptoms after which participants had to respond with a diagnosis. Diagnostic symptoms pointing more strongly to one chemical also pointed weakly to a second chemical. For example, an “Ab”-symptom would point strongly to A and weakly to B. In addition, there were unspecific symptoms, which were caused with equal probability by all four chemicals. These were denoted with “x”. Thus, an Ab-x-Ba-x symptom sequence could induce A as the initial hypothesis but was ambiguous because it contains equal support for A and B. Such a sequence is ambiguous, but it is still consistent because all symptoms are consistent with both A and B.

In Experiment 1, we presented such ambiguous symptom sequences (AB) together with sequences that more strongly supported A (AAB) or B (ABB). The A-diagnosis was strongly supported by the first symptom, which should result in a higher proportion of A- than B-diagnoses for ambiguous AB-items (primacy order effect). In Experiment 2, participants rated the current support for each of the four alternative hypotheses after each symptom. This explicit highlighting of alternative diagnoses should reduce order effects. In Experiment 3, the procedure was identical to Experiment 1, however, inconsistent symptom sequences such as Cd-Ab-x-Ba were presented that may implicitly highlight alternative diagnoses. Finally, in Experiment 4 the inconsistent symptom sequences were presented as in Experiment 2 with ratings of all alternative diagnoses after each symptom to highlight alternative diagnoses explicitly as well.

Method

Participants Forty (28 female; mean age 23.6, SD = 2.8) undergraduate students from the University of Greifswald and 39 (30 female; mean age 22.1, SD = 2.7) undergraduate students from Chemnitz University of Technology took part in experiments 1 and 2.

Forty (32 female; mean age 21.5, SD = 2.2) undergraduate students from the University of Greifswald and 39 (26 female; mean age 23.5, SD = 3.2) undergraduate students from Chemnitz University of Technology took part in experiments 3 and 4.

Material The four alternative diagnoses were introduced as chemicals that cause symptoms when they affect workers. Each chemical caused symptoms from one symptom class (e.g. Eyes) “almost always” (see Table 1). These were symptoms with a strong causal link to the respective chemical. In addition, each chemical caused symptoms from a second symptom class “occasionally”. These were weak symptoms for the respective chemical. As shown in Table 1, there were two pairs of chemicals. Within a pair, strong and weak symptoms did overlap. For example, the symptom class “Eyes” was strong for R and weak for B, “Respiration” was strong for “B” and weak for “R”. Furthermore, there were two unspecific symptom classes that each chemical could cause “occasionally”.

Each symptom class contained two symptoms. For example, the “Eyes”-symptoms were “Tears” and “Eyelid swelling”. The symptom sequences presented in the diagnostic reasoning trials consisted of four symptoms. Table 2 shows the item types and the symptom orders that they subsume. We constructed each symptom order with each chemical in the “A”-role and each possible assignment of symptoms. For example, if “W” was the “A”-chemical, and “K” was the “B”-chemical for “Ab-x-Ba-x”, one possible symptom assignment would be “Rash-Sting-Paralysis-Swoon”.

Table 1: Domain knowledge participants had to acquire at the beginning

Group	Chem.	Strong symptoms concerning	Weak symptoms concerning	Unspecific symptoms concerning
Gasi-form	R	Eyes	Respiration	Circulatory problems, Pain
Gasi-form	B	Respiration	Eyes	Circulatory problems, Pain
Fluid	W	Skin	Neurolog.	Circulatory problems, Pain
Fluid	K	Neurolog.	Skin	Circulatory problems, Pain

Note. The original materials were in German.

In each experiment ambiguous AB-items were presented. In Experiments 1 and 2, the additional item types were AAB and ABB. AB, AAB and ABB item types contain Ab- and Ba-symptoms which both could have been caused by A or B. AB items thus are ambiguous, but they are not inconsistent. In Experiments 3 and 4, the additional item types were CAB and ABC subsuming inconsistent symptom sequences (see Table 2). Inconsistent sequences confronted participants with a “Cd”-symptom that could not be caused by A or B and that was strong for C and weak for D. CAB and ABC items are inconsistent and they are ambiguous with regard to A and B.

Procedure In all four experiments, participants were first introduced to the cover story and then acquired knowledge about the chemicals and symptom classes. They studied a table of symptom classes and symptoms and were tested until they could assign symptoms to symptom classes with 100% accuracy. Then they studied a table similar to Table 1 and were tested until they could assign the correct chemical or the correct set of chemicals to a symptom-frequency combination with 100% accuracy. Then, the diagnostic reasoning task was explained and participants were told that the symptoms to be diagnosed were caused by exactly one of the four chemicals.

In each diagnostic reasoning trial in Experiments 1 and 3, four symptoms were presented serially in the center of the screen. Each symptom was shown for 2 s followed by a fixation cross shown for 1 s. After the fourth symptom, participants were prompted to enter one of the four chemicals as their final diagnosis. Then, they were asked to rate their confidence from 1 (very unsure) to 7 (very sure). In Experiments 2 and 4, the trial procedure was similar except that after each symptom participants rated for each chemical how likely it had caused the symptoms seen so far on a scale from 0 to 100. These ratings are not reported in the present paper. We just consider the effect that this procedure had on the final diagnosis.

Table 2: Orders of symptoms related to first (A) and second (B) respectively third (C) and fourth (D) chemicals; included x stands for unspecific symptoms

Experiment	Item type	Order
1 and 2	AAB Consistent	Ab-Ab-x-Ba
		Ab-Ab-Ba-x
		Ab-x-Ab-Ba
1 and 2	ABB Consistent	Ab-x-Ba-Ba
		Ab-Ba-Ba-x
		Ab-Ba-x-Ba
1 and 2	AB Consistent	Ab-x-x-Ba
		Ab-x-Ba-x
		Ab-Ba-x-x
3 and 4	AB Consistent	x-Ab-Ba-x
		x-Ab-x-Ba
		x-x-Ab-Ba
3 and 4	CAB Inconsistent	Cd-Ab-Ba-x
		Cd-Ab-x-Ba
		Cd-x-Ab-Ba
3 and 4	ABC Inconsistent	Ab-Ba-Cd-x
		Ab-Ba-x-Cd
		Ab-x-Ba-Cd

In each experiment, each participant was presented with each of nine symptom orders (see Table 2) with each of the four chemicals in the A-role resulting in 36 trials in total. The assignment of symptoms to symptom orders was chosen randomly and the trials were presented in randomized order. In addition, four training trials were presented in each experiment.

Results

Experiments 1 and 2 Mean proportions of final diagnoses are shown in the top half of Table 3 separated by item type. In both Experiment 1 and Experiment 2, the proportion of A-diagnoses decreased from AAB to AB to ABB items reflecting the decrease in relative support of A. Within-subjects contrasts confirmed this decrease in the proportion of A-diagnoses by significant linear trends, $F(1, 39) = 230.84$, $p < .001$, $\eta^2 = .86$, and $F(1, 38) = 474.09$, $p < .001$, $\eta^2 = .93$, respectively.

Focusing on ambiguous AB-items, equal proportions of A- and B-diagnoses each about 50% would be expected normatively. In Experiment 1, there was a clear bias towards A-diagnoses compared with B-diagnoses for AB-items, confirmed by a paired t-test, $t(39) = 4.54$, $p < .001$, $d = 0.72$. Thus, we obtained a clear primacy order effect for ambiguous AB-items in Experiment 1, whereas in Experiment 2, the proportion of A-diagnoses did not deviate from the proportion of B-diagnoses, $t(38) = -0.10$, $p = .924$. Mean confidence ratings are shown in the top half of Table 4. Space limitations preclude a detailed analysis but it is apparent that confidence was reduced for the ambiguous AB-items.

Table 3: Means of proportions of diagnoses

Exp.	Item type	A (SD)	B (SD)	C (SD)	D (SD)
1	AAB	.91 (.15)	.09 (.15)		
	AB	.65 (.20)	.35 (.20)		
	ABB	.14 (.21)	.86 (.21)		
2	AAB	.83 (.16)	.17 (.16)		
	AB	.50 (.21)	.50 (.21)		
	ABB	.09 (.10)	.91 (.10)		
3	AB	.62 (.26)	.28 (.19)		
	ABC	.60 (.25)	.22 (.17)	.12 (.12)	.06 (.10)
	CAB	.48 (.25)	.20 (.19)	.26 (.20)	.06 (.08)
4	AB	.50 (.21)	.44 (.19)		
	ABC	.45 (.22)	.29 (.15)	.21 (.18)	.06 (.07)
	CAB	.42 (.22)	.39 (.17)	.14 (.12)	.05 (.08)

Note. Proportions for AB items in Experiments 3 and 4 do not sum to 1 because proportions of wrong C and D diagnoses are omitted from the table.

Experiments 3 and 4 In Experiments 3 and 4, the ambiguous item type AB and inconsistent item types ABC and CAB were presented. Note that with respect to A and B, the item types ABC and CAB contain equal support as well. Thus, normatively equal proportions of A- and B-diagnoses should be elicited by all three item types. Mean proportions of final diagnoses are shown in the bottom half of Table 3.

For AB-items the results are similar to Experiments 1 and 2. Without explicit highlighting of diagnostic alternatives in Experiment 3, there was a clear bias towards A-diagnoses compared with B-diagnoses (primacy order effect) confirmed by a paired t-test, $t(39) = 4.94$, $p < .001$, $d = 0.78$, whereas with explicit highlighting in Experiment 4, the proportion of A-diagnoses did not deviate from the proportion of B-diagnoses, $t(38) = 1.02$, $p = .315$.

For ABC and CAB items, the leading strong symptom took effect in Experiment 3. The proportion of A-diagnoses was higher for ABC than for CAB items, $t(39) = 3.42$, $p = .001$, $d = 0.47$, and the proportion of C-diagnoses was higher for CAB than for ABC items, $t(39) = 4.12$, $p < .001$, $d = 0.84$. Nonetheless, A-diagnoses were more frequent than

C-diagnoses for both item types reflecting the superior support by a strong and a weak symptom as opposed to a single strong symptom. Despite equal support for A and B, A-diagnoses were also more frequent than B-diagnoses for both item types suggesting a primacy order effect in ABC and a similar order effect in CAB, in which the Ab-symptom can frame the integration of the later Ba symptom.

In Experiment 4, there was hardly any effect of the leading symptom on A- and C-diagnoses for ABC and CAB. The proportion of A-diagnoses was comparable for ABC and CAB, *t*(38) = 0.65; *p* = .520, and the proportion of C-diagnoses was even lower for CAB than for ABC items, *t*(38) = -2.02; *p* = .050; *d* = -0.46.

Table 4: Means of confidence ratings of related diagnoses

E.	Item type	A (SD)	B (SD)	C (SD)	D (SD)
1	AAB	5.66 (0.96)	4.08 (1.80)		
	AB	3.77 (1.19)	3.69 (1.29)		
	ABB	4.63 (1.11)	5.47 (1.14)		
2	AAB	5.10 (1.07)	4.06 (1.34)		
	AB	3.49 (0.94)	3.38 (1.06)		
	ABB	3.54 (1.16)	5.16 (0.95)		
3	AB	4.05 (1.45)	3.36 (1.52)		
	ABC	3.52 (1.45)	3.08 (1.40)	2.83 (1.51)	2.33 (1.35)
	CAB	3.41 (1.38)	3.23 (1.41)	3.06 (1.52)	2.29 (1.24)
4	AB	4.00 (1.75)	3.80 (1.43)		
	ABC	3.64 (1.66)	3.32 (1.38)	3.32 (1.81)	3.08 (1.60)
	CAB	3.92 (1.57)	3.39 (1.47)	2.94 (1.47)	3.25 (1.75)

On top of a decreased primacy order effect, which increased A-diagnoses compared to B diagnoses for ABC items, there was a stronger influence of the last diagnostic symptom than in Experiment 3. The proportion of B-diagnoses was higher for CAB than ABC, *t*(38) = -3.02, *p* = .004, *d* = -0.63, and the proportion of C-diagnoses was higher for ABC than CAB. B-proportions in ABC could be reduced simply because of a primacy order effect favoring A. The difference in C-proportions, however, is not open to such an alternative explanation and suggests an increased influence of the last diagnostic symptom in Experiment 4.

Mean confidence ratings are presented in the bottom half of Table 4 and show that inconsistent symptom sequences reduced confidence compared to ambiguous AB-items.

Discussion

We have investigated effects of symptom order in a diagnostic reasoning task with four candidate hypotheses. Ambiguous symptom sequences (AB-items) equally supporting two alternative diagnoses revealed a clear primacy order effect if participants only responded with a final diagnosis (Experiments 1 and 3). Consistent with the processing assumptions of construction-integration theory, the initial hypothesis suggested by the first symptom framed the integration of subsequent symptoms. An equally supported alternative diagnosis was therefore chosen less often. This order effect is in line with the notion that alternative hypotheses are typically not considered impartially in parallel. Instead, symptom processing proceeds with respect to a focal hypothesis if subsequent symptoms are consistent.

Inconsistent symptoms were not an effective means to highlight alternative diagnoses in Experiment 3. There was still a considerable primacy order effect favoring A over B in ABC- and CAB-items despite equal support. Explicit highlighting of alternative diagnoses, however, was effective. In Experiments 2 and 4, participants rated the current likelihood of each candidate hypothesis after each symptom and thus were led to consider alternative diagnoses. This eliminated the primacy order effect for AB-items. As noted in previous studies, impartial consideration of two alternative diagnoses in parallel can succeed under favorable conditions.

Yet, eliciting ratings of all alternative diagnoses after each symptom did not eliminate the primacy order effect if an inconsistent symptom pattern added a third and presumably even a fourth candidate hypothesis to the set of contenders (ABC- and CAB-items in Experiment 4). In these cases, we did not only observe an advantage for the alternative that was supported by a strong symptom before its equally supported rival (A before B), but also an effect of the last strong symptom. Forcing the participants to consider the current support for all alternatives after this last strong symptom before the final diagnosis resulted in a recency effect. The proportion of final diagnoses was increased for the alternative most strongly supported by the last strong symptom for the inconsistent items ABC and CAB in Experiment 4. Our results are consistent with process models of information integration that specify how early information can frame the processing of later information (Kintsch, 1998, Baumann et al., 2007). They are also consistent with descriptive models predicting order effects in belief updating, hypothesis testing, classification, and judgment and decision making (Hogarth & Einhorn 1992, Koehler, White, & Grondin, 2003) and are a further instance of confirmation bias (Nickerson, 1998). Studies with more than two contending alternatives are rare. Here, we have shown that the number of relevant contenders matters. The primacy order effect was overcome with two competing alternatives by explicit highlighting. With inconsistent items, more than two hypotheses had to be considered. Constrained by working-memory capacity unpacking of the set of alternatives was incomplete and rather the most likely alternatives were taken into consideration (Dougherty & Hunter, 2003).

Our results may not generalize to instances of diagnostic reasoning in everyday life, in which symptoms can be evaluated more thoroughly without time pressure and search for further information is possible. However, there are situations, in which incoming information has to be processed quickly. For example, physicians evaluating case histories are influenced by early emerging hypotheses (e.g., Kostopoulou, Mousoulis, & Delaney, 2009). The difficulties in considering more than two contenders impartially in the present experiments clearly illustrate the limits in diagnostic reasoning with multiple alternative diagnoses in similar situations.

Acknowledgments

This research was supported by German Research Foundation (DFG) Grants JA 1761/7-1 and KR 1057/17-1.

References

Baumann, M., Mehlhorn, K., & Bocklisch, F. (2007). The activation of hypotheses during abductive reasoning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Chapman, G. P., Bergus, G. R., & Elstein, A. S. (1996). Order of information affects clinical judgment. *Journal of Behavioral Decision Making, 9*, 201-211.

Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, *113*, 263-282.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Johnson, T. & Krems, J. F. (2001). Use of Current Explanations in Multicausal Abductive Reasoning. *Cognitive Science, 25,* 903-939.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, *110*, 499-519.

Koehler, D. J., White, C. M., & Grondin, R. (2003). An evidential support accumulation model of subjective probability. *Cognitive Psychology, 46,* 152-197.

Kostopoulou, O., Mousoulis, C., Delaney, B. C. (2009). Information search and information distortion in the diagnosis of an ambiguous presentation. *Judgment and Decision Making, 4(5),* 408-418.

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.

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 771-792.

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, & Cognition*, *37*, 1391-1411.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175-220.

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

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. Psychological Review, *101*, 547-567.

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