Extending TransSet: An Individualized Model for Human Syllogistic Reasoning

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

Abstract

Recently, the TransSet model for human syllogistic reasoning was introduced and shown to outperform the previous state of the art in terms of predictive performance. In this article, we pick up the TransSet model and extend it to allow for capturing individual differences with respect to the conclusion "No Valid Conclusion" indicating that no logically correct conclusion can be derived from a problem's premises. Our evaluation is based on a coverage analysis in which a model's ability to capture individuals in terms of its parameters is assessed. We show that TransSet also outperforms state-of-the-art models on the basis of individuals and provide further evidence for the existence of an NVC aversion bias in human syllogistic reasoning.

Keywords: syllogistic reasoning; transset; modeling; transitivity

Introduction

Syllogistic reasoning is one of the longest-standing domains of reasoning research persisting for over a century now (for an early investigation, see Störring, 1908). Traditionally, a syllogistic problem consists of two quantified premises (*all*, *some*, *no*, *some ... not*) interrelating three terms (e.g., A, B, C):

All A are B
Some B are C

What, if anything, follows?

The goal in syllogistic reasoning is to relate the information conveyed by both premises via the middle term (B) occurring in both of them in order to draw a conclusion about the end terms (A, C). Depending on the arrangement of terms, a syllogistic problem is said to be in one of four figures (notation taken from Khemlani & Johnson-Laird, 2012):

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

By considering all combinations of quantifiers and term orderings, a total of 64 distinct syllogistic problems are obtained all of which can possibly be concluded by eight quantified relations between A and C in either direction, or "No Valid Conclusion" (NVC) indicating that no logically valid conclusion for the pair of premises exists. This results in a total of nine possible conclusions to each syllogistic problem.

Research in the domain of syllogistic reasoning quickly came to understand that human reasoners who are confronted with syllogistic tasks do not reason in accordance to classical first order logic but commit frequent and systematic errors which require psychological explanation (e.g., Woodworth & Sells, 1935; Wetherick & Gilhooly, 1995).

Since its early beginnings, the domain has inspired countless researchers to attempt to postulate and formalize assumptions about the processes underlying human syllogistic inference which has led to a wide variety of theories and models being introduced. In a meta-analysis (Khemlani & Johnson-Laird, 2012), a list consisting of the twelve most prominent theories of syllogistic reasoning was compiled and evaluated. The authors' analysis showed that because individual theories have their distinct strengths and weaknesses it is difficult if not impossible to identify a single best account.

More recently, TransSet (Brand et al., 2019), a model focusing on transitivity-based set interpretation, was introduced and shown to outperform the state-of-the-art models in terms of its predictive power on average human reasoning behavior. Still, in their discussion of TransSet's success, the authors highlighted the fact that a lot of potential for model performance remains untapped because most approaches currently do not account for the inter-individual differentiation underlying the wide variety of inferential strategies syllogisms are known to elicit (e.g., Roberts et al., 2001).

In this article, we attempt to push the TransSet model of syllogistic reasoning one step further by extending it to adapt to the behavior of individuals. By relying on findings from the syllogistic literature, we essentially integrate processing branches into the model which enable it to vary response strategies between individuals. We evaluate the resulting model based on a prediction task and compare its performance to both state-of-the-art models and statistical baselines to measure its success. Finally, we discuss our results as well as the implications of individualization for cognitive modeling research.

Related Work

The domain of syllogistic reasoning has extensively been approached from a multitude of directions including formal logics, probabilities, and various kinds of mental representations (for a review, see Khemlani & Johnson-Laird, 2012). However, in the last decade, the traditional model evaluation paradigm based on comparisons with group data obtained from experiments yielded results suggesting that model performances had reached a plateau making differentiation based on prediction accuracies difficult if not impossible (e.g., Bacon et al., 2003; Khemlani & Johnson-Laird, 2012).

More recently, a paradigm shift concerning the evaluation of models has started to gain traction. Inspired by theoretical and

*Both authors contributed equally to this manuscript.

empirical considerations of inter-individual differences (e.g., Molenaar, 2004) and the corresponding problem of group-to-individual generalizability (Fisher et al., 2018), the focus on model evaluation has shifted from aggregate representations of data to individual response data (e.g., Riesterer et al., 2019). Evaluating the state of the art in modeling human syllogistic reasoning in terms of predictions for individual response data revealed that previous analyses had overestimated model performances considerably. While Khemlani & Johnson-Laird (2012) reported values of up to 95%, 93%, and 84% for hits, correct rejections, and correct predictions on aggregate data, the comparison with individual responses showed that the best model only accounted for 34% of participants' responses (Riesterer et al., 2019). The new analysis produced two crucial results. First, overall low accuracies on participants' responses suggest that current models are far from what can possibly be considered accurate explanation of human behavior in syllogistic reasoning experiments. Second, comparisons with data-driven neural networks illustrated the considerable potential that remains in the domain especially when actively considering inter-individual differences.

A recent analysis (Riesterer et al., in press) put the focus of attention on a different aspect of individualized modeling: model parameterization. A *coverage* task was introduced in which models are fitted to individual response patterns and assessed in terms of their ability to reproduce the observed behavior from their latent parameterization. Computing the accuracy of the fitted models in comparison with the originally observed data allows to derive a score that enables a parameterization-centric assessment of individualized model performance. The analysis included two of the most prominent models for syllogistic reasoning, mReasoner (Johnson-Laird & Khemlani, 2013) and the Probability Heuristics Model (PHM; Chater & Oaksford, 1999). Briefly summarized, mReasoner is an instance of the mental model theory (e.g., Johnson-Laird, 1983) which assumes that individuals reason by constructing mental representations of the premises from which conclusion candidates are generated and potentially revised via a search for counterexamples. PHM, on the other hand, assumes that individuals reason in accordance to probabilistic validity as opposed to logic validity and postulates a set of heuristics to simulate this behavior.

The coverage evaluation (Riesterer et al., in press) revealed that both models are lacking in their ability to account for individual behavior. Only PHM managed to outperform the statistical baseline computed from the most-frequent answer (MFA; the optimal strategy for aggregate models in this task) and thereby demonstrated a basic albeit unimpressive ability to accommodate for individual reasoning behavior in terms of its parameterization. Overall, the coverage analysis highlighted the need for an increased focus on individual differences from a different perspective than the previous prediction-oriented analyses.

The TransSet Model

TransSet (Brand et al., 2019) is a recently introduced model for syllogistic reasoning which was developed with a different goal in mind than previous models. The current state of the art has largely originated from attempts at finding comprehensive explanations of human reasoning behavior which indirectly assumes the existence of general syllogistic inference processes available to all reasoners. However, because of empirical evidence about the variety of strategies employed in the syllogistic domain (e.g., Roberts et al., 2001) this assumption has been met with skepticism in the past (e.g., Bacon et al., 2003). TransSet acknowledges the existence of distinct inferential strategies and focuses on a specific type of naive reasoner who is untrained in the task of solving syllogistic problems and therefore relies on intuitive reasoning based on the prominent surface features of syllogisms, i.e., quantifiers and term order. In particular, it expects reasoners to rely on the general concept of transitivity because of its relevance and importance in everyday reasoning (e.g., for argumentation). In doing so, TransSet reflects a single-strategy model that uses the surface features of syllogisms (e.g., quantifiers or the order of terms) to derive its predictions. Its inferential mechanisms are built on the assumption that reasoning can be defined on the basis of a set-based interpretation of premises and a transitivity-based inference scheme.

TransSet generates predictions for syllogistic problems based on a two-step process consisting of phases for conclusion direction and quantifier selection. The *direction selection phase* depends on the arrangement of terms in the premises. If the premises directly define a transitive path between the end terms (i.e., A-B;B-C or B-A;C-B), TransSet uses the positions of the end terms in the paths as the direction of the conclusion. Otherwise, it is assumed that reasoners attempt to modify the premises in order to create a transitive path. This is done by reversing one of the premises containing a universal quantifier, i.e., "All" or "No", with a preference for "All". If this is not possible, either because there is no universal quantifier or because of ties when both quantifiers are equal, NVC is returned aborting the inferential process.

The *Quantifier selection phase* uses the transitive path to infer the conclusion quantifier. The general assumption behind this phase is that individuals propagate information along the path. If the first quantifier is affirmative, both quantifiers are combined in accordance to the Atmosphere hypothesis (Woodworth & Sells, 1935). If the first quantifier is negative, information propagation is not possible directly. Here, TransSet assumes that if the second quantifier is "All", the disrupted flow of information along the path can be recovered by substituting the middle term with the last term on the path resulting in a "No" conclusion. If this is not possible, TransSet predicts NVC.

Crucially, the inferential mechanism proposed by TransSet does not incorporate traditional processes for deliberative reasoning (e.g., logics or the construction of mental representations) but focuses on a restricted mapping from syllogistic problems to specific conclusion predictions based on surface-features alone. As such, TransSet does not follow the goal to be an adequate explanation of the general behavior of human reasoners but assumes the existence of a subset of reasoners which follow the nonlogical (e.g., Evans, 1972) procedures it assumes. Still, it could be shown that TransSet outperforms the existing state of the art by a substantial margin (Brand et al., 2019). This does not necessarily mean that the processes assumed by TransSet are representative of the cognitive processes driving human inference. However, they currently give the best account of the data. It would be premature to consider TransSet generally superior to other models in its current state. Still,

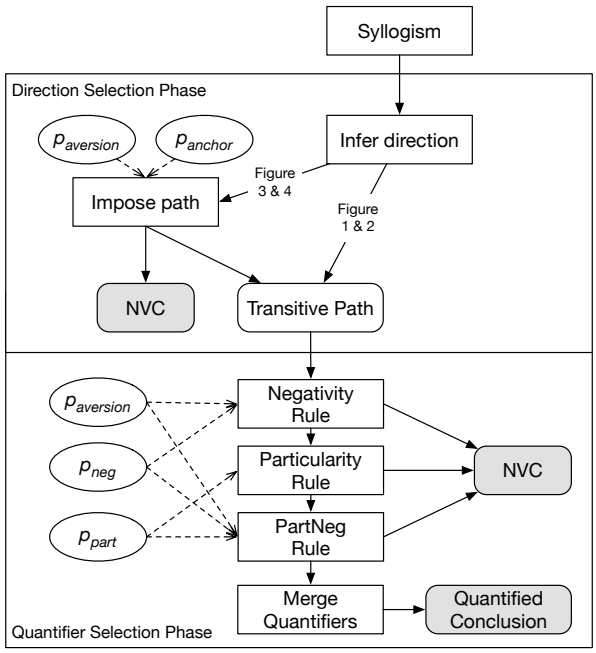

Figure 1: Overview over the individualized TransSet model.

it is a wake-up call for the proponents of the prevailing theories to justify their approaches on the level of response predictions.

Even though TransSet focuses on a single inferential strategy, it offers much potential for alteration with respect to the integration of individual differences. One of them is its handling of the NVC conclusion. Currently, NVC is generated either in cases where transitive paths cannot be formed or if information cannot be propagated along the path. However, research in syllogistic reasoning has produced two important findings with respect to NVC conclusions. First, response data suggests that reasoners are biased against producing NVC conclusions resulting in what could be considered an NVC aversion bias (e.g., Dickstein, 1976; Roberts et al., 2001). Although the reasons for this NVC aversion have not been conclusively determined as of yet, it seems reasonable to assume that individuals differ with respect to the influence it has on their reasoning making NVC aversion a promising component of an individualized model. Second, recent research has shown that certain forms of syllogism might invite NVC responses which suggests that NVC handling of individuals might not just simply be inhibited by aversion biases but also encouraged by certain combinations of premises (e.g., Galotti et al., 1986; Riesterer et al., 2020). As such, in the following sections we introduce and evaluate an extension of TransSet in terms of its NVC handling.

Individualizing TransSet

Our individualization which is summarized in Figure 1 focuses on NVC which is a peculiar conclusion in the syllogistic domain for various reasons. On the one hand, the NVC response itself might be ambiguous. Besides its intended meaning as an indication that no conclusion follows logically from the premises, NVC might also be interpreted as "giving up" signaling that a participant failed to arrive

Table 1: Parameter configurations and preconditions for NVC rules in the quantifier selection phase where Q_1 and Q_2 refer to the first and second quantifier of the transitive path, respectively.

Rule	$p_{aversion}$	p_{neg}	p_{part}	Precondition
Negativity	None	true	-	Q_1 negative
	Low	true	-	Q_1 negative, Q_2 not All
	High	true	-	Q_1,Q_2 negative
Particularity	-	-	true	Q_1,Q_2 particular
PartNeg	None	true	true	Q_1,Q_2 not All

at a quantified conclusion (Ragni et al., 2019). This interpretation can be an incentive for reasoners to "try harder" to avoid NVC responses which effectively invites illogical behavior. On the other hand, it is the logically correct answer for 37 out of the 64 problems (58%), i.e., for the majority of the domain. As there are nine possible conclusions, this imbalance might be unintuitive for some reasoners, especially since it is also unusual for riddles or puzzles, which the experimental setting might seem similar to, to be "unsolvable". This could lead to the NVC aversion phenomenon which has been discussed before (Roberts et al., 2001). However, the over-representation of NVC might also encourage the use of simple rules that can quickly derive an NVC response (e.g., Galotti et al., 1986).

One of the main concepts of TransSet is the separation of the deduction process into the direction selection and quantifier selection phases which each provide rules to check if the respective goals can be achieved. If any phase fails, NVC is concluded. However, the available rules and the likelihood to abort a phase may differ between individual reasoners which is why using them as starting points for the adaption to individual reasoners seems promising.

To allow TransSet to capture the effects of NVC, we introduced four parameters: $p_{aversion}$, p_{anchor}, p_{part}, and p_{neg}. The first parameter, $p_{aversion}$, represents the susceptibility to the NVC aversion bias of a reasoner (e.g., Dickstein, 1976) with possible values in [None, Low, High]. The parameter is used in both phases and determines the likelihood of accepting NVC responses. When NVC aversion is high, the phases of TransSet are less likely to fail since participants try to find a way around responding with NVC. For the direction selection phase, this means that a direction has to be selected, even if it is not clear if the conclusion should relate the end terms from A to C or vice versa (which can only occur for Figure 3 and Figure 4 syllogisms). In these cases, it is assumed that individual preferences decide if the end-term read first (A) is selected as an anchor point (resulting in the direction $A \rightarrow C$) or if the most recent term (C) is chosen (resulting in the direction $C \rightarrow A$). This preference is captured by the parameter p_{anchor} using the values [most-recent, first] which reflect the choice of anchor term. Note, however, that p_{anchor} is a conditional parameter which will only take effect when $p_{aversion}$ is high.

TransSet's second phase, in which the conclusion quantifier is determined, originally only had a single rule to derive NVC: When a transitive path starts with a negative quantifier ("No", "Some ... not"), the propagation of information along the path is prevented, which result in an NVC response in most cases (Brand et al., 2019). In this work, we extend the existing rule, allowing for several nuances depending on the aforementioned $p_{aversion}$ parame-

ter. In doing so, we integrate additional rules to derive NVC which have recently been shown to improve predictive performance when incorporated into various state-of-the-art models for syllogistic reasoning (Riesterer et al., 2020). In particular, we incorporate the rules related to negative quantifiers, i.e., *EmptyStart* and *Negativity*, into TransSet's original process handling negativity in the quantifier determination phase. The $p_{aversion}$ parameter is used to either strengthen or weaken the precondition starting from TransSet's original rule, which corresponds to $p_{aversion} = low$. The remaining rules proposed by Riesterer et al. (2020), i.e., *Particularity* and *PartNeg*, are integrated as additional rules. Individual availability of the above-mentioned rules is controlled via two binary parameters, p_{neg} and p_{part} which can either be set to true or to false. Table 1 summarizes the rules with the respective parameter configurations and the preconditions, that need to be fulfilled to derive NVC responses.

The proposed parameterization of TransSet is a natural extension of its original account. To match the behavior of the original TransSet model, the $p_{aversion}$ needs be set to "low" with p_{neg} being set to "true". Since $p_{aversion}$ is "low", the determination of the direction fails (resulting in a NVC response), which means that p_{anchor} has no effect. As a dedicated rule to derive NVC based on the particularity was not part of the original model, p_{part} needs to be set to "false".

It is important to note that all introduced parameters are categorical, i.e., rely on discrete values with all parameters except for $p_{aversion}$ being binary. While continuous parameter values are generally useful to describe the probabilities or relative importances of effects occurring in populations, a deterministic model with discrete parameters is a preferable description of individuals. Since the data only represents a snapshot of an individual's reasoning behavior, we cannot derive probabilities for their decisions (especially if each syllogistic problem was only solved once per individual). On an individual level, probabilities are only meaningful if each individual repeatedly provided responses to the same tasks. Thus, we have to resort to evaluating the ability of a model to reproduce exact patterns which naturally suggests deterministic model behavior and parameter usage.

Method

The core objective of the following analysis is to evaluate our extension of TransSet in terms of its ability to account for the inferential behavior of individual human reasoners. To this end, we rely on a coverage task (Riesterer et al., in press) in which the goal is to capture the response behavior of individuals in the model's parameters. By assessing the residual error, an estimate of the model's ability to account for inter-individual differences is obtained. Additionally, the parameter configurations resulting from fitting the model to individuals allows for an interpretation of the variation in the observable reasoning behavior in terms of the processes assumed by the model.

Coverage Analysis Setting

Our analysis focuses on evaluation TransSet's ability to recover individual reasoning behavior from its latent parameterization. Put differently, we assess the degree to which TransSet's parameter space *covers* individuals (Riesterer et al., in press).

Note, that the justification of coverage analyses depends on the models being included. In the case of database-like models which fit by storing the observed information, coverage will always be perfect since behavior can simply be recalled from the database. In the case of cognitive models, however, parameters usually have an associated meaning and try to capture essential properties of the assumed mental processes. As such, coverage gives a meaningful estimate of a model's ability to accommodate for individuals.

To increase the expressiveness of our analysis by providing a reference frame for the obtained coverage scores, we include the models from the previous coverage analysis (Riesterer et al., in press): mReasoner (Johnson-Laird & Khemlani, 2013) and the Probability Heuristics Model (Chater & Oaksford, 1999), as well as a random uniform model and the *Most-Frequent Answer* (MFA) model which generates predictions based on the most frequently observed response to a syllogistic problem in a dataset.

Dataset & Implementation

For our analysis, we rely on the *Cognitive Computation for Behavioral Reasoning Analysis* (CCOBRA) framework[1] for model evaluation. The dataset we use is the "Ragni2016" dataset for syllogistic reasoning which is openly available as part of the framework and has been used as benchmark data in many evaluations of syllogistic models including the previously introduced coverage analysis (Riesterer et al., in press). It consists of a total of 139 participants who were presented with the full set of 64 syllogisms and asked to select which of the nine possible responses followed from syllogistic premises. The data, model implementation, and analysis scripts developed for this article are available on GitHub[2]

Analysis & Results

Performance Analysis

Figure 2 depicts the results of the coverage evaluation obtained from CCOBRA. The box plots provide a descriptive view of coverage scores, i.e., the models' abilities to reconstruct reasoning behavior from their parameterizations, while the dots represent the scores for the 139 individuals from the dataset. When fitted to individual reasoners, TransSet significantly improves over the original model (median coverage scores of 0.50 and 0.44, respectively; Mann-Whitney U Test, $U = 7783.5$, $p = .0025$) and substantially outperforms mReasoner and PHM (median coverage scores of 0.38 and 0.45, respectively). TransSet and PHM also surpass the performance of the MFA (median coverage score of 0.45), which is the upper bound for models disregarding individual differences, showing that the concepts underlying their parameters are suited to capture the behavior of individuals. However, it is important to note that TransSet only describes a specific strategy that some individuals might use. When considering the results for specific individuals, it becomes apparent that a substantial amount is still not sufficiently covered by the model. While this might partially be due to guessing-like behavior or non-systematic mistakes, it also possible that some of these individuals are using different strategies.

The general improvement of TransSet achieved by our individualization indicates that the incorporation of NVC biases is a promising way for models to account for different individuals. This is not

[1] https://github.com/CognitiveComputationLab/ccobra

[2] https://github.com/Shadownox/iccm-transset-indiv

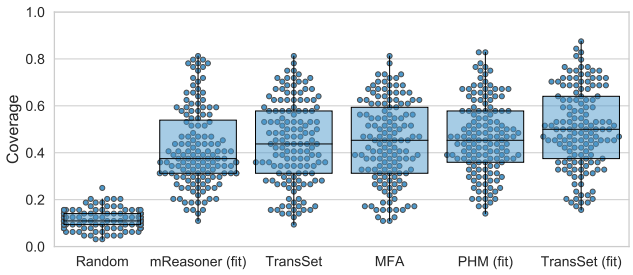

Figure 2: Accuracies of models for the coverage task. The suffix "(fit)" indicates that the model was fitted to an individual's responses.

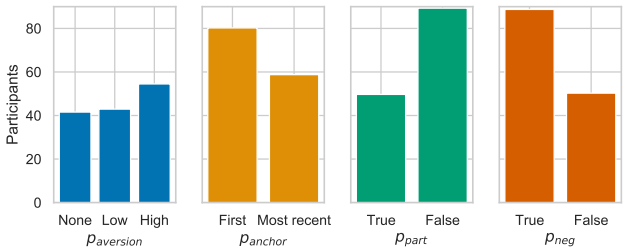

Figure 3: Parameter distributions for the individualized TransSet's parameters resulting from fitting the model to individuals from the "Ragni2016" dataset.

surprising, given the special status of the NVC due to the imbalance of the syllogistic task and ambiguity of the response. On a more general level, the improvement also shows that it is possible to significantly boost the performance of a model by focusing on rules and mechanisms that are able to differentiate between individuals. This highlights the importance of an evaluation based on individual data, as improvements beyond the performance of the MFA cannot be assessed on the basis of aggregated analyses. Additionally, these analyses provide a starting point for further investigations of individuals that are not covered sufficiently by most state-of-the-art models. For now, it is unknown to which extent this can be attributed to noise (e.g., due to guessing-like behavior). An in-depth analysis of these individuals might help to better estimate the proportion of noise and uncover additional strategies and biases in the data.

Parameter Distribution Analysis

Apart from allowing models to be fitted to individual reasoners, the utilization of parameters is also an important property of models based on which their internal integrity can be assessed. Ideally, parameters in cognitive models have specific meaning in relation to the assumed inference process. For instance, in the case of TransSet, the NVC aversion parameter $p_{aversion}$ is indicative for the model's behavior to pursue alternative conclusions to avoid NVC. Optimally, the use of parameters in cognitive models should be limited to the minimum that still enables the capturing of distinct and important differences between individuals (principle of parsimony). This, in turn, means that all parameter assignments should be relied on by the model to account for a population of reasoners. If certain parameter configurations are only used for negligible amounts of individuals, either the corresponding group of individuals was not part of the data or, more likely, the model has an inefficient use of parameters and should be revised in order to reduce its parameter complexity and increase its explainability.

Figure 3 shows the distributions for TransSets's parameters. For each possible value of a parameter, the number of participants that are described best by using the respective value is shown. When analyzing the distribution for $p_{aversion}$, we see that $p_{aversion} = high$ yields the best results for the majority of individuals, indicating that incorporating NVC aversion is indeed beneficial for individualized models of syllogistic reasoning. The particularity rule, despite being inactive for the majority of participants, still seems to be a valuable addition, as it still improved the fit for a third of the individuals. While the optimal parameterization for the majority of the data has the biggest importance for the fit, all parameter configurations still represent a substantial number of individuals. In the case of $p_{aversion}$, the majority of participants is not even represented by the most prominent value ("high"). The original TransSet model corresponds to $p_{aversion} = low$, which does not reflect the NVC aversion of most individuals, but instead describes the data better on an aggregate level. This highlights the importance of individual modeling in general: A model describing the average reasoner might not be able to reflect the most prevalent traits of reasoning.

With respect to the hypothesis of an aversion against NVC, the distribution of $p_{aversion}$ is intriguing. A substantial number of participants are described by high NVC aversion, leading to response patterns with little to no NVC responses. However, while this group does in fact seem to avoid NVC wherever possible, the majority of participants have a low or no NVC aversion at all. Therefore, the aversion against NVC conclusions seems to be a highly individual behavior that affects a substantial proportion of the participants but might not necessarily be a universal factor in human syllogistic reasoning behavior. Since only a group of individuals seems to avoid the NVC response consistently, this hints at general misunderstandings of the NVC response itself for this group.

General Discussion

In this article, we presented and evaluated an individualization of the recently introduced TransSet model for syllogistic reasoning (Brand et al., 2019). To integrate the capability to differentiate between individuals we focused our attention on the conclusion "No Valid Conclusion" (NVC) which has been in the focus of attention before for its ability to evoke aversion biases (e.g., Dickstein, 1976; Roberts et al., 2001) and for being a conclusion which was neglected by a number of models in the past (Riesterer et al., 2020).

TransSet's original specification already contained rules to derive NVC conclusions directly from surface features of syllogistic premises which were invoked when the construction of transitive paths or the propagation of information along them failed (Brand et al., 2019). Our individualization of the model extends on these rules by introducing parameters with the goal to capture individual differences in NVC behavior. We assume a total of four parameters reflecting (1) the magnitude of the aversion against NVC responses, (2) a figure anchor providing the direction of the conclusion

generated in alternative to NVC following the NVC aversion, and the susceptibility to directly conclude NVC based on (3) negativity or (4) particularity of the premises (Riesterer et al., 2020).

Our results illustrate the success of TransSet's individualization. In a coverage analysis (for an introduction to the paradigm, see Riesterer et al., in press), the model is shown to outperform not only the statistical model following a response strategy focusing on the conclusions most frequently selected by participants (most-frequent answer, MFA), but also the two state-of-the-art models mReasoner (Johnson-Laird & Khemlani, 2013) and the Probability Heuristics Model (PHM; Chater & Oaksford, 1999) which have been separately analyzed in a coverage analysis by Riesterer et al. (in press). Investigating the parameter distribution that follows from fitting TransSet to individuals illustrates the quality of the assumed factors for individualization. The parameter space is evenly distributed with no value being only assigned to a negligible number of participants. Further, the distribution of the aversion parameter adds to the evidence for such a bias in syllogistic reasoning (e.g., Dickstein, 1976; Ragni et al., 2019).

Overall, our results add to the growing corpus of modeling research on the level of individual responses. Despite the fact that TransSet is intended to only capture a distinct subset of reasoners, namely those who rely on surface-level features of the problem domain (e.g., quantifiers and term order), it currently outperforms even the most comprehensive and general models of the state of the art both on the aggregate and individual level. While we should refrain from considering it an overall superior explanation of human cognition in this task, especially given its current lack of grounding in terms of psychological/neuroscientific concepts, it should serve as a wake-up call to theorists and modelers alike. Our results demonstrate that the previous signs of a performance-based plateau were merely due to the choice of a severely restricted evaluation paradigm which can be overcome by adopting the perspective of individual responses.

Acknowledgements

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

References

Bacon, A., Handley, S., & Newstead, S. (2003). Individual differences in strategies for syllogistic reasoning. *Thinking & Reasoning*, *9*(2), 133–168.

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.

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

Evans, J. S. B. T. (1972). On the problems of interpreting reasoning data: Logical and psychological approaches. *Cognition*, *1*(4), 373–384.

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.

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General*, *115*(1), 16–25.

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

Johnson-Laird, P. N., & Khemlani, S. S. (2013). Toward a unified theory of reasoning. In *Psychology of learning and motivation* (pp. 1–42). Elsevier.

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

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., 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. (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.

Riesterer, N., Brand, D., & Ragni, M. (in press). Do models capture individuals? Evaluating parameterized models for syllogistic reasoning. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Roberts, M. J., Newstead, S. E., & Griggs, R. A. (2001). Quantifier interpretation and syllogistic reasoning. *Thinking & Reasoning*, *7*(2), 173–204.

Störring, G. W. (1908). Experimentelle Untersuchungen über einfache Schlussfolgerungsprozesse. *Archiv für die gesamte Psychologie*, *11*, 1–127.

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

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