Is Human Reasoning about Nonmonotonic Conditionals Probabilistically Coherent?

Niki Pfeifer[*]
Department of Psychology
University of Salzburg
niki.pfeifer@sbg.ac.at

Gernot D. Kleiter[*]
Department of Psychology
University of Salzburg
gernot.kleiter@sbg.ac.at

Abstract

Nonmonotonic conditionals ($A \mathrel{|\!\sim} B$) are formalizations of common sense expressions of the form "if A, normally B". The nonmonotonic conditional is interpreted by a "high" coherent conditional probability, $P(B|A) > .5$. Two important properties are closely related to the nonmonotonic conditional: First, $A \mathrel{|\!\sim} B$ allows for *exceptions.* Second, the rules of the nonmonotonic SYSTEM P guiding $A \mathrel{|\!\sim} B$ allow for *withdrawing conclusions* in the light of new premises.

This study reports a series of three experiments on reasoning with inference rules about nonmonotonic conditionals in the framework of *coherence.* We investigated the CUT, and the RIGHT WEAKENING rule of SYSTEM P. As a critical condition, we investigated basic monotonic properties of classical (monotone) logic, namely MONOTONICITY, TRANSITIVITY, and CONTRAPOSITION. The results suggest that people reason nonmonotonically rather than monotonically. We propose nonmonotonic reasoning as a competence model of human reasoning.

1 Introduction

Traditionally, psychological theories on deductive reasoning evaluate the quality of human reasoning by *classical logic* as the normative standard of reference [4, 18, 10]. Classical logic has been proposed by [12] as the surest guide towards a competence model for the psychology of reasoning. As compared to the actual reasoning performance (which is biased by memory limitations, limited information processing resources, shifts of attention, etc.), competence refers to ideal reasoning performance. This distinction, here in the domain of reasoning, is analog to the performance/competence distinction in the domain of language [5].

The *psychological plausibility* of classical logic both, as the normative standard of reference and as a competence model, is questionable on *a priori* grounds

[*]Supported by the AKTION (45p16) project, Austrian-Czech exchange program.

	Inference rule	*Propagation rule*	#
RW:	*from* $A \mid\!\sim_x B$ *and* $\models B \rightarrow C$ *infer* $A \mid\!\sim_z C$	$x \leq z \leq 1$	1,3
CUT:	*from* $A \mid\!\sim_x B$ *and* $A \wedge B \mid\!\sim_y C$ *infer* $A \mid\!\sim_z C$	$xy \leq z \leq 1-x+xy$	1,3
TR:	*from* $A \mid\!\sim_x B$ *and* $B \mid\!\sim_y C$ *infer* $A \mid\!\sim_z C$	$0 \leq z \leq 1$	2,3
MB:	*from* $B \mid\!\sim_x C$ *and* $A \mid\!\sim_y B$ *infer* $A \mid\!\sim_z C$	$0 \leq z \leq 1$	2
AN-C:	*from* $A \mid\!\sim_x B$ *infer* $\neg B \mid\!\sim_z \neg A$	$0 \leq z \leq 1$	1,2
NA-C:	*from* $\neg B \mid\!\sim_x \neg A$ *infer* $A \mid\!\sim_z B$	$0 \leq z \leq 1$	1,2

Table 1: Inference rules and propagation rules for the probabilities in the premises ($0 \leq x \leq 1$, $0 \leq y \leq 1$) to the coherent probability (z) of the conclusion. The rules above the line (RW = RIGHT WEAKENING) are valid in SYSTEM P and probabilistically informative. The arguments below the line (TR=TRANSITIVITY, MB=MODUS BARBARA, AN-C=AN-CONTRAPOSITION, NA-C= NA-CONTRAPOSITION) are neither valid in SYSTEM P nor probabilistically informative (i.e., only the unit interval, $[0, 1]$, can be inferred). $\models X \rightarrow Y$ means $X \rightarrow Y$ is a tautology. $\wedge$ ("and"), and $\neg$ ("not") are defined as usual in classical logic. Column # refers to the present experiments.

for several reasons. The two most important reasons are the monotonicity principle and the definition of the "if—then" relation. The monotonicity principle inherent in classical logic does not allow for retracting conclusions in the light of new evidence, while the studies on the suppression of conditional inferences impressively show that subjects are willing to doubt premises or to withdraw conclusions under certain circumstances [17, 19]. The "if—then" relations as defined in classical logic do not allow for dealing with exceptions and uncertainty, while exceptions and uncertainty can almost always be present in common sense reasoning. Thus, e.g., [11] observed that a high percentage (90%) of the subjects attached a probabilistic interpretation to indicative conditionals. Psychological literature reports data that indicate that subjects interpret common sense conditionals as conditional probabilities [7, 13, 16].

The present paper proposes to use a probabilistic interpretation of *nonmonotonic reasoning* [8] as the normative standard of reference for investigating human reasoning, rather than classical logic. This does not mean to abandon logic from psychology, rather then, to enrich the traditional normative standard of reference by nonmonotonic tools to handle uncertainty and the retraction of conclusions. The psychological plausibility of the proposed normative standard of reference is supported in previous studies [14, 15] and will be supported by three experiments reported in the following sections.

2 Experiment 1

Method and Procedure In Experiment 1 we investigated the CUT and the RIGHT WEAKENING rule of SYSTEM P, and two forms of CONTRAPOSITION which are not valid in SYSTEM P (cf. Table 1). CUT and RIGHT WEAKENING are the nonmonotonic versions of TRANSITIVITY, and are therefore of special interest.

Forty students of the University of Salzburg participated in the study. No students with special logical or mathematical education were included. Each subject was tested individually and received a booklet containing a general introduction, one example explaining the response modality with point percentages, and one example explaining the response modality with interval percentages. Three target tasks were presented on separate pages. Eleven additional target tasks were presented in tabular form (see Table 3 for the values of the premises of the tasks in tabular form). Twenty subjects were assigned to the CUT condition and twenty subjects were assigned to the RIGHT WEAKENING condition. In the CUT condition subjects were asked to imagine:

> *Exactly 89%* of the cars on a *big parking lot* are *blue.*
> *Exactly 91%* of *blue* cars that are on the *big parking lot* have *grey tyre-caps.*
> Imagine all the cars that are on the *big parking lot.* How many of these cars have *grey tyre-caps*?

The subjects were free to respond either by point percentages or interval percentages. The first three tasks were presented on separate pages and the eleven subsequent tasks in tabular form [14, 15]. After the CUT tasks the NA-CONTRAPOSITION task was presented (Negated premise, Affirmative conclusion):

> *Exactly 93%* of the cars that are **not** on a *big parking lot* are **not** *red.*
> Imagine all the cars that are *red.* How many of the *red* cars are on the *big parking lot*?

The RIGHT WEAKENING condition was in parallel to the CUT condition with the following two exceptions. First, the second premises of the CUT tasks were replaced by "*All blue* cars have *grey tyre-caps.*". Second, after the fourteen RIGHT WEAKENING tasks the AN-CONTRAPOSITION task was presented (Affirmative premise, Negated conclusion; see Table 1).

Results and Discussion At the end of each session the subjects rated the overall comprehensibility, how sure they were that their answers were correct, and the overall difficulty of the tasks. In both conditions, the task comprehensibility was judged to be "good", and that the subjects were intermediatly confident that their solutions are correct. The RIGHT WEAKENING tasks were judged to be easier than the CUT tasks. Comprehensibility, certainty and difficulty in both versions of the CONTRAPOSITION tasks were comparable: the task comprehensibility was judged to be between "good" and "intermediate", the subjects were intermediatly confident in the correctness of their solutions and the difficulty was intermediate.

Table 2 presents the mean upper and lower bound responses of the CUT condition and the RIGHT WEAKENING condition. 25.71% of the subjects in the CUT condition responded by point values on the average ($M = 5.14, SD = 4.75$, 14 tasks). 41.43% of the subjects in the RIGHT WEAKENING condition responded by point values on the average ($M = 8.29, SD = 3.17$, 14 tasks).

Table 3 presents the frequencies of the six possible categories of interval responses of the CUT and the RIGHT WEAKENING tasks. In the CUT condition, 55.35% of the subjects responded by coherent intervals on the average ($n = 14$

tasks). The frequency of coherent responses clearly exceeds the guessing level of 16.67% (if subjects responses were equally distributed over all six possible response categories).

In the RIGHT WEAKENING tasks 87.15% of the subjects responded coherent intervals. Practically all subjects clearly endorse the RIGHT WEAKENING rule.

To estimate the reliability of the data we calculated correlations between the responses in those tasks which have normatively equivalent bounds in the conclusions. The propagation rule of the normative lower bound of the CUT rule is commutative ($xy = yx$, see Table 1). There are two pairs of tasks in which the percentages in the premises are interchanged (tasks $B2$ and $B5$, and $B9$ and $B11$, respectively). The correlations between the lower bound responses of these pairs of tasks are $r = 0.93$ ($B2$ and $B5$, $n = 20$) and $r = 0.93$ ($B9$ and $B11$, $n = 20$). Tasks $B2$ and $B11$ have practically the same normative upper bounds, 82.36 and 81.52, respectively. The correlation between these tasks was $r = 0.95$ ($n = 20$). This indicates a high reliability of the data.

In the NA-CONTRAPOSITION task, only one subject gave a lower bound greater than 7 ($M = 6.50, SD = 20.58$). The mean upper bound was 62.39 ($SD = 46.43$). Eight subjects gave an upper bound smaller than 93 (all of these eight subjects gave an upper bound equal to 7). In the AN-CONTRAPOSITION task, 11 subjects (i.e, 55%) gave interval responses with both, lower bounds ≤ 7 and upper bounds ≥ 93.

A similar result was observed in the AN-CONTRAPOSITION. More than half of the subjects responded by lower bounds ≤ 7.00 ($M = 30.20, SD = 42.39$). More than half of the subjects responded by upper bounds ≥ 93.00 ($M = 74.70, SD = 39.09$). 9 subjects (i.e, 45%) gave interval responses with both, lower bounds ≤ 7 and upper bounds ≥ 93.

In sum, the data of Experiment 1 clearly endorse the RIGHT WEAKENING rule that is valid in SYSTEM P and clearly do not endorse the monotonic CONTRAPOSITION argument forms that are not valid in SYSTEM P. The endorsement rate of the RIGHT WEAKENING rule is very similar to that of the LEFT LOGICAL EQUIVALENCE rule of SYSTEM P [15]. Furthermore, since these rules are naturally drawn by the subjects they are attractive for mental rule theories. Mental rule theories postulate that the human inference engine is driven by basic formal rules like the MODUS PONENS [4, 18]. The majority of the intervals responses in the CUT tasks is coherent, which is comparable to the results of the CAUTIOUS MONOTONICITY tasks [14] and the the AND tasks [15]. Subjects endorse the nonmonotonic rules and do not endorse the monotonic argument forms.

The next section (Experiment 2) investigates two versions of the TRANSITIVITY argument form, namely TRANSITIVITY (or HYPOTHETICAL SYLLOGISM) and MODUS BARBARA, an argument form well known in the Aristotelian syllogistics. Both are monotonic inference argument forms and probabilistically not informative. Experiment 2 investigates whether human subjects understand to the probabilistic non-informativeness of these inference argument forms and whether the order of the premises does influence human reasoning. Furthermore, we tried to replicate our results of the CONTRAPOSITION task.

		A1L	*A1U*	*A2L*	*A2U*	*A3L*	*A3U*	*B1L*
CUT	*Mean:*	75.06 (80.99)	93.65 (89.99)	59.05 (62.37)	79.96 (99.37)	61.55 (62.72)	90.72 (64.72)	42.80 (42.00)
	SD:	21.31	5.07	13.71	17.64	16.87	13.27	11.99
RW	*Mean:*	80.10 (89.00)	95.60 (100)	94.05 (99.00)	99.55 (100)	60.80 (64.00)	83.80 (100)	59.50 (60.00)
	SD:	27.39	5.53	22.14	0.51	14.31	18.38	25.64
		B1U	*B2L*	*B2U*	*B3L*	*B3U*	*B4L*	*B4U*
CUT	*Mean:*	74.90 (72.00)	47.79 (45.36)	70.37 (82.36)	49.90 (49.50)	72.22 (94.50)	53.08 (55.44)	87.45(56.44)
	SD:	22.28	13.68	19.40	16.28	20.49	13.66	18.92
RW	*Mean:*	88.00 (100)	50.75 (63.00)	80.55 (100)	46.75 (55.00)	82.00 (100)	84.15 (56.00)	99.60 (100)
	SD:	15.08	25.17	25.17	20.15	22.62	36.27	0.50
		B5L	*B5U*	*B6L*	*B6U*	*B7L*	*B7U*	*B8L*
CUT	*Mean:*	45.20 (45.36)	74.18 (73.36)	42.85 (36.00)	70.75 (76.00)	95.05 (100)	100 (100)	41.70 (26.01)
	SD:	12.45	20.88	18.84	23.09	22.14	0.00	28.08
RW	*Mean:*	61.20 (63.00)	88.80 (100)	51.00 (60.00)	84.00 (100)	95.00 (100)	100 (100)	43.35 (51.00)
	SD:	26.38	14.07	21.98	20.10	22.36	0.00	18.68
		B8U	*B9L*	*B9U*	*B10L*	*B10U*	*B11L*	*B11U*
CUT	*Mean:*	68.16 (75.01)	68.50 (69.52)	84.41 (90.52)	43.92 (43.12)	75.36 (66.12)	67.06 (69.52)	86.30 (81.52)
	SD:	27.58	16.86	12.36	13.70	22.52	16.77	11.27
RW	*Mean:*	80.40 (100)	67.15 (79.00)	91.60 (100)	65.45 (56.00)	90.80 (100)	74.80 (79.00)	95.20 (100)
	SD:	24.63	28.94	10.56	28.21	11.56	32.24	6.03

Table 2: Subjects mean lower and upper percentage responses in the CUT ($n = 20$) and in the RIGHT WEAKENING ($n = 20$) condition of Experiment 1. *L* and *U* designate subjects lower and upper bound responses, respectively. The normative lower and upper bounds are given in round parentheses. For the percentages given in the premises see the respective values in the square brackets of Table 3. RW = RIGHT WEAKENING condition.

Frequencies of interval responses of the CUT *condition (n=20)*

	Task A1 (80.99-89.99)			*A2 (62.37-99.37)*			*A3 (62.72-64.72)*			*B1 (42.00-72.00)*			*B2 (45.36-82.36)*		
	[89 and 91]			*[99 and 63]*			*[64 and 98]*			*[60 and 70]*			*[72 and 63]*		
UA	1	6	0	0	0	0	2	13	2	1	11	0	1	4	0
UW	2	**11**	-	1	**19**	-	0	**3**	-	0	**8**	-	1	**14**	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
	Task B3 (49.50-94.50)			*B4 (55.44-56.44)*			*B5 (45.36-73.36)*			*B6 (36.00-76.00)*			*B7 (100-100)*		
	[90 and 55]			*[56 and 99]*			*[63 and 72]*			*[60 and 60]*			*[100 and 100]*		
UA	1	4	1	3	11	1	3	8	0	1	3	1	-	-	-
UW	0	**14**	-	1	**4**	-	1	**8**	-	0	**15**	-	1	**19**	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
	Task B8 (26.01-75.01)			*B9 (69.52-90.52)*			*B10 (43.12-66.12)*			*B11 (69.52-81.52)*			***Task***		
	[51 and 51]			*[88 and 79]*			*[56 and 77]*			*[79 and 88]*			*[**Pr.1** and **Pr.2**]*		
UA	1	2	3	2	4	0	2	10	1	2	9	1	***a***	***b***	***c***
UW	0	**14**	-	0	**13**	-	0	**6**	-	1	**7**	-	***d***	***e***	-
UB	0	-	-	1	-	-	1	-	-	0	-	-	***f***	-	-
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*

Frequencies of interval responses of the RIGHT WEAKENING *condition (n=20)*

	Task A1 (89.00-100)			*Task A2 (99.00-100)*			*Task A3 (64.00-100)*			*Task B1 (60.00-100)*			*Task B2 (63.00-100)*		
UW	2	**18**	-	1	**19**	-	3	**17**	-	3	**16**	-	3	**16**	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	1	-	-
	Task B3 (55.00-100)			*Task B4 (56.00-100)*			*Task B5 (63.00-100)*			*Task B6 (60.00-100)*			*Task B7 (100-100)*		
UW	3	**17**	-	3	**17**	-	3	**17**	-	3	**17**	-	3	**17**	-
UB	0	-	-	0	-	-	0	-	-	0	-	-	0	-	-
	Task B8 (51.00-100)			*Task B9 (79.00-100)*			*Task B10 (56.00-100)*			*Task B11 (79.00-100)*					
UW	1	**19**	-	3	**17**	-	3	**17**	-	3	**17**	-			
UB	0	-	-	0	-	-	0	-	-	0	-	-			
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*

Table 3: Frequencies of the interval responses in the CUT and the RIGHT WEAKENING condition of Experiment 1. The percentages presented in the premises are in the square brackets and the normative intervals are in the round parentheses. *UA*: the subjects' upper bound response is *above* the normative upper bound, *UW*: upper bound response is *within* the normative interval, *UB*: upper bound response is *below* the normative lower bound; *LA*, *LW*, and *LB*: same for the subjects' lower bound responses. The six possible interval response categories are: ***a***: too wide interval responses, ***b***: lower bound responses coherent, ***c***: both bound responses above, ***d***: upper bound responses coherent, ***e***: both bound responses coherently within ±5% (bold), ***f***: both bound responses below the normative lower bounds. In the RIGHT WEAKENING task, the percentages presented in the premises are identical to the normative lower bounds. Since the normative upper bounds are 100 in the RIGHT WEAKENING task, no violation of the upper bounds is possible.

3 Experiment 2

The method and procedure of Experiment 2 are essentially the same as in Experiment 1. Twenty subjects were assigned to the TRANSITIVITY condition and twenty subjects were assigned to the MODUS BARBARA condition. In the TRANSITIVITY condition subjects were asked to imagine the following situation:

> *Exactly 89%* of the cars on a *big parking lot* are *blue*.
> *Exactly 91%* of the *blue* cars have *grey tyre-caps*.
> Imagine all the cars that are on the *big parking lot*. How many of these cars have *grey tyre-caps*?

The rest of the instruction and procedure was in parallel to the previous experiments. After the TRANSITIVITY tasks we presented the NA-CONTRAPOSITION task.

The MODUS BARBARA condition was formulated as the TRANSITIVITY condition with two differences. First, the order of the premises of the first fourteen tasks was reversed. Second, the NA-CONTRAPOSITION task was replaced by the AN-CONTRAPOSITION task, as in Experiment 1.

Results and Discussion Table 4 presents the mean upper and lower bound responses in the TRANSITIVITY and in the MODUS BARBARA condition. 45.71% of the subjects in the TRANSITIVITY tasks responded by point values on the average ($M = 9.14, SD = 2.98$, 14 tasks, $n = 20$). 63.93% of the subjects in the MODUS BARBARA tasks responded by point values on the average ($M = 12.79, SD = 2.22$, 14 tasks, $n = 20$). The mean values of the upper and lower bound responses, and the high percentage of point value responses indicate that the subjects do not understand the probabilistic non-informativeness of these monotonic argument forms.

The mean lower bound responses of the AN-CONTRAPOSITION task was 11.30 ($SD = 28.11$) and the mean upper bound responses was 71.70 ($SD = 42.84$). Eighteen of the twenty subjects responded by a lower bound ≤ 7. Fourteen subjects responded by an upper bound ≥ 93. Twelve subjects responded by intervals which have both, lower bounds ≤ 7 and upper bounds ≥ 93. Thus, 60% of the subjects understand that the AN-CONTRAPOSITION argument form is probabilistically not informative.

In the NA-CONTRAPOSITION task the mean lower bound responses was 38.70 ($SD = 41.52$) and the mean upper bound responses was 71.60 ($SD = 39.90$). Eleven of the twenty subjects responded by a lower bound ≤ 7. Thirteen subjects responded by an upper bound ≥ 93. Seven subjects responded by intervals which have both, lower bounds ≤ 7 and upper bounds ≥ 93. Thus, 35% of the subjects understand that the NA-CONTRAPOSITION argument form is probabilistically not informative.

Compared with Experiment 1, the percentage of subjects understanding the probabilistic non-informativeness of the CONTRAPOSITION varies from 35% to 60%. Practically all of the subjects who did not infer wide intervals responded

		A1L	*A1U*	*A2L*	*A2U*	*A3L*	*A3U*	*B1L*
TRANS	*Mean:*	71.20	83.60	60.27	66.71	58.41	79.34	41.15
	SD:	25.17	20.30	6.11	15.13	15.10	20.47	18.39
MB	*Mean:*	71.45	84.95	61.05	71.95	64.79	72.00	44.15
	SD:	27.10	10.88	16.93	16.69	26.08	19.93	16.72
		B1U	*B2L*	*B2U*	*B3L*	*B3U*	*B4L*	*B4U*
TRANS	*Mean:*	67.00	45.14	63.77	45.92	58.45	50.06	78.26
	SD:	23.29	15.94	21.15	11.59	15.76	20.14	22.54
MB	*Mean:*	55.75	52.00	63.65	61.05	70.10	48.85	63.80
	SD:	22.47	20.08	24.09	24.91	23.96	16.23	19.00
		B5L	*B5U*	*B6L*	*B6U*	*B7L*	*B7U*	*B8L*
TRANS	*Mean:*	43.06	67.96	42.25	65.95	95.00	100.00	33.25
	SD:	19.24	21.70	19.31	22.80	22.36	0.00	20.00
MB	*Mean:*	47.95	60.65	49.15	58.75	95.00	100.00	39.90
	SD:	16.10	23.84	25.29	28.91	22.36	0.00	21.02
		B8U	*B9L*	*B9U*	*B10L*	*B10U*	*B11L*	*B11U*
TRANS	*Mean:*	60.35	65.63	78.25	40.68	66.97	61.23	79.45
	SD:	25.56	21.39	18.26	17.15	23.95	25.57	20.07
MB	*Mean:*	53.80	71.50	79.90	42.25	57.20	66.45	76.75
	SD:	31.75	20.50	15.88	15.68	26.43	18.18	15.42

Table 4: Subjects mean lower and upper percentage responses in the TRANSITIVITY (TRANS, $n = 20$) and in the MODUS BARBARA (MB, $n = 20$) condition of Experiment 2.

either by lower and upper bounds that are close to zero, or by lower and upper bounds that are close to one hundred.

Adams [1] stressed the probabilistic invalidity of the TRANSITIVITY and suggested to interpret TRANSITIVITY in common sense reasoning as CUT. Adams' suggestion can be justified by conversational implicatures [9]. If a speaker first utters a premise of the form $A \mid\!\sim_x B$ and then utters as the second premise $B \mid\!\sim_y C$, the speaker actually means by the second premise a sentence of the form $(\underline{A\ and}\ B) \mid\!\sim_y C$. The speaker does not mention "*A and*" to the addressat because *A and* is already conversationally implied and "clear" from the context.

This interpretation explains why subjects do not infer wide intervals close to the unit interval. If the conversational implicature hypothesis is correct, then the subjects actually interpret both forms of the TRANSITIVITY tasks as instances of the CUT rule. We analyzed the data of the TRANSITIVITY and MODUS BARBARA tasks as if they are instances of the CUT rules. Then, 62.14% of the subjects in the TRANSITIVITY tasks gave coherent interval responses on the average ($n = 14$ tasks). 50.00% of the subjects in the MODUS BARBARA tasks gave coherent interval responses on the average ($n = 14$ tasks). These results are similar to those of the original CUT condition in Experiment 1 (55.35%).

The next section (Experiment 3) investigates the CUT and the TRANSITIVITY argument forms by an improved cover story. We tried to block conversational implicatures by explicitly mentioning the universe of discourse. In the RIGHT

WEAKENING condition of Experiment 1 we used an contingent statement instead of a logical tautology. In Experiment 3 we investigated the RIGHT WEAKENING rule reformulated as a logical tautology.

4 Experiment 3

Method and Procedure is the same as in Experiment 1. Twenty subjects were assigned to the CUT condition and twenty subjects were assigned to the TRANSITIVITY condition. In the CUT condition subjects were asked to imagine a ski-resort around Christmas time, where cable cars transport the skiers up to the mountains every hour, and that

> *Exactly* 99% of all the skiers in the *Galzig-cable car* have a *blue suite.*
> *Exactly* 63% of all the skiers in the *Galzig-cable car* that have
> a *blue suite* are *ski instructors.*
> Imagine all the skiers that are in the *Galzig-cable car.* Please try to determine how many percent of the skiers in the *Galzig-cable car* are *ski instructors*?

The TRANSITIVITY condition was identical to the CUT condition with the exception that the second premise was replaced by "*Exactly* 63% of all the skiers in the Arlberg ski-resort that have a *blue suite* are *ski instructors.*". Here, "the skiers in the Arlberg ski-resort" makes the universe of discourse explicit and does not denote the subset "Galzig cable car" as in the CUT condition.

In both condition, after the tasks just described, we presented the following the RIGHT WEAKENING task,

> *All blue Volkswagen* are *blue cars.*
> *Exactly 70%* of all the *inhabitants* of a small town own a *blue Volkswagen.*
> Please imagine now all the inhabitants of this small town. Please try now
> to determine how many percent of the *inhabitants* own a *blue car.*

Results and Discussion The data of four subjects in the CUT condition and one subject in the TRANSITIVITY condition was not used in the data analysis because they did not answer all tasks.

Table 5 presents the mean upper and lower bounds of the CUT-condition ($n = 16$) and the TRANSITIVITY condition ($n = 19$). In the TRANSITIVITY condition 51.13% responded by point values and in the CUT condition 70.54% responded by point values on the average. The mean intervals in the TRANSITIVITY condition are slightly larger than in the CUT condition; t-tests comparing the mean interval sizes of TRANSITIVITY tasks and the CUT tasks did not show significant differences. As in Experiment 2, the subjects did not understand the probabilistic non-informativeness of the TRANSITIVITY argument form.

The frequencies of the interval response categories of the CUT tasks are presented in Table 5. 65.63% of the subjects in the CUT tasks responded by coherent intervals on the average (cell $\boldsymbol{e}$, $M = 10.50, SD = 2.59$, 14 tasks). This is a "better" performance compared with the 55.35% coherent interval responses in the CUT tasks of Experiment 1. The correlation between the mean lower bound responses and the normative lower bounds for all fourteen CUT tasks was $r = .95$.

As in Experiment 1, we estimate the reliability of the data by the correlations between the responses in those tasks which have normatively equivalent bounds. The correlation between the lower bound responses in the tasks $B2$ and $B5$ is $r = .97$, and in the tasks $B9$ and $B11$ the correlation is $r = .99$. The correlation between the upper bound responses of the tasks $B2$ and $B11$ is $r = .80$.

The RIGHT WEAKENING task was presented after the CUT tasks and after the TRANSITIVITY tasks. In the RIGHT WEAKENING task in CUT condition, fifteen of the sixteen subjects responded by the coherent lower bound "70" $(M = 67.50, SD = 10.00)$ and 11 responded by the coherent upper bound "100" $(M = 88.12, SD = 20.40)$. Eleven subjects responded by the optimal coherent interval (70-100%), only one subject was incoherent because of a violation of the lower bound. All except one subjects endorsed the RIGHT WEAKENING rule.

In the RIGHT WEAKENING task in the TRANSITIVITY condition, eighteen of the nineteen subjects responded coherently "70" as the lower bound. Eight subjects responded by the optimal coherent upper bound, namely "100", and eleven responded by "70" which is coherent but not optimal $(M = 82.63, SD = 15.22)$. Eight subjects responded by the optimal coherent interval. All except one subjects endorsed the RIGHT WEAKENING rule.

The formulation as a logical tautology makes the RIGHT WEAKENING task more easier for the subjects compared with the formulation of Experiment 2 where we observed 87.15% endorsement of the RIGHT WEAKENING rule.

In sum, we replicated the results on the CUT and the RIGHT WEAKENING rule of Experiment 1. The improved cover stories yielded to a better result by producing a higher rate of coherent interval responses. Explicitly mentioning the universe discourse in the TRANSITIVITY task did not help the subjects to understand the probabilistic non-informativeness of the TRANSITIVITY argument form. One explanation is, that the conversational implicatures are stronger and override the informations about the universe of discourse. Another explanation is that the TRANSITIVITY task is a proper three variable problem. Subjects thus reduce the processing demands by representing the TRANSITIVITY tasks as CUT, since the CUT can be reduced to a two variable problem by deleting the conditioning variable which is constant in all premises and the conclusion.

5 Concluding Remarks

In the present study we investigated human probabilistic reasoning about nonmonotonic conditionals. A series of three studies was investigating the understanding of elementary rules of a central formal system of nonmonotonic reasoning called SYSTEM P. For the investigation whether subjects endorse nonmonotonic inference rules and do not endorse monotonic argument forms, we investigated nonmonotonic inference rules valid in SYSTEM P and central properties of classical (monotone) logic which are not valid in SYSTEM P. While nonmonotonic inference rules are probabilistically informative (the coherent probability of the conclusion is *not* necessary the unit interval, $[0, 1]$), the monotonic argument forms are probabilistically not informative (the coherent probability of the conclusion is necessary the unit interval).

All subjects with very few exceptions (!) inferred probabilistically informative intervals in the nonmonotonic tasks. Practically all subjects perfectly endorse the RIGHT WEAKENING rule of SYSTEM P by inferring coherent intervals from the premises to the conclusion. More than 50% of the subjects inferred coherent intervals from the premises of the CUT rule of SYSTEM P. This is a rather good result. Only one out of six possible categories of interval responses contains coherent intervals and just this category contains the majority of the interval responses.

CONTRAPOSITION and TRANSITIVITY are monotonic argument forms which are not valid in SYSTEM P. A critical result of the present study is that CONTRAPOSITION was not endorsed by the subjects. The subjects understand the probabilistic non-informativeness of the monotonic argument forms. The subjects did not understand the probabilistic non-informativeness of the monotonic TRANSITIVITY. We explained this result by conversational implicatures, that the TRANSITIVITY tasks are actually interpreted by the subjects as instances of the CUT rule. We analyzed the TRANSITIVITY data under this assumption and observed a similar endorsement rate as in the CUT tasks.

Of special interest are the tasks in which all premises are sure. This is the case in those tasks in which the percentages of the lower and upper bounds in the premises are equal to 100. In the tasks with sure premises, practically all subjects endorse the SYSTEM P rules. The high endorsement rates are comparable to the endorsement rates of the non-probabilistic version of the MODUS PONENS (89–100%, [6]). In the case of the monotonic argument forms TRANSITIVITY and MODUS BARBARA the mean lower bounds are very high. As discussed above, subjects might interpret these argument forms as CUT.

In the present study we did not investigate whether subjects actually withdraw conclusions in the light of new evidence. Rather, reasoning from nonmonotonic conditionals was investigated and basic rationality postulates of nonmonotonic reasoning were corroborated. The critical condition of the monotonic argument forms indicate that most human subjects do not reason monotonically.

Adding the results of the present study to the results reported in previous studies on the probabilistic interpretation [14, 15] and on the possibilistic interpretation [3, 2] of SYSTEM P adds more evidence to the hypothesis, that the basic rationality postulates of SYSTEM P represent cornerstones in a competence theory of human reasoning.

		A1L	*A1U*	*A2L*	*A2U*	*A3L*	*A3U*	*B1L*
CUT	$M(SD)$:	72.38 (22.60)	81.75 (12.11)	57.25 (15.37)	71.12 (16.84)	58.00 (19.36)	68.12 (16.94)	37.62 (14.98)
TR	$M(SD)$:	59.42 (36.55)	86.05 (5.38)	43.70 (27.80)	72.88 (16.49)	50.25 (30.07)	74.11 (17.31)	30.05 (22.52)
		B1U	*B2L*	*B2U*	*B3L*	*B3U*	*B4L*	*B4U*
CUT	$M(SD)$:	48.19 (20.86)	40.88 (15.63)	50.69 (20.82)	44.06 (13.18)	52.44 (14.49)	45.62 (18.11)	54.56 (19.18)
TR	$M(SD)$:	52.73 (17.66)	33.30 (23.03)	55.18 (18.27)	35.03 (20.88)	59.03 (21.64)	43.34 (30.56)	61.54 (17.19)
		B5L	*B5U*	*B6L*	*B6U*	*B7L*	*B7U*	*B8L*
CUT	$M(SD)$:	40.69 (16.39)	49.81 (20.41)	43.81 (17.07)	53.75 (21.06)	93.75 (25.00)	93.75 (25.00)	36.38 (16.01)
TR	$M(SD)$:	34.55 (25.64)	58.37 (16.34)	29.05 (22.96)	48.84 (19.38)	77.95 (37.98)	94.74 (15.77)	21.55 (20.88)
		B8U	*B9L*	*B9U*	*B10L*	*B10U*	*B11L*	*B11U*
CUT	$M(SD)$:	46.25 (24.07)	56.50 (23.95)	65.31 (22.46)	37.88 (13.81)	47.44 (19.61)	55.69 (25.08)	64.12 (23.58)
TR	$M(SD)$:	39.08 (22.37)	49.92 (31.75)	76.13 (13.42)	41.74 (52.13)	53.76 (17.84)	47.76 (33.43)	76.03 (12.73)

		A1			*A2*			*A3*			*B1*			*B2*	
UA	1	0	0	0	0	0	1	2	1	1	1	0	1	1	0
UW	1	**8**	-	2	**13**	-	1	**10**	-	0	**12**	-	1	**10**	-
UB	6	-	-	1	-	-	1	-	-	2	-	-	3	-	-
		B3			*B4*			*B5*			*B6*			*B7*	
UA	1	0	0	1	1	0	1	1	0	1	1	0	-	-	-
UW	0	**13**	-	3	**7**	-	1	**10**	-	2	**12**	-	0	**15**	-
UB	2	-	-	4	-	-	3	-	-	0	-	-	1	-	-
		B8			*B9*			*B10*			*B11*				
UA	1	1	0	1	0	0	1	1	0	1	1	0			
UW	1	**13**	-	2	**7**	-	1	**10**	-	1	**7**	-			
UB	0	-	-	6	-	-	3	-	-	6	-	-			
	LB	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*	*LB*	*LW*	*LA*

Table 5: Subjects mean responses (*SD* are in the parentheses) in the CUT ($n = 16$) condition and the TRANSITIVITY condition (TR, $n = 19$), and the frequencies of the interval responses in the CUT condition of Experiment 3 (bottom table). For the abbreviations, normative intervals and the percentages given in the premises see Table 3.

References

[1] E. W. Adams. *The logic of conditionals.* Reidel, Dordrecht, 1975.

[2] S. Benferhat, J.-F. Bonnefon, and R. Da Silva Neves. An overview of possibilistic handling of default reasoning, with experimental studies. *Synthese*, 53–70, 2005.

[3] J.-F. Bonnefon and D. J. Hilton. The suppression of modus ponens as a case of pragmatic preconditional reasoning. *Thinking & Reasoning*, 8(1):21–40, 2002.

[4] M. D. S. Braine and D. P. O'Brien, editors. *Mental logic.* Erlbaum, 1998.

[5] N. Chomsky. *Aspects of the theory of syntax.* MIT Press, Cambridge, MA, 1965.

[6] J. Evans, S. Newstead, and R. Byrne. *Human Reasoning.* Erlbaum, 1993.

[7] J. Evans, S. Handley, and D. Over. Conditionals and conditional probability. *Journal of Experimental Psychology*, 29:321–355, 2003.

[8] A. Gilio. Probabilistic reasoning under coherence in System P. *Annals of Mathematics and Artificial Intelligence*, 34:5–34, 2002.

[9] H. Grice, ed. *Studies in the way of words.* Harvard University Press, 1989.

[10] P. N. Johnson-Laird. *Mental models.* Cambridge University Press, 1983.

[11] I.-M. Liu, K.-C. Lo, and J.-T. Wu. A probabilistic interpretation of 'If—Then'. *The Quarterly Journal of Experimental Psychology*, 49(A):828–844, 1996.

[12] J. Macnamara. *The place of logic in psychology.* MIT Press, 1986.

[13] K. Oberauer and O. Wilhelm. The meaning(s) of conditionals. *Journal of Experimental Psychology*, 29:680–693, 2003.

[14] N. Pfeifer and G. D. Kleiter. Nonmonotonicity and human probabilistic reasoning. In *Proceedings of the 6th WUPES*, pages 221–234, Hejnice, 2003.

[15] N. Pfeifer and G. D. Kleiter. Coherence and nonmonotonicity in human reasoning. *Synthese*, 146(1-2):93–109, 2005.

[16] N. Pfeifer and G. D. Kleiter. Towards a mental probability logic. *Psychologica Belgica*, 45(1):71–99, 2005.

[17] G. Politzer. Uncertainty and the suppression of inferences. *Thinking & Reasoning*, 11(1):5–33, 2005.

[18] L. J. Rips. *The psychology of proof.* MIT Press, 1994.

[19] K. Stenning and M. van Lambalgen. Semantic interpretation as computation in nonmonotonic logic. *Cognitive Science*, 29:919–960, 2005.