Do Models Capture Individuals? Evaluating Parameterized Models for Syllogistic Reasoning

Nicolas Riesterer* **(riestern@cs.uni-freiburg.de)**
Daniel Brand* **(daniel.brand@cognition.uni-freiburg.de)**
Marco Ragni (ragni@cs.uni-freiburg.de)
Cognitive Computation Lab, University of Freiburg
Georges-Koehler-Allee 52, 79110 Freiburg, Germany

Abstract

The prevailing focus on aggregated data and the lacking group-to-individual generalizability it entails have recently been identified as a major cause for the low performance of cognitive models in the field of syllogistic reasoning research. This article attempts to add to the discussion about the performance of current syllogistic reasoning models by considering the parameterization capabilities some cognitive models offer. To this end, we propose a model evaluation setting targeted specifically toward analyzing the capabilities of a model to fine-tune its inferential mechanisms to individual human reasoning data. This allows us to (1) quantify the degree to which models are able to capture individual human reasoning behavior, (2) analyze the efficiency of the parameters used by models, and (3) examine the functional differences between the prediction capabilities of competing models on a more detailed level. We apply this method to two state-of-the-art models for syllogistic reasoning, mReasoner and the Probability Heuristics Model, analyze the obtained results and discuss their implication with respect to the general field of cognitive modeling.

Keywords: cognitive modeling; syllogistic reasoning; mental models; probabilistic heuristics model; individualization

Introduction

Analyzing aggregated representations of data (e.g., response distributions) and interpreting the results thereof allows researchers to minimize noise and other sources of adversarial variance which may obscure the actual effects of interest (e.g., Eid et al., 2015). This is one of the reasons why aggregated, or group-based, analyses of data have traditionally been the standard procedure to investigate experimental data (e.g., in form of statistical hypothesis testing). Recent studies, however, criticized the use of aggregation in empirical sciences by demonstrating lacking generalizability of group-based results to the individual level (Fisher et al., 2018). Rekindling the long-standing controversy about group-to-individual generalizability (e.g., Molenaar, 2004), they warn that by disregarding inter-individual differences, researchers might fail to accurately describe the natural processes in question and risk arriving at misleading results (Fisher et al., 2018).

One domain in which problems with respect to group-to-individual generalizability have been surfacing recently is syllogistic reasoning (e.g., Riesterer et al., 2019), one of the core domains of deductive reasoning research. Syllogisms are quantified statements (featuring the quantifiers "All", "Some", "No", and "Some ... not") of the following form:

All A are B
All B are C

What, if anything, follows?

A syllogism consists of three terms (A, B, and C). The middle term (B) occurs in both premises and is used to connect their informational content. The end terms (A, C) only occur in one of the premises and denote the categories whose relationship is to be determined via deductive inference. In total, there are 64 distinct syllogistic problems with nine possible responses each, eight of which are obtained by connecting both end terms in either direction (A-C, C-A) via one of the four quantifiers, and "no valid conclusion" (NVC) to indicate that no logically valid response can be derived from the premises.

With a long history of research focusing on aggregations such as response distributions, syllogistic models were considered to provide reasonable explanations of human reasoning behavior (Khemlani & Johnson-Laird, 2012). More recent evaluations subjecting the traditional syllogistic models to individual response data found that individuals differ substantially from the "average reasoner", i.e., the response behavior obtained from the most frequent answers, the models tried to capture (Riesterer et al., 2019). However, by evaluating general predictions of models (as reported in Khemlani & Johnson-Laird, 2012) on individual human responses, these evaluations focused on investigating the overall group-to-individual generalizability of models and therefore disregarded the fact that at least two of the cognitive models offer parameterization capabilities. Intended to fit models to groups of reasoners (e.g., Khemlani & Johnson-Laird, 2016), the parameters should theoretically allow the models to be fine-tuned to the response behavior of individuals. This enables us to determine how well the models' assumed cognitive processes align with the inferential mechanisms employed by human reasoners.

In this article, we investigate the individualization capabilities of parameterized models for human syllogistic reasoning. To this end, we propose a model evaluation setting that aims at measuring a model's ability to "cover" the response behavior of individual reasoners, i.e., the ability to provide a specification of the reasoner in terms of its parameteriza-

*Both authors contributed equally to this manuscript.

tion (e.g., Farrell & Lewandowsky, 2018). In this setting, we exemplarily evaluate two state-of-the-art models for syllogistic reasoning, mReasoner (Khemlani & Johnson-Laird, 2013) and the Probability Heuristics Model (Chater & Oaksford, 1999), analyze the obtained results and discuss them with respect to their implications for the field of syllogistic reasoning research as well as cognitive modeling in general.

Theoretical Background

For over 100 years, researchers have contributed to the field of syllogistic reasoning research (starting with Störring, 1908). Since then, in its intent to provide accurate descriptions and formalizations for the human inferential processes, the field has generated twelve major cognitive theories (for a review, see Khemlani & Johnson-Laird, 2012) as well as various other models (e.g., Riesterer et al., 2019; Brand et al., 2019; Dietz Saldanha & Kakas, 2019).

The ability of the prevailing twelve theories of syllogistic reasoning to predict significant human response patterns was evaluated in a comprehensive meta-analysis (Khemlani & Johnson-Laird, 2012). Khemlani & Johnson-Laird (2012) collected datasets obtained from psychological experimentation and aggregated them by determining which of the conclusion options were considered "liable", i.e., given by more than 16% of the participants. From their analysis, the authors concluded that while most of the theories give fairly accurate response predictions (71%-84%), the theories differed with respect to their proportions of hits and correct rejections making it difficult to determine an overall best theory of human syllogistic reasoning.

More recent studies found that when subjecting the same models to the task of predicting individual human responses instead of aggregations, accuracies drop to values below 35% (Riesterer et al., 2018, 2019). Instinctively, one might have hoped to dismiss this issue as an effect of inconsistencies and noise that may affect the response pattern of an individual reasoner. However, the model obtained from predicting the response most frequently selected by participants (most-frequent answer, MFA) substantially outperformed the predictive capabilities of models (by 10%) suggesting general flaws in their inferential mechanisms. Moreover, by training data-driven machine learning models on syllogistic response data, it could be shown that individual patterns of syllogistic reasoning exist and can be leveraged to reach higher levels of performance surpassing the MFA (Riesterer et al., 2019). Since the MFA represents the upper bound of performance models disregarding inter-individual differences can possibly achieve, these results illustrate the potential that still remains in the field.

These past evaluations of cognitive models attempted to provide an analysis of their group-to-individual generalizability; they did not leverage the parameterization possibilities some models offer. Consequently, the results do not yet imply that the underlying theories are generally unable to accommodate individuals. They rather suggest that the iconic model predictions derived from the theories do not directly reflect the reasoning behavior of individuals. Unfortunately, though, individualization capabilities were not in the focus of most traditional approaches to modeling human syllogistic reasoning. As a result, only two of the syllogistic models are available as formalisms offering a parameterization for their underlying inferential mechanisms: mReasoner (Khemlani & Johnson-Laird, 2013) and the Probabilistic Heuristic Model (Chater & Oaksford, 1999).

mReasoner

mReasoner (Khemlani & Johnson-Laird, 2013) is a cognitive model that follows the *Mental Model Theory* (MMT) of reasoning (e.g., Johnson-Laird, 2010). At its core, MMT assumes that reasoners construct mental representations of the information contained in the premises by instantiating a set of symbolic entities, which each can be assigned to one or more of the syllogistic terms in accordance to their premise quantification. From the initially constructed model, a first conclusion candidate can be drawn and subsequently validated via a search for counterexamples that checks for inconsistencies with the input premises. If a counterexample is found, the mental model is corrected and the inferential process starts anew until a valid conclusion is found or inference is aborted by concluding NVC.

mReasoner was developed as a LISP-based program[1] that formalizes the theoretical assumptions of MMT by relying on a set of four parameters (e.g., Khemlani & Johnson-Laird, 2016): $\lambda \in (0,8]$, $\varepsilon \in [0,1]$, $\sigma \in [0,1]$, and $\omega \in [0,1]$ (for a summary, see Table 1).

λ specifies the size of the initially constructed mental model by parameterizing a Poisson distribution from which a number of entities is drawn to represent the premise information. The entities are then assigned to the syllogistic terms. For this step, ε specifies the probability to which the degree of completeness of the constructed model is determined. Low values entail that the premise information is only partially reflected by the entities in the constructed model which is one of the mechanisms making the derivation of logically invalid conclusions possible. σ denotes the propensity of mReasoner to engage its search for counterexamples and ω specifies how to continue after a counterexample is found. With the probability of ω, the conclusion quantifier is weakened and a new search for counterexamples starts. Otherwise, NVC is concluded.

Probability Heuristics Model

The *Probability Heuristics Model* (PHM; Chater & Oaksford, 1999) shifts the focus of inference away from logic validity by proposing a set of simple heuristics that approximate probabilistically valid, or *p-valid* (Adams, 1996), inferences. PHM relies on three generative heuristics (G1-G3), which are used to propose a candidate conclusion that is either accepted or rejected by two test heuristics (T1, T2).

[1] github.com/skhemlani/mReasoner

Table 1: Parameters of mReasoner (Khemlani & Johnson-Laird, 2016) and PHM (Chater & Oaksford, 1999) as per the implementations used in the analysis along with their ranges (possible values in the case of PHM) and summarizing descriptions.

Parameter	Range	Description
mReasoner		
λ	$(0,8]$	Maximum number of entities in the mental model.
ε	$[0,1]$	Completeness of the model.
σ	$[0,1]$	Propensity to engage the search for counterexamples.
ω	$[0,1]$	Likelihood to weaken conclusion candidates.
Probability Heuristics Model		
p_ent	$\{0,1\}$	Decides whether min-heuristic or p-entailment generates conclusion.
conf_A	$\{0,1\}$	Confidence based on max premise quantifier "All".
conf_I	$\{0,1\}$	Confidence based on max premise quantifier "Some".
conf_E	$\{0,1\}$	Confidence based on max premise quantifier "No".
conf_O	$\{0,1\}$	Confidence based on max premise quantifier "Some ... not".

The *min-heuristic* (G1) defines the quantifier of the conclusion to be the quantifier of the least informative premise. Chater & Oaksford (1999) define the ranking of informativeness as "All" > "Some" > "No" > "Some ... not". *P-entailment* (G2) proposes the statement that probabilistically follows (p-entails; Chater & Oaksford, 1999) from the min-heuristic candidate as an alternative conclusion candidate for the syllogistic problem. As an example, for the min-heuristic candidate "All A are C", the p-entailed conclusion would be "Some A are C". *Attachment* (G3) finally specifies the order of terms by postulating that if the least informative premise begins with an end-term, it is used as the subject of the conclusion. Otherwise, the end-term of the most informative premise is used.

After the initial conclusion candidate is determined, the *max-heuristic* (T1) postulates that a reasoner's confidence in this conclusion is proportional to the informativeness of the most informative premise. In consequence, if this premise uses uninformative quantifiers, the likelihood of the reasoner to reject it and respond with NVC is increased (Copeland & Radvansky, 2004). Similarly, the *O-heuristic* (T2) states that conclusions featuring the quantifier "Some ... not" are to be avoided because of their uninformativeness (Chater & Oaksford, 1999).

The parameters of PHM serve the purpose to provide a specification of the behavior observable in groups of reasoners (see Table 1). As such, *p_ent* specifies the probability of p-entailment use for deriving an alternative conclusion to what is proposed by the min-heuristic, and *conf_A, conf_I, conf_E*, and *conf_O* represent the confidences in the max-heuristic quantifiers "All", "Some", "No", and "Some ... not", respectively. The probability to conclude NVC instead of the candidate generated by the min-heuristic or p-entailment scales proportionally with these confidence values (Copeland & Radvansky, 2004).

Method

The goal of our analysis is to determine the degree to which cognitive models are capable of capturing the response patterns of individual human reasoners. To pursue this objective, we propose an evaluation setting for cognitive models that puts individuals into the focus of attention. It builds on the idea that parameters of cognitive models indicate properties of the assumed inferential mechanisms: given response data of a human reasoner, we search for a model parameterization minimizing its prediction error. From this, we obtain parameter values reflecting the inferential mechanisms of the reasoner. The resulting fits allow for an in-depth analysis of the model's ability to capture individual reasoning behavior.

The scores obtained from coverage analysis essentially reflect goodness-of-fit (GOF) measures which focus on the capability of accommodating for individuals in terms of model parameters. While there is consensus in the recent literature that an overreliance on GOF may adversarially affect model selection (e.g., Roberts & Pashler, 2000), coverage still adds important evidence to the discussion about model performances. If models (and their underlying assumptions) accurately reflect cognitive processes, they *must* be able to capture individual behavior in terms of their parameterization. Even if the absolute coverage scores may not allow for precise interpretation, their differences and magnitudes can be assessed: One the one hand, significant differences between models highlight substantial differences in their capability to handle individuals, which is important information for future modeling endeavors. On the other hand, comparing the coverage scores of individualized and aggregate models, i.e., models which do or do not provide parameters for capturing individual differences, respectively, allows for an assessment of the necessary condition of individualized models: If the performance of an individualized model does not exceed the performance of its aggregate competitors, its parameters are unable to properly capture the source of individual differences. Consequently, an assessment of the theoretical merit

of the model is not justified. In sum, coverage analyses can serve as an additional tool in the modeler's toolbox based on which the individualization capabilities of models can be assessed comprehensively.

Note that, realistically, we cannot expect models to perfectly predict an individual reasoner's responses because of the noise that is to be expected in human experimental data (e.g., motivational issues or misunderstandings with respect to the task). We therefore contrast the models fitted to individual reasoners with both, their aggregated counterparts (fitted to the whole dataset) and the performance obtained from the most-frequent answer (MFA), a statistical baseline representing the upper bound of performance achievable by approaches that do not leverage the potential of individualization (e.g., Riesterer et al., 2019).

Fitting Cognitive Models

For our analysis we rely on two available models for syllogistic reasoning that offer individualization capabilities via parameterization: mReasoner (Khemlani & Johnson-Laird, 2013) and the Probability Heuristics Model (PHM; Chater & Oaksford, 1999).

mReasoner mReasoner uses its parameters to specify probability distributions that configure its model generation and interpretation processes (e.g., Khemlani & Johnson-Laird, 2016). To find an optimal parameterization, we perform a grid search with 11 steps resulting in a step size of 0.1 for parameters ε, σ, ω, and step size 0.79 for λ (14641 parameter combinations in total). Due to the stochastic nature of its parameter use, we also apply repeated random sampling querying mReasoner for five predictions to obtain an estimate of the expected outcome of its probabilistic inference processes.

Probability Heuristics Model (PHM) PHM (Chater & Oaksford, 1999) is not available publicly as an implemented and parameterized model. However, in its original specification (Chater & Oaksford, 1999), the authors elaborate the various possibilities to individualize its behavior (e.g., by selecting the conclusion-generation heuristic to apply). Based on this, we developed a Python-based implementation while maintaining communication with one of PHM's authors.

The parameters PHM uses differ from mReasoner's parameters in that they specify independent Bernoulli distributions. Since PHM does not include mechanisms to additionally condition these distributions on the syllogistic problem being solved, they globally define the inferential behavior of the model. In our optimization we search for deterministic reconstructions of reasoning behavior. As such, we try to optimize for the expected outcome PHM produces with respect to its coverage accuracy allowing us to consider the parameters as binary flags. For example, even if a reasoner uses p-entailment for 40% of the syllogistic responses, the expected prediction outcome would still be maximized by setting *p_ent* to 0. Consequently, PHM features a much smaller parameter space even though its number of parameters is higher than that of mReasoner. This allows us to exhaustively search for optimal parameterizations.

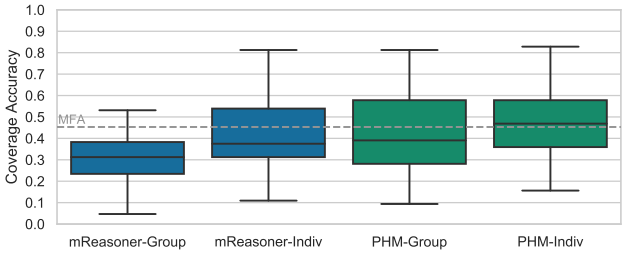

Figure 1: Coverage performances of the models. The boxplots describe the prediction accuracies of the models with boxes ranging between the quartiles, the middle line indicating the median performance, and whiskers extending to the last accuracy value within the inter-quartile range. Depicted are the performances of mReasoner (blue) and PHM (green) each fitted to both group data and individual data. The dashed line represents the median performance achieved by the most-frequent answer (MFA), i.e., the model that always predicts the response given by most participants in the dataset.

Dataset

For our analyses we use the *Ragni2016* dataset obtained from the *Cognitive Computation for Behavioral Reasoning Analysis* (CCOBRA) Framework[2] which contains data from 139 reasoners responding to all 64 syllogistic problems, each. For each problem, participants were instructed to select the conclusion following from both premises out of the nine possible response options (for more details on the dataset, see Ragni et al., 2019). The dataset we used, along with the model implementations and material for this article is publicly available on GitHub[3].

Results

Coverage Performance

Figure 1 depicts the results for the individual coverage performance of the models. It contrasts the models' predictions for the "average reasoner" which were not fitted to individual responses (mReasoner-Group, PHM-Group) with the individually parameterized variants (mReasoner-Indiv, PHM-Indiv). Comparing the predictive accuracies of the previous and new analyses, it becomes apparent that parameterization helps to boost performance slightly (median performance increases of 6% for mReasoner and 8% for PHM). This suggests that the available parameters can indeed be used to capture individual reasoning behavior better than the traditional description of an "average reasoner".

A comparison of PHM and mReasoner shows that PHM slightly outperforms mReasoner. Only PHM is able to exceed the performance of the most-frequent answer. As such,

[2] github.com/CognitiveComputationLab/ccobra

[3] github.com/nriesterer/cogsci-individualization

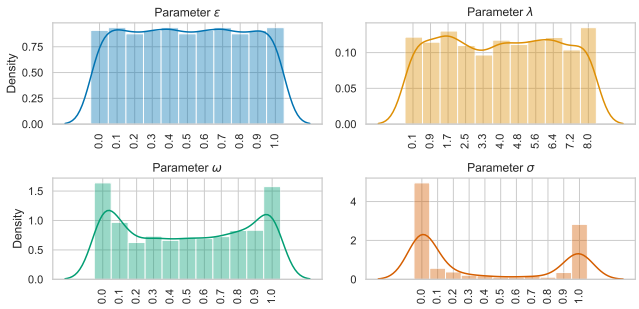

Figure 2: Distribution of the optimal parameter values for mReasoner. The values on the x-axis correspond to the values used for the applied grid search.

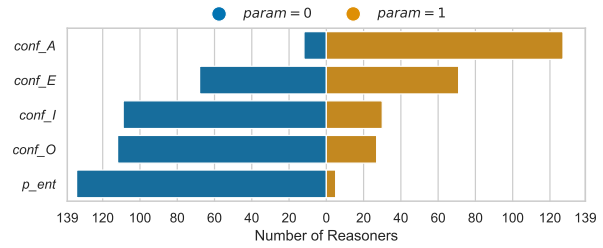

Figure 3: Distribution of PHM's parameters. The barplots depict the number of reasoners for which the optimization results in a value of 0 (blue) or 1 (orange).

it demonstrates the general ability to capture individual reasoning behavior better than purely aggregate models possibly can.

Parameter Usage

Investigating the distribution of optimal parameter values allows us to determine the efficiency of the model parameters. If a parameter is predominantly assigned similar values for a large variety of individuals, it could mean that the underlying process does not fulfill a meaningful function in representing individual properties of human reasoning. Conversely, an evenly occupied parameter space indicates an efficient use of the available parameters.

mReasoner The resulting distributions of optimal parameter values for mReasoner are depicted in Figure 2. Generally, mReasoner distributes its parameters fully across their available parameter ranges. Interestingly, parameters appear to be either uniformly (ε, λ) or bimodally (ω, σ) distributed. Reflecting the propensity to engage in counterexample search and weakening of the candidate conclusion quantifier, the bimodality of σ and ω seems to split individuals into groups of more or less deliberative reasoners. On the other hand, ε and λ appear much more continuous in their values. Theoretically, this could point to their general ability to capture the nuances of human reasonings with greater detail. However, when investigating the fit results, in many cases, the grid search yielded multiple equally performing parameterizations for the same individual. Counting the number of different parameter values, on average, individuals were assigned 4.79 values for ε, 1.74 for λ, 1.47 for ω, and 1.14 for σ. This suggests, at least for ε, that its underlying property, i.e., the completeness of the constructed mental model, is of limited importance for mReasoner's fitting capabilities with many cases in which different values lead to equivalent outcomes.

Probability Heuristics Model PHM's parameter space occupancy is presented in Figure 3. The plot depicts a stacked barplot with blue (left) and orange (right) bars indicating the number of reasoners for which the corresponding parameters were set to 0 and 1, respectively.

With respect to the max heuristic confidence parameters, the figure shows a gradual shift from *conf_A* to *conf_O*. This is to be expected, since they are internally ordered and constrained: $conf\_A \prec conf\_I \prec conf\_E \preceq conf\_O$ (Chater & Oaksford, 1999). Once a parameter is set to 0, all subsequent ones have to be 0, too. Consequently, *conf_A* is expected to show the least amount of 0 values.

p_ent exhibits the highest amount of 0 values. It appears as if only few reasoners consistently behave similar to p-entailment to generate a conclusion candidate. Revealing a weakness of PHM's p-entailment heuristic, it represents a secondary process to generate conclusion candidates serving the purpose to capture responses deviating from the min-heuristic (e.g., Chater & Oaksford, 1999). Since p-entailment use is expected to only occur occasionally, p-entailment will not be considered a consistent property of reasoning and as such is typically set to 0 in our fitting process.

Interestingly, though, despite being severely limited in its parameterization, PHM is able to outperform mReasoner which relies on a far more complex parameter space.

Performance Congruency

The overall model performances (Figure 1) suggest minor differences between mReasoner's and PHM's capabilities to capture individual syllogistic reasoning behavior. This raises the question to which degree the models differ in their predictions, i.e., whether they are capable to successfully capture different subsets of reasoners, or if they are largely congruent in their performances.

The results of this analysis are depicted in Figure 4. For each model, we split the reasoners in our dataset into four groups (quartiles) based on their associated coverage performances. In a next step, we compare the groups to determine their congruency, i.e., the overlap between mReasoner's and PHM's performance-based quartiles. For example, the first quartile indicates that there is 77% overlap between the individuals that were captured worst by the models. Overall, the quartiles of individuals identified for both models aligned nicely with an average accuracy difference of 5%.

Figure 4 shows that mReasoner and PHM agree most in the extremes, i.e., in quartiles 1 and 4 representing worst and best coverage performances, respectively. For quartile 4, this

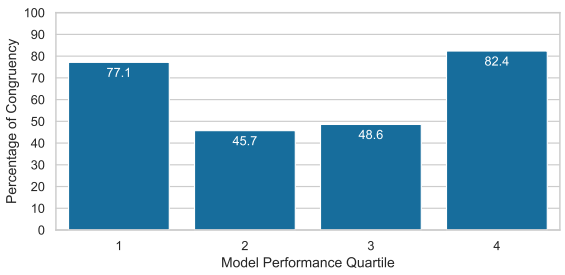

Figure 4: Overlap between mReasoner's and PHM's performance. For each model, reasoners were split into four quartiles (x-axis) in accordance to their respective model prediction accuracies in ascending order. The y-axis depicts the percentage of matching individuals between mReasoner's and PHM's quartile sets.

result is to be expected because of the aggregation-focused design of the models. It most likely contains the individuals that behave similarly to the processes targeted by the models, i.e., average reasoning behavior. For quartile 1 on the other hand, the large overlap must not necessarily correspond to agreement between the models on the level of their assumed inference mechanisms. It more likely reflects that the participants' reasoning behavior contained in this quartile cannot be captured adequately by both models' latent parameter spaces. Reasons for this could be manifold ranging from high levels of noise in the data making the response behavior unsystematic and random to the possibility that reasoners rely on processes that cannot be explained in terms of the models' assumed cognitive processes. The latter would be in line with the aggregated origin of models since it can be assumed that reasoners differing greatly from the average cannot be adequately captured by the traditional models. Finally, in quartiles 2 and 3, both models disagree more which indicates that the differences between the models' inferential mechanisms manifest most distinctly in medium performance ranges.

Discussion

Recent work evaluating the predictive performance of cognitive models for human syllogistic reasoning demonstrated substantial shortcomings when shifting the focus from aggregated representations of data to the behavior of individuals (e.g., Riesterer et al., 2019). Having been deemed a consequence of the prevailing focus on aggregation and the controversially discussed problem of lacking group-to-individual generalizability in empirical sciences (e.g., Fisher et al., 2018), this paper investigated the previously disregarded individualization capabilities of models thereby evaluating them to the fullest extent of their abilities. By fitting the two available parameterized models for syllogistic reasoning, mReasoner (e.g., Khemlani & Johnson-Laird, 2013) and the Probability Heuristics Model (PHM; Chater & Oaksford, 1999), to individual response data and evaluating their ability to successfully recreate the observed reasoning behavior from their latent parameterization alone, we were able to determine the degree to which they qualify as accurate theories of individual syllogistic reasoning. With accuracies of between 40% and 50% and the limited parameter efficiencies, especially for PHM, our results suggest substantial potential for improvement. Still, the individualizations were able to achieve a slight improvement in performance over their aggregated counterparts (6% for mReasoner, 8% for PHM).

Going beyond the general prediction accuracies of the models, our analysis allowed us to investigate the properties and efficiency of their parameterization. We found that although severely more restricted in terms of its parameters, PHM achieves similar results as mReasoner which suggests that its core principles specify a more compact representation of human reasoning behavior. Still, both models leverage the potential for parameterization to a limited degree only. PHM does not include mechanisms to dynamically select the heuristics to apply for a given syllogistic problem. As such, the parameter for selecting the alternative conclusion generation heuristic p-entailment bears almost no significance for individual reasoning. In case of mReasoner, multiple parameterizations resulting in equivalent coverage performances could be determined which indicates limited capabilities to uniquely identify latent representations of reasoners. Especially ε, which, on average, produced equivalent results for five out of the eleven possible parameter values, seems to have negligible influence.

Overall, both models were largely aligned in their ability to represent specific individuals in terms of their latent parameters. They performed best on nearly the same set of participants (82% congruency) suggesting that, on a functional level, their different approaches to formalizing human syllogistic reasoning do not differ greatly with respect to the individuals they can represent. For the case of low accuracies, a similarly high congruency can be observed suggesting that the latent parameterizations of both models are unable to accommodate a fairly distinct set of reasoners. It remains an open question for future research to investigate whether this is due to noise in the data or due to the models failing to capture the fundamentally different reasoning behaviors of some subpopulations of reasoners.

In sum, our work extends previous investigations of the problems surrounding group-to-individual generalizability in syllogistic reasoning research by focusing on the parameterization possibilities of cognitive models. Only if we develop models capable of generating accurate predictions for individual reasoners, we will reach levels of performance that truly justify treating the models' underlying theoretical assumptions as explanations for the cognitive processes of humans. Taking into consideration the inter-individual differences observable in most instances of behavioral data makes adequate parameterization of models not a commodity but a necessity. This calls for a paradigm shift not only for the development of cognitive models, but also with respect to the way models are evaluated.

Acknowledgements

This paper was supported by DFG grants RA 1934/2-1, RA 1934/3-1 and RA 1934/4-1 to MR.

References

Adams, E. W. (1996). *A primer of probability logic*. Stanford, CA: Center for the Study of Language and Information.

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.

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

Copeland, D., & Radvansky, G. (2004). Working memory and syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *57*(8), 1437–1457.

Dietz Saldanha, E.-A., & Kakas, A. (2019). Cognitive argumentation for human syllogistic reasoning. *KI - Künstliche Intelligenz*, *33*(3), 229–242.

Eid, M., Gollwitzer, M., & Schmitt, M. (2015). *Statistik und Forschungsmethoden*. Basel, CH: Beltz Verlag.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*(43), 18243–18250.

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

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

Khemlani, 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). Austin, TX: Cognitive Science Society.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218.

Ragni, M., Dames, H., Brand, D., & Riesterer, N. (2019). When does a reasoner respond: Nothing follows? In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 2640–2646). Montreal, QB: Cognitive Science Society.

Riesterer, N., Brand, D., & Ragni, M. (2018). The predictive power of heuristic portfolios in human syllogistic reasoning. In F. Trollmann & A.-Y. Turhan (Eds.), *Proceedings of the 41st German Conference on AI* (pp. 415–421). Berlin, DE: Springer.

Riesterer, N., Brand, D., & Ragni, M. (2019). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. In T. Stewart (Ed.), *Proceedings of the 17th International Conference on Cognitive Modeling* (pp. 178–183). Waterloo, Canada: University of Waterloo.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, *107*(2), 358–367. doi: 10.1037/0033-295x.107.2.358

Störring, G. (1908). *Experimentelle Untersuchungen über einfache Schlussprozesse*. Leipzig, Germany: W. Engelmann.