Model-Based Explanation of Feedback Effects in Syllogistic Reasoning

Daniel Brand[1,2] **(daniel.brand@cognition.uni-freiburg.de)**

Nicolas Riesterer[1] **(riestern@cs.uni-freiburg.de)**

Marco Ragni[1,2] **(ragni@sdu.dk)**

[1]Cognitive Computation Lab, University of Freiburg, 79110 Freiburg, Germany

[2]South Denmark University, Denmark

Abstract

In the field of syllogistic reasoning research, a significant number of models aiming at describing the human inference processes were developed. There is profound work fitting the model's parameters and analyzing each model's ability to account for the data in order to support or disprove the underlying theories. However, the model parameters are rarely used to extract explanations and hypotheses for phenomena that go beyond the original scope of the models. In this work, we apply three state-of-the-art models, PHM, mReasoner, and TransSet, to data from reasoning experiments where participants received feedback for their conclusions. We derived hypotheses based on the models' explanations for the feedback effect and putted these to test by conducting an experiment targeting the hypotheses. The work contributes to the field in three ways: (a) the feedback effect could be replicated and was shown to be a robust effect; (b) we demonstrate the use of the model parameters in order to derive new hypotheses; (c) we present possible explanations for the feedback effect based on existing theories.

Keywords: syllogistic reasoning; cognitive modeling; mReasoner; PHM; TransSet; feedback

Introduction

Routinely, psychological experiments are conducted to uncover robust effects and phenomena related to the latent processes of the human mind. Assumptions to shed light on the internals of the black box that constitutes the human mind are compiled into theories that are then corroborated or falsified based on comparison with experimental data.

Models, on the one hand, instantiate theories incorporating the knowledge about robust effects and phenomena that were found through observations in experiments. By providing measures to quantify the capabilities of a model to account for real world processes, this ultimately allows to test and verify the assumptions underlying the respective theories. On the other hand, models that have proven to be good accounts for their respective processes can also be transferred to different scenarios. In this way, models can be used to extract predictions even for hypothetical scenarios, which can subsequently be used to derive new hypotheses that fuel further investigations.

Consider for example the domain of human syllogistic reasoning, which will serve as the domain of interest throughout this article. Traditionally, syllogisms consist of two premises featuring one out of four quantifiers ("All", "Some", "No", "Some ... not") and two out of three categorical terms ("A", "B", and "C"):

All A are B.
Some B are C.

What, if anything, follows?

The goal of syllogistic reasoning is to interrelate the terms in the premises via the common *middle-term* ("B"), and derive information about the quantified relationship between the other two *end-terms* ("A", "C") or conclude "No Valid Conclusion" (NVC) to state that no quantified conclusion can be derived from the premises on logical grounds. For the sake of space and clarity, syllogisms are often abbreviated based on their structure. A syllogism is in one of four so-called figures, which represent the arrangement of terms:

Figure 1	Figure 2	Figure 3	Figure 4
A-B	B-A	A-B	B-A
B-C	C-B	C-B	B-C

Additionally, the quantifiers are encoded with A, E, I, O for "All", "No", "Some", "Some not", respectively (notation adopted from Khemlani & Johnson-Laird, 2012). Put together, the syllogism introduced before would be abbreviated with "AI1".

Due to the structural restrictions (exactly two premises, three terms, and four quantifiers), syllogistic reasoning is a well-defined domain with a total of 64 distinct problems and nine possible conclusion options. Because of this, syllogistic reasoning is one of the prime domains to study human deductive reasoning and explore hypotheses about the latent inferential processes of the human mind.

To date there exist at least twelve theories that try to explain the observable behavior of reasoners by drawing from a century worth of empirical investigation (Khemlani & Johnson-Laird, 2012). Models based on these theories provide sets of comprehensive and explanatory parameters to fine-tune the processes they assume to be operating the human mind. These parameters, in essence, are responsible for the explanatory value of theories as they provide the necessary information about the selection and strength of the processes that are responsible for the observable behavior. There is considerable work fitting parameters to data (Khemlani & Johnson-Laird, 2016; Riesterer, Brand, & Ragni, 2020a), which focuses on the ability of the models to account for the data. However, very little work focuses on the second use-case for these models, namely to go beyond the scope that they were originally created for and extract explanations for new phenomena based on the interpretation of the parameters. This, however, considerably undervalues the worth of theories and models. Reflecting embodiments of insight, theories and models are capable of providing novel insight and should do so.

In this article, we attempt to use model implementations of theories to provide insight in the syllogistic reasoning process and the changes that feedback induces to these processes. Relying on a recent dataset that introduced feedback about the logical correctness of human responses as an experimental manipulation (Dames et al., 2020), we fit three prominent models (mReasoner, PHM, and TransSet) to the data. By investigating the resulting parameter distributions, we extract explanations for the effects of feedback from the theories. We then derived hypotheses that allow to experimentally test the theories explanations. At last, we conducted a study based on a modified version of the experiment by Dames et al. (2020) which featured additional questions targeting the derived hypotheses. This allowed us to replicate the feedback effect in order to ensure its robustness and test the hypotheses derived from the model's explanations of the feedback effect.

The remainder of this paper is structured into four parts. First, we present relevant background about the theories and models for syllogistic reasoning. Second, we introduce our method of extracting explanations from the models and derive the hypotheses. Third, we describe the study and the dataset derived from it. Fourth, we present our results and discuss them with respect to the implications for the three models and the feedback effect, as well as the general implications for the field of syllogistic reasoning research.

Background

To date, syllogistic reasoning research has produced more than twelve theories attempting to explain the cognitive foundation of this form of reasoning (Khemlani & Johnson-Laird, 2012). Crucially, it was found that comparing these theories based on their ability to predict the distinctive responses of human reasoners to select an overall best explanation is difficult if possible at all (Khemlani & Johnson-Laird, 2012). Recently, however, it was found that in addition to this difficulty, predictive performances might have been overestimated due to a prevailing perspective on group analyses in the field (Riesterer, Brand, & Ragni, 2020c). If subjected to the task of predicting individual human responses instead of only the most frequently selected ones, predictive accuracies drop from above 84% (Khemlani & Johnson-Laird, 2012) to below 50% (Riesterer, Brand, & Ragni, 2020a). To see if this performance can notably be improved on—which would be clear evidence of an improved understanding of reasoning processes—or remains stuck due to high levels of noise in the data remains a crucial goal for future investigations in reasoning research.

Regardless of the questions surrounding model selection, recent results suggest that at least three accounts will play a major role in future investigations for various reasons. First, the *Mental Models Theory* (MMT; Johnson-Laird, 1983) with its model implementation *mReasoner* (Khemlani & Johnson-Laird, 2013) is one of the most comprehensive theoretical accounts of reasoning spanning multiple domains (e.g., spatial relational, conditional, modal) and persisting for almost half a century. Second, the *Probability Heuristics Model* (Chater & Oaksford, 1999) is an instance of the probabilistic paradigm of cognitive science that adopts a stance discarding logical validity in favor of probabilistic validity. Finally, *TransSet* (Brand et al., 2020) is a recently proposed account that approaches syllogistic reasoning by focusing on a set-based interpretation of quantifiers and transitivity as its core inference rule. Currently, TransSet is the most successful model of syllogistic reasoning when judged based on predictive accuracy alone (Brand et al., 2020). In the following, the functional mechanisms of the three accounts will be introduced in greater detail.

MMT & mReasoner *MMT* approaches syllogistic reasoning via a four-step procedure (e.g., Copeland, 2006). First, a mental representation, the *mental model*, is created from the first premise. This mental model consists of a number of entities that reflect the information of the premise by being associated to the categorical terms or not. Second, the second premise is integrated into the mental model by extending the entities with information about the third term. Third, the resulting mental model is inspected to extract a conclusion candidate. In the final step, this candidate is probed by constructing alternative mental model representations that are consistent to the premises but inconsistent to the conclusion candidate. If no counterexample can be found, the conclusion is accepted as the conclusion to the syllogistic problem. Otherwise a new conclusion candidate is generated and subjected to the search for counterexamples or NVC is returned.

mReasoner is a LISP-based implementation of MMT for syllogistic reasoning that follows the four-step procedure outlined above but includes four parameters to further specify details about the model's behavior (e.g., Khemlani & Johnson-Laird, 2016). First, λ specifies the maximum number of entities that are represented in the mental model. Second, ε specifies the composition of the mental model. For high values, the mental model is highly likely to exhaustively reflect the information available in the premises. For low values, it only reflects a limited canonical set of information. Third, σ reflects the propensity of the model to engage the search for counterexamples. Finally, if a counterexample is found, ω denotes the likelihood to continue the process with a weaker version of the conclusion candidate or abort the reasoning process to generate an NVC response.

PHM PHM approaches reasoning by adopting a perspective based on probabilistic validity or p-validity (Chater & Oaksford, 1999). To accomplish this without requiring computationally complex if feasible at all operations, the model is based on a set of three generation heuristics (G1-G3) and two test heuristics (T1, T2) to approximate the p-valid behavior. To generate a conclusion, the *min-heuristic* (G1) identifies the premise with minimal informativeness (min-premise) based on the order $A > I > E > O$ and uses its quantifier as the conclusion quantifier. *p-entailment* (G2) proposes the quantifier probabilistically following from the min-heuristic result as an alternative conclusion quantifier candidate. The *attachment-heuristic* then defines the direction of the conclusion. If the min-premise begins with an end-term, it is used as the subject of the conclusion. Otherwise the end-term of the max-premise, i.e., the most informative premise in accordance to the above ranking, is used. After the conclusion is generated, the *max-heuristic* (T1) assesses a reasoner's confidence in it by evaluating the informativeness of the max-premise. PHM assumes proportionality between confidence and max-premise informativeness. If confidence is low, NVC may be concluded instead (Copeland, 2006). Finally, the *O-heuristic* postulates that "Some ... not" conclusions should generally be avoided due to their extreme uninformativeness.

In a recent implementation of PHM (Riesterer, Brand, & Ragni, 2020a), a set of five binary parameters were used to further specify the model's behavior. *p_ent* decides whether to use the min-heuristic or p-entailment to generate the conclusion quantifier. In addition, *A_conf*, *I_conf*, *E_conf*, *O_conf* are used to specify the confidence in the corresponding max premise quantifier.

TransSet TransSet is based on two phases: direction selection and quantifier selection (Brand et al., 2019, 2020). In *direction selection*, TransSet attempts to construct a transitive path from the premises. If this is not possible, the model returns NVC. Otherwise, it enters the *quantifier selection phase* in which the quantifier information is propagated along the transitive path. This procedure fails and leads to NVC if the first quantifier on the path is negative and the second quantifier is not all. Otherwise, the conclusion quantifier is obtained and can be combined with the direction to create the full conclusion.

TransSet uses four parameters to further specify its inferential mechanisms. First, *nvc_aversion* defines its susceptibility to the NVC aversion bias that might prevent reasoners from acknowledging the importance of this conclusion (e.g., Brand et al., 2020). In the direction selection phase, NVC aversion forces the model to create a transitive path regardless of the premises. *anchor_set* determines which term to start the transitive path from in this case. Third and fourth, *particularity* and *negativity* specify the availability of additional rules to directly derive NVC in the quantifier selection phase (Riesterer, Brand, Dames, & Ragni, 2020).

Method

Objective

The goal of the analyses presented in the following is to leverage the current understanding of human reasoning in form of available model implementations in order to investigate the effects of feedback. By providing feedback about the logical correctness to reasoners, it is expected that the reasoning behavior changes. These changes should be reflected by different parameterizations resulting from fitting the models to the data. By interpreting the difference in parameter values, the effects of feedback on reasoning behavior can be analyzed and compared to the theoretical assumptions postulated previously (Dames et al., 2020; Riesterer, Brand, & Ragni, 2020b).

Dataset

To investigate the effects of feedback, we rely on the three datasets collected by Dames et al. (2020). First, `control` ($N = 39$) contains the control group of reasoners who were not provided with feedback about the correctness of their responses. Second, `1s` ($N = 146$) contains the group of reasoners who were presented with a feedback screen stating either "correct" or "incorrect" after each given response. Finally, `10s` ($N = 29$) contains the group of reasoners who were presented with feedback for a duration of 10s after each response. Regardless of the feedback condition, all participants were presented with the full set of 64 distinct syllogistic problems and tasked to select which of the nine possible conclusion options (including "No Valid Conclusion") followed from the presented premises. There was a time limit for each task, which forced participants to respond within 1.5 minutes.

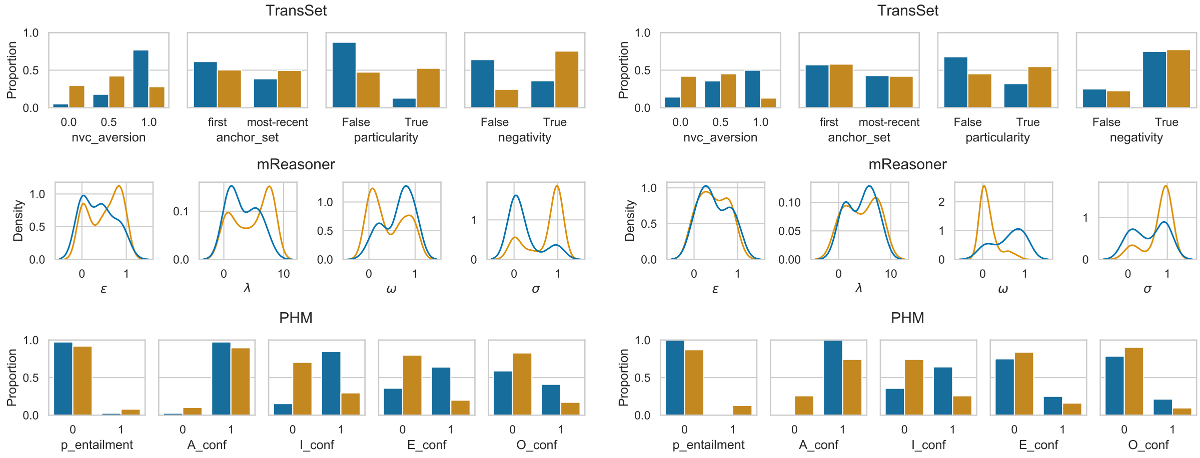
Performing traditional statistical (Dames et al., 2020) and data-driven modeling analyses (Riesterer, Brand, & Ragni, 2020b), feedback was shown to predominantly affect the propensity of reasoners to conclude NVC, a conclusion option that has previously been hypothesized to elicit aversion biases (Dickstein, 1976). Presenting feedback provides reasoners with the opportunity to realize the importance of the NVC response (correct in 37 out of the 64 syllogistic problems, i.e., 58%). As such, we expect models to reflect this increase in NVC usage in terms of their parameterizations. As the overall differences between the 10s condition and the 1s condition were rather small compared to the control group, we combined both feedback conditions for the following analysis.

Model Fitting

The analyses presented in the following rely on the *Cognitive Computation for Behavioral Reasoning Analysis* (CCOBRA) framework[1]. CCOBRA facilitates the evaluation of computational models in a well-defined and structural manner and provides implementations for the three models considered in the analyses: *mReasoner* (Riesterer, Brand, & Ragni, 2020a), *PHM* (Riesterer, Brand, & Ragni, 2020a), and *TransSet* (Brand et al., 2020). Each model was fitted to each individual in the dataset separately. The resulting fits were then aggregated and broken down by the feedback condition.

The core results of our analysis are summarized in Figure 1a. The figure contains separate plots for each of the mod-

[1] `github.com/CognitiveComputationLab/ccobra`



(a) Dataset by Dames et al. (2020)   (b) Our Study

Figure 1: Parameter value distributions resulting from fitting the models to individual reaoners based on data from the original feedback study (left) and the study conducted in this work (right). Control and feedback are depicted in blue and orange, respectively.

els' parameters. Each plot visualizes the distribution of the resulting values in terms of their proportions of occurrence (TransSet and PHM due to the parameters being discrete) or distribution (mReasoner due to being continuous). The different feedback conditions are represented by color with control and feedback in blue and orange, respectively.

On a high level, the plots reveal the obvious: the feedback manipulation of the experimental setting has an influence on human reasoning behavior that is reflected by differences in the fit results. To work out the explanatory meanings from the fits, the following sections inspect the results of each model separately and derive a hypothesis from the possible explanation.

TransSet TransSet shows distinct differences between control and the feedback condition for *nvc_aversion*, *particularity*, and *negativity*. The parameter *anchor_set* is ignored in the following due to its technical purpose and relative uniformity between the different conditions.

The value of the *nvc_aversion* parameter is substantially higher for the control condition than for feedback. In the case of *particularity*, control exhibits a strong skew in favor of *False* with the feedback condition leaning slightly towards *True*. For *negativity*, a similar skew can be observed but to a minor degree, at least for control. The feedback condition are skewed stronger towards *True*.

To summarize, TransSet attributes the effects of feedback to NVC handling, which is not surprising as it is TransSet's main method of distinguishing individuals. The reduced value of *nvc_aversion* suggests that feedback incentivizes reasoners to accept NVC more leniently when compared to the behavior of naive reasoners (control). A similar interpretation is suggested by *particularity* and *negativity*, which control the availability of rules to abort the reasoning process in favor of NVC. With control leaning more towards *False* and the feedback conditions to *True*, TransSet suggests that feedback allows reasoners to find and leverage heuristic rules to easily derive NVC, the response that naive reasoners (control) try to avoid.

As it is assumed that participants in the feedback condition use fast detection methods allowing them to identify NVC responses early, it is expected that the difference between the time needed for NVC responses and non-NVC responses is lower for the feedback group compared to the control group (**H1.1**). Although NVC is usually important in more difficult tasks, the NVC-specific heuristics could outweigh the difficulty and lead to overall lower times for NVC responses in the feedback condition compared to the control group (**H1.2**).

mReasoner Interestingly, mReasoner's parameter distributions are bimodal between control and the feedback conditions. Perhaps most crucially, the σ parameter is substantially affected by feedback. As the parameter controls the propensity to engage in a search for counterexamples, which the prerequisite to derive NVC responses, this was to be expected. Less distinctly, λ and ε show similar behavior with control and feedback being skewed towards lower and higher values, respectively. For ω, which only plays a role within the search for counterexamples and therefore dependent on σ, feedback is mainly skewed towards the lower spectrum, while the control condition yields higher values.

To summarize, mReasoner seems to attribute the effects of feedback to a switch from a more intuitive reasoning to a more thought-out process incorporating a search for counterexamples. The propensity to rigorously evaluate the mental model via the search for counterexamples is increased (σ), and the likelihood to weaken the conclusion (which in turn allows to avoid an NVC response) is reduced. Additionally, a

Table 1: Syllogisms selected for the test phase of the study. The encoding is in line with Khemlani & Johnson-Laird (2012).

Valid	AA4, AE2, AO3, EA1, EI1, IA4, IE4, OA3
Invalid	EE2, EO3, II4, IO3, OA2, OE1, OI3, OO1

learning component can be identified: When confronted with feedback, reasoners realize the importance of correctly interpreting the premise information resulting in more comprehensive (λ) and complete (ε) mental models. However, it is important to note that the effects of λ and ε have shown to have little impact on the general behavior of mReasoner in comparison to σ (Riesterer, Brand, & Ragni, 2020a). Due to the expensive search for counterexamples, it is expected, that participants in the feedback group should be substantially slower when deriving NVC responses (**H2**).

PHM PHM's parameterization is special because of the dependencies between the confidence parameters *A_conf*, *I_conf*, *E_conf*, and *O_conf*, which control the behavior of the max-heuristic. Since this heuristic states that confidence in the conclusion is *proportional* to the max-premise quantifier's informativeness (Chater & Oaksford, 1999), the corresponding parameters are ordered. As soon as one parameter is set to 0, all proceeding ones must necessarily be 0 as well, indicating that confidences are so low that the conclusion is abandoned in favor of NVC.

The order of confidences is reflected by the model parameters with proportions of 1s decreasing from *A_conf* through *O_conf*. Importantly, across the board, control elicits the highest proportions of 1s with both feedback conditions eliciting similar results. *p_entailment*, the only truly independent parameter, is dominated by 0s regardless of the condition.

To summarize, PHM suggests that feedback results in an overall decrease of confidence in conclusions potentially caused by the importance of the NVC response. Therefore, the confidence in non-NVC responses should be lower for the feedback group (**H3**).

Study design

Based on the hypotheses described above, we conducted an online-study via Prolific, in which participants were instructed to give conclusions to all 64 syllogistic problems. The study had a single-choice design, where the participants selected the conclusion by clicking on the respective button. In order to avoid a bias due to content effects, hobbies and professions were used as content for the syllogisms. The order of the response options was randomized. Participants were randomly assigned to the control condition or to the feedback condition. The experiment was divided into two parts: The first 48 syllogism (presented in random order) were regarded as a training phase, where feedback was shown for 1 second for the feedback condition. As in the original experiment, the feedback only stated if the selected answer was correct. After the first 48 syllogisms, both groups received no feedback as we assumed that the feedback effect would be apparent after training. In the second phase (test phase), the participants were asked to not only select an answer, but also to estimate their confidence in the selected option by choosing values from 0% to 100% on a slider. A predefined set of syllogism was used, which featured the 8 valid and 8 invalid syllogisms that had the most differences with respect to the response behavior between feedback and non-feedback in the dataset from Dames et al. (2020), with the constraint that only unique quantifier-combinations were in the set. This was done to increase the variability and to minimize the effect of single strategies and biases (e.g., the Atmosphere effect Wetherick & Gilhooly, 1995). The selected tasks are shown in Table 1. We did not include a time limit, which allows us to disentangle the effect of feedback from the effect that the short time-frame might have had in the original study.

After excluding participants which did not take the experiment seriously (i.e., needed less than 10 minutes for all tasks, performed worse than chance, or interrupted the study for more than 5 minutes; $N = 6$), there were $N = 59$ participants, with $N = 28$ in the control group and $N = 31$ in the feedback condition. The dataset, all materials and scripts are openly available on GitHub[2].

Analysis

First, we compared the dataset from our study with the dataset by Dames et al. (2020). In particular, we investigated if, and to which extend, the feedback effect is apparent without a time limit. Second, we re-fitted the models to the new data in order to verify that the main predictions still hold. Subsequently, the hypotheses derived from the model's explanations were tested based on the results of the second phase of the study. In the following section, the results are presented and discussed.

Results

The feedback effect in the original study mainly manifested in the number of NVC responses (Dames et al., 2020). This effect was also apparent in our data, as the average percentage of NVC responses (control: $mean = 0.21$, $std = 0.17$; feedback: $mean = 0.41$, $std = 0.23$) showed higher values for the feedback condition. However, the correctness (control: $mean = 0.46$, $std = 0.50$; feedback: $mean = 0.48$, $std = 0.50$) was not affected, which differs from the results by Dames et al.. This is likely the effect of the time limit, which caused the control group to perform worse (0.326 without feedback; 0.434 with feedback), while they achieved a closer result in our data.

Figure 1b shows the results of the model fits on the data from our study. Overall, the model parameters still show the same pattern, clearly showing the feedback effect. However, the effect is not as dominant as in the original study, which is

[2] github.com/Shadownox/iccm-feedbackexplanation

Table 2: Results of Mann-Whitney U tests (p-values and U statistic) for the hypotheses H1.1, H1.2, H2, and H3. p_{cor} shows the Bonferroni-corrected p-values.

Hypothesis	Median		U	p	p_{cor}
	Control	Feedback			
H1.1	0.92	-5.1	178.0	.006	.017
H1.2 / H2	21.26	12.06	228.0	.06	.18
H3	71.45	55.5	228.0	.001	.003

especially prominent for PHM the *E_conf* parameter, where the control condition is almost identical to the feedback condition in our data, while there was still a substantial difference in the original study. This is likely due to the missing time limit, which might have strengthened existing biases and pushed participants more towards intuitive responses. For mReasoner, this gets also apparent for the more subtle parameters λ and ε, which are differing substantially between the datasets with control and feedback showing almost not difference without a time limit while having distinct patterns when a time limit is present. Despite these differences, the parameter distributions between both experiments are comparable with respect to the extracted explanations, allowing a evaluation of the derived hypotheses. In the following, we compare the two conditions in order to test the hypotheses. To correct for multiple comparisons, we use the Bonferroni correction and also the corrected p-values. The results of the comparisons between control and feedback conditions for the hypotheses are shown in Table 2.

At first, we discuss the hypothesis of TransSet (H1.1). According to TransSet's mechanism, NVC responses should be derived faster compared to non-NVC responses. As the feedback effect is explained by participants being less hesitant to derive NVC and also the utilization of NVC-specific rules, it is expected that this time difference is higher in the feedback condition compared to the control group. In fact, there is a significantly bigger difference in the feedback condition compared to the control group: participants in the feedback condition are substantially faster when deriving NVC, while the control group even needs more time for NVC responses. Additionally, the overall time of participants for NVC responses was lower in the feedback condition compared to the control group (H1.2), although significance was not reached. The hypothesis for mReasoner (H2) directly contradicts hypothesis H1.1 by predicting participants in the feedback condition to take more time, as they are more likely to engage in the expensive search for counterexamples. Since the data is even leaning towards H1.1, it is not supported by the data.

Finally, the prediction of PHM is tested. PHM predicts the confidence in non-NVC responses to be lower in the feedback group. This prediction was indeed supported by the data, showing a significantly lower confidence in the feedback condition compared to the control group. This indicates that NVC could in fact be an option that is selected if participants have low confidence in other response options.

Discussion

The present work has three main contributions: First, we fitted three cognitive models that are state-of-the-art, PHM, mReasoner and TransSet, to each individual participant and used the resulting parameter distributions to extract explanations based on the assumed processes of the respective model. At last, these were used to derive new hypotheses that allowed us to test the models' explanatory capabilities. Second, the feedback effect in syllogistic reasoning which was reported by Dames et al. (2020), was replicated by our study without the time limit imposed in the original study. This indicates that the effect is in fact robust and not only an interaction effect induced by the time limit. Third, by testing the hypotheses derived from the models, we were able to assess the models' capabilities to account for the feedback effect.

Regarding the feedback effect for syllogistic reasoning, the explanations extracted from PHM and TransSet were supported by our study. Both explanations are also compatible, as it is possible that feedback at first has the effect of lowering the confidence in non-NVC-responses and later helps to develop fast and frugal detection strategies for NVC once the importance of NVC responses is realized. In contrast, the prediction of mReasoner was not supported and the data seems to even contradict its explanation. Based on our findings, we conclude that the feedback effect is best described as a heuristic process, where participants learn that NVC is a viable response option and therefore adapt their general judgment of the other response options. An explanation based on the assumption that feedback improves the reasoning process (e.g., by shifting away from intuitive responses) could not be supported. In summary, our findings indicate that PHM and TransSet are more probable accounts for the feedback effect in syllogistic reasoning. While they provide differing explanations, they might describe different parts of the same process.

Generally, our work successfully applied cognitive models for syllogistic reasoning to a new phenomenon in order to derive new hypotheses by interpreting the parameters, which is rarely done in this field. Instead, it is often the other way round: New findings were first integrated in theories and then into the respective models. While this is a valid approach to formalize the current knowledge about human reasoning into models, it does not utilize the predictive capabilities of the models. We hope that future modeling endeavors will test models more based on predictions outside their original scope, which will not only improve model selection, but also advance the field as a whole by fueling further investigations with new hypotheses.

Acknowledgements

Support to MR by the Danish Institute of Advanced Studies and the DFG (RA 1934/4-1, RA1934/8-1) is gratefully acknowledged.

References

Brand, D., Riesterer, N., & Ragni, M. (2019). On the matter of aggregate models for syllogistic reasoning: A transitive set-based account for predicting the population. In T. Stewart (Ed.), *Proceedings of the 17th International Conference on Cognitive Modeling* (pp. 5–10). Waterloo, Canada: University of Waterloo.

Brand, D., Riesterer, N., & Ragni, M. (2020). Extending TransSet: An individualized model for human syllogistic reasoning. In T. C. Stewart (Ed.), *Proceedings of the 18th International Conference on Cognitive Modeling* (pp. 17–22). University Park, PA: Applied Cognitive Science Lab, Penn State.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*(2), 191–258.

Copeland, D. E. (2006). Theories of categorical reasoning and extended syllogisms. *Thinking & Reasoning*, *12*(4), 379–412.

Dames, H., Schiebel, C., & Ragni, M. (2020). The role of feedback and post-error adaptations in reasoning. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 3275–3281). Cognitive Science Society.

Dickstein, L. S. (1976). Differential difficulty of categorical syllogisms. *Bulletin of the Psychonomic Society*, *8*(4), 330–332.

Johnson-Laird, P. N. (1983). *Mental models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.

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

Khemlani, S. S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*(1), 4–20.

Khemlani, S. S., & Johnson-Laird, P. N. (2016). How people differ in syllogistic reasoning. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2165–2170). Cognitive Science Society.

Riesterer, N., Brand, D., Dames, H., & Ragni, M. (2020). Modeling human syllogistic reasoning: The role of "No Valid Conclusion". *Topics in Cognitive Science*, *12*(1), 446–459.

Riesterer, N., Brand, D., & Ragni, M. (2020a). Do models capture individuals? Evaluating parameterized models for syllogistic reasoning. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3377–3383). Toronto, ON: Cognitive Science Society.

Riesterer, N., Brand, D., & Ragni, M. (2020b). Feedback influences syllogistic strategy: An analysis based on joint nonnegative matrix factorization. In T. C. Stewart (Ed.), *Proceedings of the 18th International Conference on Cognitive Modeling* (pp. 223–228). University Park, PA: Applied Cognitive Science Lab, Penn State.

Riesterer, N., Brand, D., & Ragni, M. (2020c). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. *Topics in Cognitive Science*, *12*(3), 960–974.

Wetherick, N. E., & Gilhooly, K. J. (1995). 'Atmosphere', matching, and logic in syllogistic reasoning. *Current Psychology*, *14*(3), 169–178.