Effect of Response Format on Syllogistic Reasoning

Daniel Brand (daniel.brand@psychologie.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Abstract

Comprehensive datasets used for modeling endeavours in syllogistic reasoning research usually contain only a single conclusion per task for each subject. However, this means that no information about the other conclusions is provided, preventing the distinction between conclusions that were rejected, conclusions deemed to be valid but not the preferred conclusion and conclusions that were not considered at all. In this work, we present a multiple choice dataset containing all conclusions that participants considers valid. The data is compared to datasets with other response designs, and an extensive evaluation is performed to assess the impact of the response design on the predictive performance of cognitive models. Finally, our results are discussed and put into perspective.

Keywords: Syllogistic reasoning; Response design; Cognitive modeling; Jaccard coefficient

Introduction

Syllogistic reasoning, as one of the oldest domains of human reasoning research, is actively researched for over a century (e.g., Störring, 1908). Traditionally, syllogistic problems consist of two quantified statements (premises) with the first-order logic quantifiers *All*, *Some*, *No*, and *Some not*, which interrelate three terms as shown in the following example:

All A are B.
Some B are C.

What, if anything, follows?

The arrangement of the terms defines the so-called figure of a syllogism, where each syllogism is one of four figures. Throughout this paper, we will use the notation from Khemlani and Johnson-Laird (2012), which is shown in the table below:

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

For the sake of space, quantifiers are often abbreviated by letters (*All*: A, *No*: E, *Some*: I and *Some not*: O). Combined with the figure, the syllogism in the example can be abbreviated with AI1. The conclusions can be abbreviated in a similar fashion, using the quantifier and the direction (*ac* or *ca*). For example, *All A are C* would be abbreviated by *Aac*. In case that no valid conclusion exists, it is often abbreviated by *NVC*.

A comprehensive meta-analysis by Khemlani and Johnson-Laird (2012) describes twelve theories and models for the domain. In an extensive analysis, seven of the theories are tested with respect to their ability to account for the aggregated human response patterns, concluding that the models offer different strengths and drawbacks. Later, Riesterer, Brand, and Ragni (2020) proposed an analysis procedure on the individual level, showing that the performance differs substantially from aggregated analyses. However, despite large amount of effort that had gone into models and their evaluation, most of the work is based on datasets with a common trait: Participants are asked for a single response, either by allowing them to freely formulate their conclusion or by selecting it from a set of possible candidates. The former approach forces subjects to actively construct conclusions on their own, whereas the latter poses the problem that subject might just evaluate the given responses instead (e.g., Dickstein, 1978). However, freely formulated responses, if not guided further (e.g., by using a restricted entry box only allowing to enter valid inputs; see Dames, Klauer, & Ragni, 2022), may require additional interpretation, potentially leading to a loss of responses (e.g., as miscellaneous errors; Khemlani & Johnson-Laird, 2012). Finally, no format sheds light on the reasons why a conclusion was not selected: It remains unclear if a conclusion was either considered to be valid, but not selected due to a bias, or rejected as invalid or even not considered at all. To our knowledge, no comprehensive dataset exists that contains *all* conclusions that a participant considers to be valid.

This paper aims to make a first step towards this by presenting such data for all 64 syllogisms and all 9 possible response options. The data is then compared to datasets with both single response formats. Furthermore, the impact that different response formats have on the predictive performance of cognitive models is investigated: An aggregate analysis is performed based on the procedure used in the meta-analysis by Khemlani and Johnson-Laird (2012), as well as an evaluation on the individual level adopting the method used by Riesterer, Brand, and Ragni (2020). To avoid misunderstanding, we will briefly clarify our terminology: While sometimes referred to as multiple choice, we use *single choice* for a response format, where a single response has to be selected from a set of candidates, whereas *multiple choice* will be used when one or more candidates could be selected. A format with freely formulated responses will be referred to as *free response*.

Background

In their meta-analysis, Khemlani and Johnson-Laird (2012) compiled prediction tables for the models for all 64 syllogisms, which they compared against the relevant responses in the aggregated dataset with respect to accuracy, hits and rejections. The evaluated models comprised the *Atmosphere Hypothesis* (Woodworth & Sells, 1935), the *Conversion Hypothesis* (Revlis, 1975), the *Matching Hypothesis* (Wetherick & Gilhooly, 1995), *Psychology of Proof model* (PSYCOP; Rips, 1994), *Verbal Models* (Polk & Newell, 1995) as well as the *Mental Model Theory* (MMT; Johnson-Laird, 1975) and the *Probability Heuristics Model* (PHM; Chater & Oaksford, 1999; Oaksford & Chater, 2001). For our evaluations, we relied on the tables compiled by Khemlani and Johnson-Laird (2012) for most models. For the sake of space, we will only introduce PHM and MMT, as they were adapted to the multiple choice task for the individual analysis (see Riesterer, Brand, & Ragni, 2020) while the other models remained unchanged (for a detailed description, see Khemlani & Johnson-Laird, 2012).

The *Mental Model Theory* (Johnson-Laird, 1975) is a cognitive theory based on the assumption that inferential mechanisms operate on a mental model constructed from the premises. It consists of several phases: Model construction, conclusion generation and a search for counterexamples. If a counterexample is found, the model is updated and a new conclusion is created or, finally, NVC is concluded. For MMT, a computational implementation exists with mReasoner (Khemlani & Johnson-Laird, 2013). mReasoner allows to control the phases via parameters and was successfully used to predict individual reasoners (Riesterer, Brand, & Ragni, 2020).

The *Probability Heuristics Model* (Chater & Oaksford, 1999; Oaksford & Chater, 2001) is based on a set of simple heuristics approximating p-validity with two phases: First, generative heuristics create a conclusion candidate. Thereby, the min-heuristic selects the least informative quantifier (according to the order $A > I > E \gg O$). Then, probabilistic entailment (*p-entailment*) might generate alternative conclusions that are entailed (e.g., *All* entails *Some*, and *Some* and *Some not* entail each other). The attachment-heuristic then determines the direction of the conclusion, if possible. The second phase consists of the test-heuristics: The max-heuristic checks the confidence in the quantifier of the conclusion candidate. The lower the confidence, the higher the probability of returning NVC instead of the conclusion candidate. The O-heuristic states that O-responses should be avoided, possibly by returning NVC (Copeland, 2006). Note that the prediction table by Khemlani and Johnson-Laird (2012) does not include the test-heuristics and is therefore missing NVC responses (Baratgin et al., 2015). PHM was also successfully used to predict individual reasoners (Riesterer, Brand, & Ragni, 2020), mainly relying on the max-heuristic to represent individual reasoners.

Datasets

We acquired the data of 100 participants a web experiment on the platform Prolific[1]. Each participant was presented with one syllogism at a time and was asked to select all conclusions that follow from the respective premises. The conclusions were thereby presented below the premises and could be selected/deselected by clicking on them. Participants had to explicitly select *No valid conclusion*, if they thought that nothing follows from the premises. After selecting the responses, they had to click on a *Continue* button in order to get to the next task. The syllogisms were presented in a random order and contained professions and hobbies as contents to avoid biases. In the end, participants were asked about their interpretation of the quantifier *some*, i.e., if it also includes *all*.

For comparison, two openly available datasets that contain all 64 syllogistic tasks were included: As a dataset with free responses, the aggregated dataset compiled for the meta-analysis by Khemlani and Johnson-Laird (2012) was used. The dataset consists of the data from six experiments with a combined number of 156 participants which were asked to generate a single response freely for each syllogism. As some responses could not be interpreted, the percentages for the conclusions don't add up for all tasks (i.e., some proportion is lost to *miscellaneous errors*). For a better comparison with the other datasets, we normalized the percentages for each task.

For single choice, the Ragni-2016 dataset was used, which is part of the Cognitive Computation for Behavioral Reasoning Analysis (CCOBRA) Framework[2]. The dataset was obtained from a web experiment on Amazon Mechanical Turk and contains the responses of 139 participants to all 64 syllogisms. The participants were asked to select one of the nine possible response options.

Data Analysis

We will first assess some distinctive features of the multiple choice dataset. Afterwards, a detailed comparison between the three datasets is presented. All materials and scripts are publicly available on GitHub[3]. The response distributions of the multiple choice dataset and the most frequently selected patterns are shown in Figure 1. It is important to note that the presented pattern is unweighted and shows the total number of selections of each response, thereby weakening the relative strength of NVC which is mutually exclusive with other conclusions.

Based on first-order logic it would be expected that if an universal quantifier was selected, responses with the respective particular quantifier would also be chosen as it is implied. Based on first-order logic, it would be expected that universal quantifiers would imply the respective particular quantifiers.

[1] https://www.prolific.co/

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

[3] https://github.com/Shadownox/cogsci-2023-multiplechoice

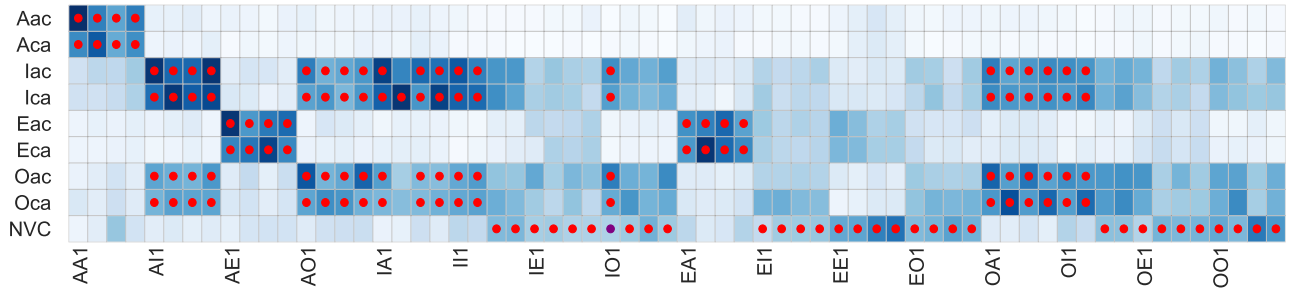

Figure 1: Unweighted response patterns to all 64 syllogisms with a multiple-choice task. Darker shades of blue denote a higher number of selections for the respective response option. Red dots denote the most frequently selected response combinations for each syllogism (purple is used in case of a tie).

However, the pattern shows a distinct separation between universal (*A* and *E*) and particular quantifiers (*I*, *O*). Only in 23.2% of the cases where *A* was selected, *I* was also included in the responses (and only 8.8% for $E \rightarrow O$). While this can be explained by a preference of universal over particular quantifiers (e.g., the gricean maxim of quantity; Grice, 1975) for single response formats, this explanation does not apply for the multiple choice scenario. Instead, the general understanding of the quantifiers appears to be the reason: 88% of the participants stated that *Some A are B* does not include the possibility that *All A are B*. In line with this, *I* and *O* in fact are chosen together (58.3% and 61.7% for $I \rightarrow O$ and $O \rightarrow I$, respectively), with all four particular conclusions being the most frequent pattern for all tasks except of IA2.

Another peculiarity of the most frequent pattern is its invariance to the direction of the response: in most cases, both directions (*ac* and *ca*) are selected together, even if it is logically not warranted (e.g., AA1). At first glance, this seems to contradict the figural effect (e.g., Dickstein, 1978; Johnson-Laird & Bara, 1984), which is well established and commonly replicated in syllogistic reasoning and predicts a bias towards *ac* for figure 1 syllogisms and towards *ca* for figure 2, respectively. When quantifying the effect by using the difference between the percentages of responses given in line with the *effect* (*ac* for figure 1 and *ca* for figure 2) and the percentage of responses in the other direction (*no-effect*), it showed that the effect is significantly weaker than for the other datasets (Differences: free response: $mean = .42$; single choice: $mean = .32$; multiple choice: $mean = .1$, Mann-Whitney-U test between single choice and multiple choice: $U = 905.0$, $p < 0.001$). However, although the effect was weaker than in the other datasets and not represented in the most frequently selected response combinations, it was still significant (Mann-Whitney-U test between *effect* and *no-effect*: $U = 835.0$, $p < 0.001$).

In order to compare the patterns of the multiple choice dataset to the other datasets, we discounted each selected response given by the participants by the total number of selected responses for the task. This step is important to make the NVC response comparable: While a non-NVC response for single choice and free responses gets potentially distributed across all other options (thereby relatively weakening specific non-NVC responses), this would not occur for multiple choice. Additionally, it allows to directly interpret percentages and compare them between the datasets. To determine responses that can be considered *reliable*, we apply the same criterion as used by Khemlani and Johnson-Laird (2012): By using a binomial test against the guessing probability of $1/9$, a threshold of 16% can be used. Due to the discounting, it also can be applied to the multiple choice data. Figure 2 shows the patterns for all three datasets with reliable responses and the most frequent answer (MFA) highlighted. The mean number of reliable responses is thereby similar between all datasets with 2.05 for free responses, 1.95 for single choice and 2.27 for multiple choice.

On the first glance, the patterns of the datasets seem to be rather similar. To gain a deeper insight into the similarities, we compare all datasets pairwise relying on several metrics: First, we calculate the root mean squared error (RMSE) between the normalized patterns of the respective datasets. Second, we compare the congruency of the most-frequent answer patterns (i.e., the percentage of matches between both MFA-patterns). Finally, we calculate the Jaccard coefficient, which is defined as follows (Aggarwal, 2016):

$$jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

As a set-based metric, we argue that the Jaccard coefficient is well-suited for comparing selected responses. As participants only selected responses, but did not actively reject each other option (especially in the single choice or free response datasets), the interpretation of the Jaccard coefficient is intuitive: A value of 0.75 means that out of the reliable selected responses in both datasets, 75% are reliable in both datasets, while the other 25% are only present in one of the datasets. The number of reliable responses is sparse, making it beneficial to exclude other responses from the equation, as the score would be skewed by a high (but rather meaningless) congruency based on the large set of not chosen responses.

Table 1 shows the results for the four metrics for all pairs out of the three datasets. Based on the RMSE, it can be seen that the datasets are overall comparable, which also transfers to the most frequent answer patterns. With the Jaccards

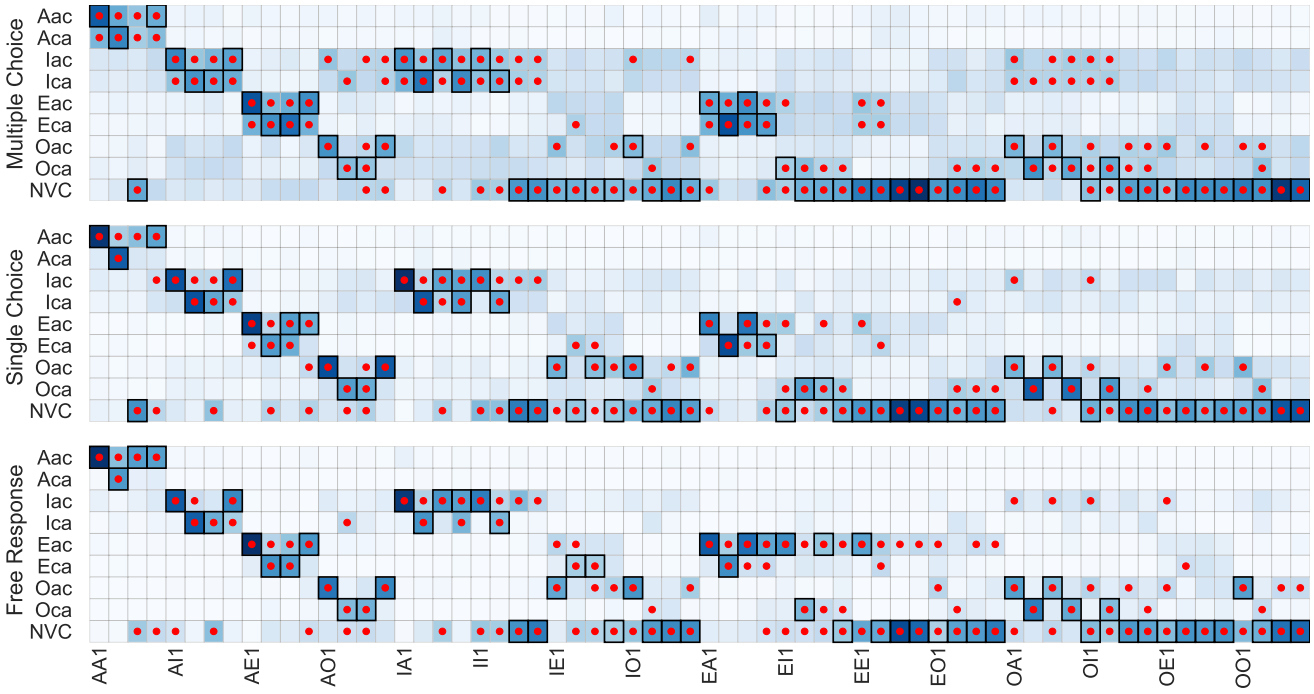

Figure 2: Comparison of the response patterns between multiple choice (normalized), single choice and free response data. Red dots denote the responses that are considered *reliable* (account for $>= 16\%$ of responses), the most frequent response is highlighted with a black border.

coefficient, the differences become more visible, especially between the free response and the multiple choice dataset. Single choice thereby seems to be a *middle ground* between the other response types, which is not surprising as it shares traits with both types. However, the values of the Jaccard coefficient still show that the majority of reliable responses are transferable across datasets. The results illustrate the advantage of the Jaccard coefficient over the RMSE (or similar metrics that compare datasets directly) for this analysis: While the RMSE indicates that no substantial difference exists across the whole distribution, the set-based Jaccard coefficient provides a better resolution for the points of interest.

The overall high similarity between the datasets allows for the assumption that most findings and effects found in experiments asking for single responses can still be expected to be transferable to multiple choice. In the following section, this will be investigated by an evaluation of the predictive performance of cognitive models.

Model Evaluation

Aggregate Level

For the aggregate analyis, the general evaluation approach by Khemlani and Johnson-Laird (2012) was adopted. The authors compiled prediction tables for the seven cognitive models introduced before and evaluated the predictions in terms of accuracy, the percentage of hits and the percentage of correct rejections based on the set of reliable responses on each dataset of the meta-analysis. For our analysis, we also rely on

Table 1: Jaccard coefficient, root mean squared error (RMSE) and congruency between the most-frequent-answer patterns from a pairwise comparison between response formats.

Response Type	Jaccard	RMSE	MFA-Congruency
Free vs Single	.78	.06	.97
Free vs Multi	.66	.10	.96
Single vs Multi	.76	.06	.98

the compiled prediction tables. However, unlike Khemlani and Johnson-Laird (2012), we use the resulting meta-dataset (with the normalization described in the *Data* section) instead of the separate datasets that it was compiled from. Therefore, the exact values differ from the results reported by Khemlani and Johnson-Laird. Additionally, we included the Jaccard coefficient again as a metric as an alternative for accuracy, which comes with a substantial downside that can be seen from the following thought experiment:

Given that there are approximately two reliable responses per task and assuming a model that selects its responses randomly, the accuracy depends highly on the number of predictions. For a single randomly selected response (i.e., the minimum number of predictions), the expected accuracy is 71.6%, while the accuracy when predicting all nine responses would only achieve an accuracy of 22.2%. A model that has the correct number of predictions, but still selects them randomly

Table 2: Evaluation of the models ability to account for reliable responses with respect to accuracy and jaccard coefficient for free responses, single choice and multiple choice. The best values for each metric and dataset are marked in bold.

Model	Multiple Choice				Single Choice				Free Response			
	Jaccard	Acc.	Hits	Rej.	Jaccard	Acc.	Hits	Rej.	Jaccard	Acc.	Hits	Rej.
Atmosphere	.51	.82	.56	.91	.41	.8	.55	.87	.41	.79	.55	.87
Conversion	.5	.83	.51	**.96**	.5	.86	.54	**.95**	.43	.84	.49	**.95**
Matching	.43	.74	.58	.78	.28	.67	.51	.72	.34	.7	.59	.74
MMT	**.59**	.8	**.91**	.78	.56	.81	**.98**	.77	**.57**	.81	**.96**	.77
PHM	.34	.73	.47	.81	.3	.72	.49	.78	.33	.73	.53	.79
PSYCOP	.39	.75	.48	.86	.43	.78	.55	.86	.39	.77	.51	.86
Verbal Models	.53	**.84**	.64	.93	**.61**	**.87**	.73	.93	.55	**.85**	.67	.92

would achieve 65.4%, which is over 6% worse than predicting just a single answer, which means that smaller differences in the actual predictive capabilities could be outweighed by a slightly lower number of predictions. The Jaccard coefficient is not invariant of the number of responses ($1/9$ for a single prediction, $2/9$ if all responses are predicted), but is not influenced as substantially.

The results of the evaluation are shown in Table 2. With respect to accuracy, hits and rejections, the results support the findings of the analysis by Khemlani and Johnson-Laird (2012), with Verbal Models having the best overall accuracy with MMT being close behind. Based on the Jaccard coefficient, MMT performs better on the multiple choice dataset and the free responses, however, the differences are also slim. PHM is one of the worst performance models, potentially suffering from the missing ability to conclude NVC, which is the most frequently chosen response overall (Riesterer, Brand, Dames, & Ragni, 2020). The two remaining metrics, hits and rejections, appear to be less suited for gaining further insight. Mainly, there are two problems: First, they are tightly coupled as a model predicting too many responses will likely score higher on hits and lower on rejections. Second, overall, the metrics are overly sensitive to the number of predicted responses. Especially the rejections highly correlate with the number of predicted responses (Spearman's rank correlation: $r = -.95$, $p < .001$), therefore providing little information beyond the number of predicted responses itself.

Individual Level

It is important to note that a mismatch between the task that the models were designed for and the task that the participants solved exists in the analysis above. Most models were designed to predict the outcome of a single response task, which means that the sets of responses that the models predicted refer to a population and not a single participant (which is especially important for NVC, which is mutually exclusive to any other response for participants). In a similar fashion, the interpretation for the two single response datasets is prone to mixing up preferences for a certain response with rejections of another. To overcome this problem, we evaluate the predictive performance of the models on the individual level.

For the evaluation, the Cognitive Computation for Behavioral Reasoning Analysis (CCOBRA) framework was used, which was built to facilitate such analyses. We performed a *coverage* evaluation (see Riesterer, Brand, & Ragni, 2020), where each model is presented with the full set of responses for a specific participant and allowed to fit it's parameters accordingly. As the models used in the previous analysis are represented by tables, they lack the ability to fit their responses to a specific participant. Riesterer, Brand, and Ragni (2020) provided a CCOBRA-implementation of mReasoner as well as an implementation of PHM, which we then adapted for the multiple choice task. As mReasoner natively supports queries for a given conclusion (*Is it necessary that...?*), the predictions for multiple choice were realized by querying for every possible conclusion (with NVC being the result in case of rejection). For PHM, we utilized the p-entailment to obtain multiple predictions by adding a three-valued parameter controlling the selection of the prediction candidates: (1) only use conclusion candidates from min-heuristic, (2) only use the candidates from p-entailment and (3) use both conclusion candidates. For comparison, the most-frequent answer (MFA) is included as a baseline model, which always predicts the most common combination and thereby serves as the upper bound for all models that do not adapt to individual participants. Again, the Jaccard coefficient was used to compare the predictions to the responses participants selected and both models were fitted to optimize the Jaccard coefficient. The table-based versions of PHM and MMT are included for comparison, whereas the other models are not included as their implementation does support neither multiple choice nor individual predictions. Figure 3 shows the results with respect to the Jaccard coefficient. For PHM and MMT, the results differ substantially from the aggregate results. This illustrates the effect of the information loss happening in the aggregation process by the determination of reliable responses, which leads to the same treatment for all responses above 16% (e.g., predicting a response with just above 16% has the same worth as predicting a response that over 50% of the participants selected). Furthermore, the results indicate that the fitting of the models to individual participants worked, as both, mReasoner and PHM (individualized), outperform the table-based

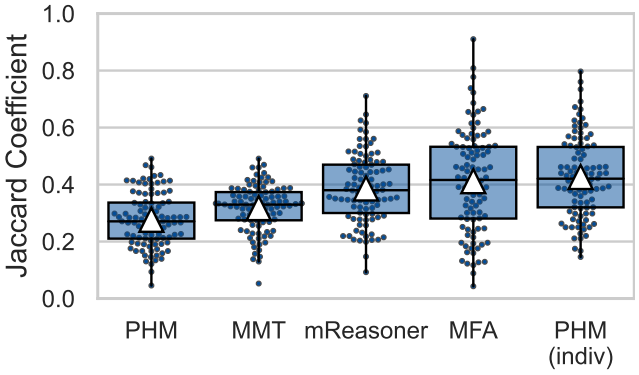

Figure 3: Jaccard coefficient for all models when predicting individual multiple-choice patterns. Boxplots denote medians and inter-quartile ranges. Mean values are denoted by the white triangle and points show results for specific individuals.

models. However, only PHM managed to surpass the most-frequent answer, while mReasoner remained at a level that an (hypothetical) ideal model also could reach without considering the individual reasoners.

Discussion

Most datasets used for modeling in the domain of syllogistic reasoning were conducted using a response format asking for single conclusions. Thereby, it is argued that free responses have an advantage over single choice, because they ensure that participants create the conclusions on their own (Dickstein, 1978). With this paper, we looked at the impact that the response format has on the results of cognitive model, including a neglected perspective: When asking for single conclusions, only an incomplete picture of the participants' understanding of the valid inferences is provided. The reasons why a specific conclusion is not selected remains unclear. To shed light on this, we conducted a study to obtain a dataset containing the full set of conclusions that a participant considers to be valid and used it to re-assess state of the art models of syllogistic reasoning. A finding indicating that multiple choice data was indeed helpful to disentangle certain effects was that the figural effect, while still significant, was significantly weaker for multiple choice tasks. This could indicate that certain effects are in fact combinations of reasoning effects finding possible conclusions with preferences that choose certain conclusions over others (despite considering both to be valid). Overall, however, the comparison between the datasets showed that the main patterns (e.g., the most frequently selected conclusions) were robust with respect to the response format. This also carried over to the aggregate analysis of the cognitive models, which supported the general results by Khemlani and Johnson-Laird (2012) across all response formats, with MMT and Verbal Models achieving the highest predictive performance. We further introduced the Jaccard coefficient as a means to compare sets of selected responses, which is useful for comparing the patterns in different datasets or evaluating the predictive performance of cognitive models. Thereby, we argued that the metric is well-suited for the task due to its close relation to the task formulation itself (i.e., participants do not actively reject conclusions, which makes it debatable if conclusions that were not selected should be treated in the same way as the selected conclusions). Furthermore, its values can be interpreted intuitively and it is less affected by the total number of predictions in the case of model evaluation.

Finally, we performed an evaluation on the individual level following the proposed approach by Riesterer, Brand, and Ragni (2020) on our multiple choice data. The implementations of PHM and mReasoner (Riesterer, Brand, & Ragni, 2020) were adapted to select multiple conclusions instead of generating a single response and then fitted to each individual participant at a time. Thereby, mReasoner natively supported queries for the acceptance of a given conclusion, which made the adaption straightforward. PHM, via the p-entailment, already offered a possibility to generate an alternative set of conclusion candidates, which we could use to predict multiple responses. The individual analysis showed a substantial drop in performance compared to the aggregate results (especially for MMT), which is in line with other results of individual analyses (e.g., Riesterer, Brand, & Ragni, 2018) and highlights the loss of detail that can come with aggregating. As expected, the fitting allowed PHM and mReasoner to outperform their table-based versions. However, only PHM surpassed the level of the most frequent answer, which represents the upper bound achievable with a model not adapting to individual reasoners.

Conclusion

The comparison between the different response types showed that the effects and patterns found in syllogistic reasoning research are robust and not easily susceptible by different response formats. This can allow the combination of data from different response formats for general modeling endeavors, especially for approaches from machine learning which usually require larger datasets. For the multiple choice dataset, the results indicated that some frequently given responses found in the typically used single response datasets might actually be boosted by preference effects (e.g., the figural effect). Additionally, most cognitive models were designed to only generate single responses, thereby not modeling syllogistic reasoning but rather a specific task design of syllogistic reasoning. It is important for future research in the field of syllogistic research to include multiple choice to the standard repertoire for future modeling endeavours, in order to obtain a more complete understanding of the processes behind the human ability to deduce conclusions in syllogistic tasks. Finally, our work illustrated the impact that the model evaluation procedure itself has on the results. For future model evaluation, focusing on the individual level is key, not only to account for the inter-individual differences, but also to stick as close as possible to the collected data.

Acknowledgements

This research was supported by the German Research Foundation, DFG (Grant RA1934/5-1 and RA1934/8-1) within the SPP 1921 "Intentional Forgetting" and by the Saxony State Ministry of Science and Art (SMWK3-7304/35/3-2021/4819) excellence initiative "Productive Teaming" on the basis of the budget passed by the deputies of the Saxony state parliament.

References

Aggarwal, C. C. (2016). *Recommender systems: The textbook* (1st ed.). Springer Publishing Company, Incorporated.

Baratgin, J., Douven, I., Evans, J. S. T., Oaksford, M., Over, D., & Politzer, G. (2015). The new paradigm and mental models. *Trends in Cognitive Sciences*, *19*(10), 547–548.

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., Klauer, K. C., & Ragni, M. (2022). The stability of syllogistic reasoning performance over time. *Thinking & Reasoning*, *28*(4), 529-568. doi: 10.1080/13546783.2021.1992012

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics. Vol. 3 : Speech Acts* (pp. 41–58). New York: Academic Press.

Johnson-Laird, P. N. (1975). Models of deduction. In R. J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 7–54). New York, US: Psychology Press.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, *16*(1), 1–61.

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.

Oaksford, M., & Chater, N. (2001, aug). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*(3), 533–566.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior*, *14*(2), 180–195.

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. (2018). The predictive power of heuristic portfolios in human syllogistic reasoning. In *Ki 2018: Advances in artificial intelligence: 41st german conference on ai, berlin, germany, september 24–28, 2018, proceedings 41* (pp. 415–421).

Riesterer, N., Brand, D., & Ragni, M. (2020). 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.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Mit Press.

Störring, G. (1908). *Experimentelle untersuchungen über einfache schlussprozesse*. W. Engelmann.

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.