UNIVERSITY OF TECHNOLOGY
IN THE EUROPEAN CAPITAL OF CULTURE
CHEMNITZ

Effect of Response Format on Syllogistic Reasoning

Daniel Brand & Marco Ragni

Professorship of Predictive Analytics

Contact: daniel.brand@psychologie.tu-chemnitz.de
https://www.tu-chemnitz.de/hsw/pva

Syllogistic Reasoning and Response Type

All **Architects** are **Bankers**
Some **Bankers** are **Cooks**
What, if anything, follows?

- A Syllogism consists of **two quantified statements** interrelating three terms via a common term
- The task is to **conclude** what the relation between the other terms is (Architects and Cooks)
- Syllogistic reasoning is one of the oldest domains in reasoning research [1]
- A variety of **cognitive models & theories** exist that try to account for human syllogistic reasoning [2]

▸ **Participants can usually only respond with a single conclusion**
- **Free Response:** Participants generate their conclusion freely (e.g., via free text responses)
- **Single Choice:** Participants select a conclusion from a list of possible options

▸ However, this can mix **preferences** with actual **reasoning effects**

▸ What changes, if participants can select **multiple conclusions**?
- Do typical effect still occur (e.g., Figure Effect [3])?
- What are the implications for modeling and model evaluation?

Datasets

- Datasets with three different response formats were used:
  - For Multiple-Choice, a study was conducted
  - Two openly available datasets were used for single-choice and free responses

Free Response

- Aggregated dataset compiled for a meta-analysis [2]
- Dataset is openly available
- Contains the responses of **156 participants** from six experiments
- Participants were asked to **freely generate a single response**
- Due to the free responses, not **all responses could be interpreted**, leading to percentages **not adding up** to 100%
- We **normalized** the percentages for better comparability

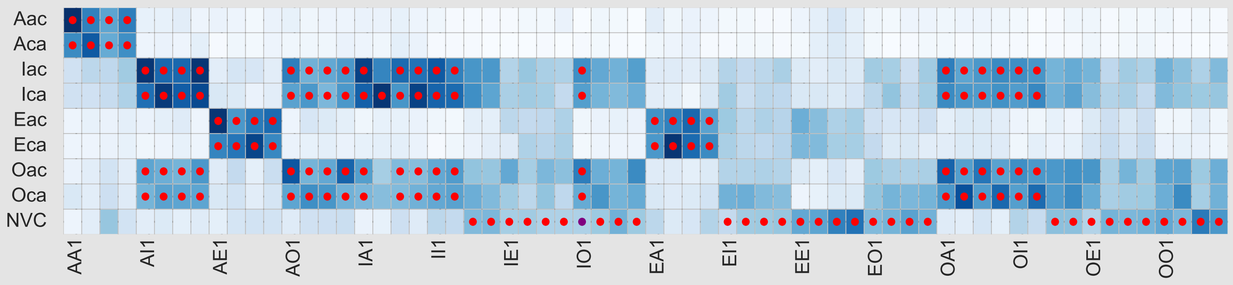
Single-Choice

- Dataset is part of the CCOBRA-Framework [4]
- Dataset is openly available
- Contains responses from **139 participants**
- Dataset was obtained from a web-experiment on Amazon Mechanical Turk
- Participants were asked to **select one** of the nine possible **response options**

Multiple-Choice

- We conducted a web-experiment on the platform Prolific
- The dataset contains responses from **100 participants**
- Participants were asked to **select all conclusions** that follow from the premises
- Alternatively, they could select that no valid conclusion exists (NVC)
- After selecting the responses, participants **had to lock their response in** to continue with the next syllogism
- On average, **1.9 conclusions** were selected (2.2 when excluding NVC)

Behavior with Multiple-Choice Response Format



Unweighted response patterns for multiple-choice. Darker shades of blue denote a higher occurrence of the respective response option. Red dots denote the most frequently selected response combinations for each syllogism (column-wise; purple in case of a tie).

- **NVC** seems to be **unaffected**
- Most common combinations **ignore direction**:
  - Except for NVC, **both directions** (ac and ca) are selected (except for *IA2*)
  - **This holds even in cases where it is logically not warranted**! (e.g., *AA1*)
  - **Contradicts** the **figural effect**!
- Overall, the figural effect is **significantly weaker** compared to single responses ($M=.32$ for single-choice vs $M=.1$ for multiple-choice; Mann-Whitney U: $U=905.0$, $p<0.001$)
- However, the figural effect was **still significant** (MWU between effect/no effect: $U=835.0$, $p<0.001$)
- Against expectations [5], selection of **universal quantifiers** did **not imply** the selection of **particular quantifiers** (i.e., All → Some and No → Some not)
- This is likely due to the **interpretation** of the quantifiers: 88% of the participants stated that *"Some A are B"* **does not include** the possibility that *"All A are B"*

Comparison of Response Patterns

Response Format	RMSE	MFA	Jaccard
Free Response - Single Choice	.06	.97	.78
Free Response - Multiple Choice	.10	.96	.66
Single Choice - Multiple Choice	.06	.98	.76

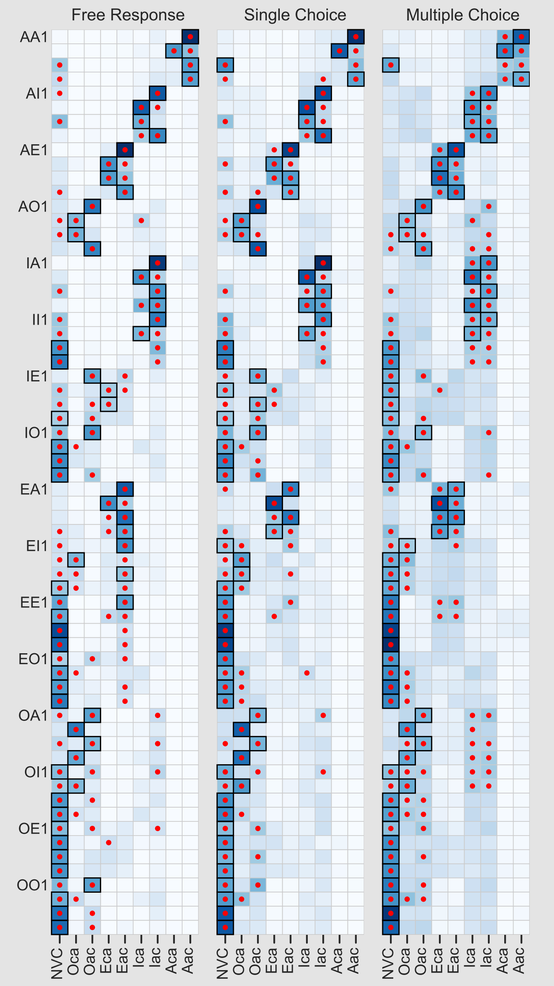
- Comparison of the patterns of each dataset based on
  - **Root Means Squared Error (RMSE)**
  - **Jaccard Coefficient**
  - Congruency of the **most-frequent answers (MFA)**
- The **Jaccard Coefficient** [6] is a metric commonly used to compare sets:

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

- **Intuitive interpretation:** A value of 0.75 means that 75% of the selected responses were are selected in both datasets
- Well-suited for comparing selected responses, since participants selected a **set of conclusions**, but did not **actively not select** the remaining options

- No substantial differences with respect to the **RMSE**
- **MFA patterns** not affected by response type
  - ▸ Important for models, which usually reflect the MFA
- Jaccard Coefficient indicates that **Single-Choice** is in **between** Free Response and Multiple-Choice
- Overall **high similarity** between datasets indicates:
  - Impact of response format is not substantial

▸ **Most findings and effects should be transferable**



Comparison of the (normalized) response patterns: Red dots denote reliable responses (≥ 16% of responses), a black border highlights the most-frequent answer (MFA).

Model Evaluation

- Jaccard Coefficient is **less dependent** on the number of responses than Accuracy
- Jaccard Coefficient can be interpreted **intuitively** for multiple choice
  - ▸ Only selected responses are considered (Accuracy has same weight for not-selected conclusions)

▸ For model evaluations on multiple-choice, Jaccard Coefficient is preferable

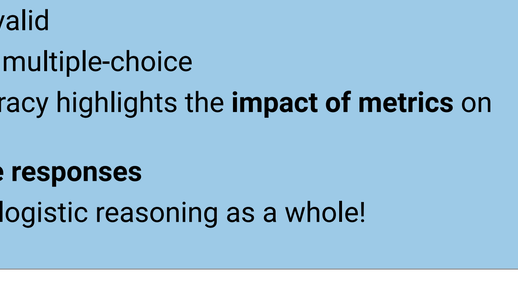
Aggregate Level

Model	Free Response		Single Choice		Multiple Choice	
	Acc.	Jaccard	Acc.	Jaccard	Acc.	Jaccard
Atmosphere	.79	.41	.80	.41	.82	.51
Conversion	.84	.43	.86	.50	.83	.50
Matching	.70	.34	.67	.28	.74	.43
MMT	.81	**.57**	.81	.56	.80	**.59**
PHM	.73	.33	.72	.30	.73	.34
PSYCOP	.77	.39	.78	.43	.75	.39
Verbal Models	**.85**	.55	**.87**	**.61**	**.84**	.53

▸ **No substantial differences** between the datasets

▸ Best model **depends on the metric**:
- Based on Accuracy: **Verbal Models**
- Based on Jaccard coefficient: **MMT**

Individual Level



Boxplots show medians and inter-quartile ranges, triangles denote the mean.

▸ **Individual predictions** [7] only from MMT & PHM [8]
- ▸ mReasoner [9] was used as a model for MMT

▸ Models were adapted to multiple choice

▸ **PHM outperforms MMT** when fitted to individuals

Conclusions

- **Effects and patterns** found in syllogistic reasoning research are **robust**
  - **No substantial differences between the response formats** (especially for single responses)
  - Allows the combination of different datasets for modeling endeavors
  - Error-prone interpretation of free response is not worth it for investigating general patterns
- The **figural effect** could be a **combination of reasoning and preference effects**:
  - Most participants deem **both directions** to be valid
  - However, the **figural effect is still present** with multiple-choice
- Difference between Jaccard Coefficient and Accuracy highlights the **impact of metrics** on evaluation
- Most **models** are only designed to generate **single responses**
  - Thereby, a specific task is modeled, but not syllogistic reasoning as a whole!

References

[1] Störring, G. (1908). *Experimentelle Untersuchungen über einfache Schlussprozesse.* W. Engelmann.

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

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

[4] Riesterer, N., & Brand, D., CCOBRA Framework (2021), GitHub Repository, https://github.com/CognitiveComputationLab/ccobra

[5] 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.

[6] Aggarwal, C. C. (2016). *Recommender systems: The text-book* (1st ed.). Springer Publishing Company, Incorporated.

[7] Riesterer, N., Brand, D., & Ragni, M. (2020). Do models capture individuals? Evaluating parameterized models for syllogistic reasoning. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3377–3383). Toronto, ON: Cognitive Science Society.

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

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

See Poster & Materials online