Thinking & Reasoning

ISSN: 1354-6783 (Print) 1464-0708 (Online) Journal homepage: https://www.tandfonline.com/loi/ptar20

Quantifying disablers in reasoning with universal and existential rules

Lupita Estefania Gazzo Castañeda & Markus Knauff

To cite this article: Lupita Estefania Gazzo Castañeda & Markus Knauff (2018) Quantifying disablers in reasoning with universal and existential rules, Thinking & Reasoning, 24:3, 344-365, DOI: 10.1080/13546783.2017.1401000

To link to this article: https://doi.org/10.1080/13546783.2017.1401000

THINKING & REASONING, 2018
VOL. 24, NO. 3, 344–365
https://doi.org/10.1080/13546783.2017.1401000

Routledge
Taylor & Francis Group

Quantifying disablers in reasoning with universal and existential rules

Lupita Estefania Gazzo Castañeda and Markus Knauff

Experimental Psychology and Cognitive Science, Justus Liebig University Giessen, Giessen, Germany

ABSTRACT

People accept conclusions of valid conditional inferences (e.g., if *p* then *q*, *p* therefore *q*) less, the more disablers (circumstances that prevent *q* to happen although *p* is true) exist. We investigated whether rules that through their phrasing exclude disablers evoke higher acceptance ratings than rules that do not exclude disablers. In three experiments we re-phrased content-rich conditionals from the literature as either universal or existential rules and embedded these rules in Modus Ponens and Modus Tollens inferences. In Experiments 2 and 3, we also used abstract rules. The acceptance of conclusions increased when the rule was phrased with "all" instead of "some" and the number of disablers had a higher impact on existential rules than on universal rules. Further, the effect of quantifier was more pronounced for abstract rules and when tested within subjects. We discuss the relevance of phrasing, quantifiers and knowledge on reasoning.

ARTICLE HISTORY Received 2 August 2016; Accepted 31 October 2017

KEYWORDS Reasoning; quantifiers; disablers; disabling conditions

Introduction

For a long time, human reasoning has been investigated by testing people's capacity to reason according to the principles of classical logic. Participants were confronted with a set of premises, including an if–then conditional rule and a fact, and asked to infer what necessarily follows from these premises. As part of the instructions, participants had to assume the premises as true. For example:

If a person goes to bed late, then the person will be tired.
A person goes to bed late.
———————————————————————————
The person will be tired.

CONTACT Lupita Estefania Gazzo Castañeda Estefania.Gazzo@psychol.uni-giessen.de

This inference is called Modus Ponens (MP) and is valid according to classical logic: whenever the if-part (i.e., the antecedent *p*) of the conditional is true, then the when-part (i.e., the consequent *q*) is also true. In other words, in classical logic, the antecedent is sufficient for the occurrence of the consequent (e.g., Hilton, Jaspars, & Clarke, 1990; Thompson, 1994, 1995). Another valid – but more difficult – inference is Modus Tollens (MT):

If a person goes to bed late, then the person will be tired.
A person will not be tired.
———————————————————————————
The person did not go to bed late.

The validity of MT can be also explained by the sufficiency relation of *p* and *q*: since *p* is sufficient for *q*, when *q* is false, then *p* has to be false, too. In classical logic, such a conclusion can only be true or false, nothing in between. Moreover, no additional information can make such a conclusion false, because as long as the premises are true, the conclusion is necessarily true. This property of classical logic is referred to as *monotonicity*.

Insightful findings and theories on human reasoning have emerged from studies investigating people's capacity to reason deductively (e.g., Braine & O'Brien, 1991; Johnson-Laird & Byrne, 1991). However, in recent years, researchers' attention has moved to the investigation of everyday reasoning. Everyday reasoning differs in two main aspects from classical logic. First, in our daily life, the information we get is often true only to a certain degree. Thus, the prerequisite of classical logic to assume the premises as true is often difficult to fulfil. Second, many conclusions we draw in our daily life can be withdrawn in light of additional information, thus conflicting with the principle of monotonicity. Consequently, people often refuse to draw logically valid conclusions because they know that the antecedent is not always sufficient for the occurrence of the consequent. Instead, the perceived sufficiency of *p* for *q* depends on the availability of disabling conditions in long-term memory (e.g., Thompson, 1994, 1995).

Different accounts exist to explain the effect of disablers. In one account, disablers (or disabling conditions) are interpreted as circumstances that prevent *q* to happen although *p* is true. For instance, in the example above, such disablers can be coffee in the morning, that the person was able to sleep in, or that the person only needs a few hours to feel rested. The more disablers exist for a given conditional, the less sufficient *p* is perceived to be for *q*, and the less participants accept otherwise valid conclusions (Cummins, 1995; Cummins, Lubart, Alksnis, & Rist, 1991; De Neys, Schaeken, & d'Ydewalle, 2003a). Cummins et al. (1991), for example, asked one group of participants to generate as many disablers as possible for a set of conditionals. Depending on how many disablers they generated, she divided the conditionals as having many or few disablers. In a later experiment, conducted with another

group of participants, she embedded these conditionals in inference tasks and found that acceptance rates varied according to the number of disablers, even though the actual disablers were never presented explicitly as part of the task (for similar findings, see also, e.g., Cummins 1995; De Neys, Schaeken, & d'Ydewalle, 2002, 2003a).

Another account is to explain the effect of disablers through probabilities. According to the "new psychology of reasoning" (Evans, 2012), the proper norm for human reasoning is the (Bayesian) probability calculus. In this framework, disablers can lower the conditional probability (cf. Evans, 2012; Over, 2009; see also Weidenfeld, Oberauer, & Hörnig, 2005) which is the probability that *q* actually follows from *p* (see e.g., Oaksford & Chater, 2007; Oaksford, Chater, & Larkin, 2000). The conditional probability can be calculated by performing the so-called Ramsey test (1929/1990; see also Evans, Handley, & Over, 2003). For this, people first assume that *p* holds. On the basis of this hypothetical belief, they then calculate how probable it is that *q* follows. For instance, for the conditional probability of "If a person goes to bed late, then the person will be tired," reasoners first assume that a person goes to bed late. Then they start thinking how probable it is that a person actually feels tired after going to bed late and how probable it is that this is not the case. The higher the probability of *P* (*p* and *q*) is relative to *P* (*p* and not-*q*), the higher the conditional probability *P* (*q* | *p*) is (Evans & Over, 2004; Over & Evans, 2003) – and the more sufficient *p* is perceived to be for *q*. Or, the other way around, the higher *P* (*p* and not-*q*) is relative to *P* (*p* and *q*), the *less* a logically valid conclusion is drawn.

A third approach – that we advocate here – is that people's acceptance of conclusions is also influenced by contextual factors, such as the phrasing of the conditional. In Gazzo Castañeda and Knauff (2016), we created legal conditionals describing offences and their corresponding punishments, either with the modal auxiliary *should* ("If a person downloads child pornography, then the person should be punished for possession of child pornography") or *will* ("If a person downloads child pornography, then the person will be punished for possession of child pornography"). We found that when the conditional contained a severe offence and asked for what *should* happen, participants concluded that the offender should be punished, ignoring potential disablers. However, when the legal conditional asked about what *will* happen, then participants did consider disablers – even for morally severe offences. Similar effects of the phrasing of conditionals on inferences were also found by Thompson (1994). She compared normal conditionals ("If *p*, then *q*") with reversed conditionals ("If *q*, then *p*") and embedded them in inferences were the second premise either asked for *p* (e.g., "If the car runs out of gas, then it stalls. The car runs out of gas. Therefore it stalls") or for *q* (e.g., "If the car stalls, then it has run out of gas. The car runs out of gas. Therefore it stalls"). Although both tasks share the same second premise and

conclusion, reasoners more likely accepted the first version than the second version of the task. This shows that the same content of an inference can play a different role, depending on the syntax of the "if–then" relationship (Thompson, 1994).[1]

The aim of this study is to continue investigating the importance of phrasing in everyday reasoning. So far, the consideration of disablers has been investigated with conditionals, describing an if–then relationship between antecedent and consequent. However, many of the relationships between antecedent and consequents can be also described with quantifiers. For instance, the relationship described in the conditional "If a person goes to bed late, then the person will be tired" can be also phrased as "All persons that go to bed late will be tired" or as "Some persons that go to bed late will be tired." Now imagine these quantified statements embedded in MP and MT inferences[2] :

All persons that go to bed late will be tired.
A person goes to bed late.
———————————————————
The person will be tired.
Vs.
Some persons that go to bed late will be tired.
A person goes to bed late.
———————————————————————————
The person will be tired.

Which conclusion would you accept more? Our assumption is that the rephrasing of conditional relationships with different quantifiers can either inhibit or enhance people's acceptance of conclusions: while universal quantifiers negate the existence of disablers, existential quantifiers suggest that exceptions do exist. Two consequences result from this hypothesis. First, participants should accept conclusions from universal rules to a higher extent than conclusions from existential rules. Second, by inhibiting disablers, the difference in acceptance ratings between rules with many and few disablers should be smaller for universal than for existential rules. Only when paired with existential rules, reasoners are "allowed" to consider the number of disablers. This should result in higher acceptance rates for rules with few disablers as compared to rules with many disablers. However, this effect of the number of quantifiers should be less pronounced for universal rules, since universal quantifiers encourage reasoners to ignore disablers.[3]

[1]We thank Valerie Thompson for this helpful suggestion.
[2]Formally speaking, the logical form of existential quantifiers is a conjunction and not a conditional, so that phrasing MP and MT inferences with existential quantifiers does not result in genuine MP and MT inferences, but rather in invalid inferences. We will address this point in the discussion of Experiment 2 and argue why this does not conflict with our results.
[3]Again, we thank Valerie Thompson for this helpful suggestion.

These hypotheses are strengthened by Cruz and Oberauer (2014). They showed that although "all" and "if" are equivalent in predicate logic ("All A are B" corresponds to "For all *x* it holds that if *x* is an A, then *x* is B"), people understand both differently. While the interpretation of "if" depends on the conditional probability – allowing thus for different degrees of belief – the quantifier "all" is interpreted as $P(B \mid A) = 1$. The statement "All As are Bs" is thus considered false as soon as this probability is below 1, that means, as soon as one disabler is considered. This in turn implies that the quantifier "all" should inhibit the consideration of disablers. However, this does not apply for the existential quantifier "some". Similar to Cruz and Oberauer, Chater and Oaksford (1999) also argued that "All As are Bs" is understood as $P(B \mid A) = 1$. Yet, Chater and Oaksford also make claims for the quantifier "some". They argue that "Some As are Bs" only implies that the conditional probability is above 0 ($P(B \mid A) > 0$) and that some things are both As and Bs ($\exists A, B$), permitting thus the consideration of disablers. However, Chater and Oaksford (1999) tested these assumptions for syllogistic reasoning in the strict sense. That is, they tested whether their probabilistic interpretation of quantifiers predicts people's inferences (and people's usage of heuristics) when reasoning from quantified premises to quantified conclusions (e.g., "All As are Bs; Some Cs are As; therefore, some Cs are Bs"). In our paper, we are instead interested in understanding how different quantifiers can affect people's acceptance of conclusions in inference tasks following the structure of MP and MT inferences.

We now present three experiments on reasoning with quantifiers. In the first experiment, we used a between-subjects design and re-phrased conditionals from the literature as either universal or existential rules by putting the corresponding quantifier in front of each statement. For universal statements, we used the quantifier "all", and for existential statements, we used the quantification "there is at least one", which is a logical equivalent to "some". In the second experiment, we used abstract problems alongside content rich problems. In addition, we changed the phrasing of the existential quantifier to "some". Finally, in Experiment 3, we changed our experimental paradigm to a within-subjects design to further understand the effects found in Experiment 2. The paper ends with a discussion on the importance of phrasing, content and experimental designs in reasoning.

Experiment 1

Methods

Participants. Sixty participants (41 female) took part in the experiment. Their mean age was $M = 22.78$ years ($SD = 3.45$).

Table 1. Structure of the problems used in Experiment 1 illustrated by a rule with many disablers (though rules with few disablers were used as well).

	Quantifier	
	All	There is at least one
MP	R: All that study hard will do well in tests. F: Person X studies hard. C: Person X will do well in tests.	R: There is at least one person that studies hard and will do well in tests. F: Person X studies hard. C: Person X will do well in tests.
MT	R: All that study hard will do well in tests. F: Person X does not do well in tests. C: Person X did not study hard.	R: There is at least one person that study hard and will do well in tests. F: Person X does not do well in tests. C: Person X did not study hard.

Note: R: quantified rule; F: fact; C: conclusion.

Materials and design. We created our problems by taking 12 conditionals from the existing literature and phrasing them either as universal or existential rules. Eight of the 12 conditionals came from De Neys et al. (2002; some of them also used by Cummins, 1995), and 4 from Verschueren, Schaeken, and d'Ydewalle (2005). According to the authors, half of the conditionals have many disablers, the other half few.[4] We rephrased these 12 conditionals either as universal or existential rules by adding either "all" (e.g., "All that jump into the pool will get wet"; "All apples that are ripe will fall from the tree") or "there is at least one" (e.g., "There is at least one person that jumps into the pool and gets wet"; "There is at least one apple that is ripe and will fall from the tree") in the beginning of each statement. Each quantified rule was presented twice, once embedded in an MP inference and once embedded in an MT inference, creating a total of 24 problems. The person or object described in the fact was always labelled "X" (e.g., Person X, girl X, apple X) to emphasise that we are referring to one particular person or object. The problems consisted thus of (1) a quantified statement, (2) a fact (p for MP, or not-q for MT) and (3) a conclusion (q for MP or not-p for MT). The conclusion was followed by a 5-point Likert scale, where participants had to indicate to which degree they accept of the conclusion (1 = *not at all* to 5 = *fully*; the order of the extremes was counterbalanced). For an illustration, see Table 1.

The experiment followed a 2 (disablers: many vs. few) × 2 (inference: MP vs. MT) × 2 (quantifier: all vs. there is at least one) mixed design. The kind of quantifier was varied between individuals: 30 participants were confronted with the universal quantifier "all" and 30 with the existential quantifier "there is at least one". The number of disablers and the kind of inference was varied within individuals.

[4]Besides the number of disabling conditions, De Neys et al. (2002) and Verschueren et al. (2005) also consider the number of alternatives, which is the number of alternative situations which also bring about q, without the necessity of p. We selected our items only on the basis of disablers, because the number of alternatives does not usually influence MP and MT inferences (cf. Cummins, 1995).

In addition to the inference task, we also included a generation task. Similar to Cummins et al. (1991) and De Neys et al. (2002, 2003a), participants had 1.5 minutes to generate as many disablers as possible for the different rules we used in the inference task ("A person jumps into the pool but does not get wet" [Why?]). The generation task aimed to test whether our German translation of the rules affected the number of disablers participants can generate.

Procedure. The experiment was programmed with Superlab 4.5 by Cedrus Cooperation. Participants were tested individually. During the instructions, participants were told that they would be presented with statements containing some general rule and that their task was to indicate, considering the provided information, how strongly they accept the conclusion on the basis of the before mentioned rule. Participants gave their acceptance ratings on the 5-point Likert scale. As in previous experiments, participants were told to answer intuitively and that right or wrong answers do not exist (cf. Cummins, 1995; De Neys, Schaeken, & d'Ydewalle, 2003b). We did therefore not tell the participants to reason logically. Each statement (the quantified rule, the fact, and the conclusion) was presented on a separate screen. Participants could switch to the next screen by pressing the space bar. The conclusion was written in red font and was followed – on a separate screen – by the image of the 5-point Likert scale. The 24 inference problems were presented in a random order after a short practice trial consisting of one denial of the antecedent inference (not-*p*, therefore not-*q*). After the inference task, participants completed the generation task.

Results

Generation task. Participants generated more disablers for rules classified as having many disablers ($M = 4.55$, $SD = 1.07$) than for those classified as having few ($M = 2.88$, $SD = 0.94$), $t(58) = 16.01$, $p < .001$, $d = 1.642$.[5] Our manipulation check was thus effective and the translation process did not affect the classification of the conditionals as having many or few disablers.

Inference task. We analysed the acceptance ratings of the different conclusions by conducting a 2 (disabler: many vs. few) × 2 (inference: MP vs. MT) × 2 (quantifier: all vs. there is at least one) analysis of variance (ANOVA).[6] Acceptance ratings ranged from 1 to 5, higher numbers indicating higher acceptance ratings. Descriptive data can be found in Figure 1.

[5]Standardised mean differences (*d*) were computed as described by Borenstein (2009).

[6]We also computed mixed-effects models for our analyses (Experiments 1–3) and they show similar results to the ones reported here.

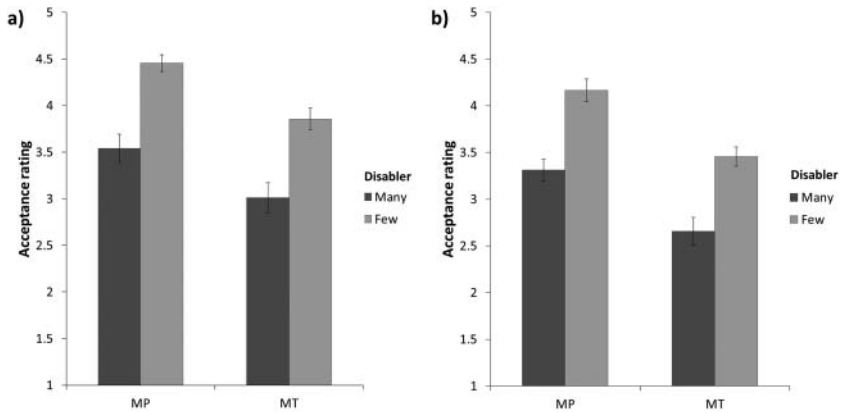

Figure 1. Acceptance ratings (1–5) for MP and MT inferences for statements with the quantifiers: (a) "All" and (b) "There is at least one" in Experiment 1. Error bars show standard errors.

The ANOVA revealed a main effect of the number of disablers, $F(1, 58) = 136.26$, $p < .001$, $\eta_p^2 = .701$, and a main effect of inference, $F(1, 58) = 127.03$, $p < .001$, $\eta_p^2 = .687$. Acceptance ratings were higher for rules with few disablers ($M = 3.98$, $SD = 0.55$) compared to those with many disablers ($M = 3.13$, $SD = 0.76$). And acceptance ratings were higher for MP ($M = 3.87$, $SD = 0.60$) than for MT inferences ($M = 3.24$, $SD = 0.66$). In addition, we could also find a main effect of quantifier, $F(1, 58) = 4.49$, $p = .038$, $\eta_p^2 = .072$. Acceptance ratings were higher for problems with the quantifier "all" ($M = 3.72$; $SD = 0.61$) than with the quantifier "there is at least one" ($M = 3.40$; $SD = 0.55$). All other effects were not significant (F's ≤ 1.17, p's $\geq .285$).

Discussion

The aim of Experiment 1 was to show that people's acceptance of MP and MT conclusions can be enhanced by phrasing rules in inference tasks with the universal quantifier "all" instead of the existential quantifier "there is at least one" and that the consideration of the number of disablers will be thus larger for existential than for universal rules. On the one hand, our results replicate previous findings showing that the number of disablers and the type of inference affects conclusions, even when they are phrased with quantifiers. On the other hand, we also showed that existential quantifiers lower the overall acceptance ratings, but universal quantifiers elevate participants' acceptance of conclusions. Yet, this effect of quantifier was smaller than expected and the quantifier did not interact with the number of disablers. There are two possible explanations for these results. The first reason may be the content of our problems. The conditionals we re-phrased to quantified statements are

content rich and describe everyday situations. Thus people can rate the conclusions on the basis of their background knowledge and do not need the initial rule or the concrete quantifier to do so. For instance, in the example "All that jump into the pool will get wet", participants can conclude that it is very likely that someone gets wet after jumping into the swimming pool without having to consider the concrete rule describing this relationship between jumping into a pool and getting wet as being universal or existential; this "rule" is already part of their general knowledge. It may thus be that participants did not pay much attention to the concrete quantifier in the rule because it was not necessary for rating the conclusion or perhaps even conflicted with background knowledge.

Another reason for the small difference between "all" and "at least one" might be the way the existential rules were phrased. It might be the case that the phrasing "there is at least one" is not appropriate to enhance the consideration of disablers. Evidence in this direction can be found in Schmidt and Thompson (2008). According to classical logic, the existential quantifications "there is at least one" and "some" have the same logical meaning, namely that "at least one A is B" but that maybe also "all As are Bs". People usually have problems in assigning this logical meaning to "some" because in everyday language, "some" is usually understood as "some but *not* all" (e.g., Begg & Harris, 1982; Newstead, 1989; see also Grice, 1975) – which was actually the reason why we expected the existential quantifier to enhance the consideration of disablers. However, Schmidt and Thompson (2008) showed that the usage of "there is at least one" instead of "some" diminishes this discrepancy between logic and everyday language in the meaning of the existential quantification. Probably, because "at least one" only refers to a particular case and does not exclude the possibility of "all As are Bs" for the remaining cases of A (Schmidt & Thompson, 2008). Consequently, it might be the case that participants did not simply ignore the initial quantified rule as previously suggested. Instead, it could also be that participants understood "there is at least one" as including the possibility of "all As are Bs". As a consequence, "there is at least one" could not enhance much the consideration of disablers, resulting only in a small difference in acceptance ratings between "there is at least one" and "all."

Due to these potential problems of Experiment 1, in Experiment 2, we decided to make two main changes to our experimental paradigm. First, we decided to change the phrasing of the existential quantifier into "some". Following Schmidt and Thompson (2008), this should allow us to find the expected effects of quantifiers on reasoning, because "some" does not have the connotation of including "all". The idea is similar to Moxey and Sanford (1993; see also Oaksford, Roberts, & Chater, 2002), who argued that "some" is pragmatically speaking *positive*, meaning that it used to tell how many is the case, relative to *all* (p. 48). In other words, it suggests that its corresponding (logical) "all" interpretation is false (Oaksford et al., 2002). The second change

we made in Experiment 2, was to add also abstract problems to the experiment. In case the interference with background knowledge is still too strong to allow effects of quantifiers, we should at least be able to find an effect of quantifiers for abstract problems, where people do not have prior knowledge about the relation between *p* and *q* and need to consider the initial quantifier in order to know if *q* follows necessarily from *p* or not.

Experiment 2

Methods

Participants. We tested 64 participants, but 4 participants were excluded from the computations because they reported having prior knowledge on formal logic after the experiment. The remaining 60 participants (38 female) were $M = 23.99$ years old ($SD = 4.62$).

Materials and design. The problems followed the same structure as the problems from Experiment 1, but with three important changes. First, in addition to content-rich quantified rules with many and few disablers, we also created "abstract" problems. These abstract problems still had concrete content, but did not describe any relationship that can be reasonable assumed to already be stored in people's memory as a general rule. For example, "All/ Some persons that participate in the contest will have to sing." Without the concrete quantifier in this rule, it is difficult to decide whether *q* ("A person has to sing") actually follows from *p* ("A person participates in the contest"). In other words, participants need the quantifier to rate the conclusion and should thus consider it during reasoning. Second, we now phrased the existential rules with "some" instead of "there is at least one". And third, we expanded the rating scale to a 7-point Likert scale and presented this scale together with the conclusion on the same slide. In addition, we also made our universal and existential rules more homogenous, by adding the word "persons" also to the universal rules (e.g., "All persons that jump into the pool will get wet").

We created our problems by taking three content-rich rules with many disablers and three with few disablers from Experiment 1, and by creating additionally three abstract rules. Each rule was presented twice: once embedded in an MP and once embedded in an MT inference. As in Experiment 1, the kind of quantifier was varied between participants, resulting in 18 problems per participant. Examples of the problems used in Experiment 2 can be found in Table 2.

Procedure. The procedure was the same as Experiment 1, presenting each premise on a separate screen and the conclusion in red font. This time, however, the conclusion was presented together with the rating scale (7-point)

Table 2. Example problems of Experiment 2.

	Inference	
	MP	MT
Content rich (many disablers)	R: All/ Some persons that study hard will do well in tests F: Person X studies hard C: Person X will do well in tests	R: All/ Some persons that study hard will do well in tests F: Person X does not do well in tests C: Person X did not study hard
Content rich (few disablers)	R: All/ Some apples that are ripe will fall from the tree F: Apple X is ripe C: Apple X will fall from the tree	R: All/ Some apples that are ripe will fall from the tree F: Apple X does not fall from the tree C: Apple X was not ripe
Abstract	R: All/ Some balls that roll down will fall into the box F: Ball X rolls down C: Ball X will fall into the box	R: All/ Some balls that roll down will fall into the box F: Ball X does not fall into the box C: Ball X did not roll down

Note: R: quantified rule; F: fact; C: conclusion. Participants got all problems either with the quantifier "all" or with the quantifier "some".

on the same screen. In addition, given that the content-rich problems were already tested in Experiment 1, we also decided to omit the generation task. As in Experiment 1, participants were told that no right or wrong answer exists and that they should answer intuitively.

Results

Acceptance ratings were analysed by conducting a 3 (content: abstract vs. many disabler vs. few disablers) × 2 (inference: MP vs. MT) × 2 (quantifier: all vs. some) ANOVA. The acceptance rating ranged from 1 to 7. Descriptive data can be found in Figure 2.

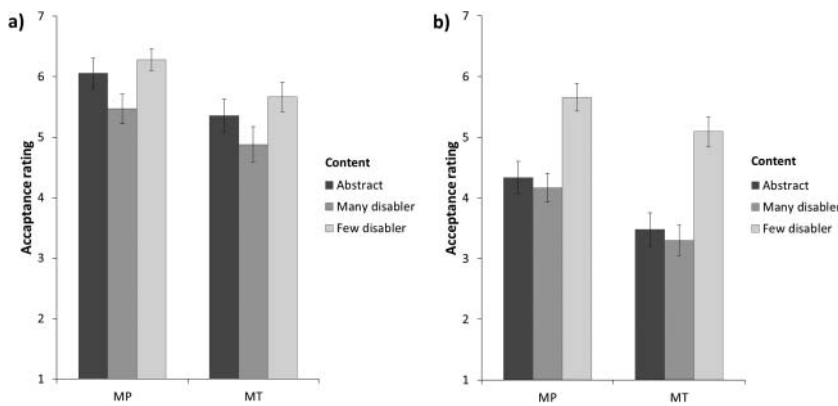

Figure 2. Acceptance ratings (1–7) for MP and MT inferences for statements with the quantifiers: (a) "All" and (b) "Some" in Experiment 2. Error bars show standard errors.

The ANOVA showed a main effect of inference, $F(1, 58) = 35.69$, $p < .001$, $\eta_p^2 = .381$, revealing that acceptance ratings were higher for MP inferences ($M = 5.33$; $SD = 1.20$) than for MT inferences ($M = 4.63$; $SD = 1.29$). In addition, the ANOVA also showed a main effect of content, $F(2, 116) = 30.52$, $p < .001$, $\eta_p^2 = .345$, a main effect of quantifier, $F(1, 58) = 26.12$, $p < .001$, $\eta_p^2 = .311$, and a significant interaction between both factors, $F(2, 116) = 7.35$, $p = .001$, $\eta_p^2 = .112$. The main effect of content indicates that rules with few disablers ($M = 5.67$; $SD = 1.10$) received overall higher acceptance ratings than rules with many disablers ($M = 4.45$; $SD = 1.46$), $t(59) = 6.69$, $p < .001$, $d = 0.929$, and also higher acceptance ratings than rules with abstract content ($M = 4.81$; $SD = 1.55$), $t(59) = 5.08$, $p < .001$, $d = 0.626$ (abstract problems and problems with many disablers did not differ, $t(59) = 2.31$, $p = .024$, $d = 0.234$; Bonferroni-adjusted alpha 0.0167). The main effect of quantifier indicates that acceptance ratings were as expected higher for problems with the quantifier "all" ($M = 5.62$; $SD = 1.00$) compared to problems with the quantifier "some" ($M = 4.34$; $SD = 0.93$). However, the interaction between these two factors indicates that the effect of quantifier depended on the content of the problem. The decrease in acceptance ratings from "all" to "some" was highly significant for abstract problems ($M_{all} = 5.71$; $M_{some} = 3.91$), $t(58) = 5.52$, $p < .001$, $d = 1.425$, and for problems with many disablers ($M_{all} = 5.17$; $M_{some} = 3.73$), $t(58) = 4.34$, $p < .001$, $d = 1.121$, but did not reach the Bonferroni-adjusted alpha of 0.0167 for problems with few disablers ($M_{all} = 5.97$; $M_{some} = 5.37$), $t(58) = 2.19$, $p = .033$, $d = 0.565$. As a consequence, the effect of the number of disabler (i.e., the difference in acceptance ratings between the few- and many-disabler conditions) was as expected higher for "some" ($M_{few-many} = 1.64$; $SD = 1.56$) than for "all" ($M_{few-many} = 0.80$; $SD = 1.12$), $t(58) = 2.39$, $p = .020$, $d = 0.617$. All other effects were not significant (F's ≤ 0.40; p 's $\geq .581$).

Discussion

Experiment 2 shows that by phrasing the existential quantifier with "some" instead of "there is at least one", we were indeed able to find the expected differences between the acceptance ratings of inferences phrased with universal and existential quantifiers. Acceptance ratings were lower for "some" than for "all", not only for abstract problems but also for content-rich problems with many disablers (and marginally for problems with few disablers). This indicates that the reason for the results of Experiment 1 was not that the participants ignored the quantifier (if this would have been the case, then we would not have found an effect of quantifier in the content-rich problems), but rather that our specific phrasing was responsible for the mild effects of existential quantifiers. Another result of Experiment 1 was that the difference between many and few disablers was indeed significantly higher for existential than for universal rules. This indicates that "all" inhibited the consideration

of disablers and "some" enhanced their consideration. This result supports our assumption that the pragmatic implications of quantifiers such as "some" and "all" can affect people's consideration of background knowledge (e.g., disablers) and thus their inferences. We return to that in the General Discussion.

Of course, an alternative explanation for our findings is that universal and existential MP and MT inferences differ in logical validity: while universal MP and MT inferences are valid, existential MP and MT inferences are not. Formally speaking, existential MP and MT inferences are not valid because existential rules are conjunctions and not conditional relationships. As a consequence, "some" may have received lower acceptance ratings than "all", not because of a difference in the consideration of disablers, but because existential MP and MT inferences are invalid. However, this alternative explanation is not critical for our findings. In fact, our participants were never asked to reason deductively. Instead, they were told to answer spontaneously and that no right or wrong answer exists. Moreover, the logical explanation cannot explain three observations we made: first, it does not explain why the effect of quantifiers was not that pronounced in Experiment 1 with the existential quantification "there is at least one". From a logical point of view, "there is at least one" and "some" are equivalent. If the effect of quantifier would be caused by the logical invalidity of existential MP and MT inferences, then changing the existential quantification to "some" should not have enhanced the effect of quantifier, since both ("some" and "at least one") are equally invalid. Second, the difference in logical validity between universal and existential MP and MT inferences does not explain the interaction between content and quantifier (higher effects of quantifier for abstract problems than for content-rich problems). A confound with validity should only result in a main effect of quantifier, but in no interactions. Finally, differences in validity do not explain why the effect of the number of disablers was higher for "some" than for "all". Again, we would only expect a main effect of quantifier if the invalidity of existential MP and MT inferences is the reason for its lower acceptance ratings. The higher effect of the number of disablers for "some" than for "all" can only be explained if the quantifier "some" did indeed encourage participants to consider disablers.

In sum, all these findings indicate that quantifiers have different pragmatic implications which encourage or inhibit the consideration of disablers. However, we still thought it was necessary to replicate the effects of Experiment 2 in order to draw reliable conclusions from our findings. Such a replication would also help to further understand how the quantifier and the content of problems interact. This was done in Experiment 3 by converting Experiment 2 to a within-subjects experiment.

Experiment 3

Methods

Participants. Thirty-two participants took part in the experiment, but two were excluded from the computations because after the experiment one reported to already be familiar with conditionals and formal logic and the other reported to have not understood the task. The remaining 30 participants (18 female) were $M = 22.2$ years old ($SD = 3.9$).

Materials, design and procedure. Experiment 3 was a replication of Experiment 2, with the only difference that the factor "quantifier" was varied within and not between subjects (i.e., all participants were confronted with both kinds of quantifiers). For this, the "all" and the "some" versions of Experiment 2 were merged together as one single experiment, resulting in 36 problems embedded in a 3 (content: abstract vs. many disabler vs. few disablers) × 2 (inference: MP vs. MT) × 2 (quantifier: all vs. some) within-subjects design. Everything else (instructions, response modality and content of the problems) was kept constant. Only the rule in the instructions and in the practice trial was changed by phrasing it without any quantifier to avoid learning effects (i.e., "On roads that are slippery there will be accidents").

Results

Acceptance ratings were analysed by conducting a 3 (content: abstract vs. many disablers vs. few disablers) × 2 (inference: MP vs. MT) × 2 (quantifier: all vs. some) repeated-measures ANOVA. The acceptance rating ranged from 1 to 7. Descriptive data can be found in Figure 3.

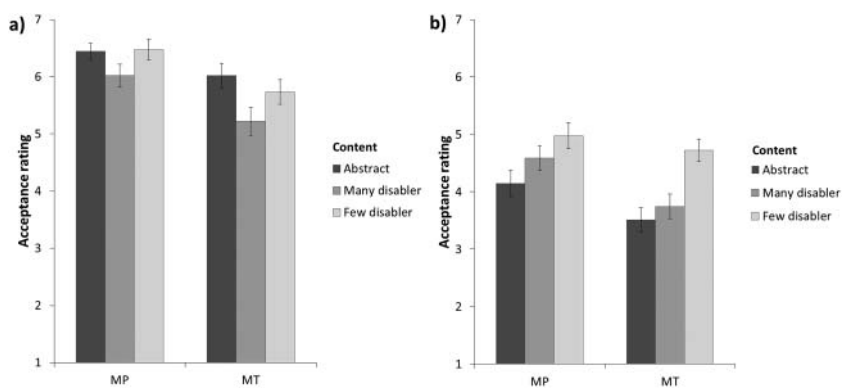

Figure 3. Acceptance ratings (1–7) for MP and MT inferences for statements with the quantifiers: (a) "All" and (b) "Some" in Experiment 3. Error bars show standard errors.

The ANOVA revealed main effects of inference, $F(1, 29) = 26.86$, $p < .001$, $\eta_p^2 = .481$, of quantifier, $F(1, 29) = 71.23$, $p < .001$, $\eta_p^2 = .711$, and of content, $F(2, 58) = 13.28$, $p < .001$, $\eta_p^2 = .314$. Acceptance ratings were higher for MP inferences ($M = 5.44$; $SD = 0.65$) than for MT inferences ($M = 4.83$; $SD = 0.79$) and also higher for problems with the quantifier "all" ($M = 5.99$; $SD = 0.75$) compared to problems with the quantifier "some" ($M = 4.28$; $SD = 0.94$). In addition, acceptance ratings were overall higher for rules with few disablers ($M = 5.48$; $SD = 0.73$) compared to those with many disablers ($M = 4.89$; $SD = 0.85$), $t(29) = 4.17$, $p < .001$, $d = 0.733$, and compared to those with abstract content ($M = 5.03$; $SD = 0.65$), $t(29) = 4.46$, $p < .001$, $d = 0.645$. Abstract problems and problems with many disablers did not differ, $t(29) = 1.22$, $p = .233$, $d = 0.173$ (Bonferroni-adjusted alpha 0.0167). Also, the interaction between content and quantifier was significant, $F(1.67, 48.46) = 19.59$, $p < .001$, $\eta_p^2 = .403$. The decrease in acceptance ratings from "all" to "some" was higher for abstract problems ($M_{\text{all–some}} = 2.41$; $SD = 1.27$) than for problems with many disablers ($M_{\text{all–some}} = 1.46$; $SD = 1.20$), $t(29) = 4.50$, $p < .001$, $d = 0.767$, and also than for problems with few disablers ($M_{\text{all–some}} = 1.26$; $SD = 1.33$), $t(29) = 5.18$, $p < .001$, $d = 0.882$; problems with many and few disablers did not differ in this respect, $t(29) = 1.36$, $p = .185$, $d = 0.156$ (Bonferroni-adjusted alpha 0.0167). As a consequence, the difference between the many- and few-disabler conditions was only descriptively higher for "some" ($M = 0.68$; $SD = 0.83$) than for "all" ($M = 0.48$; $SD = 0.90$), $t(29) = 1.36$, $p = .185$, $d = 0.231$. All other effects were not significant (F's ≤ 2.08; p's $\geq .134$).

Discussion

As in Experiment 2, we were again able to find an effect of quantifier on acceptance ratings. Participants gave lower acceptance ratings for inferences with the quantifier "some" than for inferences with the quantifier "all". Interestingly, however, the effect of quantifier was much more pronounced than in Experiment 2. The effect of quantifier grew considerably for problems with abstract content, obtaining over two scale points difference between "all" and "some". At the same time – but into a lesser extent –the difference between "all" and "some" for problems with few disablers also grew, no longer differing from the difference found for problems with many disablers. The difference between "all" and "some" is probably the highest for abstract problems, because in abstract problems, the only information about the sufficiency of the relation between *p* and *q* people can refer to is the quantifier, which in one case says that "all" As are Bs (high acceptance ratings) and in the other case that only "some" As are Bs (lower acceptance ratings). This is not the case for content-rich problems, where in addition to the sufficiency relation proposed by the quantifier, people also have their own beliefs about the sufficiency relation between *p* and *q* from their background knowledge.

In the case of problems with many disablers, this background knowledge makes it easy for participants to accept the quantifier "some" but probably hinders the acceptance of "all". The same happens for content-rich problems with few disablers. Here, the small number of disablers encourages the acceptance of the quantifier "all" but is in conflict with the quantifier "some". This conflict between the quantifier and the number of disablers does not exist for abstract problems, which explains why the effect of quantifier was higher for abstract than for content-rich problems.

But why the effect of the number of disablers was only slightly higher for "some" than for "all"? We think that the main reason for the lack of significance is that participants gave – in comparison to Experiment 2 – relatively low ratings to existential rules with few disablers (see Figure 3). The reason might be that the within-subjects design of Experiment 3 made participants more aware of the different quantifiers. By allowing comparisons between universal and existential rules, the quantifier became more salient. Thus participants may have guided their responses more strongly on the quantifier, giving overall high acceptance ratings to "all", but also low acceptance ratings to "some" – even when the rule has actually only a few disablers. This was not the case in Experiment 2, where participants were only confronted with one quantifier and had not the possibility to compare both quantifiers directly. In fact, this design-dependent weighting of quantifiers also explains why the effect of quantifier was larger in Experiment 3 than in Experiment 2. This effect of the experimental design can have important consequences for how people understand the problems in reasoning experiments (cf. Kahneman, 2000; Stanovich & West, 2008, see also Charness, Gneezy, & Kuhn, 2012; Grice, 1966). We discuss that in the General Discussion.

General discussion

The aim of this work was to show that (1) quantifiers can affect people's acceptance of conclusions in inference tasks, and (2) that this effect is related to pragmatic aspects of the used quantifiers in natural language. We therefore re-phrased conditionals as either universal or existential rules and embedded them in MP and MT inferences. We supposed that the quantifier "all" should inhibit the consideration of disablers since "all" implies that $P(B \mid A) = 1$. However, contrary to "all", the quantifier "some" should allow the consideration of disablers since it only implies that $P(B \mid A) > 0$ (and that some things are both As and Bs). The results support our hypotheses. Inferences were accepted more when the rule was phrased with the quantifier "all" than when it was phrased with the quantifier "some", this was especially the case for abstract problems but also for content-rich problems. Furthermore, given that "some" allowed the consideration of disablers and "all" inhibited it, also the difference in acceptance ratings between inferences with many and few disablers was

higher for "some" than for "all", significantly in Experiment 2 and descriptively in Experiment 3.

Our findings have several implications for reasoning research. One corollary is that the consideration of disablers – and people's engagement in uncertain reasoning and non-monotonicity – depends on how inference rules are phrased. People's usual consideration of uncertainties during reasoning can be attenuated when rules are phrased universally, even for problems people are familiar with. The inferences people draw do thus not only depend on their prior knowledge about the sufficiency relation between *p* and *q*, but also on the rule with which they are confronted. Klauer, Beller, and Hütter (2010) already showed how the actual presence of rules can affect peoples' inferences. They asked participants to estimate how probable it is that *q* follows from *p*, once after confronting them with the corresponding "if–then" conditional rule and once by asking them directly, without showing them an if-then rule before. Participants gave higher ratings when they were first confronted with the corresponding conditional rule than otherwise (see also Liu, 2003). Our results go even a step beyond the findings by Klauer et al. (2010). We do not just show that the presence of a rule influences inferences, but even that it is important *which* rule is presented and *how* this rule is phrased. In particular, if the phrasing of the rule suggests to exclude *p* and not-*q* cases, this inhibits the consideration of disablers and people are thus more inclined to accept the conclusion.

In future studies, it may be also interesting to compare how phrasing interacts with the trustworthiness of a source. For instance, Wolf, Rieger, and Knauff (2012) showed that conditionals are believed more when they are uttered by a trustworthy source. As a result, it is possible that trustworthy sources also inhibit the consideration of disablers. For example, when a doctor utters the conditional "If a person goes to bed late, then the person will be tired", then disablers may be considered less than when this conditional is uttered by a lay person. But can the trustworthiness of the source also inhibit the consideration of disablers for universal and existential rules? Maybe the effect of quantifiers is diminished for untrustworthy sources, because in those cases, reasoners might try to reason from their background knowledge and discard the information about quantifiers given by this untrustworthy source.

Another consequence from our studies is that the meaning of quantifiers in classical logic does not agree with how such words are used in natural language. Of course, this is not new (see e.g., Braine & O'Brien, 1991; Stenning & van Lambalgen, 2008). However, our results add a new component to these approaches by linking them to the role of disablers on reasoning. For instance, as already described in the introduction, in predicate logic, "all" and "if" are treated as equivalent. Therefore, one might also expect that people interpret both similarly in reasoning: such as conditionals are interpreted with differing degrees of belief, so should also rules with the quantifier "all" be

interpreted with different degrees of belief. However, our findings corroborate the findings of Cruz and Oberauer (2014) and Chater and Oaksford (1999), showing that in everyday reasoning "all" is not equivalent to "if". Statements with the universal quantification "all" seem to be considered false as soon as one disabler is known. Consequently, when embedded in inference tasks, "all" inhibits the effect of the number of disablers known from the literature, evoking a high acceptance of MP and MT inferences even for rules with many disablers (although not as high as for rules with abstract content). This mismatch between the meaning of quantifiers in classical logic and everyday language and reasoning is also corroborated by the different effects of the quantifiers "some" and "there is at least one". According to classical logic both – "some" and "there is at least one" – are equivalent because both are two different ways to express that "at least one A is B" but that maybe also "all As are Bs". However, our results suggest that these two quantifiers are also understood differently by naïve reasoners.

It is also worth noticing that our experimental paradigm varied in different aspects from that of classical reasoning research. In fact, our paradigm did not ask participants to draw a logically valid inference, as we did in previous experiments (e.g., Knauff & Johnson-Laird, 2002). In the present study, our instructions told participants that no right or wrong answer exists and also our Likert scale asked for acceptability ratings (cf. Douven & Verbrugge, 2010). In this respect, our paradigm is similar to many experiments in the "new paradigm of reasoning" (Evans, 2012; see also Elqayam & Evans, 2011), which considers Bayesian probability theory as the appropriate norm for human reasoning and assumes that disablers affect conclusions by lowering the conditional probability of *q* given *p* (e.g., Evans, 2012; Over, 2009; Weidenfeld et al., 2005). However, in our experiments, we did not ask our participants for probabilities, which is usually done in the Bayesian framework. In future experiments, we will consider to bring these paradigms even closer together, for instance, by studying the pragmatics of quantifiers and by asking for probability ratings. We believe that this can provide more insights into the relationship between quantifiers, pragmatics, disablers and probabilities.

Our experiments also have important consequences for the design of further experiments. It is important to be cautious when deciding which experimental design to use in experiments. Many times researchers prefer within-subject designs due to the smaller standard errors and the possibility of controlling for sampling effects. However, our study suggests that within-subject designs can also sensitise people to the different experimental conditions and thus inflate possible effects (cf. Stanovich & West, 2008). This problem of within-subject designs has already been discussed in the literature (e.g., Grice, 1966; Poulton, 1973) – sometimes under the labels of sensitisation (e.g., Greenwald, 1976) or demand effects (e.g., Charness et al., 2012; see also Orne, 1962) – and has important implications for future research on everyday

reasoning. As already Kahneman (2000) noticed, reasoning in real-life situations more resembles a between-subject design: people are confronted with one rule and have to make inferences only on the basis of this single rule. They are not confronted with two or more rules so that they can compare them and realise the distinctions of each rule. So, if within-subject designs indeed boost effects by making comparisons between conditions possible, then some of the findings in the literature on reasoning may not be as strong in real life as suggested by experiments. For instance, Stanovich and West (2008) argued that correlations between cognitive ability and thinking biases are more pronounced in within-subject designs, because those designs make highly intelligent participants aware that some bias has to be overridden. Also, the interpretation of some problems can change according to with which other problems it is presented. Oaksford et al. (2002) argued that the quantifier "few" is interpreted differently if it is presented in the same experiment with or without the quantifiers "all" and "none". When "few" is presented as a response option together with "all", then participants will interpret "few" to mean "a few" which they will not take to imply "all" since they have already the quantifier "all" to do so. But, when "few" is presented as a response option together with "none", then they will not use "few" to imply "none".

We think that our study opens an interesting new field of research. To our knowledge, we are the first group to investigate how phrasing MP and MT inferences with quantifiers affects people's acceptability of conclusions and their consideration of the number of disablers. This allowed us to discover mismatches between the meaning of quantifiers in classical logic and the corresponding words in everyday reasoning and language, and also helped us to understand more thoroughly the role of phrasing on reasoning. We encourage researchers to conduct further studies on the role of pragmatics in reasoning with quantifiers, for instance, by testing the impact of other quantifiers like "most" or "few" (see Chater & Oaksford, 1999; Johnson-Laird, 1983, 1994). Further, instead of quantified statements, it could also be interesting to add frequency information to the rules, such as "always" or "sometimes" (cf. Stevenson & Over, 1995). Comparisons between the impact of such *frequentist rules* and *quantified rules* can shed some light on the debate about the relative importance of the number of disablers and frequency of "*p* and not-*q*" cases in general (De Neys, 2010; De Neys et al., 2003a; Geiger & Oberauer, 2007). Geiger and Oberauer (2007), for instance, argued that although the number of disablers and the frequency of "*p* and not-*q*" cases often correlate (i.e., the more disablers exist, the more "*p* and not-*q*" cases there are), when both are disentangled, the frequency of "*p* and not-*q*" cases is the better predictor for inferences. By comparing the impact of "always" and "sometimes" with the impact of "all" and "some", we could thus gain insights into the cognitive mechanisms behind the consideration of disablers. For instance,

if frequency information such as "always" and "sometimes" has a higher impact on inferences than quantified statements such as "all" and "some", then this result would speak in favour of Geiger and Oberauer's (2007) findings.

In sum, this study shows that quantifiers can enhance or inhibit the consideration of disablers and affect people's acceptability of conclusions. However, the effect of quantifiers also depends on the content of the inference, on how quantifiers are phrased, and on the experimental design. We hope future studies will continue investigating the factors influencing people's consideration of disablers and their acceptance of conclusions in everyday reasoning.

Acknowledgment

We thank Jessica Ewerhardy for data collection. We are also very thankful to Valerie Thompson, Mike Oaksford, Sangeet (Sunny) Khemlani, and one anonymous reviewer for their helpful suggestions and comments on this paper.

An earlier version of Experiment 1 has been presented at the 37th Annual Conference of the Cognitive Science Society.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by DFG grant [grant number KN 465/10-1], [grant number KN 465/10-2] within the Priority Program "New Frameworks of Rationality" (SPP 1516) to Markus Knauff.

References

Begg, I., & Harris, G. (1982). On the interpretation of syllogisms. *Journal of Verbal Learning & Verbal Behavior, 21*, 595–620.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). New York, NY: Russell Sage Foundation.

Braine, M. D. S., & O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98*, 182–203.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*, 1–8.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38*, 191–258.

Cruz, N., & Oberauer, K. (2014). Comparing the meanings of "if" and "all". *Memory & Cognition, 42*, 1345–1356.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition, 23*, 646–658.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19*, 274–282.

De Neys, W. (2010). Counterexample retrieval and inhibition during conditional reasoning: Direct evidence from memory probing. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinkin*g (pp. 197–206). Oxford: Oxford University Press.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory & Cognition, 30*, 908–920.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003a). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition, 31*, 581–595.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003b). Causal conditional reasoning and strength of association: The disabling condition case. *European Journal of Cognitive Psychology, 15*, 161–176.

Douven, I., & Verbrugge, S. (2010). The Adams family. *Cognition, 117*, 302–318.

Elqayam, S., & Evans, J. St. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences, 34*, 233–290.

Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning, 18*, 5–31.

Evans, J. St. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning Memory and Cognition, 29*, 321–355.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Gazzo Castañeda, L. E., & Knauff, M. (2016). When *will* is not the same as *should*: The role of modals in reasoning with legal conditionals. *The Quarterly Journal of Experimental Psychology, 69*, 1480–1497.

Geiger, S. M., & Oberauer, K. (2007). Reasoning with conditionals: Does every counterexample count? It's frequency that counts. *Memory & Cognition, 35*, 2060–2074.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83*, 314–320.

Grice, G. R. (1966). Dependence of experimental laws upon the source of experimental variation. *Psychological Bulletin, 66*, 488–498.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Hilton, D. J., Jaspars, J. M. F., & Clarke, D. D. (1990). Pragmatic conditional reasoning: Context and content effects on the interpretation of causal assertion. *Journal of Pragmatics, 14*, 791–812.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition, 50*, 189–209.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove: Erlbaum.

Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences, 23*, 681–683.

Klauer, K. C., Beller, S., & Hütter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 298–323.

Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition, 30*, 363–371.
Liu, I.-m (2003). Conditional reasoning and conditionalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 694–709.
Moxey, L., & Sanford, A. (1993). *Communicating quantities*. Hove: Erlbaum.
Newstead, S. E. (1989). Interpretational errors in syllogistic reasoning. *Journal of Memory and Language, 28*, 78–91.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 26*, 883–899.
Oaksford, M., Roberts, L., & Chater, N. (2002). Relative informativeness of quantifiers used in syllogistic reasoning. *Memory & Cognition, 30*, 138–149.
Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–783.
Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking & Reasoning, 15*, 431–438.
Over, D. E., & Evans, J. St. B. T. (2003). The probability of conditionals: The psychological evidence. *Mind & Language, 18*, 340–358.
Poulton, E. C. (1973). Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin, 80*, 113–121.
Ramsey, F. P. (1929/1990). General propositions and causality (1929). In D. H. Mellor (Ed.), *Philosophical papers by F. P. Ramsey* (pp. 145–163). Cambridge: Cambridge University Press.
Schmidt, J., & Thompson, V. A. (2008). "At least one" problem with "some" formal reasoning paradigms. *Memory & Cognition, 36*, 217–229.
Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672–695.
Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.
Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *The Quarterly Journal of Experimental Psychology, 48A*, 613–643.
Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory & Cognition, 22*, 742–758.
Thompson, V. A. (1995). Conditional reasoning: The necessary and sufficient conditions. *Canadian Journal of Experimental Psychology, 49*, 1–58.
Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning, 11*, 239–278.
Weidenfeld, A., Oberauer, K., & Hörnig, R. (2005). Causal and noncausal conditionals: An integrated model of interpretation and reasoning. *The Quarterly Journal of Experimental Psychology, 58A*, 1479–1513.
Wolf, A. G., Rieger, S., & Knauff, M. (2012). The effects of source trustworthiness and inference type on human belief revision. *Thinking & Reasoning, 18*, 417–440.