Are Logical Intuitions Only Make-Believe?
Reexamining the Logic-Liking Effect

Constantin G. Meyer-Grant[1], Nicole Cruz[2,3], Henrik Singmann[4,5], Samuel Winiger[6], Spriha Goswami[3], Brett K. Hayes[3], and Karl Christoph Klauer[1]

[1]University of Freiburg, Freiburg, Germany

[2]University of Innsbruck, Innsbruck, Austria

[3]University of New South Wales, Sydney, Australia

[4]University College London, London, United Kingdom

[5]University of Warwick, Coventry, United Kingdom

[6]University of Zurich, Zurich, Switzerland

Author Note

Constantin G. Meyer-Grant https://orcid.org/0000-0002-5991-6596

Nicole Cruz https://orcid.org/0000-0001-7354-7785

Henrik Singmann https://orcid.org/0000-0002-4842-3657

Samuel Winiger https://orcid.org/0000-0002-3339-913X

Brett K. Hayes https://orcid.org/0000-0003-1415-0088

Constantin G. Meyer-Grant received support from the research training group Statistical Modeling in Psychology (SMiP; grant GRK 2277), funded by the German Research Foundation (DFG). Henrik Singmann and Samuel Winiger were supported by the Swiss National Science Foundation (SNSF; grant 100014_165591). This work was also supported by Australian Research Council Discovery Grant DP190102160 awarded to Brett K. Hayes.

We like to thank Dries Trippas for providing his materials. We also thank André Aßfalg for technical support with the online data collection.

Materials, participant data, and analysis scripts with complete output for all reported experiments can be found in the Open Science Framework archive https://osf.io/9avjc/.

Correspondence concerning this article should be addressed to Constantin G. Meyer-Grant, Abteilung Sozialpsychologie und Methodenlehre, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Engelbergerstraße 41, 79085 Freiburg, Germany. Email: constantin.meyer-grant@psychologie.uni-freiburg.de

Abstract

An ongoing debate in the literature on human reasoning concerns whether or not the logical status (valid vs. invalid) of an argument can be intuitively detected. The finding that conclusions of logically valid inferences are liked more compared to conclusions of logically invalid ones – called the logic-liking effect – is one of the most prominent pieces of evidence in support of this notion. Trippas et al. (2016) found this logic-liking effect for different kinds of inferences, including conditional and categorical syllogisms. However, all invalid conclusions presented by Trippas et al. (2016) were also impossible given the premises and had a particular structure of surface features – that is, an incongruent atmosphere. We present new data from five preregistered experiments in which we replicate the effect reported by Trippas et al. (2016) for conditional and categorical syllogisms, but show that this effect is eliminated when controlling for confounds in surface features. Moreover, we present evidence that there is a demand effect at play, which suggests that people are deliberately considering atmosphere cues of an argument to inform their liking ratings. Taken together, the findings of the present study cast doubt on the existence of logical intuitions.

Keywords: reasoning, liking ratings, logical intuition, demand effect, atmosphere effect

Are Logical Intuitions Only Make-Believe? Reexamining the Logic-Liking Effect

It is well known that people's judgments about whether an argument is logically valid can be tainted by vague supposition or gut feelings driven by content and context (e.g., Evans, 2002; Evans et al., 1983; Johnson-Laird & Byrne, 1991; Klauer et al., 2000; Tversky & Kahneman, 1974). A well-established explanation for such phenomena is that people tend to rely on a fast, heuristic evaluation of encountered arguments (Evans, 2008, 2009, 2018; Evans & Stanovich, 2013; Kahneman, 2011). In this context, it is often assumed that explicitly evaluating the validity of inferences is a "resource-demanding and effortful cognitive process that requires goal-directed manipulation and coordination of multiple mental representations" (Singmann et al., 2014, p. 1).

Dual-Process Models of Reasoning and DP 2.0

In traditional dual-process models of reasoning (e.g., Evans, 2008, 2018), logical processing of this kind is ascribed to analytic "Type 2" processes characterized as slow, controlled, context independent, goal-directed, and resource-demanding. These are complemented by "Type 1" processes described as fast, heuristic, context dependent, and making few demands on processing resources. Although Type 1 processes can sometimes deliver normatively correct responses, they do so for the wrong reasons; that is, they do not apply or respect logical and other normative constraints.

More recently, however, various studies suggested that normatively correct responses can be detected and produced in an intuitive, implicit way (*logical intuitions*; De Neys, 2012; De Neys & Pennycook, 2019; Thompson & Newman, 2018) by processes that are traditionally considered Type 1. For example, in the *conflict-detection paradigm* (De Neys, 2012), reasoners are presented problems that present cues of two kinds. One kind of cue (e.g., the believability of a conclusion) is believed to trigger a response via a heuristic Type 1 process, a second type of cue (e.g., the logical structure of the problem) is believed to trigger a response via a process that respects and applies logical or statistical rules. In conflict problems, both cues suggest different responses and a typical finding is that responses to conflict problems, whether normatively correct or not, are associated with increased response latencies and decreased confidence (e.g., De Neys & Glumicic, 2008; Thompson & Johnson, 2014). This suggests that both responses are elicited, resulting in a response conflict the resolution of which requires time and costs confidence. Such effects occur even under cognitive load and when strict response deadlines are imposed (e.g., Bago et al., 2020; Bago & De Neys, 2017), which is difficult to reconcile with the idea that processing according to logical or statistical rules is the exclusive domain of Type 2 processing (but see Klauer, in press).

This and related findings (see, e.g., Bago & De Neys, 2017) therefore question the assumption of traditional dual-process models that logical processing, characterized as a Type 2 process, needs to be slow and effortful. Other lines of research have questioned the assumption that logical processing is elicited only when the task demands logical analysis and thus, in a strategic, goal-dependent fashion. For example, Handley et al. (2011) asked participants to judge the believability of conclusions of logically valid and invalid problems. They found that conclusions of valid problems were judged more believable than conclusions of invalid problems. Similarly, effects of logical structure were found when participants were asked to rate how much they liked the conclusion (Morsanyi & Handley, 2012) as elaborated on below. Findings of this kind suggest that logical structure is spontaneously processed even though it is not relevant to the task at hand. In the automaticity literature (Bargh, 1994; Moors & De Houwer, 2006), unintentional processing of this kind is referred to as goal-independent processing, and goal independence is at odds with the idea that logical analysis is a Type 2 process that as such is strategically recruited and engaged with the goal to meet task instructions and demands. Instead, it suggests a more spontaneous, intuitive access to logicality.

Such considerations led to the development of second-generation dual-process models of reasoning – often referred to as “Dual-Process 2.0” (DP 2.0; e.g., De Neys, 2018; De Neys & Pennycook, 2019; Handley & Trippas, 2015). Like traditional dual-process models of reasoning (e.g., Evans, 2008, 2018), DP 2.0 theories distinguish between two distinct cognitive processes. However, DP 2.0 theories diverge from previous accounts by allowing for more flexibility in the role of each type of processing. Although they differ in detail, all DP 2.0 theories share the assumption that intuitive Type 1 processes are sensitive to both the content *and* the logical structure of text arguments, which is why – according to DP 2.0 – Type 1 processes underlie both logical intuitions and traditional heuristic-based intuitions.

One possible rationale for this phenomenon is that the application of simple logical principles will be automatized to a certain degree through consistent overlearning throughout one’s lifespan, which we refer to as the *automatization hypothesis* (De Neys, 2012; De Neys & Pennycook, 2019). According to the classical literature on automaticity (for a review, see Moors & De Houwer, 2006), automatization would be expected to lead to a decrease in processing resources required for logical analysis as well as to an increase in the speed of logical processing, and it might lead to a decrease in the dependence on explicit goals to process logical structure, that is to increased goal independence.

The Logic-Liking Effect

As already mentioned, a prominent finding supporting the existence of such intuitions is that people appear to take into account logicality of arguments in tasks that do not require logical analysis, such as when asked to judge the likability of a conclusion statement (e.g., Ghasemi et al., 2021; Morsanyi & Handley, 2012; Nakamura & Kawaguchi, 2016; Trippas et al., 2016). We follow Hayes et al. (2020) and henceforth refer to the sensitivity to argument validity in liking ratings as the *logic-liking effect.* At this point, we also want to introduce the superordinate term *structure effect* to describe any effect of inference structure on liking ratings. Thus, the logic-liking effect is one specific structure effect that describes an effect of logical necessity on liking ratings.

One explanation of the effect stems from the automatization hypothesis. In the course of automatization, simple logical analyses become automatized acquiring the classical automaticity feature of goal independence and thus, logical analysis is increasingly conducted in the absence of intentions to evaluate logicality. The outcome of goal-independent logical analysis is experienced as a logical intuition that has the power to color liking ratings such that a feeling of truth facilitates a positive rating.

Morsanyi and Handley (2012; see also Trippas et al., 2016) proposed another explanation of the logic-liking effect – the so-called *conceptual fluency hypothesis* – that differs from the automatization hypothesis outlined above in that it assumes that logical validity elicits changes in affect which in turn mediate the logic-liking effect. More precisely, Morsanyi and Handley (2012) suggested that people automatically construct a mental model (Johnson-Laird, 1983) representing the state of affairs when reading the premises of an argument. They further argued that a valid conclusion is processed with higher *conceptual fluency*, as it can be more readily integrated with the premises into a coherent model. According to Morsanyi and Handley (2012) and Trippas et al. (2016), a higher conceptual fluency elicits a slightly more positive affect, which should be reflected in higher liking ratings (but see Hayes et al., 2020). Importantly, “logical arguments should give rise to feelings of conceptual fluency even when the task does not explicitly call for reasoning” (Trippas et al., 2016, p. 1449). This implies that logical intuitions should be goal-independent and nondeliberate – that is, “at least partly opaque to conscious understanding or introspection” (Trippas et al., 2016, p. 1448).

Confounds in Studies of the Logic-Liking Effect

Morsanyi and Handley (2012) also conducted a series of experiments in which they presented categorical syllogisms to participants and found higher liking ratings for valid inferences compared to invalid ones. However, as they themselves pointed out, the syllogisms they used are prone to correlations of superficial features with logical status. The logic-liking effect found in Morsanyi and Handley's (2012) Experiments 1 and 3 might arise because of a *figural bias* (e.g., Johnson-Laird & Byrne, 1991; Johnson-Laird & Bara, 1984), since syllogistic figure and conclusion direction was confounded with logical validity in the used materials. More specifically, the position in which the propositions appeared in the premises on the one hand and in the conclusion on the other hand was concordant for valid syllogisms (e.g., "all S are M; all M are P; therefore, all S are P") and discordant for invalid ones (e.g., "all S are M; all M are P; therefore, all P are S").

Another issue with Morsanyi and Handley's (2012) study was raised by Klauer and Singmann (2013; see also Singmann et al., 2014), who pointed out that in the materials of Experiments 2 and 4, logical validity was accidentally confounded with other surface features of the syllogisms as well as with the material's content. The results by Klauer and Singmann (2013) as well as Singmann et al. (2014) suggest that there might in fact be no logic-liking effect when content is properly counterbalanced between conditions. However, Trippas et al. (2016) were able to replicate a logic-liking effect across arguments based on different logical forms (e.g., categorical syllogisms, conditional syllogisms, and disjunctions) with counterbalanced content, creating new confidence in the existence of the logic-liking effect (see Hayes et al., 2020 as well as Ghasemi et al., 2021 for replications of these effects).

Yet, certain features are still confounded with logical status in the materials used by Trippas et al. (2016). For example, they presented arguments for which all invalid conclusions were also impossible given the premises (i.e., they were *determinately invalid*). That means that there is no state of affairs in which both the conclusion and the premises are true. However, certain invalid inferences (viz., *indeterminately invalid* inferences) can also describe a state of affairs in which conclusion and premises are possible although the premises do not necessitate the conclusion. Thus, if possible conclusions are liked more than impossible ones, this could have been the source of the supposed logic-liking effect reported by Trippas et al. (2016). In other words, what participants might do when reading the statements is not intuitive reasoning but merely the attempt to build a coherent model of premises and conclusion as an automatic part of normal reading and text-comprehension processes. Building such a model is possible for both valid as well as indeterminately invalid arguments, but not for determinately invalid arguments and success in model construction may lead to better liking than failure.

Furthermore, the inferences in Trippas et al. (2016) experiments all confound logical validity with certain surface features. For example, the well-known *atmosphere effect* in syllogistic reasoning (Sells, 1936; Woodworth & Sells, 1935) was characterized by Begg and Denny (1969) as follows: “Whenever at least one premise is negative, the most frequently accepted conclusion will be negative; whenever at least one premise is particular [(i.e., including “some”)], the most frequently accepted conclusion will likewise be particular; otherwise the bias is towards affirmative and universal [(i.e., not including “some”)] conclusions.” (as cited in Johnson-Laird and Steedman, 1978, pp. 86-87; see also Khemlani and Johnson-Laird, 2012). All valid syllogisms in Trippas et al. (2016) were congruent with the atmosphere effect (e.g., “all S are M; no M are P; therefore, no S are P”), whereas all invalid syllogisms did not conform to it (e.g., “all S are M; no M are P; therefore, some S are P”). In the following, we will extend the use of the term “atmosphere effect” to describe an effect of the structure of surface features in general.

An atmosphere effect (with regard to the negation structure) is therefore also found for conditional inferences: Given the major premise “if p then q”, the most frequently accepted conclusion is positive when the minor premise is positive and negative when the minor premise is negative. This is a strong effect that is revealed when the inferences traditionally studied are contrasted with what Oaksford et al. (2000) called the converse inferences that alter the negation structure by switching the polarity of the proposition in the conclusion of the original inferences (e.g., “if p then q; p; therefore, not-q” instead of "if p then q; p; therefore, q"; see also Klauer et al., 2010).[1] Again, all valid conditional inferences in Trippas et al. (2016) were congruent with this atmosphere effect; all invalid conditional inferences did not conform to it.

Finally, considering disjunctive syllogisms, it is possible that atmosphere would take a different form: For the major premise "either p or q", the preferred conclusions might be positive when the minor premise is negative and negative when the minor premise is positive. Again, all valid disjunctive inferences in Trippas et al. (2016) conform to this atmosphere, whereas all invalid ones are incongruent with it.[2] However, other than for conditional and categorical syllogisms, these particular atmosphere conditions are inextricably tied to logical validity for disjunctive inferences. We therefore disregard disjunctive inferences in the following, as we believe that their investigation would not be diagnostic for the research question at hand.

In summary, atmosphere (indicated by a certain structure of surface features, such as negations or quantifiers) was always congruent for logically valid inferences and never congruent for logically invalid inferences in Trippas et al. (2016). This entails that such atmosphere effects could also be responsible for the observed emergence of a supposed logic-liking effect; ergo, it is possible that what appears to be intuitive sensitivity to logic is in fact sensitivity to the surface structure of the text arguments. That is, people may like certain arguments not because they are valid but because their surface features makes

[1] Note that "positive" and "negative" here refer to the propositions p and q as they occur in the conditional statement. The propositions p and q may themselves be phrased as negations in which case "positive" means that the respective proposition from the conditional premise occurs with the same polarity as minor premise or conclusion and "negative" means that its negation is presented as minor premise or conclusion.

[2] We refrain from speculating on the exact causes of such an atmosphere effect for disjunctions, although plausible explanations (e.g., differences in familiarity with certain surface features in disjunctive arguments and – as a consequence – facilitated or deteriorated comprehensibility or readability of the conclusion) are not very difficult to conceptualize. Rather, the point here is that simple heuristics based on surface features of disjunctive syllogisms might be sufficient to account for this particular logic-liking effect as well.

them, for example, easier to read or comprehend. The converse may also be true, certain surface features that, for example, make a text argument more structurally complex may be disliked, regardless of their logical status.[3]

The Present Research

Here we address those issues by reexamining the logic-liking effect. Besides trying to replicate the findings by Trippas et al. (2016), we aim at evaluating alternative accounts in terms of the confounds outlined above that could explain the ostensible effect of validity on liking ratings in Trippas et al. (2016). In doing so, we want to clarify whether the mechanisms specified by both the automatization hypothesis and the conceptual fluency hypothesis respond to logical validity or are driven by other features of the argument (viz., possibility and/or atmosphere-congruency). To this end, we investigate whether an effect of logicality on liking ratings can still be observed when confounds in terms of possibility and atmosphere are held constant between logically valid and logically invalid arguments. Our first research question thereby assesses the alleged logicality of logical intuitions. A second research question that we pursue addresses the alleged intuitive, non-strategic nature of logical intuitions by assessing their possible dependence on task demands.

Experiment 1

Experiments 1 to 3 focus on conditional inferences. As stated above, all invalid arguments in Trippas et al. (2016) were determinately invalid and had an incongruent atmosphere. However, indeterminately invalid arguments are in fact easily constructed for conditional inferences and can exhibit both a congruent or incongruent atmosphere.

[3] We acknowledge that the question of why and in which facets atmosphere effects arise is an interesting research question (see, e.g., Begg & Denny, 1969; Chater & Oaksford, 1999; Oaksford et al., 2000; Wetherick & Gilhooly, 1995, for promising starting points). Our research question here is, however, a different one; namely, whether or not intuitive processes are sensitive to logicality per se.

As in Experiment 1 by Trippas et al. (2016), we used valid *modus ponens* (MP) and *modus tollens* (MT) arguments, as well as determinately invalid MP' and MT' converse arguments, which were generated by switching the polarity of the proposition in the conclusion of MP and MT inferences, respectively. Additionally, we augmented the design by Trippas et al. (2016) by adding further types of indeterminately invalid arguments. More precisely, we included arguments *affirming the consequent* (AC) and *denying the antecedent* (DA) as well as AC' and DA' converse arguments, which were likewise generated by switching the polarity of the proposition in the conclusions of AC and DA inferences, respectively. An overview of the inference types used can be found in Table 1. The indeterminately invalid AC and DA inferences are similar to the valid MP and MT inferences in that the minor premise and conclusion either both have the same polarity with respect to the propositions in the conditional (MP and AC) or are both negated (MT and DA). That is, they are congruent with respect to the above-described atmosphere effect. On the other hand, AC' and DA' are similar to MP' and MT' in that one and only one of minor premise and conclusion is negated with respect to the conditional; hence they run counter the atmosphere effect. As far as we know, it is impossible to generate valid conditional syllogisms which are atmosphere incongruent or determinately invalid conditional syllogisms which are atmosphere congruent. Therefore, all arguments we used were either valid with congruent atmosphere, indeterminately invalid with congruent or incongruent atmosphere, or determinately invalid with incongruent atmosphere. The affiliation of an argument to one of those four categories will henceforth be called its *conclusion status* (see Table 1).

We expect to replicate the finding reported by Trippas et al. (2016) that in terms of liking ratings, conclusions of valid problems should receive on average higher values than determinately invalid conclusions. If only validity is responsible for the effect, the liking ratings should be highest for valid inferences, while there should be no difference between the remaining conditions. If on the other hand, the possibility of constructing a coherent

Table 1

The inferences types for conditional syllogisms

Type	Form (exemplary)	Conclusion status	
		Validity	Atmosphere
MP	If p then q; p; therefore q	Valid	Congruent
MT	If p then q; not-q; therefore not-p	Valid	Congruent
AC	If p then q; q; therefore p	Indet. invalid	Congruent
DA	If p then q; not-p; therefore not-q	Indet. invalid	Congruent
MP'	If p then q; p; therefore not-q	Det. invalid	Incongruent
MT'	If p then q; not-q; therefore p	Det. invalid	Incongruent
AC'	If p then q; q; therefore not-p	Indet. invalid	Incongruent
DA'	If p then q; not-p; therefore q	Indet. invalid	Incongruent

Note. Indet. = indeterminately; det. = determinately.

model (i.e., whether or not the conclusion is possible given the premises) is the decisive factor, there should be no difference in liking ratings between valid and indeterminately invalid inferences. If surface features relating to the congruency of atmosphere (i.e., negation structures) play a role, then we expect to find the main differences between original and converse inferences (i.e., MP, MT, AC, and DA arguments receiving on average higher ratings than MP', MT', AC', and DA' arguments).

In addition to these main hypotheses, we also expected to observe an effect of believability as found in previous studies. Note that we followed Trippas et al. (2016) such that believability for conditional inferences only refers to whether minor premise and conclusion describe a believable versus unbelievable state of affairs (e.g., "The child is happy. Therefore, the child is laughing." vs. "The child is happy. Therefore, the child is crying."). However, believability is not of major concern for answering the current research question and is included mainly for comparability of the present study with Trippas et al.

(2016).

Methods

Experiment 1 was a preregistered lab-study (see Open Science Framework registration https://osf.io/j4xp3/ for further details).[4]

Participants and ethics statement

Fifty-two participants (36 females, 16 males) aged between 16 and 36 ($M_{age} = 23.44$, $SD_{age} = 3.69$), fifty-one of which were undergraduates of the University of Freiburg with diverse majors, took part in the lab-study in exchange for either partial course credit or a small monetary compensation. People with expertise regarding logical reasoning were not permitted to participate.

In Germany no ethics approval is required if the research objectives do not refer to issues regulated by medical law. Since none of our studies has such objectives, no approval was required. Participation was voluntary, informed consent was obtained from each participant prior to the study, and all collected data were anonymized.

Design

The inference type (MP, MT, AC, DA, MP', MT', AC', and DA'), determined by crossing the two factors conditional type (MP/MP' vs. MT/MT' vs. AC/AC' vs. DA/DA') and negation structure (original = MP/MT/AC/DA vs. converse = MP'/MT'/AC'/DA'), as well as argument believability (believable vs. unbelievable) were manipulated within subjects.

[4] Note that we deviate partially from some of the analysis strategies outlined in the Open Science Framework registrations in order to adhere to a consistent analysis strategy across all of our experiments. The points of deviation are described in the analysis scripts provided in the respective folders in the Open Science Framework archive https://osf.io/9avjc/, which additionally presents the preregistered analyses (analysis scripts and complete outputs) for all experiments.

Materials

We used 64 different arguments for each participant (eight arguments per inference type). Half of the arguments (four arguments of each inference type) comprised a believable combination of minor premise and conclusion (e.g., “The child is happy. Therefore, the child is laughing.”), while the other half did not (e.g., “The child is happy. Therefore, the child is crying.”). In accordance with Trippas et al. (2016), we used only implicit negations. The four replicates resulted from the fact that equivalent inference types and believability conditions arise when either the direction of the argument is reversed (e.g., “If a child is laughing, then it is happy. The child is laughing. Therefore, the child is happy.” vs. “If a child is happy, then it is laughing. The child is happy. Therefore, the child is laughing.”) or the polarities of all propositions are reversed (e.g., “If a child is laughing, then it is happy. The child is laughing. Therefore, the child is happy.” vs. “If a child is crying, then it is sad. The child is crying. Therefore, the child is sad.”).

Only MP and MT inferences are valid. MP’ and MT’ inferences, on the other hand, were determinately invalid – that is, invalid and impossible. AC, DA, AC’, and DA’ inferences were indeterminately invalid – that is, invalid but possible. Moreover, the converse inferences (MP’, MT’, AC’, DA’) have an incongruent atmosphere regarding the negation structure of the conditional statement on the one hand and minor premise and conclusion on the other while the original inferences (MP, MT, AC, DA) have a congruent atmosphere. Recall that an incongruent atmosphere in this context means that if the two terms in the first premise have the same polarity (i.e., being either both negated or both not negated), the two terms in the second premise and conclusion have opposite polarities (i.e., one being negated and the other one not) or vice versa. Conversely, a congruent atmosphere means that if the two terms in the first premise have the same polarity (or opposite polarities) then so do the two terms in the second premise and conclusion.

We used 32 different German-language contents modeled after the contents used by Trippas et al. (2016). These contents were randomly assigned to each of the 64 arguments for each participant individually (see Open Science Framework archive https://osf.io/9avjc/ for copies of all materials as well as their translation into English). Hence, each specific item content was equally likely to appear in each inference type and believability condition. Moreover, we presented each of the 64 arguments twice, but with different content; thus, participants saw a total of 128 unique trials and each content was presented exactly four times.

Procedure

The procedure closely followed Experiment 1 by Trippas et al. (2016). Hence, we instructed participants to read the sentences carefully and then rate how much they like the final sentence on a 6-point Likert scale from 1 ("dislike it very much") to 6 ("like it very much"). The instructions stated that "when you make the liking judgment focus on your feelings about the statement. Don't think about why you like or dislike the statement, just go with your intuition and gut feelings" (Trippas et al., 2016, p. 1451).

In each trial, participants were first presented with the major premise for 2.25 s, then with the minor premise for 2.25 s, followed by the conclusion and the response scale. We choose a presentation duration of 2.25 s (instead of 2 s presentation intervals used by Trippas et al., 2016) because our materials were approximately 12.5% longer than the materials of Trippas et al. (2016; mean number of characters for the conditionals is 47.8 for Trippas et al. and 53.7 for our materials). The difference is accounted for by differences in the English and German language.

The trials were presented in randomized order. After each quarter of trials, participants were given the chance for a short break. We additionally presented another MP argument as a warm-up based on a different content prior to the 128 experimental trials.

Results

Analysis approach

We used linear mixed model analyses with crossed random effects for participants and material contents (Judd et al., 2012).

Model selection regarding the random-effect structure was addressed by a backwards selection approach. We first conducted two separate backwards model selection procedures including only one of the two random-effect factors (i.e., either participants or material contents). Each of those two selection procedures started with the respective maximal random-effect structure. Given the complexity of the random effects structure and the comparatively limited data, we omitted the correlations among random effect parameters from all models. If a model failed to converge or showed a singular fit, we reduced the random-effect structure by excluding the random effect with the smallest estimated variance. Exclusion did not violate the principle of marginality. We stopped at the first random-effect structure for each of the two random-effect factors that converged and led to a nonsingular fit (Barr et al., 2013; cf. Matuschek et al., 2017). These random-effect structures were then combined and served as a starting point for a final model selection procedure containing both random-effect factors. This was accomplished by another backwards selection approach akin to the two previous ones – that is, the random-effect structure was iteratively reduced until a converging model without singular fit emerged.

The *p*-values for fixed effects in the final model as well as the *p*-values for linear contrasts were computed using the Satterthwaite approximation for degrees of freedom, since the Kenward-Roger approximation for degrees of freedom was computationally infeasible (see, e.g., Singmann & Kellen, 2019, for a brief commentary on this issue).

Liking ratings

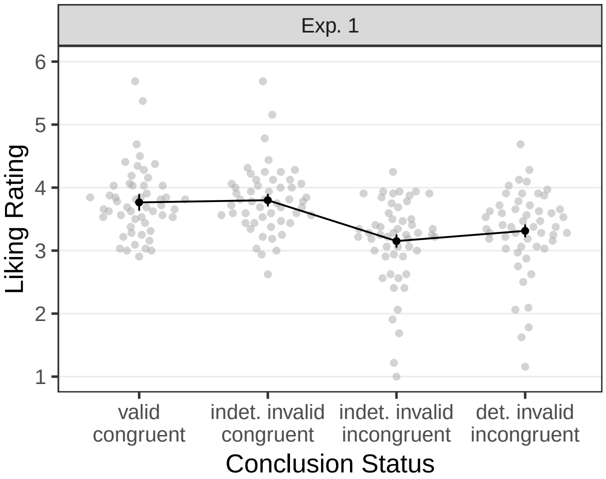
The liking ratings were first submitted to an analysis in which we only included the fixed-effect within-subjects factor conclusion status (valid vs. indeterminately invalid with congruent atmosphere vs. indeterminately invalid with incongruent atmosphere vs. determinately invalid).[5] This allowed us to visualize the relevant patterns in the data in a simple fashion. The existence of a main effect of conclusion status was strongly supported by our data, $F(3, 104.48) = 16.87$, $p < .001$.

Figure 1 shows the mean and individual liking ratings as a function of conclusion status. The ratings are clearly higher for arguments with congruent atmosphere and lower for arguments with incongruent atmosphere, whereas there seems to be no noticeable difference between atmosphere-congruent, indeterminmately invalid and valid arguments as well as between atmosphere-incongruent, indeterminately invalid and determinately invalid inferences.

Figure 1

Mean (black symbols) and individual (gray symbols) liking ratings in Experiments 1 as a function of conclusion status. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

Indet. = indeterminate; det. = determinate.



To further investigate specific contrasts of interest, we conducted an analysis in

[5] The final random-effect structure included random intercepts for participants and material contents as well as by-participant random slopes for conclusion status.

terms of the full study design in which we included the within-subjects factors conditional type (MP/MP' vs. MT/MT' vs. AC/AC' vs. DA/DA'), negation structure (original vs. converse), and believability (believable vs. unbelievable) as fixed effects.[6] A depiction of the liking ratings from Experiment 1 broken down by inference type can be found in the appendix (see Figure A1). To see whether we replicate greater liking of conclusions of valid relative to conclusions of determinately invalid arguments as reported by Trippas et al. (2016), we calculated a linear contrast comparing these two types of inferences. Results ($d = 0.45$,[7] $t(62.30) = 3.23$, $p = .002$) indicate that the replication was successful. To see whether we also replicate greater liking of believable than unbelievable conclusions, another linear contrast juxtaposed these two types of inferences. Results ($d = 0.51$, $t(51.00) = 5.43$, $p < .001$) again indicate a successful replication. A third linear contrast addressed the question whether there was an effect of logical validity per se when the confoundings in terms of possibility and atmosphere are held constant. The contrast juxtaposes valid inferences (MP and MT) and the indeterminately invalid inferences DA and AC, all of which have a congruent atmosphere. Results ($d = -0.03$, $t(6436.70) = -0.64$, $p = .523$) indicate that there is no effect of validity per se (see also Table A1 in the appendix for a summary of these effects across all experiments). A contrast comparing atmosphere-congruent and atmosphere-incongruent inferences suggests the presence of a strong atmosphere effect ($d = 0.55$, $t(53.70) = 4.08$, $p < .001$). This effect is also apparent when validity and possibility are held constant by juxtaposing indeterminately invalid, atmosphere-congruent inferences (AC and DA) and indeterminately invalid, atmosphere-incongruent inferences (AC' and DA'; $d = 0.65$, $t(62.40) = 4.63$, $p < .001$).

[6] The final random-effect structure included random intercepts for participants and material contents as well as by-participant random slopes for negation structure and believability and by-content random slopes for negation structure.

[7] Note that for each linear contrast, we always report the simple effect size d which represents the estimated difference on the response scale (Baguley, 2009; Pek & Flora, 2018). For example, $d = 0.45$ indicates that there was a difference of almost half a point on the response scale from 1 to 6.

Finally, we assessed the role of possibility versus impossibility while holding logical validity and atmosphere-congruency constant by contrasting indeterminately invalid inferences with incongruent atmosphere (AC' and DA') and determinately invalid inferences (MP' and MT'). This contrast seems to suggest an effect of possibility that is the opposite of the hypothesized effect ($d = -0.16$, $t(6436.70) = -3.05$, $p = .002$); that is, possible inferences appear to be liked less than impossible ones.

Discussion

First, we replicated the structure effect reported by Trippas et al. (2016). More specifically, valid inferences were liked more compared to determinately invalid ones. Hence, when not controlling for the confounds in Trippas et al.'s (2016) study, conclusions of valid inferences appear to be liked more compared to conclusions of invalid ones. However, when controlling for a confounding by atmosphere, it becomes apparent that this effect is not a logic-liking effect, but rather a different structure effect (viz., an atmosphere effect). Arguments with a negation structure corresponding to a congruent atmosphere are liked more than arguments with a different negation structure (i.e. with an incongruent atmosphere). In contrast, if we compare liking ratings for valid inferences to those for indeterminately invalid inferences with congruent atmosphere, we fail to find convincing evidence of there being any difference.

Our results also suggest that the confound in terms of possible and impossible inferences is not responsible for the structure effect observed by Trippas et al. (2016), since the effect is opposite to what we had hypothesized (see the contrast between determinately invalid inferences MP'/MT' and indeterminately invalid inferences AC'/DA'). This implies that possibility attenuates liking ratings, which is surprising. We are cautious, however, in embracing this conclusion, because this effect of possibility on liking ratings did not replicate in Experiments 2 to 5.

Taken together, Experiment 1 suggests that Trippas et al.'s (2016) structure effect is not a logic-liking effect, but rather an atmosphere effect, reflecting surface features of the presented argument.

Experiment 2 & 3

Although the results were relatively clear-cut, our previous experiment shares one of the shortcomings of the study by Trippas et al. (2016); namely, the lack of explicit ratings of logical validity. Recent research on the topic suggests that liking judgments are in fact related to explicit reasoning. Nakamura and Kawaguchi (2016) demonstrated for example that reasoners who performed better in an explicit reasoning task also gave higher liking ratings to valid inferences. Hayes et al. (2020) recently found that working memory capacity could predict both explicit logic and affect rating tasks. This notion received further support by Ghasemi et al. (2021), who found that higher cognitive ability led to better performance in explicit logic ratings and a stronger logic-liking effect. Therefore, it seems that "the logic effect for liking and the logic effect for validity are strongly correlated and predict one another" (Ghasemi et al., 2021, p. 9). As acknowledged by Ghasemi et al. (2021), the simplest explanation for this phenomenon is that the decision makers are – at least partially – resorting to rate logical validity when asked to rate likability of the conclusion. We agree with this assessment. It seems that when instructed to rate the likability of a sentence, people face a somewhat vague task. Thus, they might deliberately choose to rate a more objective criterion (viz., logical validity) instead.

Additionally, the experimental materials and procedures make it unlikely that participants do not notice and acknowledge the logical structure of the presented inferences as well as variations therein in a conscious manner. Being asked to rate only the likability of the conclusion, while being consistently and obtrusively administered the premises preceding it, constitutes a gross violation of the Gricean maxim of quantity (Grice, 1989). According to the maxim of quantity, communications should give enough, but not too much information. Violations of the Gricean maxims in turn trigger Gricean implicatures on the part of the recipient of the communication, implying in the present case that the premises must be relevant for the task at hand (Sperber & Wilson, 1986; Wilson & Sperber, 1986) and that the experimenter expects participants to consider them for their judgments. This demand characteristic may thereby lead participants to attempt to assess cues to logical validity of the presented arguments and to let these cues influence their liking ratings. In other words, we suspect that a conscious evaluation of logical validity rather than logical intuitions factor into a person's liking ratings. This would imply that a congruent atmosphere simply constitutes an easily accessible heuristic cue for logical validity.

Let us emphasize, however, that in our view such a mechanism does not necessitate logic and liking ratings to be identical. Decision makers may very well be able to consider multiple characteristics of the presented arguments and integrate the available information into a final verdict when asked to judge a relatively vague aspect of the presented materials, such as likability. On the other hand, they might invest some extra effort that goes beyond merely using the atmosphere heuristic to assess logical validity, if rating logical validity is explicitly required.

In Experiment 2 and 3, we want to address these issues directly. Therefore, we employed a design which in many aspects resembles the previous one, but with the addition of a second block of trials, in which participants will be asked to explicitly rate logical validity. We suspect that any structure effect might simply be the result of a demand effect caused by an unclear instruction and/or by suggestive design choices leading to the liking rating responses being effectively performed – at least in part – as a logic rating. If such effects are indeed caused by a deliberate response strategy, they should be malleable by a manipulation of the task's demand characteristics. If, on the other hand, implicit (i.e., nondeliberate and/or automatic) processes are responsible for the occurrence of structure effects within liking ratings as proposed by both the conceptual fluency hypothesis (Morsanyi & Handley, 2012; Trippas et al., 2016) and the automatization hypothesis (De Neys & Pennycook, 2019), these effects should be goal-independent; that is, they should be independent of the task's demand characteristics.

Hence, we implemented two different instruction conditions, which were used in Experiment 2 and 3, respectively. In Experiment 2, we did not tell the participants in advance that there would be two different tasks. In Experiment 3, on the other hand, we informed the participants at the beginning of the experiment that there would be two different tasks, the first of which only concerns their feelings toward the conclusion, while the second only focuses on the logical structure of the whole inference. This instruction manipulation aims at reducing demand characteristics by implying that the inference structures will be relevant later on, which might prevent Gricean implicatures. Thus, we expect to observe response patterns in the liking ratings of Experiment 2 which match the ones observed in Experiment 1. In contrast, we expect to observe less pronounced structure effects in Experiment 3 compared to Experiments 1 and 2 if demand characteristics do in fact influence how participants approach rating likability.

We also decided to deviate from the design used by Trippas et al. (2016) as well as in our previous experiment in one additional aspect; that is, both studies used implicit negation throughout the whole experiment. We see a severe problem with this approach that arises when considering an MT inference as implicit negations are usually contraries while explicit negations are contradictions. An MT argument with only implicit negations would for example be, "If a child cries, then it is sad. The child is happy. Therefore, the child laughs". This is not a valid inference, since we are dealing with an inferential structure that is less akin to a modus tollens – that is, "if p then q; not-q; therefore, not-p" – than to something of the form "if p then q; q'; therefore, p'" (where p' and q' are implicit negations of p and q). However, the latter is clearly not a valid inference (although q' may imply not-q, not-p need not imply p'), while the former is. Since it is essential for our research question that supposedly valid conclusions are actually valid, we will only use explicit negations (e.g., "the child is not happy" instead of "the child is sad") in Experiments 2 and 3, which eliminates this problem.

Methods

Experiments 2 and 3 are both preregistered online-studies (see Open Science Framework registration https://osf.io/ws5yp/ for further details; see also Footnote 4).

Participants

Forty-nine participants (23 females, 26 males) aged between 18 and 68 ($M_{age} = 30.51$, $SD_{age} = 10.71$) completed Experiment 2 and fifty-one participants (18 females, 33 males) aged between 18 and 61 ($M_{age} = 28.84$, $SD_{age} = 10.53$) completed Experiment 3.[8] All participants were recruited via Prolific (Peer et al., 2017) and participated in exchange for a monetary compensation (£15.00). Inclusion criteria were an age between 18 and 80 and fluency in German. Participation in both experiments was not possible.

Design

Both experiments each followed a within-participant design with task as a blocked variable (first, the judgement of conclusion likability, followed by the judgement of logical validity). The inference type (MP, MT, AC, DA, MP', MT', AC', and DA'), determined by crossing the two factors conditional type (MP/MP' vs. MT/MT' vs. AC/AC' vs. DA/DA') and negation structure (original vs. converse) as well as argument believability (believable vs. unbelievable) were manipulated within subjects. The two different instruction conditions, on the other hand, were manipulated between subjects – that is, between the two experiments.

Materials

The materials were mostly identical to the materials of Experiment 1. However, as mentioned previously explicit negations were used instead of implicit ones (see Open

[8] We initially collected data from fifty participants for Experiment 2; however, one participant withdrew consent.

Science Framework archive https://osf.io/9avjc/ for copies of all materials as well as their translation into English).

Procedure

Both experiments consisted of two parts. The first part (henceforth also called liking task) was mostly identical to Experiment 1, while in the second part (henceforth also called logic task) participants were instead asked to rate whether the conclusion followed necessarily from the previously shown premises. For each participant, the second part contained exactly the same 128 trials as the first, although in a different randomized order. Since the experiments were carried out online and we had no direct control over the exact experimental setting, we decided to make the presentation of the sentences self-paced. However, each sentence was displayed for a minimum of 2 seconds. Moreover, participants were given the option to review the previous two sentences before they had to give an answer. Morsanyi and Handley (2012), for example, used a similar procedure in their Experiment 1.

For the logic task we instructed participants to read the sentences carefully and then rate how much they believe the argument to be a logically valid inference on a 6-point Likert scale from 1 (“definitely not logically valid”) to 6 (“definitely logically valid”). The instructions also stated that “logically valid” means that the state of affairs described by the last sentence necessarily follows from the two previous sentences. We asked participants to very carefully consider this fact for their responses during the logic task.

The only difference between Experiment 2 and 3 was – as mentioned earlier – a change in the instructions given to the participants at the beginning of the study. That is, in Experiment 3, participants were informed about there being two parts with two different tasks prior to the liking task. On this occasion it was also pointed out that they are supposed to rate only likability of the conclusion in the first part and only logical validity of the inference in the second part. Contrary to this, participants of Experiment 2 were initially left completely ignorant about there being two different tasks.[9]

At the end of both experiments, participants were asked to indicate whether they actually considered likability of the last statement, logical validity of the inference, or both for their responses during the first part of the study (i.e., during the liking task).

Results

Analysis approach

We again used linear mixed model analyses with crossed random effects for participants and material contents to analyze participants' liking and logic ratings. Model selection regarding the random-effect structure was addressed as for Experiment 1. We also included participants' reported response behavior as a fixed-effect factor in one of the mixed model analyses to see whether it affected their liking ratings. To this end we created a between-subjects factor with two levels, participants that only rated likability versus participants that rated only validity or used both likability and validity.

We, additionally, analyzed the response behavior self-reports itself with a Wilcoxon-Mann-Whitney test. The ranks were assigned according to their reported response behavior (1 = rated likability, 2 = rated likability and logical validity, 3 = rated logical validity). This approach was chosen since the different response options indicate different degrees of perceived demand. In other words, the stronger the demand effect, the more one is drawn to rate logical validity of the inference instead of likability of the conclusion in the liking task. Thus, someone who stated rating only logical validity of the inference in the liking task can be assumed to have experienced a stronger demand effect than someone who considered both aspects for their liking rating.

[9] Note, however, that the instructions for both the logic and the liking tasks themselves, which included asking participants to carefully read all consecutively presented sentences, were identical in both instruction conditions.

Response behavior self-report

In Experiment 2, five participants reported that they had rated only logical validity of the inference in the liking task while seventeen reported that they had considered both logical validity of the inference and likability of the conclusion. In Experiment 3, six participants reported that they had considered both logical validity of the inference and likability of the conclusion in the liking task. All remaining participants reportedly rated only likability of the conclusion. A Wilcoxon-Mann-Whitney test suggests that these ordinal rank distributions are different between the two experiments ($W = 1665.00$, $p < .001$).

Liking ratings

The liking ratings of both experiments were first submitted to a joint analysis in which we only included the within-subjects factor conclusion status (valid vs. indeterminately invalid with congruent atmosphere vs. indeterminately invalid with incongruent atmosphere vs. determinately invalid) as well as the between-subjects factors instruction condition (Experiment 2 vs. Experiment 3) and self-reported response behavior during the liking task (rated only likability vs. rated only validity or both) as fixed effects.[10] There was strong evidence for a main effect of conclusion status, $F(3, 117.17) = 31.60$, $p < .001$. Besides that, the analysis revealed interaction effects between conclusion status and instruction condition, $F(3, 117.19) = 8.54$, $p < .001$, as well as between conclusion status and response behavior, $F(3, 117.17) = 12.47$, $p < .001$. All remaining effects had p-value equal to or greater than .085 ($p = .085$ was observed for the main effect of self-reported response behavior).

Figure 2 shows the mean and individual liking ratings as a function of conclusion

[10] The final random-effect structure included random intercepts for participants and material contents, by-participant random slopes for conclusion status and instruction condition, and by-content random slopes for response behavior.

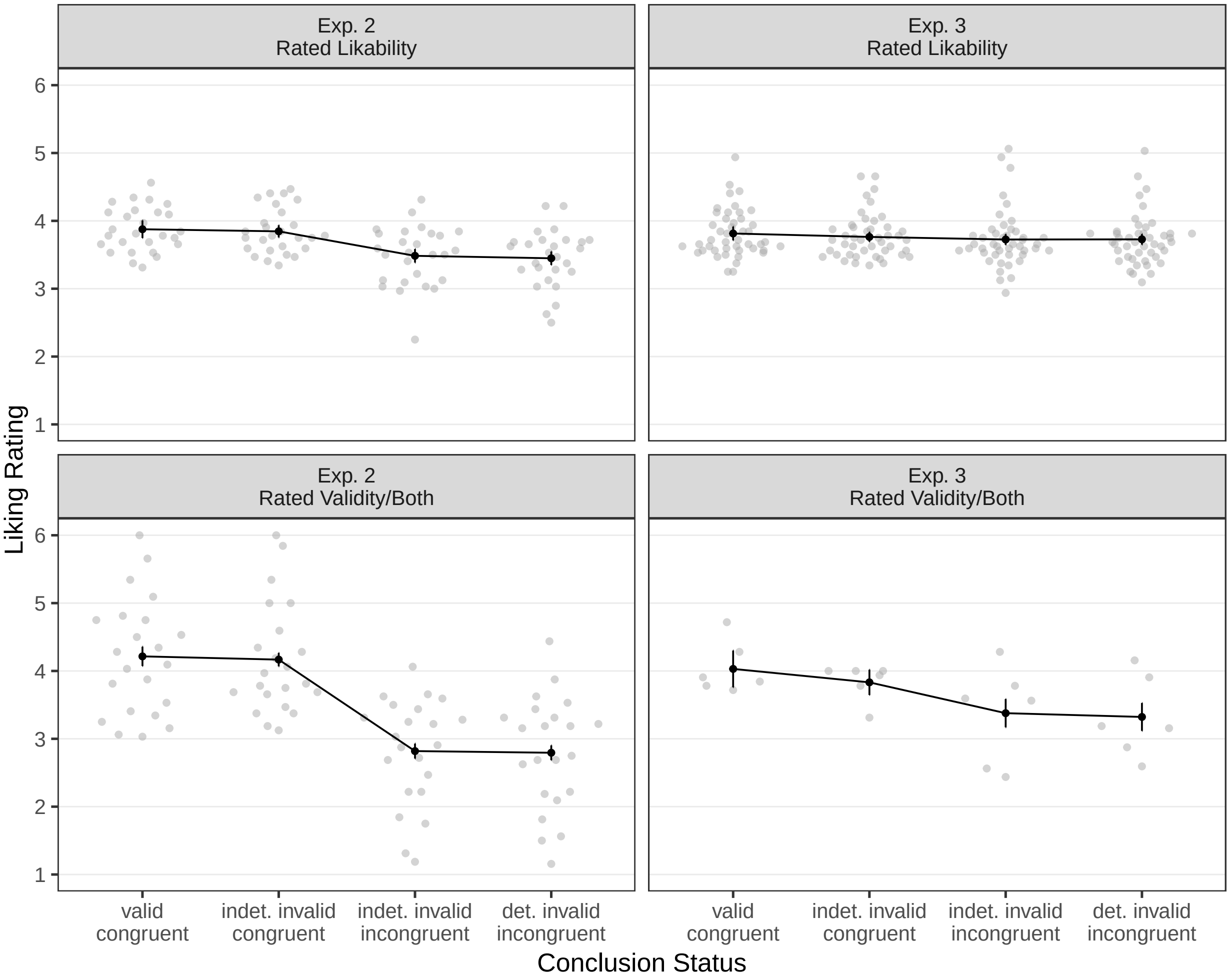
status separately for different groups defined by self-reported response behavior (only likability vs. only validity or both) and instruction condition (Experiment 2 vs. Experiment 3). The patterns mirror the ones observed in Experiment 1. That is, the ratings tend to be higher for valid and indeterminately invalid arguments with congruent atmosphere and lower for determinately invalid and indeterminately invalid arguments with incongruent atmosphere, whereas there seems to be no noticeable difference between either the first two or the last two conditions. Moreover, we can see clearly that this difference is more prominent in Experiment 2 compared to Experiment 3 as well as for those participants who reported that they additionally (or exclusively) considered logical validity of the inference during the liking task. The effect almost completely vanishes for those participants of Experiment 3 who reported that they only considered likability of the conclusion in their liking ratings.

To investigate the contrasts of interest, we analyzed the liking ratings for each experiment in two separate analyses in terms of the full design. Hence, we included the within-subjects factors conditional type (MP/MP' vs. MT/MT' vs. AC/AC' vs. DA/DA'), negation structure (original vs. converse), and believability (believable vs. unbelievable) as fixed effects.[11] Depictions of the liking ratings from Experiment 2 and 3 broken down by inference type can be found in the appendix (see Figures A2 and A3). To assess whether we still replicate greater liking of conclusions of valid relative to conclusions of determinately invalid arguments as reported by Trippas et al. (2016), we again calculated a linear contrast comparing these two types of inferences. Results (Exp. 2: $d = 0.87$, $t(51.40) = 5.17$, $p < .001$; Exp. 3: $d = 0.16$, $t(103.00) = 2.53$, $p = .013$) indicate that the replication was successful. However, the difference is more pronounced in Experiment 2

[11] The final random-effect structure for both analyses included random intercepts for participants and material contents as well as by-participant random slopes for negation structure and believability. The final random-effect structure for Experiment 2 additionally included a by-participant random slope for the interaction between negation structure and believability.

Figure 2

Mean (black symbols) and individual (gray symbols) liking ratings of Experiment 2 (left panels) and 3 (right panels) as a function of conclusion status. Liking ratings of participants who reported rating only likability of the conclusion are displayed in the two upper panels, while liking ratings of participants who reported rating also (or exclusively) logical validity of the inference are displayed in the lower panels. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based). Indet. = indeterminate; det. = determinate.



than in Experiment 3. To see whether we also replicate greater liking of believable than unbelievable conclusions, we also juxtaposed these two types of inferences. Results (Exp. 2: $d = 0.75$, $t(48.00) = 6.73$, $p < .001$; Exp. 3: $d = 0.35$, $t(50.00) = 3.26$, $p = .002$) again indicate a successful replication. The effect is likewise more pronounced for Experiment 2 than for Experiment 3. Another contrast addressed the question whether there was an effect of logical validity per se when the confoundings in terms of possibility and atmosphere are held constant. The contrast juxtaposes valid inferences (MP and MT) and indeterminately invalid inferences with congruent atmosphere (DA and AC). Results (Exp. 2: $d = 0.04$, $t(6064.40) = 0.82$, $p = .411$; Exp. 3: $d = 0.07$, $t(6359.30) = 1.37$, $p = .172$) indicate that there is no effect of validity per se (see also Table A1 in the appendix). Contrasting atmosphere-congruent and incongruent inferences suggests the presence of an atmosphere effect (Exp. 2: $d = 0.84$, $t(48.00) = 5.05$, $p < .001$; Exp. 3: $d = 0.13$, $t(50.10) = 2.37$, $p = .022$). Again, this effect is more pronounced in Experiment 2 where it is still detectable even when validity and possibility are held constant by juxtaposing indeterminately invalid, atmosphere-congruent inferences (AC and DA) and indeterminately invalid, atmosphere-incongruent inferences (AC' and DA'; $d = 0.80$, $t(51.40) = 4.75$, $p < .001$). However, the same contrast does not reach statistical significance in Experiment 3 ($d = 0.09$, $t(103.10) = 1.42$, $p = .160$). We again assessed the role of possibility versus impossibility while holding logical validity and atmosphere-congruency constant by contrasting indeterminately invalid, atmosphere-incongruent inferences (AC' and DA') and determinately invalid inferences (MP' and MT'). These contrast provided essentially no evidence for a role of possibility in either experiment (Exp. 2: $d = 0.03$, $t(6064.10) = 0.77$, $p = .442$; Exp. 3: $d = 0.00$, $t(6360.40) = 0.07$, $p = .941$).

Logic Ratings

As with the liking ratings, we first analyzed the logic ratings of Experiment 2 and 3 together. We therefore included the within-subjects factor conclusion status (valid vs. indeterminately invalid with congruent atmosphere vs. indeterminately invalid with incongruent atmosphere vs. determinately invalid) as well as the between-subjects factor instruction condition (Experiment 2 vs. Experiment 3) as fixed effects.[12] This analysis clearly revealed a main effect of conclusion status, $F(3, 177.97) = 301.65$, $p < .001$. All remaining effects had p-values equal to or greater than .407 ($p = .407$ was observed for the interaction effect of conclusion status with instruction condition).

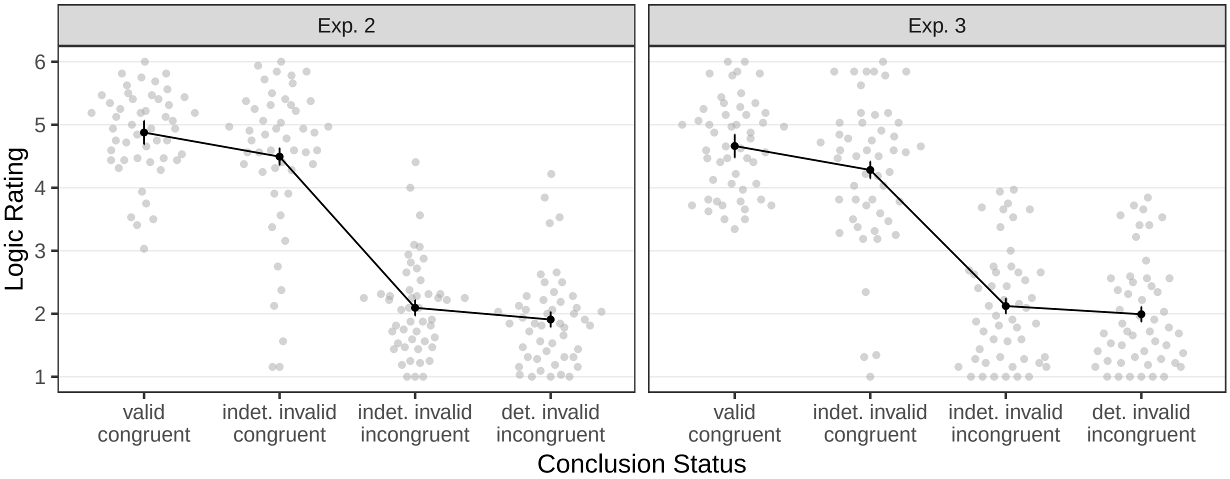
Figure 3 shows the mean and individual logic ratings as a function of conclusion status separately for different groups defined by the instruction condition (Experiment 2 vs. Experiment 3). The patterns are qualitatively similar to the ones observed in the liking task. That is, the ratings are clearly higher for valid and indeterminately invalid, atmosphere-congruent arguments and lower for determinately invalid and indeterminately invalid, atmosphere-incongruent arguments. However, we can see that the ratings for valid inferences are even higher than for indeterminately invalid inferences with congruent atmosphere, although this difference appears to be considerably smaller compared to the effect of surface features. In other words, there seems to be a strong atmosphere effect as in the liking ratings, but also a small effect of logical validity per se.

Mirroring the analysis of the liking ratings, we analyzed the logic ratings for each experiment in two separate analyses, in which we included the within-subjects factors conditional type (MP/MP' vs. MT/MT' vs. AC/AC' vs. DA/DA'), negation structure

[12] The final random-effect structure included random intercepts for participants and material contents as well as by-participant and by-content random slopes for conclusion status and instruction condition.

Figure 3

Mean (black symbols) and individual (gray symbols) logic ratings of Experiment 2 (left panel) and 3 (right panel) as a function of conclusion status. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based). Indet. = indeterminate; det. = determinate.



(original vs. converse), and believability (believable vs. unbelievable) as fixed effects.[13] Depictions of the logic ratings from Experiment 2 and 3 broken down by inference type can be found in the appendix (see Figures A4 and A5). We calculated the same linear contrasts for the logic ratings as we did for the liking ratings. Thus, to evaluate whether valid inferences were endorsed more strongly than determinately invalid arguments, we calculated a contrast which compared these two types of inferences. Results (Exp. 2: $d = 2.97$, $t(65.30) = 16.34$, $p < .001$; Exp. 3: $d = 2.66$, $t(60.20) = 13.39$, $p < .001$) indicate that this was indeed the case. To see whether believable inferences were endorsed more than unbelievable ones, we juxtaposed these two types of inferences. Results (Exp. 2:

[13] The final random-effect structure for both analyses included random intercepts for participants and contents as well as by-participant random slopes for conditional type, negation structure, believability, and for the interaction between conditional type and negation structure.

$d = 0.49$, $t(48.00) = 5.40$, $p < .001$; Exp. 3: $d = 0.45$, $t(50.00) = 5.78$, $p < .001$) indicate that this was the case as well. To address the question whether there was an effect of logical validity per se when the confoundings in terms of possibility and atmosphere are held constant we juxtaposes valid inferences (MP and MT) and indeterminately invalid inferences with congruent atmosphere (DA and AC). Results (Exp. 2: $d = 0.38$, $t(151.20) = 4.31$, $p < .001$; Exp. 3: $d = 0.38$, $t(146.20) = 4.56$, $p < .001$) indicate that there is an effect of validity per se (see also Table A2 in the appendix). Comparing atmosphere-congruent and incongruent inferences suggests the presence of an atmosphere effect (Exp. 2: $d = 2.68$, $t(48.00) = 16.02$, $p < .001$; Exp. 3: $d = 2.41$, $t(50.00) = 12.72$, $p < .001$). This effect is also apparent when validity and possibility are held constant by juxtaposing indeterminately invalid, atmosphere-congruent inferences (AC and DA) and indeterminately invalid, atmosphere-incongruent inferences (AC' and DA'; Exp. 2: $d = 2.40$, $t(65.30) = 13.20$, $p < .001$; Exp. 3: $d = 2.16$, $t(60.20) = 10.85$, $p < .001$). Finally, we also assessed the role of possibility versus impossibility while holding logical validity and atmosphere-congruency constant by contrasting indeterminately invalid inferences with incongruent atmosphere (AC' and DA') and determinately invalid inferences (MP' and MT'). Although there is a significant difference in Experiment 2 ($d = 0.19$, $t(151.20) = 2.14$, $p = .034$), this is not the case for Experiment 3 (Exp. 3: $d = 0.13$, $t(146.20) = 1.52$, $p = .130$) and both effect sizes are comparatively small.

Discussion

In Experiment 2 and 3, we replicated the structure effect on liking ratings observed in Experiment 1. That is, surface feature atmosphere accounts for an apparent difference of liking ratings between valid and invalid inferences.

Moreover, the formal structure effect on liking ratings seems to be moderated by perceived demand, since there was a pronounced difference in the strength of the structure effect for liking ratings between both experiments (i.e., between the instruction conditions).

This suggests that requesting a liking rating of the conclusion, while always presenting the full argument with premises, triggers the Gricean implicature – accounting for the violation of the maxim of quantity – that formal structure should be considered in one's judgement. Thus, participants resort to salient cues for logical validity (i.e., atmosphere) to inform their rating. Such a demand effect is countered to some extent by partially resolving the violation of the maxim of quantity by the instruction given in Experiment 3 informing participants that the full formal structure is relevant for the subsequent, second task of assessing logical validity and hence, by implication, not in the first.[14]

This notion is further backed up by the fact that a considerable number of participants in both experiments (but even more so in Experiment 3) actually explicitly stated that they had rated logical validity of the inferences exclusively or in addition to likability of the last statement during the liking task. Furthermore, the atmosphere effect is much stronger for those participants who indicate that they rated logical validity (exclusively or in addition to likability), thereby rendering their response patterns more similar to the responses observed in the logic task.

Importantly, we also found a difference between valid inferences and invalid inferences with congruent atmosphere for logic ratings, but not for liking ratings. In other words, there appears to be an effect of logical validity per se in the logic ratings. The size of this effect found within logic ratings was notably smaller than the size of the atmosphere effect. This could be interpreted as evidence that an assessment of logical necessity beyond congruent atmosphere indeed requires mental effort and thus was only attempted when explicitly requested – that is, during the logic task.

The data do not suggest that the distinction between possible and impossible

[14] An alternative explanation for this observation could be that participants may not have read or attended to the premises if there was no implicit task demand to consider logicality for their liking ratings. However, this appears to be rather unlikely given the explicit instructions to read the premises carefully and the sequential presentation regime in force in our studies.

inferences has noteworthy influence on the liking ratings.

Experiment 4 & 5

Trippas et al. (2016; see also Hayes et al., 2020; Ghasemi et al., 2021) did not limit their investigation to conditional inferences, but also presented categorical syllogisms and disjunctive inferences. Earlier studies by Morsanyi and Handley (2012; see also Klauer and Singmann, 2013; Singmann et al., 2014) also used syllogisms to investigate the logic-liking effect. Hence, it is desirable to replicate our findings for syllogisms as well. We therefore had to construct arguments that are analogous to the ones used for the previous experiments regarding their surface-feature atmosphere and whether the conclusion is necessary, possible, or impossible given the premises.

A syllogism has a major premise (e.g., “all guitars are mips”) introducing a subject (S; e.g., “guitars”) as well as a middle or distributed term (M; e.g., “mips”) that is always a nonword in our study (following Trippas et al., 2016). The minor premise (e.g., “some mips are fruits”) introduces the predicate (P; e.g., “fruits”). The conclusion (e.g., “therefore, some fruits are guitars”) combines predicate and subject. Furthermore, there can be different syllogistic figures (describing different directions of major and minor premise) as well as two additional conclusion directions. As previously mentioned in the introduction, quantifiers in categorical syllogisms (similar to the negation structure in conditional inferences) determine the atmosphere of the inference.

We used the quantifier “all” (A) for the major premise and “some” (I) and “no” (E) for minor premise and conclusion, resulting in four different possible quantifier structures (A-I-I, A-I-E, A-E-I, and A-E-E). When “some” (“no”) is used in the minor premise, syllogisms with “some” (“no”) conclusions are atmosphere-congruent, and syllogisms with “no” (“some”) conclusion atmosphere-incongruent. Different figures (with the major premise directions S-M and M-S) were used within these quantifier constellations to obtain valid, determinately invalid and indeterminately invalid syllogisms as shown in Table 2.

Note again that the valid and invalid syllogisms used by Trippas et al. (2016) confounded validity with atmosphere-congruency as well as possibility by contrasting valid syllogisms with determinately invalid (atmosphere-incongruent) syllogisms.

Table 2

The inferences types for categorical syllogisms

Type		Form (exemplary)	Conclusion status	
Quant.	Dir.		Validity	Atmosphere
A-I-I	S-M	All S are M; some M are P; therefore, some S are P	Indet. invalid	Congruent
A-I-I	M-S	All M are S; some M are P; therefore, some S are P	Valid	Congruent
A-I-E	S-M	All S are M; some M are P; therefore, no S are P	Indet. invalid	Incongruent
A-I-E	M-S	All M are S; some M are P; therefore, no S are P	Det. invalid	Incongruent
A-E-I	S-M	All S are M; no M are P; therefore, some S are P	Det. invalid	Incongruent
A-E-I	M-S	All M are S; no M are P; therefore, some S are P	Indet. invalid	Incongruent
A-E-E	S-M	All S are M; no M are P; therefore, no S are P	Valid	Congruent
A-E-E	M-S	All M are S; no M are P; therefore, no S are P	Indet. invalid	Congruent

Note. Indet. = indeterminately; det. = determinately. The type is determined by the quantifier structure (quant.) and the major premise direction (dir.).

As for Experiments 2 and 3, we manipulated instructions across experiments. Participants in Experiment 4 were only informed about the logic task after they completed the liking task (i.e., right before the logic task), whereas participants in Experiment 5 were informed about both tasks prior to the first task – that is, prior to the liking task.

Methods

Experiments 4 and 5 are both preregistered online-studies (see Open Science Framework registration https://osf.io/9h6np/ and https://osf.io/94mdj/ for further details; see also Footnote 4).

Participants

Fifty participants (18 females, 32 males) aged between 19 and 59 ($M_{age} = 30.54$, $SD_{age} = 10.30$) completed Experiment 4 and fifty-one participants (17 females, 34 males) aged between 19 and 52 ($M_{age} = 29.98$, $SD_{age} = 8.06$) completed Experiment 5. One of the participants of Experiment 5 reported not to have participated seriously. This participant was excluded from all subsequent analyses. All participants were recruited via Prolific and participated in exchange for a monetary compensation (£15.00). Inclusion criteria were an age between 18 and 80 and fluency in German. Participation in both experiments was not possible.

Design

Both experiments followed a within-participant design with task as a blocked variable (the liking task followed by the logic task). The inference type (A-E-E/S-M, A-E-E/M-S, A-E-I/S-M, A-E-I/M-S, A-I-E/S-M, A-I-E/M-S, A-I-I/S-M, and A-I-I/M-S), determined by crossing the three factors major premise direction (S-M vs. M-S), minor premise quantifier (I vs. E), and conclusion quantifier (I vs. E) as well as argument believability (believable vs. unbelievable; note that this only refers to the believability of the conclusion) were manipulated within subjects. The two different instruction conditions, on the other hand, were manipulated between subjects – that is, between the two experiments.

Materials

We used 64 different arguments for each participant (eight arguments for each of the eight unique combinations of quantifier structure and major premise direction). Half of the arguments (four arguments of each inference type) comprised a matching content pair – that is, subject and predicate standing in a subset-superset relation (as, e.g., apples and fruits), while the other half comprised a mismatching content pair – that is, subject and predicate denote a disjoint pair (as, e.g., guitars and fruits). The four remaining replicates with matching (mismatching) content pair resulted from the fact that for each of our quantifier structures, equivalent inference types and believability conditions arise when either the direction of the minor premise is reversed (P-M vs. M-P) or the direction of the conclusion is reversed (P-S vs. S-P).

Only A-E-E/S-M and A-I-I/M-S inferences are valid. A-E-I/S-M and A-I-E/M-S inferences, on the other hand, are determinately invalid (i.e., invalid and impossible). The remaining inferences are indeterminately invalid (i.e., invalid but possible). Moreover, A-E-E and A-I-I inferences have a congruent atmosphere with respect to the quantifier structure, while A-E-I and A-I-E inferences do not (see Table 2). Conclusion believability was manipulated by assigning either a matching content pair to a condition with an affirmative conclusion quantifier or a mismatching content pair to a negative conclusion for believable syllogisms and vice versa for unbelievable ones. Thus, for example, “some fruits are apples” as well as “no fruits are guitars” are both believable, whereas “some fruits are guitars” as well as “no fruits are apples” are both unbelievable.

We used 32 different German-language predicates with four different subset designators as matching subjects for each predicate, as well as 64 non-words (see Open Science Framework archive https://osf.io/9avjc/ for copies of all materials as well as their translation into English). For every participant each predicate was randomly paired with a non-word and two matching subjects as well as with a different non-word and two mismatching subjects (i.e., subjects belonging to a different predicate). This resulted in 128 different contents that were generated for each participant. We therefore presented each of the 64 arguments twice, but with different contents. Thus, participants saw a total of 128 unique trials. Each predicate was presented exactly four times, each non-word was presented exactly two times, and each subject was presented only once. A specific item content was equally likely to appear for each inference type.

Procedure

In the instructions given to the participants, we made clear that the nonwords we presented were arbitrary category names subsuming some existing entities. For subjects and predicates this was self-evident, as the respective materials denoted real-world sets. Thus, all sets referred to in the arguments (S, P, and M) are to be assumed to be non-empty, thus ensuring existential import. The procedures of Experiment 4 and 5 were otherwise identical to the procedures of Experiment 2 and 3, respectively. This included the same instruction manipulation. That is, instructions given prior to the first task were identical for Experiment 2 and 4 as well as for Experiment 3 and 5.

Results

Analysis approach

We once more used linear mixed model analyses with crossed random effects for participants, predicate content, subject content, and non-words to analyze participants' liking and logic ratings. Model selection regarding the random-effect structure was addressed as for the previous experiments. Note, however, that we had to conduct four separate preliminary model selections now, one for every random-effect factor.

The response behavior self-reports were also again analyzed by means of a Wilcoxon-Mann-Whitney test.

Response behavior self-report

In Experiment 4, three participants reported that they had rated only logical validity of the inference in the liking task while twelve participants reported that they had considered both logical validity of the inference and likability of the conclusion. In Experiment 5, seven participants reported that they had considered both logical validity of the inference and likability of the conclusion in the liking task. All remaining participants reportedly rated only likability of the conclusion. A Wilcoxon-Mann-Whitney test suggest that these ordinal rank distributions are different between the two experiments ($W = 1460.50$, $p = .044$).

Liking rating

As with Experiment 2 and 3, we first jointly analyzed the liking ratings of Experiment 4 and 5. The liking ratings of both experiments were thus submitted to an analysis in which we only included the within-subjects factor conclusion status (valid vs. indeterminately invalid with matching atmosphere vs. indeterminately invalid with mismatching atmosphere vs. determinately invalid) as well as the between-subjects factors instruction condition (Experiment 4 vs. Experiment 5) and self-reported response behavior during the liking task (rated only likability vs. rated only validity or both) as fixed effects.[15] There was strong evidence for a main effect of conclusion status, $F(3, 109.89) = 40.09$, $p < .001$. Besides that, the analysis revealed interaction effects between conclusion status and instruction condition, $F(3, 109.89) = 9.90, p < .001$, between conclusion status and response behavior, $F(3, 109.89) = 19.04$, $p < .001$, and between conclusion status, instruction condition, and response behavior $F(3, 109.89) = 5.77$, $p = .001$. All remaining effects had p-values equal to or greater than .217 ($p = .217$ was observed for the main effect of instruction condition).

Figure 4 shows the mean and individual liking ratings as a function of conclusion status separately for different groups defined by response behavior (rated only likability vs. rated only validity or both) and instruction condition (Experiment 4 vs. Experiment 5). The patterns mirror the ones observed for the liking ratings of all previous experiments. That is, the ratings tend to be higher for valid and indeterminately invalid, atmosphere-congruent arguments and lower for determinately invalid and indeterminately invalid, atmosphere-incongruent arguments, whereas there seems to be no noticeable

[15] The final random-effect structure included random intercepts for participants, by-participant random slopes for all three main effects, as well as all three two-way interactions.

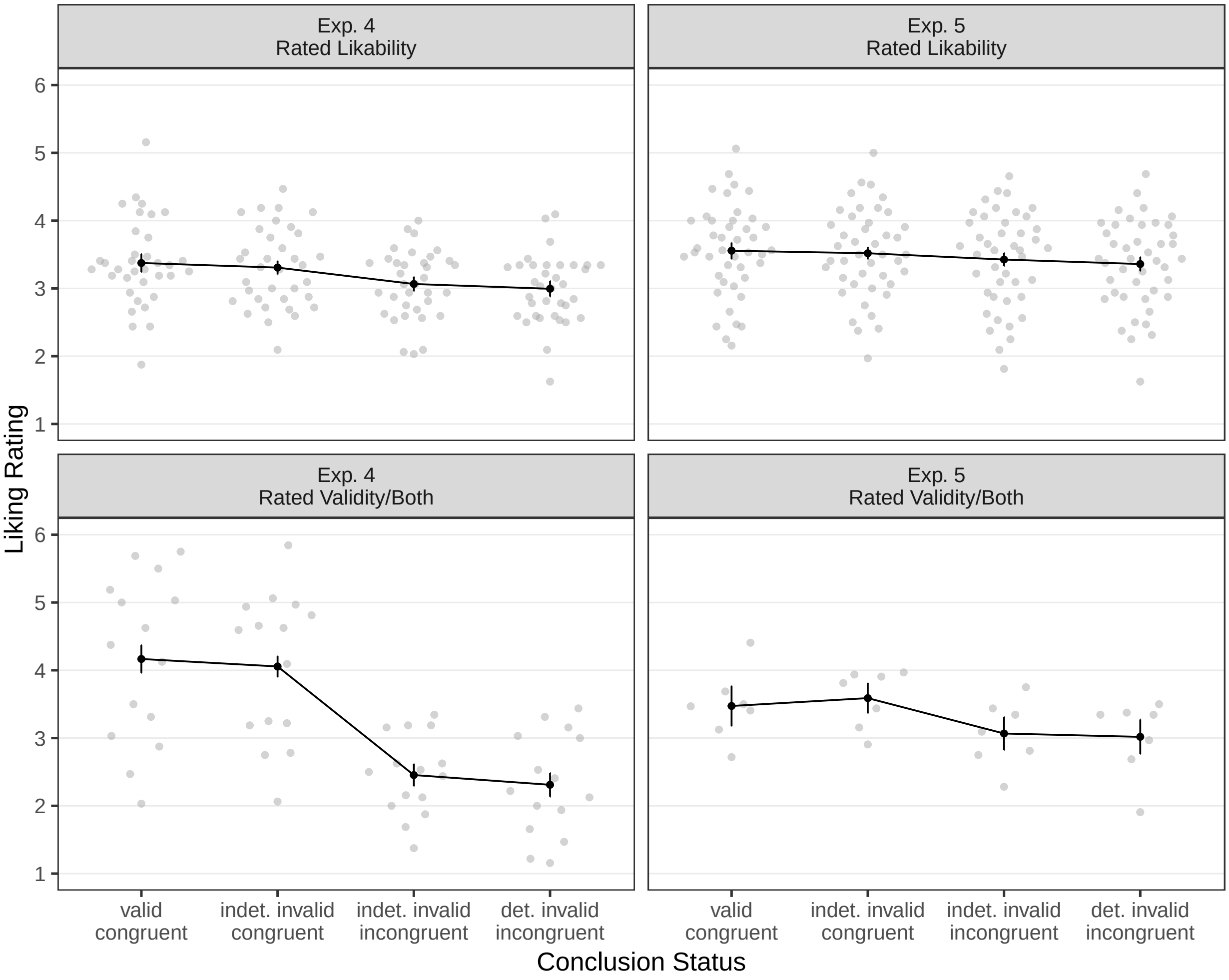
difference between valid and indeterminately invalid inferences with congruent atmosphere or between determinately invalid and indeterminately invalid inferences with incongruent atmosphere. Analogous to Experiments 2 and 3, we can clearly see that this difference is more prominent in Experiment 4 compared to Experiment 5 as well as for those participants who reported that they additionally (or exclusively) considered logical validity of the inference during the liking task. The effect almost completely vanishes for participants of Experiment 5 who reported that they only considered likability of the conclusion in their liking ratings.

We then again analyzed the liking ratings for each experiment individually by conducting two separate analyses in terms of the full design. Hence, we included the within-subjects factors major premise direction (S-M vs. M-S), minor premise quantifier (I vs. E), conclusion quantifier (I vs. E), and conclusion believability (believable vs. unbelievable) as fixed effects.[16] Depictions of the liking ratings from Experiment 4 and 5 broken down by inference type can be found in the appendix (see Figures A6 and A7). Since Trippas et al. (2016) reported greater liking of conclusions of valid relative to conclusions of determinately invalid arguments for categorical syllogisms, we calculated a linear contrast comparing these two types of inferences to assess whether we also replicate this effect. Results (Exp. 4: $d = 0.82$, $t(51.60) = 4.71$, $p < .001$; Exp. 5: $d = 0.23$, $t(62.40) = 2.70$, $p = .009$) indicate that the replication was successful. The difference is

[16] The final random-effect structure for both analyses included random intercepts for participants, subject contents, and predicate contents as well as by-predicate random slopes for conclusion quantifier, by-participant random slopes for the main effects of minor premise quantifier, conclusion quantifier, and conclusion believability as well as for the two-way interactions between minor premise quantifier and conclusion quantifier and between conclusion quantifier and conclusion believability. The final random-effect structure for Experiment 4 additionally included by-predicate random slopes for conclusion believability and the two-way interaction between conclusion believability and conclusion quantifier, while the final random-effect structure for Experiment 5 additionally included a by-subject random slope for conclusion quantifier.

Figure 4

Mean (black symbols) and individual (gray symbols) liking ratings of Experiment 4 (left panels) and 5 (right panels) as a function of conclusion status. Liking ratings of participants who reported rating only likability of the conclusion are displayed in the two upper panels, while liking ratings of participants who reported rating also (or exclusively) logical validity of the inference are displayed in the lower panels. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based). Indet. = indeterminate; det. = determinate.



more pronounced in Experiment 4 than in Experiment 5. To see whether we also replicate greater liking of believable than unbelievable conclusions of categorical syllogisms, we also juxtaposed these two types of inferences. Results (Exp. 4: $d = 1.33$, $t(51.10) = 8.84$, $p < .001$; Exp. 5: $d = 0.96$, $t(49.00) = 6.94$, $p < .001$) again indicate a successful replication. Once more, the effect is more pronounced in Experiment 4 than in Experiment 5. Another contrast addressed the question whether there was an effect of logical validity per se when the confoundings in terms of possibility and atmosphere are held constant by juxtaposing valid inferences (A-E-E/S-M and A-I-I/M-S) and indeterminately invalid, atmosphere-congruent inferences (A-E-E/M-S and A-I-I/S-M). Results (Exp. 4: $d = 0.08$, $t(6059.00) = 1.96$, $p = .051$; Exp. 5: $d = 0.00$, $t(6073.30) = 0.11$, $p = .915$) indicate that there is no effect of validity per se (see also Table A1 in the appendix). A comparison between atmosphere-congruent and atmosphere-incongruent inferences suggests the presence of an atmosphere effect (Exp. 4: $d = 0.74$, $t(49.00) = 4.30$, $p < .001$; Exp. 5: $d = 0.19$, $t(49.00) = 2.44$, $p = .018$). Again, this effect is more pronounced in Experiment 4 where it is still detectable even when validity and possibility are held constant by juxtaposing indeterminately invalid, atmosphere-congruent inferences (A-E-E/M-S and A-I-I/S-M) and indeterminately invalid, atmosphere-incongruent inferences (A-E-I/M-S and A-I-E/S-M; $d = 0.66$, $t(51.60) = 3.78$, $p < .001$). The same contrast does not reach statistical significance in Experiment 5 ($d = 0.16$, $t(62.50) = 1.90$, $p = .063$). We also once more assessed the role of possibility versus impossibility while holding logical validity and atmosphere-congruency constant by contrasting indeterminately invalid, atmosphere-incongruent inferences(A-E-I/M-S and A-I-E/S-M) and determinately invalid inferences (A-E-I/S-M and A-I-E/M-S). Although there is a significant difference in Experiment 4 ($d = 0.08$, $t(6065.70) = 2.13$, $p = .033$), this is not the case for Experiment 5 ($d = 0.06$, $t(6071.50) = 1.58$, $p = .116$) and both effect sizes are comparatively small.

Logic ratings

The logic ratings of both experiments were again first submitted to an analysis in which we only included the within-subjects factor conclusion status (valid vs. indeterminately invalid with congruent atmosphere vs. indeterminately invalid with incongruent atmosphere vs. determinately invalid) as well as the between-subjects factor instruction condition (Experiment 4 vs. Experiment 5) as fixed effects.[17] This analysis revealed a strong main effect of conclusion status, $F(3, 179.08) = 285.60$, $p < .001$. All remaining effects had p-values equal to or greater than .300 ($p = .300$ was observed for the main effect of instruction condition).

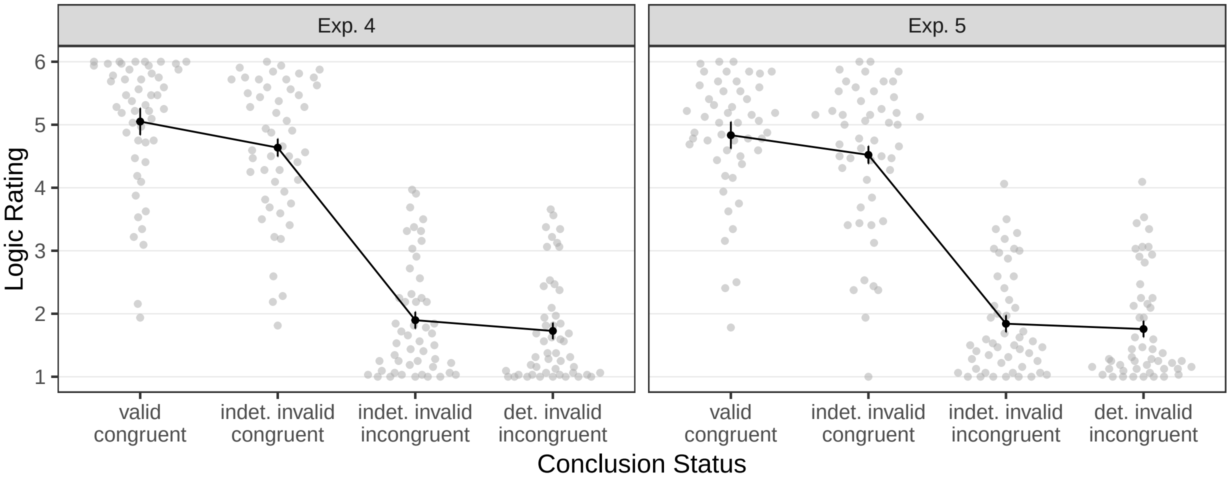
Figure 5 shows the mean and individual logic ratings as a function of conclusion status separately for different groups defined by the instruction condition (Experiment 4 vs. Experiment 5). The patterns match the ones observed for the logic ratings of Experiment 2 and 3. That is, the ratings are clearly higher for valid and indeterminately invalid arguments with congruent atmosphere and lower for determinately invalid and indeterminately invalid arguments with incongruent atmosphere. Furthermore, we can see that the ratings for valid inferences are higher compared to indeterminately invalid, atmosphere-congruent inferences, although this difference is once more comparatively small.

We then also analyzed the logic ratings for each experiment separately. Both analyses included the within-subjects factors major premise direction (S-M vs. M-S), minor premise quantifier (I vs. E), conclusion quantifier (I vs. E), and conclusion believability

[17] The final random-effect structure included random intercepts for participants as well as by-participant random slopes for conclusion status and instruction condition.

Figure 5

Mean (black symbols) and individual (gray symbols) logic ratings of Experiment 4 (left panel) and 5 (right panel) as a function of conclusion status. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based). Indet. = indeterminate; det. = determinate.



945 (believable vs. unbelievable) as fixed effects.[18] Depictions of the logic ratings from 946 Experiment 4 and 5 broken down by inference type can be found in the appendix (see 947 Figures A8 and A9). We again calculated the same linear contrast for the logic ratings as 948 we did for the liking ratings. Thus, to evaluate whether valid inferences were endorsed

[18] The final random-effect structure for both analyses included random intercepts for participants, by-participant random slopes for all main effects and interactions including major premise direction, minor premise quantifier, and conclusion quantifier as well as for the main effect of conclusion believability and the two-way interaction between conclusion believability and conclusion quantifier. The final random-effect structure for Experiment 4 additionally included by-participant random slopes for the two-way interaction between minor premise quantifier and conclusion believability and the three-way interaction between minor premise quantifier, conclusion quantifier, and conclusion believability. The final random-effect structure for Experiment 5 additionally included random intercepts for predicate contents and by-predicate random slopes for conclusion quantifiers.

more relative to determinately invalid arguments, we compared these two types of inferences. Results (Exp. 4: $d = 3.32$, $t(64.70) = 13.70$, $p < .001$; Exp. 5: $d = 3.08$, $t(60.10) = 13.32$, $p < .001$) indicate that this was indeed the case. To see whether inferences with believable conclusions were endorsed more than inferences with unbelievable ones, a linear contrast juxtaposed these two types of inferences. Results (Exp. 4: $d = 0.27$, $t(49.00) = 3.38$, $p = .001$; Exp. 5: $d = 0.37$, $t(49.00) = 3.09$, $p = .003$) indicate that this was the case as well. Another contrast addressed the question whether there was an effect of logical validity per se when the confoundings in terms of possibility and atmosphere are held constant. The contrast juxtaposes valid inferences (A-E-E/S-M and A-I-I/M-S) and indeterminately invalid, atmosphere-congruent inferences (A-E-E/M-S and A-I-I/S-M). Results (Exp. 4: $d = 0.41$, $t(77.50) = 3.95$, $p < .001$; Exp. 5: $d = 0.31$, $t(95.70) = 3.20$, $p = .002$) indicate that there is an effect of validity per se (see also Table A2 in the appendix). Contrasting atmosphere-congruent and atmosphere-incongruent inferences suggests the presence of an atmosphere effect (Exp. 4: $d = 3.03$, $t(49.00) = 13.48$, $p < .001$; Exp. 5: $d = 2.88$, $t(49.00) = 13.17$, $p < .001$). This effect is also apparent when validity and possibility are held constant by juxtaposing indeterminately invalid, atmosphere-congruent inferences (A-E-E/M-S and A-I-I/S-M) and indeterminately invalid, atmosphere-incongruent inferences (A-E-I/M-S and A-I-E/S-M; Exp. 4: $d = 2.74$, $t(64.70) = 11.29$, $p < .001$; Exp. 5: $d = 2.68$, $t(60.10) = 11.62$, $p < .001$). The last contrast once more assessed the role of possibility versus impossibility while holding logical validity and atmosphere-congruency constant by comparing the logic ratings for indeterminately invalid, atmosphere-incongruent inferences (A-E-I/M-S and A-I-E/S-M) and for determinately invalid inferences (A-E-I/S-M and A-I-E/M-S). The contrast provided little evidence for a role of possibility (Exp. 4: $d = 0.17$, $t(77.50) = 1.63$, $p = .108$; Exp. 5: $d = 0.08$, $t(95.70) = 0.84$, $p = .401$).

Discussion

We found a structure effect on liking ratings for the conclusions of categorical syllogisms which mirrors the one observed for conditional inferences in our previous experiments. That is, there is once more no logic-liking effect, but rather an atmosphere effect.

This structure effect on liking ratings seems again to be moderated by perceived demand, since there was a clear difference in the strength of the effect between both experiments (i.e., between the instruction conditions). This supports the notion that presentation of a logical argument like a syllogism has a suggestive character that implies to rate – at least partially – logical validity of the inference during the liking task.

Analogous to the previous experiments, there was again a considerable number of participants who stated that they had considered logical validity of the inference during the liking task and for those participants the structure effects are much stronger. We also observed that more participants reported doing so in Experiment 4 than in Experiment 5, indicating that our instruction manipulation indeed affected perceived demand to consider logical validity of the inference during the liking task. This is perfectly in line with the interpretation in terms of Gricean implicatures, which are mitigated by the instructions used for Experiment 5, as outlined previously.

Once more, convincing evidence for an unconfounded effect of logical validity was only present for logic ratings but not for liking ratings. As in the previous experiments, we found that this effect is rather small compared to the effect of atmosphere.

Results regarding the influence of possibility on liking ratings were mixed at best.

General Discussion

In the present work, we identified two major confounds (viz., possibility and atmosphere-congruency) that might have been responsible for the supposed logic-liking effect reported by Trippas et al. (2016; see also Ghasemi et al., 2021). This raises the question whether such an effect can still be found when the confounds are properly controlled for. When doing so for conditional and categorical syllogisms,[19] we failed to find convincing evidence of any structure effect on liking ratings beyond an effect of atmosphere-congruency (regarding certain surface features). Hence, our results challenge the notion of there being a logic-liking effect and instead suggest that the supposed effect of logical validity on liking ratings reported by Trippas et al. (2016) was caused by an atmosphere confound rather than by logical validity per se.

Even more problematic for the notion of logical intuitions affecting liking ratings are our results regarding the demand effect. We found that any effect of inference structure on liking ratings is heavily susceptible to a manipulation of the instructions. When given only a vague instruction, participants tend to use the presented inference structure (more precisely, certain surface features associated with atmosphere-congruency) as guidance for their decision. This seems to indicate that there is a considerable amount of perceived demand to consider heuristic cues for logical validity, perhaps because the Gricean maxim of quantity is violated during the liking task. That is, when presented with the complete argument while being asked to rate only the conclusion, Gricean implicatures are likely triggered and suggest that cues to logical validity are to be taken into account in one's ratings.

This notion is further supported by the participants' self-reports regarding their

[19] While Trippas et al. (2016; see also Hayes et al., 2020; Ghasemi et al., 2021) also used disjunctive inferences to assess the logic-liking effect, we decided to omit disjunctions for the present study, since it is not straightforward to disentangle surface-feature atmosphere from logical validity for that kind of arguments. We want to point out, however, that the same confoundings are also present within the disjunctive materials used by Trippas et al. (2016), taking into account that the atmosphere effect must be defined differently for disjunctive syllogisms as discussed earlier. Thus, we do not see any good reason why the structure effect should be qualitatively different for disjunctive arguments. However, if one finds a way to disentangle atmosphere from logical validity for disjunctions, future research might aim to confirm this conjecture.

response behavior. Not only was the tendency to consider logical validity during the liking task influenced by the instruction condition, but that tendency was also accompanied by a stronger atmosphere effect. We also want to point out that these self-reports are given after the second task, that is, after participants learned that they were in actual fact not supposed to rate logical validity during the first task. Consequently, we suspect some degree of desirability bias to factor into these self-reports. Hence, the demand effect might be even stronger than can be inferred from the self-report data.

Importantly, our findings regarding the influence of demand characteristics challenge only the goal-independent nature of the processes underlying effects of inference structure. However, the present research was not designed to investigate other possible automaticity features of the processes underlying effects of inference structure besides goal independence such as whether they are fast and/or effortless. And thus, we are only questioning the lines of research suggesting that logical intuitions are elicited independently of a goal to evaluate logical structure and that logical intuitions in these paradigms are sensitive to logical validity per se. We do not address the lines of research that suggest that the underlying processes are fast and effortless (but see Hayes et al., 2020; cf. Bago and De Neys, 2017; Thompson and Johnson, 2014) – nor do we think that whether or not this is the case affects our conclusion.

Taken together, the processes underlying the supposed logic-liking effect neither appear to be intuitive (in the sense that they are elicited independently of a goal to evaluate logicality), nor do they appear to be logical (in the sense that they would respect logical validity per se).

Moreover, other than for the liking ratings, we did find a consistent unconfounded effect of logical validity on logic ratings for both conditional and categorical syllogisms, which suggests that the logic task, but not the liking task, to some extent also recruits analytic Type 2 processes that respect logical validity per se.[20] Interestingly, this effect was small relative to the effect of atmosphere-congruency. It is well known, however, that this atmosphere effect accounts for ample variance in logic judgments for categorical syllogisms (e.g., Khemlani & Johnson-Laird, 2012). The present results are consistent with these earlier observations and, furthermore, imply that an atmosphere heuristic affects logic judgments for conditional syllogisms in a very similar manner.

In many respects, the current work therefore complements the findings and conclusions of Hayes et al. (2020), who also examined the basis for the logic-liking effect. They applied signed difference analysis (Stephens et al., 2018) to test computational models of liking and logic ratings of the same stimuli and concluded that a model based on a single latent processing dimension could account for both data sets. However, their analysis was silent on the details of this processing dimension. The current work suggests that one dimension that influences responses on both liking and logic tasks is sensitivity to atmosphere cues. Crucially, the current work shows that, when these cues are dissociated from logical validity, they are the key factor driving liking ratings, and exert a strong influence on logic ratings. This has interesting implications, as it suggests that differentiating logical validity from those surface features responsible for atmosphere-congruency is difficult. However, further research is certainly required to

[20] It should be noted, however, that although the present evidence does not favor the possibility of there being an unconfounded effect of logical validity on liking ratings as proposed by Morsanyi and Handley (2012), we have only null effects to base our conclusion on. Therefore, it might be imprudent to rule out that such an effect might exist after all, albeit being small. However, the mere presence of demand effects renders the hypothetical occurrence of an unconfounded logic-liking effect inconclusive for answering the question if there exists something like logical intuition. Some participants might experience such a strong demand to base their liking rating on logical validity of the inference that they deliberately invest the mental effort to evaluate the latter during the liking task. In other words, they would not only use atmosphere cues, but also engage in deeper analyses evaluating logical necessity. We argue that this would be a simple and parsimonious explanation of such a hypothetical effect, assuming it exists at all.

investigate the underlying mechanisms in more detail.

Possible Explanations of the Atmosphere Effect

The goal of the present research was not to contribute to explanations of such atmosphere effects (but see Begg & Denny, 1969; Chater & Oaksford, 1999; Oaksford et al., 2000; Wetherick & Gilhooly, 1995). Yet, we note that atmosphere and validity are often confounded in arguments that reasoners encounter. In fact, atmosphere-incongruent arguments are always logically invalid, whereas a substantial proportion of atmosphere-congruent arguments are logically valid. Consequently, atmosphere is a diagnostic, though fallible heuristic cue to logical validity. Reasoners may have learned to rely on atmosphere cues as a fast and frugal heuristic in judging logical validity (Gigerenzer & Todd, 1999). This also supports an interpretation of the results from liking and – to a certain degree – logic tasks as both being affected by *perceived* logical validity as the experiential outcome of an atmosphere heuristic operating in both tasks to the extent to which reasoners intend to evaluate logicality.

Although such heuristic accounts of atmosphere effects are now widely accepted (Khemlani, in press), there have been attempts to reconcile atmosphere effects with reasoning that adheres to normative principles. In the present case, for example, it could be argued that atmosphere effects are effects of logical validity after all if one assumes that all conditional premises in our study were always interpreted biconditionally (e.g. "if a child cries, then it is happy" is interpreted to mean that "if and only if a child cries, then it is happy") and all syllogistic premises involving the quantifier "all" were interpreted as indicating that the two sets involved are in fact identical (e.g., "All guitars are mips" are interpreted as "All guitars are mips and all mips are guitars"). Given these assumptions, atmosphere-congruency and logical validity would coincide for all arguments that we used.

Considering conditional syllogisms, the idea that the conditional premises of such arguments are sometimes interpreted biconditionally has a long tradition in the reasoning literature (e.g., Johnson-Laird & Byrne, 1991), accounting, for example, for the fact that AC inferences are frequently endorsed as logically valid. Under a conditional interpretation, only MP and MT inferences are valid inferences; whereas under a biconditional interpretation MP, AC, DA, and MT are valid inferences. There are, however, several lines of research speaking against the idea that the biconditional interpretation of conditionals is a widespread phenomenon.

For example, with abstract or arbitrary rule contents, endorsement rates for MP are typically close to 100%, whereas the AC (and DA, and MT) inference rates show wide variability across studies (Schroyens et al., 2001) although MP and AC should be treated equivalently under a biconditional interpretation. In another line of research, conditional arguments with everyday contents as used in the present research are presented twice, once with the conditional rule present, the other time without it (i.e., only minor premise and conclusion are presented; e.g., Klauer et al., 2010; Liu, 2003) and the task is in both cases to assess the plausibility or probability of the conclusion. This allows one to disentangle content-based, pragmatic contributions as captured in ratings of conclusions presented without the rule from contributions that are genuinely rule-driven. It turns out that introducing a rule boosts acceptability of the different inferences to varying degrees. Consistent with a conditional, but not a biconditional interpretation of the rule, MP receives a major boost, followed by MT, with lower contributions to DA and AC (Klauer et al., 2010; Singmann et al., 2016). As another example, in the truth-table evaluation task, reasoners treat the cases in which the two propositions p and q of a conditional rule of the form “If p then q” are both true very differently from cases in which both are false (e.g., Evans & Over, 2004) although both should be treated equivalently under a biconditional interpretation.

Considering categorical syllogisms, the idea that premises such as “All guitars are mips” are sometimes seen as implying that “All mips are guitars” likewise has a long history in the reasoning literature where it is known as the *conversion hypothesis*

(Chapman & Chapman, 1959). It is, however, generally agreed upon that conversions of this kind do not occur consistently and pervasively. If they did, they would, for example, eliminate effects of the syllogisms' figure (Khemlani & Johnson-Laird, 2012), and figural effects are one of the most robust effects found in studies of syllogistic reasoning.

Perhaps more convincing than these findings based on previous empirical and theoretical work is the fact that the present data themselves are neither consistent with a biconditional interpretation of conditional premises nor with the conversion hypothesis: As reported above, we observe effects of logical validity in the logic tasks for both conditional and categorical syllogisms when atmosphere and possibility are held constant – that is, over and above atmosphere effects – which should not be the case if biconditional interpretations or conversions were consistently adopted (see also Figures A4, A5, A8, and A9 as well as Table A2 in the appendix).

Finally, note that these alternative accounts do not jeopardize the conclusiveness of the finding that atmosphere effects are strongly dependent on demand characteristics nor its interpretation that the logic-liking effect does not reflect an *intuitive* logicality (in the sense of being driven by a non-strategic, goal-independent process), as we have already discussed above.

Implications for Related Research

Ghasemi et al. (2021) recently argued that ratings of physical brightness manipulated by changing the contrast of the black text against a white background (see also Trippas et al., 2016) are a more appropriate measure of intuitive reasoning, since demand effects allegedly are a less plausible alternative explanation. However, this line of argument might be questionable in the light of the Gricean analysis outlined in the present work. While rating brightness is arguably a more objective and less ambiguous task than rating likability, the maxim of quantity is still violated. Hence, it is doubtful that brightness ratings are free from demand effects in general.

In fact, recent research by Hayes et al. (in press) did reexamine brightness ratings for conclusions of various arguments. They found that the effect of logical validity on brightness rating was susceptible to a manipulation of difficulty, disappearing when brightness conditions were easy to discriminate. These results seem to confirm our hypothesis that demand characteristics – and thus deliberate response behavior on the part of the participants – are critical for an effect of logical validity to emerge in tasks unrelated to the assessment of logical status.

Although an evaluation of brightness ratings was beyond the scope of the present study, we also want to point out that the studies that used brightness ratings to argue in favor of logical intuitions (Ghasemi et al., 2021; Trippas et al., 2016) still suffer from the same confoundings we targeted in the present study. Thus, the results of those studies should only be interpreted with caution until verified by a more informative design.

From a practical perspective, we therefore advise that – at a minimum – the above considerations must be taken into account when employing perceptual and affective ratings tasks to investigate possible logical intuitions. In order to avoid spurious conclusions, two design factors seem indispensable: Problems should be designed so that effects of logical validity can be disentangled from atmosphere effects, and instructions should be designed so as to block demand effects suggesting that logical structure is relevant for the task at hand. However, it is plausible that completely eliminating demand effects is impossible in this context. This issue critically limits the informational value provided by such rating tasks. Therefore, we are skeptical that conclusive evidence in favor of logical intuitions can be derived from them in general.

Theoretical Implications and Conclusion

Overall, we conclude that the present study provides strong support for the notion that implicit affective reactions and intuitions are not sensitive to logical validity per se and for the hypothesis that their activation is dependent on a context in which raters strategically intend to evaluate logical structure due to instructed or perceived task demands.

These conclusions have important theoretical implications – especially for DP 2.0 theories. As reviewed in the introduction, there exist quite a number of results from a range of diverse paradigms that support the central claim of DP 2.0. theories (see, e.g., Bago et al., 2020; Bago & De Neys, 2019; De Neys, 2012, 2014; De Neys et al., 2011; De Neys & Glumicic, 2008; De Neys & Pennycook, 2019; Johnson et al., 2016; Newman et al., 2017). Nevertheless, previous findings of (supposedly intuitive) sensitivity to logical validity in perceptual and affective ratings tasks – as, for example, the logic-liking effect – have been one key source of evidence motivating their development. Our finding that no such sensitivity exists in affective ratings therefore represents a challenge to such theories.

The finding is particularly difficult to reconcile with the conceptual fluency hypothesis, because conceptual fluency is seen as an automatic experiential byproduct of reading and understanding the premises translating directly into graded feelings of liking or disliking. Logic-liking effects generated via this route should be independent of a goal to evaluate logicality.

The automatization hypothesis, on the other hand, can be specified in different ways, some of which are compatible with the absence of goal-independent effects of logical structure. For example, it could be argued that the learning episodes that lead to automatization consistently occur in the context of goals to arrive at normatively correct responses so that a goal context becomes part of what is learned. In this view, logical intuitions would indeed not arise independently of a goal to arrive at the normatively correct response and hence, no effects of logical structure would be expected in tasks that do not elicit such goals. In this spirit, De Neys (2014) explicitly states that “the logical principles need to be activated at some level. The logical intuition suggestion boils down to the claim that this knowledge is implicit in nature and is activated automatically *when* *people are faced with a reasoning task.* [emphasis added]” (De Neys, 2014, p. 175).[21]

Alternatively, it could be argued that logical intuitions are activated whenever perceivers are confronted with a logical argument irrespective of current goals, but they can only interfere with responses to unrelated tasks to the extent to which there is some overlap between features of the logical intuitions and task-relevant features (Kornblum & Lee, 1995). For example, in the context of the Stroop task, word reading is believed to be overlearned to such an extent that a word is read in many contexts in which this is not required by or even relevant for the task at hand (Lindsay & Jacoby, 1994). Nevertheless, the overlearned reading of words interferes with naming the word’s print color only to the extent to which the word itself evokes a color (MacLeod, 1991). And thus, by analogy, even if logical intuitions arise independently of current goals, they might have the capacity to color liking ratings only to the extent to which overlap is assumed to exist between a like-dislike dimension or categorization and a valid-invalid dimension or categorization. If such overlap is denied, logical intuitions would again not be expected to have the power to affect liking ratings.

Whereas some of these theoretical implications remain within the DP 2.0 framework, a more radical possibility is that logical intuitions as conceptualized by DP 2.0 theories do not exist after all. We believe to have provided evidence questioning their existence in the logic-liking paradigm. Future work may consider other paradigms as reviewed in the introduction that support the idea of logical intuitions implementing similar design features and controls as the present work to assess this possibility.

[21] Note, however, that De Neys and Pennycook (2019) discuss the automatization hypothesis as consistent with the logic-liking effect and similar effects suggesting goal independence reviewed in the introduction (but see De Neys, 2021; De Neys & Franssens, 2009). Note also that automatization is frequently assumed to result in unintentional, goal-independent processing (Bargh, 1994; Posner & Snyder, 1975a, 1975b).

References

Bago, B., Bonnefon, J.-F., & De Neys, W. (2020). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*, *150*(6), 1081–1094. https://doi.org/10.1037/xge0000968

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Bargh, J. A. (1994). The four horsemen of automaticity: Intention, awareness, efficiency, and control as separate issues. In R. S. J. Wyer & T. K. Srull (Eds.), *Handbook of social cognition: Vol. 1: Basic processes* (2nd ed., pp. 1–44). Lawrence Erlbaum Associates.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Begg, I., & Denny, J. P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology*, *81*(2), 351–354. https://doi.org/10.1037/h0027770

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology*, *58*(3), 220–226. https://doi.org/10.1037/h0041961

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*(2), 191–258. https://doi.org/10.1006/cogp.1998.0696

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W. (Ed.). (2018). *Dual process theory 2.0*. Routledge.

De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on Psychological Science*, *16*(6), 1412–1427. https://doi.org/10.1177/1745691620964172

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLOS ONE*, *6*(1), e15954. https://doi.org/10.1371/journal.pone.0015954

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition*, *113*(1), 45–61. https://doi.org/10.1016/j.cognition.2009.07.009

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Evans, J. S. B. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*(6), 978–996. https://doi.org/10.1037/0033-2909.128.6.978

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. (2009). How many dual-process theories do we need? one, two, or many? In J. S. B. Evans & K. Frankish (Eds.), *Two minds: Dual processes and beyond* (pp. 33–54). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0002

Evans, J. S. B. (2018). Dual process theories: Perspectives and problems. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 137–155). Routledge.

Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*(3), 295–306. https://doi.org/10.3758/bf03196976

Evans, J. S. B., & Over, D. E. (2004). *If: Supposition, pragmatics, and dual processes.* Oxford University Press.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Ghasemi, O., Handley, S., & Howarth, S. (2021). The bright homunculus in our head: Individual differences in intuitive sensitivity to logical validity. *Quarterly Journal of Experimental Psychology.* Advance online publication. https://doi.org/https://doi.org/10.1177/17470218211044691

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 3–34). Oxford University Press.

Grice, P. (1989). *Studies in the way of words.* Harvard University Press.

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 28–43. https://doi.org/10.1037/a0021098

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 33–58). Elsevier. https://doi.org/10.1016/bs.plm.2014.09.002

Hayes, B. K., Stephens, R. G., Lee, M. D., Dunn, J. C., Kaluve, A., Choi-Christou, J., & Cruz, N. (in press). Always look on the bright side of logic? Testing explanations of intuitive sensitivity to logic in perceptual tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Hayes, B. K., Wei, P., Dunn, J. C., & Stephens, R. G. (2020). Why is logic so likeable? A single-process account of argument evaluation with logic and liking judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 699–719. https://doi.org/10.1037/xlm0000753

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting system 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64. https://doi.org/10.1016/j.actpsy.2015.12.008

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

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

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology*, *10*(1), 64–99. https://doi.org/10.1016/0010-0285(78)90019-1

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, *16*(1), 1–61. https://doi.org/10.1016/0010-0277(84)90035-0

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Khemlani, S. (in press). Psychological theories of syllogistic reasoning. In M. Knauff & W. Spohn (Eds.), *Handbook of rationality*. MIT Press.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457. https://doi.org/10.1037/a0026841

Klauer, K. C. (in press). Dual-process theories of deductive reasoning. In M. Knauf & W. Spohn (Eds.), *Handbook of rationality*. MIT Press.

Klauer, K. C., Beller, S., & Huetter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 298–323. https://doi.org/10.1037/a0018705

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, *107*(4), 852–884. https://doi.org/10.1037/0033-295x.107.4.852

Klauer, K. C., & Singmann, H. (2013). Does logic feel good? Testing for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1265–1273. https://doi.org/10.1037/a0030530

Kornblum, S., & Lee, J.-W. (1995). Stimulus-response compatibility with relevant and irrelevant stimulus dimensions that do and do not overlap with the response. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 855–875. https://doi.org/10.1037/0096-1523.21.4.855

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 219–234. https://doi.org/10.1037/0096-1523.20.2.219

Liu, I.-m. (2003). Conditional reasoning and conditionalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 694–709. https://doi.org/10.1037/0278-7393.29.4.694

MacLeod, C. M. (1991). Half a century of research on the stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Morsanyi, K., & Handley, S. J. (2012). Logic feels so good—I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 596–616. https://doi.org/10.1037/a0026099

Nakamura, H., & Kawaguchi, J. (2016). People like logical truth: Testing the intuitive detection of logical value in basic propositions. *PLOS ONE*, *11*(12), 1–22. https://doi.org/10.1371/journal.pone.0169166

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 883–899. https://doi.org/10.1037/0278-7393.26.4.883

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208–225. https://doi.org/10.1037/met0000126

Posner, M. I., & Snyder, C. R. R. (1975a). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The loyola symposium* (pp. 55–85). Lawrence Erlbaum Associates.

Posner, M. I., & Snyder, C. R. R. (1975b). Facilitation and inhibition in the processing of signals. In P. M. A. Rabbit & S. Dornič (Eds.), *Attention and performance v* (pp. 669–682). Academic Press.

Schroyens, W. J., Schaeken, W., & d'Ydewalle, G. (2001). *A meta-analytic review of conditional reasoning by model and/or ruule: Mental model theory revised* [Unpublished manuscript]. University of Leuven.

Sells, S. B. (1936). The atmosphere effect: An experimental study of reasoning. *Archives of Psychology*, *29*, 3–72.

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (1st ed., pp. 4–31). Routledge. https://doi.org/10.4324/9780429318405

Singmann, H., Klauer, K. C., & Beller, S. (2016). Probabilistic conditional reasoning: Disentangling form and content with the dual-source model. *Cognitive Psychology*, *88*, 61–87. https://doi.org/10.1016/j.cogpsych.2016.06.005

Singmann, H., Klauer, K. C., & Kellen, D. (2014). Intuitive logic revisited: New data and a bayesian mixed model meta-analysis. *PLOS ONE*, *9*(4), e94223. https://doi.org/10.1371/journal.pone.0094223

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Blackwell.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018). Are there two processes in reasoning? the dimensionality of inductive and deductive inferences. *Psychological Review*, *125*(2), 218–244. https://doi.org/10.1037/rev0000088.supp

Thompson, V. A., & Newman, I. R. (2018). Logical intuitions and other conundra for dual process theories. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 121–136). Routledge. https://doi.org/10.4324/9781315204550

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1448–1457. https://doi.org/10.1037/xlm0000248

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wetherick, N., & Gilhooly, K. (1995). ‘atmosphere’, matching, and logic in syllogistic reasoning. *Current Psychology*, *14*(3), 169–178. https://doi.org/10.1007/BF02686906

Wilson, D., & Sperber, D. (1986). On defining relevance. In R. E. Grandy & R. Warner (Eds.), *Philosophical grounds of rationality: Intentions, categories, ends* (pp. 243–258). Clarendon Press.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451–460. https://doi.org/10.1037/h0060520

Appendix

Figure A1

Mean (black symbols) and individual (gray symbols) liking ratings in Experiments 1 as a function of inference type. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

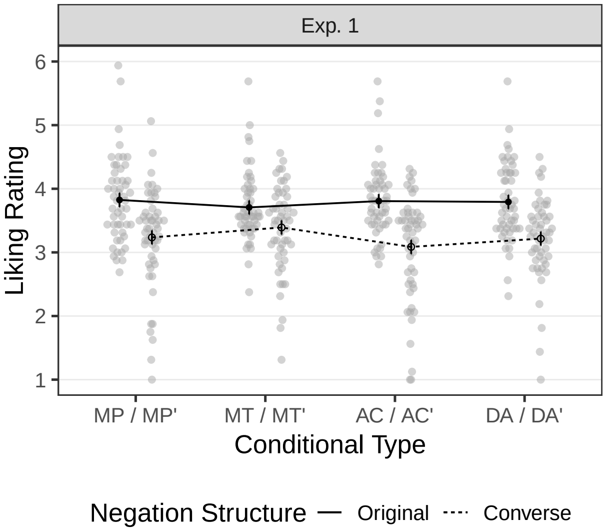


Figure A2

Mean (black symbols) and individual (gray symbols) liking ratings in Experiments 2 as a function of inference type. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

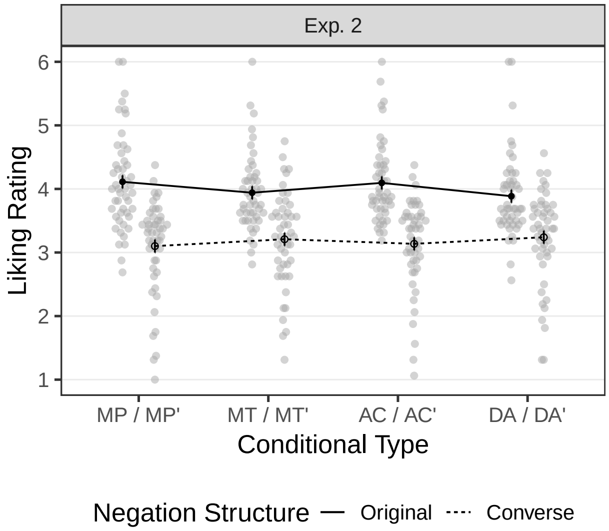


Figure A3

Mean (black symbols) and individual (gray symbols) liking ratings in Experiments 2 as a function of inference type. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

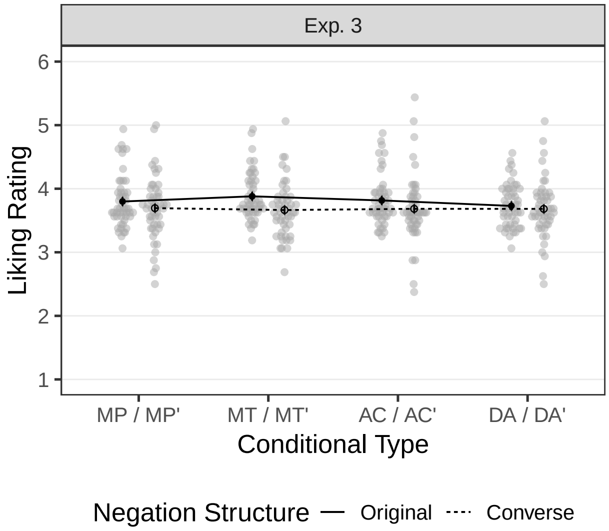


Figure A4

Mean (black symbols) and individual (gray symbols) logic ratings in Experiments 3 as a function of inference type. Vertical jitter was added to individual logic ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

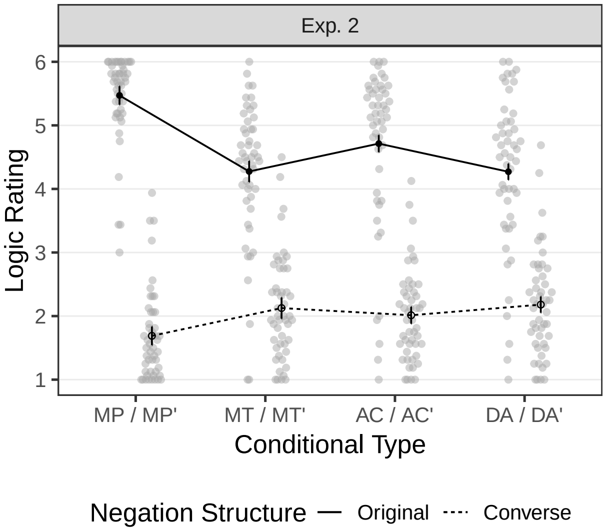


Figure A5

Mean (black symbols) and individual (gray symbols) logic ratings in Experiments 3 as a function of inference type. Vertical jitter was added to individual logic ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

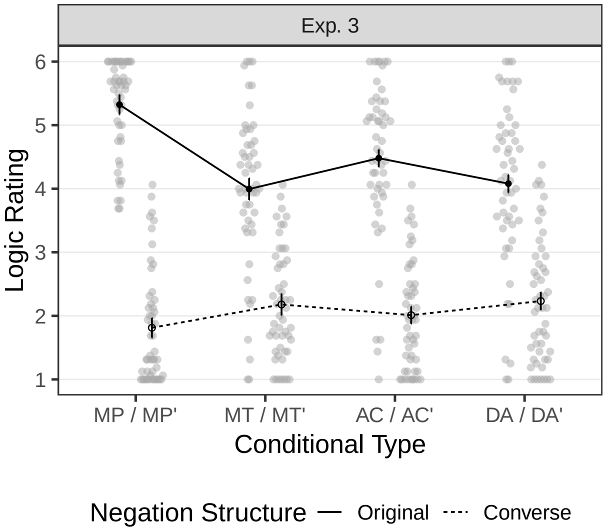


Figure A6

Mean (black symbols) and individual (gray symbols) liking ratings in Experiments 4 as a function of inference type. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

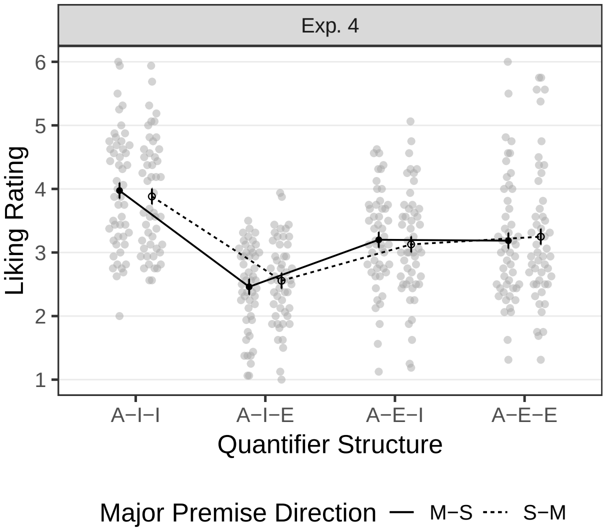


Figure A7

Mean (black symbols) and individual (gray symbols) liking ratings in Experiments 4 as a function of inference type. Vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

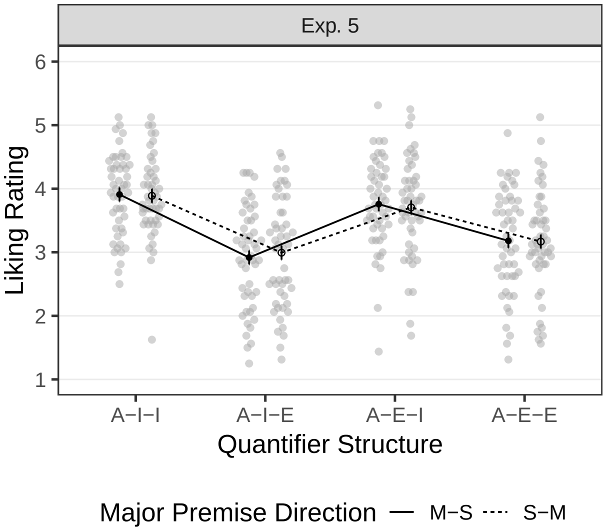


Figure A8

Mean (black symbols) and individual (gray symbols) logic ratings in Experiments 4 as a function of inference type. Vertical jitter was added to individual logic ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

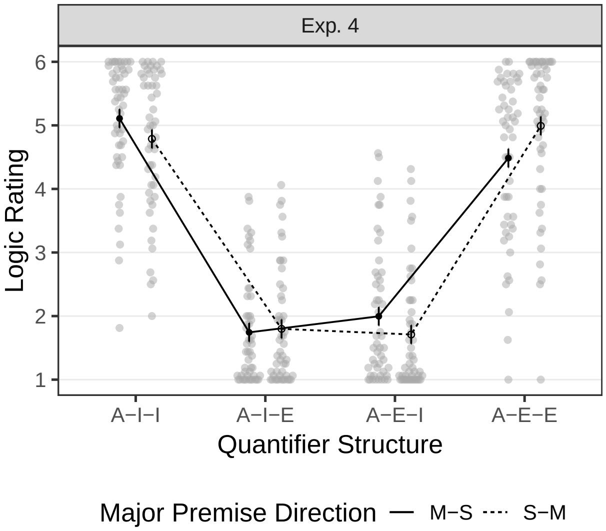


Figure A9

Mean (black symbols) and individual (gray symbols) logic ratings in Experiments 5 as a function of inference type. Vertical jitter was added to individual logic ratings to avoid perfect overlap of two ratings. Error bars show ±1SE (model based).

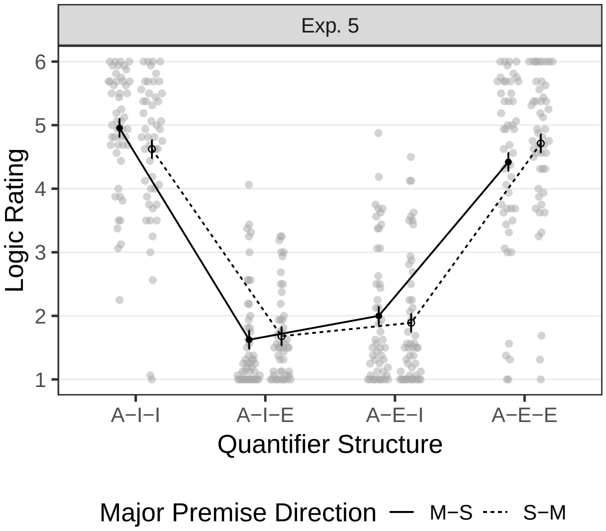


Table A1

The simple effect sizes (d) and p-values for the structure effect on liking ratings between valid and invalid arguments when controlling for different confounds

	Valid					
	vs. inv.		vs. indet.		vs. cong.	
	d	*p*	*d*	*p*	*d*	*p*
Exp. 1	0.34	< .001	0.29	< .001	−0.03	.523
Exp. 2	0.58	< .001	0.44	< .001	0.04	.411
Exp. 3	0.13	.008	0.11	.018	0.07	.172
Exp. 4	0.54	< .001	0.41	< .001	0.08	.051
Exp. 5	0.13	.029	0.08	.100	0.00	.915

Note. Inv. = invalid; indet. = indeterminately invalid; cong. = atmosphere-congruent and indeterminately invalid.

Table A2

The simple effect sizes (d) and p-values for the structure effect on logic ratings between valid and invalid arguments when controlling for different confounds

	Valid					
	vs. inv.		vs. indet.		vs. cong.	
	d	*p*	*d*	*p*	*d*	*p*
Exp. 2	2.04	< .001	1.58	< .001	0.38	< .001
Exp. 3	1.85	< .001	1.46	< .001	0.38	< .001
Exp. 4	2.30	< .001	1.78	< .001	0.41	< .001
Exp. 5	2.13	< .001	1.65	< .001	0.31	.002

Note. Inv. = invalid; indet. = indeterminately invalid; cong. = atmosphere-congruent and indeterminately invalid.