Journal of Experimental Psychology:
Learning, Memory, and Cognition
2013, Vol. 39, No. 4, 1265–1273

© 2012 American Psychological Association
0278-7393/13/$12.00 DOI: 10.1037/a0030530

RESEARCH REPORT

Does Logic Feel Good? Testing for Intuitive Detection of Logicality in Syllogistic Reasoning

Karl Christoph Klauer and Henrik Singmann
Albert-Ludwigs-Universität Freiburg

Recent research on syllogistic reasoning suggests that the logical status (valid vs. invalid) of even difficult syllogisms can be intuitively detected via small changes in affective state (Morsanyi & Handley, 2012). In a series of 6 experiments, we replicated effects of logical status on liking ratings of difficult syllogisms (although their shape differs from that reported by Morsanyi and Handley), and we tested 2 alternative accounts of our and Morsanyi and Handley's findings in terms of surface features accidentally confounded with logical status: the partial-repetition hypothesis and the content-effects hypothesis. The results support the content-effects hypothesis, according to which the effects of logical status reflect differences in mean liking for the presented conclusions rather than effects of logical status itself.

Keywords: dual-process theories, fluency, intuition, liking ratings, syllogistic reasoning

In many situations, judgments are based on hunches or gut feelings that we cannot justify rationally. Such intuitions, Topolinski (2011) argued, may underlie a number of puzzling findings such as hidden covariation detection (Lewicki, 1986), implicit grammar learning (Pothos, 2007), and unconscious thought (Dijksterhuis, 2004).

Topolinski (2011) distinguished intuitions from analytic judgments, a distinction that resonates with current dual-process and dual-system theories of human reasoning (e.g., Evans, 2009). In these theories, human reasoning is believed to involve at least two dissociable systems or classes of processes. Only one of them is tuned to following normative prescriptions (System II), whereas the other one reflects a more heuristic mode of reasoning (System I). System I heuristics are fast and relatively effortless, and they rely on extra-logical characteristics of the logical problems (e.g., on the believability of the conclusion) to generate intuitive responses that are often logically correct. They are logically correct to the extent to which the extra-logical characteristics are correlated with logical validity. In this framework, it seems natural to consider intuitions as the experiential output of System I heuristics.

On the other hand, Morsanyi and Handley (2012), henceforth referred to as MH, recently argued that logicality of even difficult syllogisms can be detected in an intuitive manner via slight changes in affective state. Sensitivity of intuitions for logicality would blur a central defining distinction between System I and System II, between heuristic and analytic processes: Only the latter are considered sensitive to the normative prescriptions of logic, whereas the former capitalize on extra-logical characteristics of the problems.

For example, in their Experiment 4, MH presented valid and invalid syllogisms such as

1. *Valid:* Some snakes are poisonous. No poisonous animals are obbs. Some snakes are not obbs.

2. *Invalid:* No ice creams are vons. Some vons are hot. Some ice creams are not hot.

Premises (the first two sentences) and conclusion (the last sentence) of each syllogism were presented sequentially. Participants were asked to read the three sentences carefully and to indicate how much they liked the last statement by clicking on one of five smileys/sad faces arranged in a 5-point Likert scale ranging from 1 (*don't like it at all*) to 5 (*like it very much*). Instructions stated that "when you make the liking judgment focus on your feeling about the statement. Don't think about why you like or dislike the statement, just go with your intuition and gut feelings" (MH, p. 609).

The syllogisms varied in logical validity and, orthogonally, in the believability of the conclusion: There were syllogisms with believable conclusion (see Example 2 above) and syllogisms with unbelievable conclusions. In addition, some syllogisms' conclusions contained a nonword (see Example 1); these were termed abstract. Across several experiments, MH observed effects of believability so that believable conclusions were liked more than unbelievable conclusions. More important, there was an effect of validity: Conclusions of valid syllogisms were generally liked slightly better than conclusions of invalid ones.

This article was published Online First October 22, 2012.

Karl Christoph Klauer and Henrik Singmann, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

The research reported in this paper was supported by Grant Kl 614/33-1 from the Deutsche Forschungsgemeinschaft to Karl Christoph Klauer.

Correspondence concerning this article should be addressed to Karl Christoph Klauer, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany. E-mail: christoph.klauer@psychologie.uni-freiburg.de

The syllogisms employed by MH in their Experiments 2 and 4 (see Table 1) are among the most difficult ones for human reasoners to evaluate correctly (e.g., Dickstein, 1978). Classical theories of syllogistic reasoning cast the evaluation of the logical validity of these syllogisms as an effortful, resource-demanding process, requiring the intentional and goal-directed manipulation and coordination of multiple mental representations (see Khemlani & Johnson-Laird, 2012, for an overview). Given this theoretical background, it is quite surprising that reasoners should be sensitive to the syllogisms' logical status in an intuitive and nonintentional manner. As already mentioned, this would also pose a challenge for dual-process and dual-system accounts of reasoning.

According to MH, the effects of validity are mediated by *conceptual* fluency. In this view, the premises are more likely to prime a valid syllogism's conclusion conceptually than that of an invalid syllogism, leading to greater perceived fluency for valid than for invalid syllogisms. Perceived fluency in turn drives affective ratings. Taken together, logical validity is argued to be the cause of the effect on liking ratings mediated by conceptual fluency. As acknowledged by MH, it is, however, possible that the effects of validity are instead mediated by surface features of the syllogisms that are accidentally confounded with logical status. In fact, this seemed a likely possibility for the simple syllogisms used in their Experiments 1 and 3 and in part motivated the use of the more complex syllogisms used in their Experiments 2 and 4.

Yet, surface features are still confounded with logical status. Table 1 shows the syllogistic forms employed in MH's Experiments 2 and 4. All syllogisms have a conclusion of the form "Some A are not C." In valid syllogisms, the end term A occurs in the "some" premise and is again associated with "some" in the conclusion, whereas the end term C occurs in the "no" premise and is again associated with negation in the conclusion. In contrast, in invalid syllogisms, these associations are reversed in going from the premises to the conclusion. Given the well-known links between repetition and liking (e.g., Zajonc, 1980), between repetition and processing fluency (e.g., Butler, Berry, & Helman, 2004), and between fluency and liking (e.g., Winkielman & Cacioppo, 2001), it is perhaps plausible to assume that the validity effect on liking ratings is driven by these surface characteristics rather than by a deep, underlying conceptual variable, logical status. Let us refer to the hypothesis that the particular confound just described causes the validity effect on liking ratings as the *partial-repetition hypothesis.*

Table 1
The Eight Syllogistic Forms

Syllogism	Valid	Invalid
Premise	No C are B.	No A are B.
Premise	Some B are A.	Some B are C.
Conclusion	Some A are not C.	Some A are not C.
Premise	Some A are B.	Some C are B.
Premise	No B are C.	No B are A.
Conclusion	Some A are not C.	Some A are not C.
Premise	Some B are A.	Some B are C.
Premise	No C are B.	No A are B.
Conclusion	Some A are not C.	Some A are not C.
Premise	No B are C.	No B are A.
Premise	Some A are B.	Some C are B.
Conclusion	Some A are not C.	Some A are not C.

Experiments 1a, 1b, and 1c

MH were the first to demonstrate the surprising and theoretically challenging validity effect on liking ratings. It is desirable to see whether this initial demonstration can be replicated, given that replicability is a basic criterion for admitting empirical findings to further scientific debate.

A second purpose in the present experiments was to evaluate a number of alternative accounts of the effects such as the partial-repetition hypothesis. We focus on the above syllogisms because MH acknowledged that these provide stronger support for their claims than the easier syllogisms they used in Experiments 1 and 3.

The present Experiments 1a, 1b, and 1c implemented replications of the intuitive condition of MH's Experiment 4 in which liking ratings for the syllogisms' conclusions were obtained.[1] Furthermore, we introduced a new control condition in order to disentangle the effects of logical status and confounded surface features. MH used a nonword as one of the A, B, or C terms (see Table 1) in each syllogism. Each nonword occurs in two of the three sentences making up the syllogism. In the control condition, we simply used different nonwords in the first and second sentence. For example, the control versions for the above two example syllogisms were the following:

- *Pseudo-valid:* Some snakes are poisonous. No poisonous animals are obbs. Some snakes are not ubbs.
- *Pseudo-invalid:* No ice creams are vons. Some vens are hot. Some ice creams are not hot.

Using two similar nonwords renders the syllogisms logically invalid but should leave surface features (such as partial repetitions) largely intact. We refer to the thus-generated sets of sentences as pseudo-syllogisms and classify a given pseudo-syllogism as pseudo-valid or pseudo-invalid depending upon whether it was generated from a valid or invalid syllogism, respectively.

One group of participants received (German translations of) MH's syllogisms, and a second group of participants saw the corresponding pseudo-syllogisms. The questions were (a) whether we could replicate the validity effect on liking ratings and (b) whether pseudo-validity would have the same effect on liking ratings for the pseudo-syllogisms. A pseudo-validity effect on liking ratings cannot reflect effects of logical status (because all pseudo-syllogisms are invalid) but must go back to effects of confounded surface features. If the pseudo-validity effect mirrors the validity effect, this would in turn suggest that the same surface features, rather than logical status, are responsible for the validity effect.

Experiments 1a, 1b, and 1c differed only in minor details, so it is efficient to describe them together. For Experiment 1a, we

[1] Pursuing the possibility of an evaluative-priming effect, MH also presented pleasant or unpleasant faces between the conclusion and the liking rating in this experiment (but not in the other ones). We omitted presentation of the faces, because our question was whether the basic validity effect could be replicated and not whether an evaluative-priming effect can additionally be obtained.

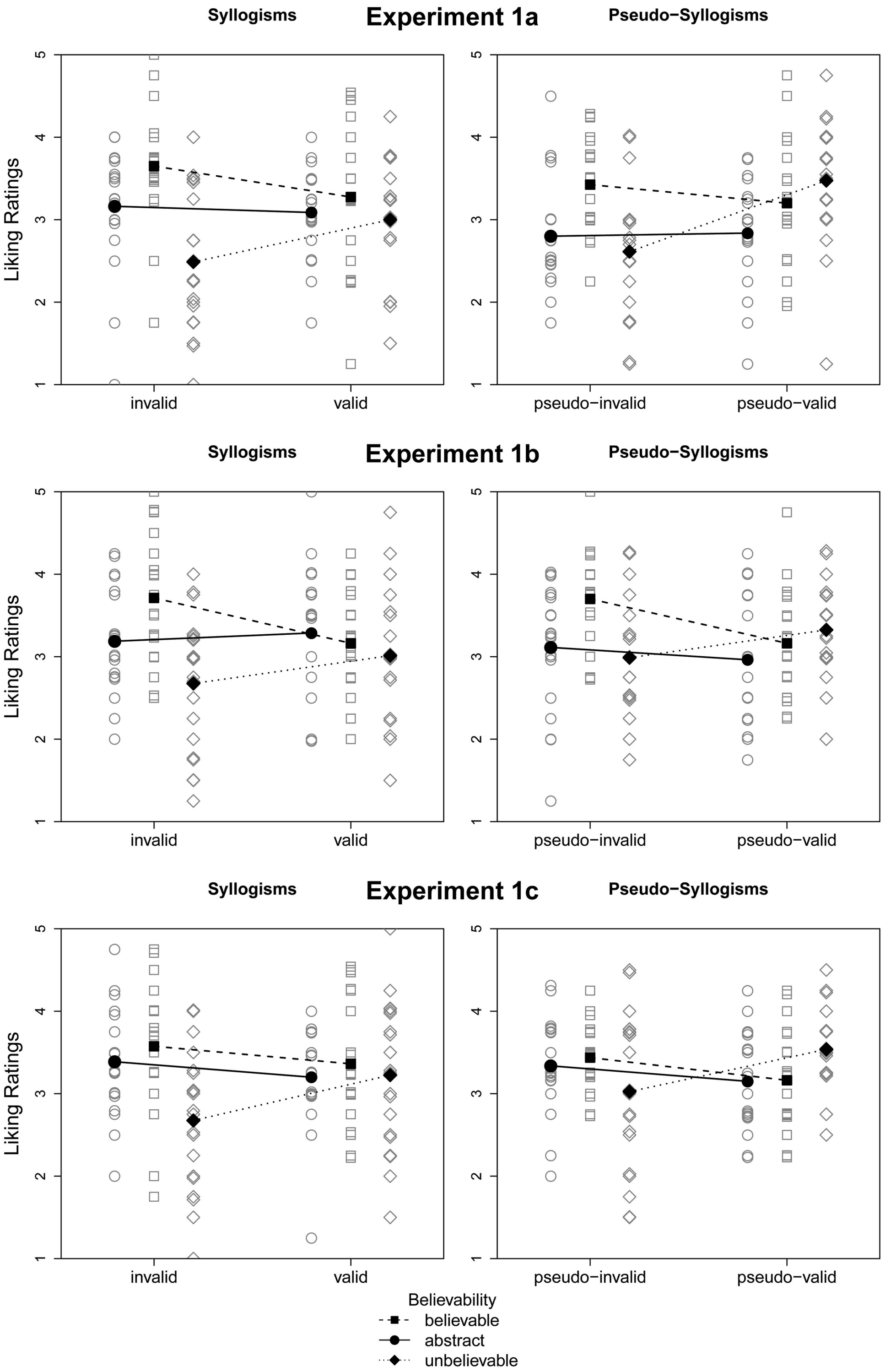

Figure 1. Mean (filled symbols) and individual (nonfilled symbols) liking ratings in Experiments 1a (upper panels), 1b (middle panels), and 1c (lower panels) for the groups with syllogisms (left panels) and those with pseudo-syllogisms (right panels) as a function of validity/pseudo-validity and conclusion believability. A small amount of vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings.

slightly emphasized the above-described partial repetitions by using "no" in the premise with negation (as in MH) and in the conclusion (instead of "not" as in MH or instead of negative adjectives such as "inedible"). Because the results differed somewhat from those reported by MH, we used literal translations of MH's syllogisms for Experiments 1b and 1c. The German translations also contained more letters than the original sentences (an average of 61 and 73 letters per English and German syllogism, respectively). To compensate, we increased presentation times from 2 s (as in Experiment 1a and in MH) to 2.5 s per sentence in Experiment 1b. As pointed out by an anonymous reviewer, the presentation schedule in MH's Experiment 4 may have been faster than that implied in their other experiments, and some participants may not have noticed the change in nonwords as a consequence of the fast presentation schedule. In Experiment 1c we stepped up presentation times to 3.5 s per sentence (pretests had established that even longer presentation times were experienced as inducing boredom by our participants). We also explicitly mentioned in Experiment 1c that nonwords would be shown and that each individual nonword would occur exactly two times and exactly one time in the group with syllogisms and in the group with pseudo-syllogisms, respectively. This was to ensure that even shallow readers would be aware of the fact that two different nonwords were employed per pseudo-syllogism in the group with pseudo-syllogisms.

Method

Participants. In each of Experiments 1a, 1b, and 1c, there were 40 participants, who were randomly assigned to one of the two groups so that there were 20 participants per group in each experiment. Mean ages were, in order, 23.3 years (*SD* = 3.4), 21.1 years (*SD* = 2.7), and 24.1 years (*SD* = 4.3) in Experiments 1a, 1b, and 1c. Most participants (in order, 38, 37, and 36) were University of Freiburg students with different majors. In all experiments reported in this paper, participants received either partial course credit or a small monetary compensation for participating.

Materials. Syllogisms were converted into pseudo-syllogisms by replacing a random one of the two occurrences of a nonword by a similar nonword, generated by exchanging a vowel in the original nonword and another vowel.

Procedures. The procedures closely followed MH's Experiment 4, intuitive condition. A syllogism or pseudo-syllogism was presented sentence by sentence with a presentation time of 2 s (2.5 s and 3.5 s in Experiments 1b and 1c, respectively) and a blank interval of 0.5 s between sentences. The last statement was followed by the above-described Likert scale for the liking rating of that statement. Following MH, we instructed participants to read the sentences carefully, to rely on their intuition and gut feelings in rating the last statement, and not to think about why they liked or disliked the statement. The 24 syllogisms or pseudo-syllogisms were presented in random order. We presented an additional warm-up syllogism (or pseudo-syllogism) based on a different content prior to the 24 experimental syllogisms.

Design. Each experiment followed a design with between-participants factor group (group with syllogisms vs. group with pseudo-syllogisms) and within-participants factors (pseudo-) validity (valid/pseudo-valid vs. invalid/pseudo-invalid) and believability (unbelievable, abstract, and believable).

Results

Figure 1 shows the mean and individual liking ratings for the groups with syllogisms (left panels) and the groups with pseudo-syllogisms (right panels) as a function of conclusion believability and (pseudo)validity for each experiment. As can be seen, the pattern of ratings is similar in all groups.

The ratings were submitted to analyses of variance with factors believability, (pseudo)validity, and group (group with syllogisms vs. group with pseudo-syllogisms) with repeated measures on the first two factors. This revealed a main effect of believability in both Experiments 1a and 1b,[2] $F(1.88, 71.59) = 7.26$, $\eta^2_G = .09$, $p = .002$, and $F(1.65, 62.60) = 6.60$, $\eta^2_G = .06$, $p = .004$, respectively, but not in Experiment 1c, $F(1.91, 72.53) = 2.08$, $\eta^2_G = .03$, $p = .13$. In all three experiments, there was an interaction of believability and (pseudo)validity: In order, $F(1.73, 65.78) = 15.52$, $\eta^2_G = .08$, $p < .001$; $F(1.80, 68.48) = 12.34$, $\eta^2_G = .06$, $p < .01$; and $F(1.96, 74.40) = 11.38$, $\eta^2_G = .06$, $p < .01$, for Experiments 1a, 1b, and 1c. Group entered no significant effects or interactions in any of the experiments, largest $F = 2.55$, largest $\eta^2_G = .03$, smallest $p = .09$ (for the interaction of group and believability in Experiment 1a), whereas the interaction of believability and (pseudo-)validity was individually significant in each of the six groups, smallest $F = 5.14$, smallest $\eta^2_G = .05$, largest $p = .02$. As can be seen in the figure, unbelievable conclusions of (pseudo-)valid syllogisms were liked more than those of (pseudo-)invalid syllogisms and vice versa for believable conclusions.

Discussion

Results were relatively clear cut. We did not replicate the main effect of validity reported by MH, whereas we replicated the main effect of believability, with believable conclusions liked more than abstract and unbelievable ones, in two of three experiments. Nevertheless, logical validity exerted significant (interactive) effects on the liking ratings. However, exactly the same effects were observed for pseudo-validity. This strongly suggests that surface features shared by the syllogisms and the pseudo-syllogisms are responsible for the parallel effects of logical validity and of pseudo-validity.

The pattern of findings is thus not supportive of the idea that logical validity can be intuitively detected, but it is also not consistent with the partial-repetition hypothesis. That hypothesis would predict the same pattern of effects as reported by MH and in particular a main effect of validity and pseudo-validity.

In Experiments 2 and 3, we therefore focused on another alternative hypothesis in terms of extra-logical features that we term the *content-effects* hypothesis. Remember that participants were instructed to rate how much they like the conclusions. Thus, liking ratings might in large part reflect just this, degree of liking of the conclusions considered in isolation. For example, like MH we found that believable conclusions were liked somewhat better overall than unbelievable and abstract ones.

Different conclusions must be used to manipulate conclusion believability, because the same conclusion cannot be both believ-

[2] Degrees of freedom are Greenhouse–Geisser corrected, and effect sizes are specified in terms of generalized eta-squared as per Bakeman's (2005) recommendations.

able and unbelievable. But in MH's Experiments 2 and 4, different sets of conclusions (and premises) were also used for the valid and invalid syllogisms. That is, for each believability by validity condition, a different set of contents was used without counterbalancing of sets across validity conditions. Thus, content is confounded with validity within each believability condition. What is more, MH's premises and conclusions frequently employed affectively laden words such as *snake, rich, millionaire, diamond, shark, friendly, unhealthy, dangerous, ice cream, criminal,* and so forth, rendering it likely that the different contents by themselves provoked different degrees of liking.

The content-effects hypothesis states that MH's and our effects of logical validity reflect an accidental confound between logical validity and mean degree of liking of the different sets of conclusions. Because syllogisms and pseudo-syllogisms in our Experiment 1 used the same conclusions (up to an occasional exchange of the nonword and a similar nonword in the conclusions of abstract syllogisms), the hypothesis accounts for the absence of differences between the groups with syllogisms and pseudo-syllogisms. In addition, the differences between MH's and our effect pattern would flow from ubiquitous differences in explicit and implicit attitudes that exist between substantially different samples of participants. For example, the present German sample and MH's English sample are likely to differ in many cultural values and norms, leading to somewhat different degrees of liking for statements about fast cars, rich people and millionaires, sweet drinks, snakes, and so forth as used as conclusions.

Experiments 2 and 3 tested two predictions of the content-effects hypothesis. Experiment 2 examined whether the interaction of validity and believability consistently found in Experiments 1a to 1c would emerge even if only the conclusions are presented without premises. Experiment 3 tested the prediction that the effects of validity should be erased when the covariation of conclusions and validity is disrupted.

Experiment 2

In Experiment 2, we presented the conclusions without any premises. Here, logical status is nominally assigned on the basis of the syllogism of which the conclusion was a part in MH's experiments.

Method

The procedure was the same as in Experiment 1b with the exception that the presentation of the premises was omitted. The 31 participants (mean age = 23.0 years, *SD* = 4.8) were mostly (27 of 31) University of Freiburg students with different majors.

Results and Discussion

Figure 2 shows the mean and individual liking ratings as a function of nominal logical status and believability. There was an interaction of believability and validity, $F(1.84, 55.27) = 3.57$, $\eta^2_G = .03$, $p = .04$, of the same shape as that in Experiment 1. The main effect of believability fell short of significance, $F(1.79, 53.71) = 2.96$, $\eta^2_G = .04$, $p = .07$. Descriptively, believable conclusions received somewhat more positive ratings than abstract and unbelievable conclusions, as before.

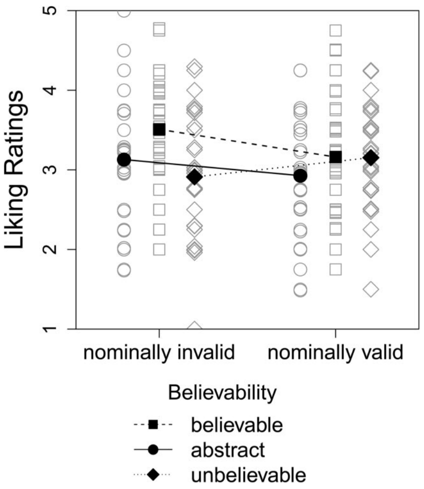

Figure 2. Mean (filled symbols) and individual (nonfilled symbols) liking ratings in Experiment 2 as a function of nominal validity and conclusion believability. A small amount of vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings.

The results were relatively clear cut. Presenting conclusions without premises is sufficient to reproduce the interactive effect of logical validity on liking ratings found in Experiment 1. This strongly suggests that the content-effects hypothesis contributes to accounting for the pattern of results observed in Experiment 1.

Experiment 3

Considering effect sizes, the interactive effect in Experiment 2 ($\eta^2_G = .03$) appeared to be somewhat smaller than those observed in Experiment 1 ($\eta^2_G \geq 0.5$). Experiment 3 tested whether content effects were sufficient to account for the interactive effects of validity in Experiment 1. Two groups were contrasted. One group, the group with fixed conclusions, was a replication of the group with syllogisms of Experiment 1b. The second group, the group with randomized conclusions, saw syllogisms with conclusions randomly assigned to the validity conditions within each believability condition.

For this purpose, we created eight syllogisms, four valid, four invalid, from each of MH's syllogisms by rearranging premises and terms. The eight syllogisms shared the conclusion of the original syllogism and implemented all eight forms shown in Table 1. An example is shown in Table 2 for the above syllogism with the ice-cream content (i.e., for Example 2 from the introduction). This made it possible to rotate conclusions across validity conditions across participants. If the content-effects hypothesis is true, randomizing conclusions across logical status should eliminate any effects of logical validity.

Method

The procedure was the same as in the group with syllogisms in Experiment 1b with the following exceptions.

Participants. The 60 participants (mean age = 22.1 years, *SD* = 3.1) were randomly assigned to one of the two groups so that

Table 2
Eight Syllogisms Generated From the Ice-Cream Content

Syllogism	Valid	Invalid
Premise	No hot things are vons.	No ice creams are vons.
Premise	Some vons are ice creams.	Some vons are hot.
Conclusion	Some ice creams are not hot.	Some ice creams are not hot.
Premise	Some ice creams are vons.	Some hot things are vons.
Premise	No vons are hot.	No vons are ice creams.
Conclusion	Some ice creams are not hot.	Some ice creams are not hot.
Premise	Some vons are ice creams.	Some vons are hot.
Premise	No hot things are vons.	No ice creams are vons.
Conclusion	Some ice creams are not hot.	Some ice creams are not hot.
Premise	No vons are hot.	No vons are ice creams.
Premise	Some ice creams are vons.	Some hot things are vons.
Conclusion	Some ice creams are not hot.	Some ice creams are not hot.

there were 30 participants per group. Most of the participants (57 of 60) were University of Freiburg students.

Materials. The syllogisms in the group with fixed conclusions were the ones used in Experiment 1b in the group with syllogisms. For the group with randomized conclusions, we generated 24 tables such as Table 2, one for each of the 24 original syllogisms.[3]

In the group with randomized conclusions, the 24 syllogisms presented to a participant were randomly sampled from these 24 tables with the restrictions (a) that one syllogism was drawn from each table, so that all 24 original conclusions were presented, and (b) that all eight forms shown in Table 1 were presented in each of the three believability conditions. Note in particular that this implies that logical status and believability are crossed orthogonally. This randomization was performed for each participant anew.

Results and Discussion

Figure 3 shows the mean and individual liking ratings in the group with fixed conclusions (left panel) and the group with randomized conclusions (right panel) as a function of logical validity and believability. An analysis of variance with factors group, believability, and validity revealed a three-way interaction of these factors, $F(1.85,\ 107.52) = 3.58$, $\eta^2_G = .01$, $p = .03$. Separate analyses were run for each group to elucidate the nature of the interaction. In the group with fixed conclusions, there were a main effect of believability, $F(1.58, 45.92) = 4.61$, $\eta^2_G = .06$, $p = .02$, and an interaction of believability and validity, $F(1.75, 50.88) = 8.66$, $\eta^2_G = .05$, $p < .001$, of the same shape as that in Experiment 1. In the group with randomized conclusions, there was only a main effect of believability, $F(1.87, 54.28) = 3.94$, $\eta^2_G = .05$, $p = .03$, whereas neither the main effect of validity nor its interaction with believability reached significance: $F(1, 29) = 0.45$, $\eta^2_G = .003$, $p = .51$, and $F(1.94, 56.31) = 2.25$, $\eta^2_G = .01$, $p = .12$, respectively.[4]

Results were thus relatively clear cut. Randomizing conclusion content across valid and invalid logical forms eliminated all effects associated with logical status, as predicted by the content-effects hypothesis. In contrast, the interactive effects of validity were again significant in the group with fixed conclusions. Importantly, this difference between groups was significant, so that the elimination of the effects of validity is not just a null effect. Instead, the effect of logical validity appears to depend critically on the confound between conclusions and logical validity in the group with fixed conclusions.

Experiment 4

Although both MH and ourselves observed significant effects of validity, the effects differ in shape. On the basis of the content-effects hypothesis, we attribute these differences to ubiquitous differences in evaluative attitudes between samples of participants with different backgrounds. It is, however, difficult to test this account of the differences directly, because we do not have access to MH's samples of participants and cannot run a replication of our experiments on them. In consequence, it is possible (a) that MH's participants would have expressed identical degrees of liking for the different conclusion sets when they were presented without premises or as part of pseudo-syllogisms and (b) that the effects of logical validity reported by MH were indeed genuine effects of validity.

If so, we should, however, have seen a main effect of logical validity as reported by MH when we controlled for the effects of conclusions on liking ratings in our Experiment 3 in the group with randomized conclusions. But all effects of validity were eliminated in that group. The effect size of the main effect of validity in MH's Experiment 4 was $\eta^2_G = .14$. This implies a power of $1 - \beta = .93$ for detecting that effect in the group with randomized conclusions in our Experiment 3, rendering the null finding relatively strong.

Nevertheless, there must be many small differences in procedure between our and MH's experiments, and there are probably substantial differences between the samples of participants. In consequence, the main effect of logical validity may be smaller in our case than in MH's situation. Experiment 4 implemented a powerful test for it based on randomizing conclusions as in Experiment 3. We left out the abstract syllogisms because of the difficulties

[3] Note, however, that for the abstract syllogisms, this creates highly unbelievable premises in some cases, such as that no poisonous animals are snakes in the first example syllogism above. To avoid this, we permitted the four valid and the four invalid syllogisms to be associated with different conclusions of the form "Some X are not Z" and "Some Z are not X," respectively (i.e., with the order of end terms interchanged). Note that either X or Z is a nonword in abstract syllogisms, and we therefore did not expect the two conclusions with different order of end terms to differ substantially in degree of liking associated with them. Exchanging the order of end terms does, however, introduce a new confound with validity for the abstract syllogisms that we accepted as the lesser evil relative to presenting syllogisms with highly invalid premises.

[4] In the group with randomized conclusions, a new confound for the abstract syllogisms was permitted to avoid presenting highly unbelievable premises (see Footnote 3). When the abstract syllogisms were left out of the analyses, all effects just reported as significant remained significant, and all effects involving validity in the group with randomized conclusions were associated with F values smaller than one, largest $\eta^2_G = .0008$.

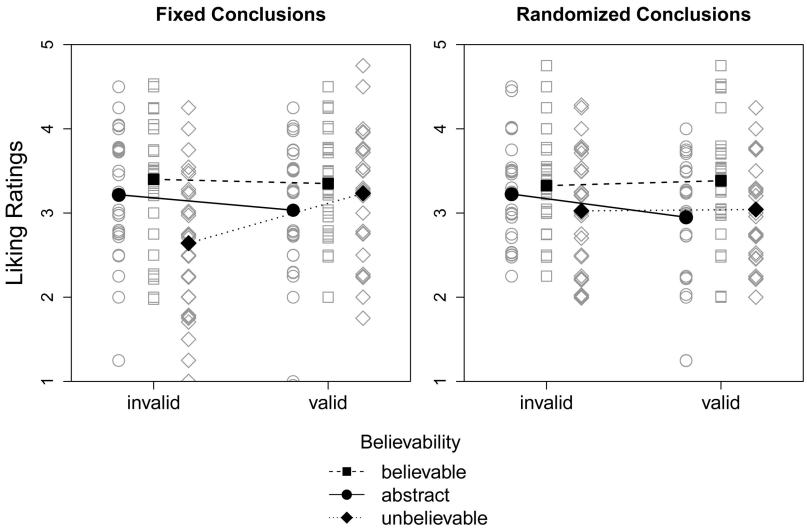

Figure 3. Mean (filled symbols) and individual (nonfilled symbols) liking ratings in Experiment 3 for the group with fixed contents (left panel) and the group with randomized contents (right panel) as a function of validity/pseudo-validity and conclusion believability. A small amount of vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings.

associated with randomizing their conclusions across validity conditions, as described in Footnote 3.

Method

The procedure was the same as in the group with randomized conclusions in Experiment 3 with the exception that we did not present abstract syllogisms.

The 200 participants (mean age = 24.0 years, SD = 4.3) were mostly (185 of 200) University of Freiburg students with different majors.

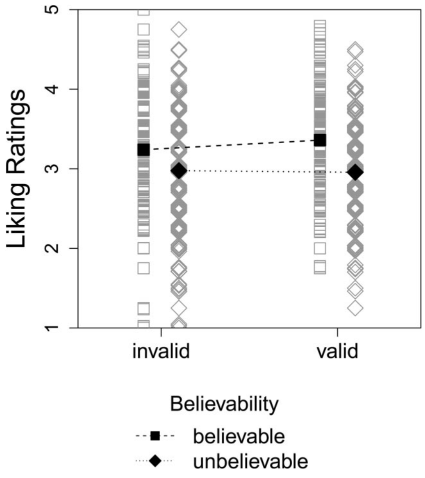

Figure 4. Mean (filled symbols) and individual (nonfilled symbols) liking ratings in Experiment 4 as a function of validity and conclusion believability. A small amount of vertical jitter was added to individual liking ratings to avoid perfect overlap of two ratings.

Results and Discussion

Figure 4 shows the mean and individual liking ratings as a function of logical status and believability. There was a main effect of believability, $F(1, 199) = 43.17$, $\eta^2_G = .06$, $p < .001$. Neither the main effect of validity, $F(1, 199) = 1.40$, $\eta^2_G = .001$, $p = .24$, nor the interaction of believability and validity, $F(1, 199) = 3.07$, $\eta^2_G = .003$, $p = .08$, reached significance.

Experiment 4 was sufficiently large to permit the estimation of a mixed linear model with both participants and conclusions as random factors. As explained by Judd, Westfall, and Kenny (2012), substantial biases are inherent in analyses that ignore one or the other of these random factors. We fitted the model that Judd et al. (2012, shooter data, pp. 62–63) recommended for the design of Experiment 4. Accordingly, participants and conclusions are treated as random effects, with random error components for the intercept, believability effect, validity effect, and believability by validity interaction for participants and random error components

[5] Considering the variance components in this mixed model, none of the variances and covariances involving validity were significantly different from zero, meaning that effects involving validity are the same for all participants and conclusions ($\chi^2[df = 9] = 8.60$, $p = .47$, in a log-likelihood ratio test of the full model against a model with validity removed as random factor). Thus, these analyses do not encourage one to search for individual differences in possible effects of validity. There was, however, strong evidence for heterogeneity across participants in the effects of believability ($\chi^2[df = 4] = 26.75$, $p < .001$, in a contrast of the full model and a model without random error components for the believability effect for participants), suggesting that participants differ in the degree to which believability of the conclusions affects their liking ratings. Consistent with the content-effects hypothesis, a significant variance component in the liking ratings was due to conclusions ($\chi^2[df = 1] = 136.11$, $p < .001$, in a contrast of the full model without random component for the validity effect for conclusions and a model without a random component for the intercept for conclusions).

for the intercept and validity effect for conclusions, as well as the covariances between these random effects, using the Kenward–Roger approximation for the tests of fixed effects. For the fixed effects, this more rigorous analysis revealed a main effect of believability, $F(1, 15.21) = 4.69$, $p = .046$. Neither the main effect of validity, $F(1, 13.91) = 0.46$, $p = .46$, nor the interaction of believability and validity, $F(1, 12.40) = 2.27$, $p = .16$, reached significance.[5]

General Discussion

The present research sought to replicate and extend MH's surprising finding that people can intuitively detect logicality of even difficult syllogisms. Although we did not replicate the precise effects pattern reported by MH, we did observe (interactive) effects of logical status on intuitive liking ratings in Experiment 1.

A new control condition introduced a small change in the wording of the syllogisms, creating what we called pseudo-syllogisms, all of which are invalid but share formal surface features and contents with the original syllogisms. We observed the same effects for the control problems as for the syllogisms, which suggests that the effects are caused by shared surface features.[6]

One of these surface features is a formal one that we termed partial repetition. The hypothesis that partial repetitions drive the effects did not, however, find support. Another confound is in terms of contents, because content was not counterbalanced in MH's Experiments 2 and 4. It turns out that the effects of logical validity are eliminated when content is controlled for via randomization (Experiment 3). Furthermore, presenting the conclusions without premises produces much the same effects as in Experiment 1 with premises present. We conclude that the effects of logical validity observed in the conditions with confounded contents in large part reflect preexisting differences in the mean liking of the different conclusions.

We attribute the differences in the precise effect patterns of our Experiment 1 and MH's Experiment 4 to differences between our samples of participants in the mean liking of the different conclusions. We acknowledge, however, that it is difficult to test this account of the differences directly, given that we do not have access to MH's samples of participants. However, regardless of this issue, randomizing conclusions across validity conditions should have allowed us to reproduce the main effect of logical validity reported by MH, if there is an intuitive sensitivity for logicality expressed as shifts in liking. In fact, randomizing conclusions did not achieve this either in Experiment 3 or in the powerful Experiment 4.

Taken together, the present findings suggest that it may be premature to postulate an intuitive grasp of logicality for even difficult syllogisms expressed as shifts in momentary affect. As discussed in the introduction, an intuition for logical status would have been difficult to square with classical theories of syllogistic reasoning that reserve the detection of validity to effortful, slow, and analytic processes. Moreover, it would have posed a challenge for current dual-system theories of reasoning in that it would question a central defining distinction between heuristic System I and analytic System II processes: Only the latter are considered sensitive to the normative prescriptions of logic, whereas the former capitalize on extra-logical characteristics of the problems that may or may not correlate with logical validity.

For such reasons, it is in our opinion worthwhile to continue searching for intuitive detection of logicality. There are many procedural variations that could be explored, and conditions may exist that do generate a genuine effect of logical validity on liking ratings. For example, it is possible that a genuine effect of validity would occur under self-paced presentation conditions, as used in MH's Experiments 1 to 3, or for simple syllogisms, as used in MH's Experiments 1 and 3. In pursuing this interesting line of research, it would be helpful to rely on the protective value of extensive randomization and counterbalancing for avoiding artifactual effect patterns and to use control problems such as the present pseudo-syllogisms, which share surface characteristics with the logical problems but are all equally invalid.

The new control technique may also be helpful in related lines of research. For example, De Neys (2012) has argued that people have an intuitive access to the normatively correct response in Kahneman and Tversky's (1973) famous base-rate neglect problems. In these problems, base-rate information is normatively relevant because the instances to be judged are said to be *randomly* sampled from a population with specified base rates. Pseudo-versions of these problems can be obtained simply by stating instead that the instances to be judged are *not randomly* sampled from the population, thereby disrupting the normative link between the base-rate information and the instances to be judged while leaving surface features of the problems unchanged. Use of such pseudo-problems as a control condition could be helpful in determining whether De Neys and colleagues' effects are driven by formal or content-related surface features that accidentally covary with the normative implications or whether the normative implications themselves are causally responsible.

[6] One criticism of this control condition might be based on Ball and Quayle (2009), who contrasted syllogisms in which all terms were phonologically nondistinctive (e.g., Some bubs are bebs. No bebs are babs. Therefore, some bubs are not babs) with syllogisms with phonologically distinctive content (e.g., Some zaps are toks. No toks are yugs. Therefore, some zaps are not yugs). Ball and Quayle found logical performance to be enhanced in the distinctive relative to the nondistinctive condition. The two nonwords occurring in pseudo-syllogisms were nondistinctive, so that the syllogisms and the pseudo-syllogisms might differ somewhat in distinctiveness. In consequence, differences in the effects pattern for syllogisms and pseudo-syllogisms could have reflected distinctiveness effects, but it seems difficult to account for the observed absence of differences in terms of differences in distinctiveness.

References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37,* 379–384. doi:10.3758/BF03192707

Ball, L. J., & Quayle, J. D. (2009). Phonological and visual distinctiveness effects in syllogistic reasoning: Implications for mental models theory. *Memory & Cognition, 37,* 759–768. doi:10.3758/MC.37.6.759

Butler, L. T., Berry, D. C., & Helman, S. (2004). Dissociating mere exposure and repetition priming as a function of word type. *Memory & Cognition, 32,* 759–767. doi:10.3758/BF03195866

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

Dickstein, L. S. (1978). The effect of figure on syllogistic reasoning. *Memory & Cognition, 6,* 76–83. doi:10.3758/BF03197431

Dijksterhuis, A. (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology, 87,* 586–598. doi:10.1037/0022-3514.87.5.586

Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 33–54). New York, NY: Oxford University Press.

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,* 54–69. doi:10.1037/a0028347

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251. doi:10.1037/h0034747

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

Lewicki, P. (1986). Processing information about covariations that cannot be articulated. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 135–146. doi:10.1037/0278-7393.12.1.135

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,* 596–616. doi:10.1037/a0026099

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin, 133,* 227–244. doi:10.1037/0033-2909.133.2.227

Topolinski, S. (2011). A process model of intuition. *European Review of Social Psychology, 22,* 274–315. doi:10.1080/10463283.2011.640078

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation leads to positive affect. *Journal of Personality and Social Psychology, 81,* 989–1000. doi:10.1037/0022-3514.81.6.989

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35,* 151–175. doi:10.1037/0003-066X.35.2.151

Received July 13, 2012
Revision received September 10, 2012
Accepted September 12, 2012 ■