Using Cross-Domain Data to Predict Syllogistic Reasoning Behavior

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

Humans can reason across multiple domains (e.g., syllogistic, conditional, relational reasoning). Previous research has often investigated these domains in isolation, hence often identifying domain-specific strategies. A fundamental question in cognitive science is, however, if we apply rather a more general reasoning process or more domain-specific mechanisms. To approach this question and allow for analyses and modeling across domains, we first present a general data set that is well-grounded in the state-of-art of reasoning research and covers not only syllogistic, conditional and spatial reasoning, but also includes a test battery (e.g., memory). Second, we investigate relationships between the domains and present a preliminary step towards cross-domain modeling by predicting an individual's syllogistic conclusion based on behavior observed in the other domains. Our results show that domains are heavily interrelated with subtle differences between domains, highlighting the need for explanations that integrate subtle domain-specific strategies in a general theory of human reasoning.

Keywords: Syllogistic reasoning; Spatial reasoning; Conditional reasoning; Cognitive modeling; Cross-domain modeling

Introduction

The human mind is capable of reasoning across multiple domains, including spatial relations, conditional statements, and syllogisms. While each of these domains has been extensively studied, but often only in isolation, leading to the identification of domain-dependent strategies and mechanisms (Manktelow, 1999). However, an ongoing debate in cognitive science questions whether reasoning relies on "specialized learning systems, domain-specialized rules of inference, default preferences that are adjusted by experience" (Tooby & Cosmides, 2015), or whether a general reasoning mechanism, such as the general factor *g* (Spearman, 1904) underlies cognitive ability or if a more mediated position accounts better (Sternberg & Grigorenko, 2002). While this research question has been discussed little in the past (potentially due to the lack of a comprehensive data basis), the different fields have developed quite different modeling approaches. One of the core modeling paradigms in conditional reasoning is a Bayesian approach (Oaksford & Chater, 2007), for spatial reasoning most are model-based (Ragni & Knauff, 2013), and for syllogistic reasoning there are about 11 theories including heuristic, model-based and logic based ones (Khemlani & Johnson-Laird, 2012). Hence, in the past, reasoning research has led to accounts for isolated domains via optimized domain-dependent models, forgetting to raise the question for the need of a general cognitive reasoning model or being assured that this question can only be answered domain-specific. Our key contribution is twofold: In a first step we aim to provide a comprehensive data set that does not only capture the three major domains, but includes several additional tests for different cognitive resources, including working memory and tests like the Cognitive Reflection Test. Second, we analyze the potential factors by a classical correlational analysis and make a step towards predictive modeling across domains, using a recommender system to predict reasoning behavior based on observed behavior from other domains.

Domains, Study and Dataset

To lay a foundation for future analyses and cross-domain modeling in reasoning, we conducted a study covering most of the major domains and tasks in the field. Importantly, to allow modeling and analyses on an individual level, all participants solved every task. In total, the dataset offers the data of 95 participants, and was conducted over the course of three sessions as a web-experiment on Prolific[1]. The participants were required to be native English speakers. In the spirit of open science, dataset and the analyses in this paper are made openly available on GitHub[2]. In the following, the different tasks and domains in the study will be introduced.

Syllogistic Reasoning

As one of the oldest researched domains with over a century of research (e.g., Störring, 1908), the domain of syllogistic reasoning has a large variety of theories and models accounting for the observed effects (for an overview, see Khemlani & Johnson-Laird, 2012). Consider the following example:

All cooks are golfers.
Some golfers are monks.

What, if anything, follows?

A syllogism thereby consists of two quantified statements (*premises*) connecting the three terms via a middle term (*golfers*). The task is usually to derive a conclusion about the remaining terms (*cooks* and *monks*), or deduce that *no valid conclusion* (NVC) exists. Traditionally, the statements

[1] https://www.prolific.com/
[2] https://github.com/brand-d/cogsci-2025-crossdomain

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

use one of the four quantifiers from first-order logic, abbreviated with a single letter: *All* (A), *Some* (I), *No* (E) and *Some not* (O). The arrangement of the terms is referred to as the *figure*, and can be abbreviated with a number. In this paper, we use the notation presented by Khemlani and Johnson-Laird (2012), which is shown in the table below:

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

Using this notation, the syllogism above would be abbreviated with *AI1*. In a similar fashion, responses are abbreviated by the quantifier and *ac* and *ca* to denote the direction: "All C are A" would therefore be abbreviated with *Aca*. With all combinations of figures and quantifiers, the traditional syllogisms are a set of 64 tasks. We covered all of them in the study in randomized order with hobbies and professions as contents to avoid biases. Furthermore, we used a multiple choice response format, in which participants had to select all conclusions that follow from the premises, instead of selecting only one, as it does provide more information without losing the most important effects (Brand & Ragni, 2023). Participants additionally had to state for each task if they were confident in their response or made an informed guess. To control for known problems with the interpretation of traditional quantifiers (e.g., Ceraso & Provitera, 1971; Grice, 1975), participants were asked at the end if their interpretation of the quantifier *Some* (i.e., "Some A are B") also entails the possibility that *All A are B*.

Conditional Reasoning

In order to test conditional inferences, we included four tasks with normal argument types and four tasks with counterfactual argument types. Structurally, the tasks consist of a major premise with a conditional rule (e.g., "if p then q") and a minor premise (only p, q, $\neg p$ or $\neg q$).

Contents for the tasks were adapted from commonly used tasks in the field for the normal (Castañeda & Knauff, 2021; Singmann, Klauer, & Beller, 2016; Singmann & Klauer, 2011; Neys, Schaeken, & d'Ydewalle, 2002) and counterfactual (Byrne & Tasso, 2019; Byrne & Johnson-Laird, 2020; Espino & Byrne, 2020; Lewis, 1973) argument types. Participants got all four inference types with the respective contents in randomized order, with counterfactuals being tested in a later session than the normal conditionals. In order to avoid biases due to believability, the tasks were adapted to feature specific names instead of general rules. As an example, consider the following task, which is in modus ponens:

If Sara steps on a banana peel, then she slips.
Sara stepped on a banana peel.

What, if anything, follows?

Depending on the inference type (i.e., modus ponens (MP), modus tollens (MT), affirming the consequent (AC) or denying the antecedent (DA)), a potential conclusion (p or q) is

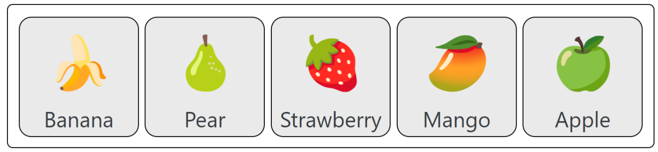

Figure 1: Example fruit arrangement shown to participants.

suggested in positive or negated form. Additionally, participants could select that *nothing follows*. The latter option is rarely used in conditional reasoning research, but allows comparability with syllogistic reasoning.

Furthermore, we included the abstract Wason selection task (Wason, 1968) as a related task to the classical conditional reasoning.

Spatial Reasoning

Spatial-relational reasoning, for simplicity referred to as only *spatial reasoning* throughout this paper, is another fundamental domain of reasoning research. We tested two-dimensional tasks featuring cardinal directions with two premises, as well as one-dimensional tasks with four-premises using only the relations *left* and *right* (for an overview over both task types, see Ragni, Brand, & Riesterer, 2021).

For cardinal direction tasks, participants were presented with two premises, relating three objects (buildings and places, e.g., church, bank and market) using cardinal and ordinal directions analogue to the structure of syllogisms. Participants were then asked for the relation between the two end-terms and had to select one of the eight directions or deduce that nothing can be concluded. However, note that in most of the tasks, there is no clear logically valid conclusion, making the task useful for preferences, but not for assessing logical correctness.

The four-premise tasks use fruits as contents and introduce 5 fruits per task, of which the positions are described with the premises. They were tested in two different ways: First, the more commonly tested way is referred to as *Spatials* throughout the paper. When presented with the premises, participants were asked whether a relation between two objects is guaranteed to hold. In 8 tasks, they were asked if one object is left or right the other. In another set of 8 tasks, they were instead asked whether the two objects are next to each other or if other objects are in between. Participants can answer with *yes*, *no*, or conclude that it is impossible to determine (the equivalent of *NVC* in this task). Indeterminate and determinate tasks were balanced and presented in randomized order. The 16 tasks are tested with premises visible while the question is present and with participants having to memorize them before the question is shown (referred to as *no memory* versus *memory*).

Second, as another form of four premise tasks, we presented participants with the premises and asked them to memorize the arrangements they describe. Afterwards, they were presented with an arrangement (see Figure 1) and had to decide if it is consistent with the premises before. After they made their decision, they were asked to either correct it, or rearrange the fruits to resemble the arrangement they had in mind. For that, they could use drag & drop with the mouse. Arrangements explicitly tested the *preferred mental models* as well as alternatives (see Ragni & Knauff, 2013). Indeterminate and determinate tasks were balanced and presented in randomized order.

Supplementary Tasks

In addition to the tasks covering the major reasoning domains, we added several short tasks that test additional cognitive abilities and tendencies that could shed light onto the processes behind the inter-individual differences. To this end, we used the 4-item short scale of the *Need for Cognition* (NFC; Beißert, Köhler, Rempel, & Beierlein, 2015) as well as a 7-item version of the *Cognitive Reflection Test* (CRT; Toplak, West, & Stanovich, 2014) to assess the cognitive style in terms of heuristic and deliberate reasoning. Since the ability to manipulate and memorize (spatial) mental representations is an intuitive predictor for spatial reasoning performance, we included the CORSI block tapping task (Brunetti, Del Gatto, & Delogu, 2014) to measure the spatial working memory as well as the Mental Rotation task by Shepard and Metzler (1971), which we implemented for the use in web-experiments. Finally, we included a custom test aimed at measuring the ability to perform mental operations on verbal information, referred to as *Verbal Substitution Task*. This was included to include an equivalent of the CORSI and the Mental Rotation tasks for verbal reasoning. In this task, participants are presented with sequences of consonants that they are asked to memorize. When continuing, they are presented with a chain of replacement rules, which are presented one after the other (e.g., “Replace all occurrences of Q by K”). At the end, they are asked for the final sequence. The task consists of 8 trials with increasing difficulty.

Analysis

In the following, we provide a brief general overview of the results of the reasoning tasks with respect to correctness. For the sake of space, we only report results for the main reasoning domains: Conditional, spatial and syllogistic reasoning.

Cardinal Directions Since the cardinal direction tasks usually do not have a unique logically correct response, they are not considered for the correctness analysis. However, the specific patterns can provide insights into preferences participants might have, e.g., the propensity to rely on direct transitive conclusions or their likelihood to conclude that there is *no valid conclusion*.

Spatial Reasoning Overall, participants achieved a mean correctness of 55.0%, with only little differences between the *memory* and the *no memory* group (53.3% and 56.6%, respectively). In the memory group, participants stated in 8% of the cases that they forgot parts of the premises instead of selecting an answer. Similarly, there is little difference between tasks asking if two objects are left/right of each other (53.1%) and tasks asking if they are directly adjacent/have other objects in between (56.7%). The only substantial difference was between determined and indetermined tasks (70.0% and 40.3%, respectively). This is in line with accounts based on the mental model theory (e.g., Johnson-Laird & Byrne, 1991; Ragni & Knauff, 2013), which explain the higher difficulty by the need to construct more models.

Spatial Arrangements The spatial arrangement tasks consist of two phases: the *verification phase*, where participants had to say if the shown arrangement of fruits is in line with the premises (e.g, Figure 1) and an *arrangement phase*, where they have to bring the arrangement in the correct order they had in mind. In the *verification phase*, participants achieved a mean correctness of .708. Again, indeterminate tasks are harder (.664), while determinate tasks are solved better (.774). In the *arrangement phase*, participants changed the ordering in 61.4% of the cases, but only provided a correct arrangement in 46.6% of the cases when they did. However, when they provided a correct ordering, it was predominantly (72.6%) the preferred mental model (PMM; see Ragni & Knauff, 2013), following a first-free-fit strategy.

Conditional Reasoning The overall correctness for conditional reasoning was 54.8%. For *modus ponens (MP)* and *modus tollens (MT)* they correctly decided if the conclusion necessarily holds. Due to the overall high correctness for those inference rules, there is little difference between normal and counterfactual argument types. For the invalid conclusions presented using *Affirming the consequent (AC)* and *Denying the antecedent (DA)*, participants would have to conclude that there is *No valid conclusion*. For normal argument types, only 29.8% and 27.2% (for AC and DA, respectively) did this correctly, showing no substantial difference between the two inference rules. For counterfactuals, *DA* the correctness dropped substantially to $11-7\%$, which is in line with the effect reported by Spiegel, Kern-Isberner, and Ragni (2019). Note that our results are not directly comparable due to the slightly different response format, where participants could always choose between a positive and a negative conclusion, or select the option that nothing follows. In the Wason selection task, only 3% of the participants solved the task correctly. In order to obtain a less dichotomous interpretation, we also calculated the Jaccard coefficient, obtaining a mean value of 0.394. This allows to also consider responses that selected only one card (e.g., p) or the biconditional interpretation ($p \wedge q$), yielding coefficients of $\frac{1}{2}$ and $\frac{1}{3}$, respectively. In the following analyses, the Jaccard coefficient is used.

Syllogistic Reasoning Since we used a multiple choice response format in our experiment, it is no longer trivial to determine the correctness, as it is for single choice response formats. Since participants were asked to select a set of answers, we need to compare two sets to determine the correspondence with the set of logically valid conclusions. For this, we use the Jaccard coefficient, which was already used by Brand and Ragni (2023) in the context of syllogistic responses. It is defined as the ratio between the intersection and the union of two sets (e.g., see Aggarwal, 2016):

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

Since it is a metric considering only values present in both sets, it is intuitive for a task where sets of responses are to be given: Both, omissions of logically correct conclusions and incorrect conclusions are considered, but it is invariant to conclusion candidates that are neither selected nor correct. As an example, consider the syllogism *AA1* (i.e., "All A are B, all B are C") and a participant selecting $\{Aac, Aca\}$, but the logical correct responses are $\{Aac, Iac, Ica\}$. In this case, the intersection would only be *Aac*, while all four other options are in the union, leading to a Jaccard coefficient of $\frac{1}{4}$.

Overall, participants achieved a mean Jaccard coefficient of .308. Furthermore, we calculated the proportion of logically correct responses out of the selected conclusions, thus ignoring when participants missed a conclusion. When using the latter, participants achieved a mean correctness of .356. However, this value is slightly lower than what is typically achieved with single choice response formats (e.g. 0.414 in the Ragni-2016 dataset provided with the Cognitive Computation for Behavioral Reasoning Analysis (CCOBRA) Framework[3]). This is to be expected, since multiple choice introduces a new source of errors, for example, when a preferred conclusion can be given with high confidence, but subsequent conclusions are more difficult to determine. The biggest source of mistakes is again *No Valid Conclusion (NVC)*, which has an even greater impact since it is the only response where a partially correct response is not possible.

Domain Relationships

Given the correctness values for the different tasks, we can investigate first relationships between the domains. To this end, we use a correlation-heatmap, which is shown in Figure 2. In the heatmap, the Spearman rank correlations between the different tasks with respect to correctness are shown. Significant results are shown in bold. Note that the significance is not corrected using any form of alpha-correction, since we use it in order to cluster the different domains in an exploratory sense. The goal is hereby to guide future modeling endeavors, rather than derive a definitive statement.

First, it becomes apparent that all three directly premise-based reasoning domains (spatial, conditional and syllogistic reasoning) correlate with each other. This shows that there is likely at least some transferrable capability that allows participants to either integrate information from premises well or not. Interestingly, neither conditionals nor syllogisms correlate with the spatial arrangement tasks, though. Since at least the verification part is comparable in terms of being directly based on premises, this is unexpected at first. We would assume that this is likely an effect of the presence of a *No valid conclusion (NVC)* option, which is not an option in verification tasks and has a substantial impact on the correctness, since roughly half of all tasks are indeterminate/invalid,

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

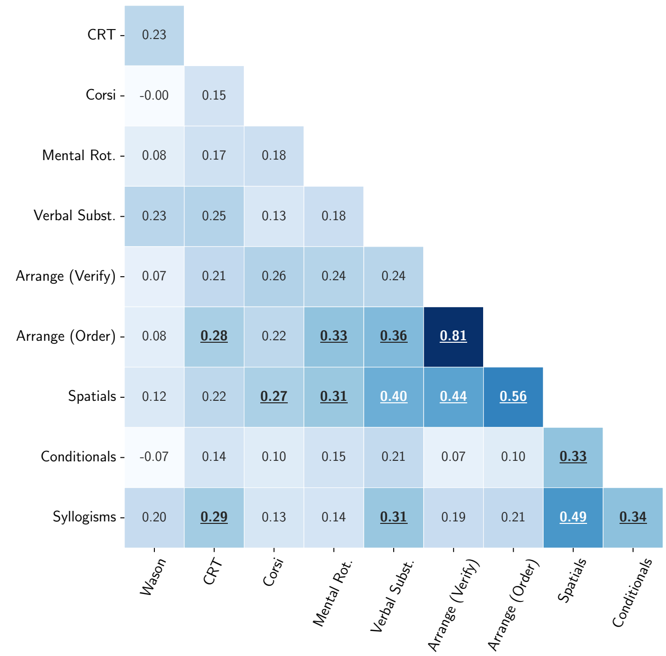

Figure 2: Spearman's rank correlations between the correctness of participants' responses across the different domains. Significant effects ($p < .01$) are bold and underlined. Text color only changes for better readability.

i.e., having NVC as the logically correct answer. Instead of simply checking consistency, determining if NVC is the correct answer requires an understanding of the concepts of *necessity* and *possibility* (for the differences between those concepts in syllogistic reasoning, see Brand, Todorovikj, & Ragni, 2024). Additionally, deriving NVC can require additional processes (e.g., a search for counterexamples; see Khemlani & Johnson-Laird, 2013).

Besides that general relation between the reasoning tasks, tasks with a spatial component seem in fact connected. Spatial reasoning correlates with the arrangement task as well as with the Corsi block tapping task and the mental rotation task, indicating that it relies on some ability to create, store and manipulate mental spatial representations. Conditionals, on the other hand, do not seem to correlate with any of the supplementary tasks. We argue that this is due to their simplicity, making them more about knowledge of logics than reasoning ability.

Finally, for syllogistic reasoning, the Cognitive reflection test (CRT), which hints at a participant's propensity to use heuristics or deliberative reasoning, is found to be related. This is in line with other investigations (e.g., Brand, Riesterer, & Ragni, 2023). Additionally, as we intended for tasks strongly relying on mental operations based on premises, the verbal substitution task is related to syllogistic reasoning correctness. However, against our expectations, it seems to be more general, since it also correlates to both spatial tasks. This makes the task interesting for predictive modeling and future investigations.

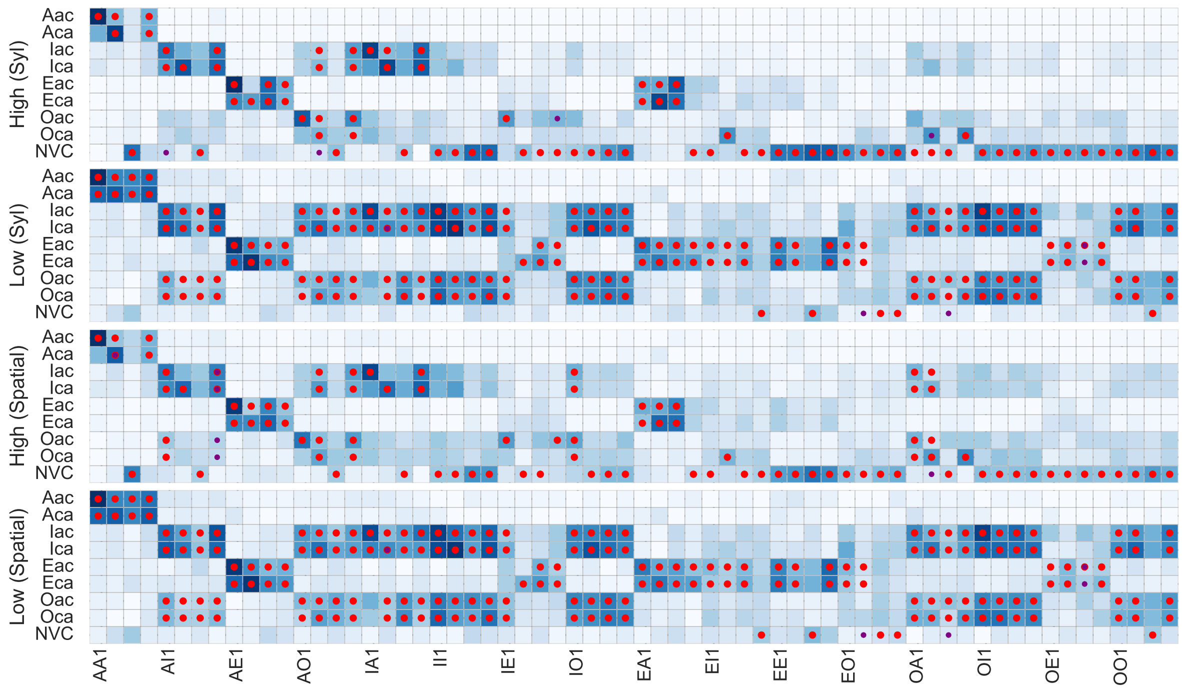

Figure 3: Syllogistic response patterns when divided by correctness using the Jaccard coefficient with the set of logically valid responses (top) as well as when divided by correctness in spatial reasoning tasks (bottom). Groups are divided by median correctness. Darker shades of blue denote a higher rate of the respective response. Red dots (column-wise) denote the set of responses that was most frequently selected together. Purple circles are used in case of a tie. For clarity, only responses participants stated to be confident in were considered.

Towards Predictive Modeling

In the following investigation, we focus on the domain of syllogistic reasoning, using it as an example. Thereby, we use the relationships between the domains we found. Since an in-depth evaluation of all of them with respect to their predictive power is out of scope for this paper, we focus on spatial reasoning as a predictor for syllogistic reasoning, as it had the strongest relationship based on the correlations. As a first step, we divide the participants into a high- and low-performing group by using the correctness in both domains, based on the respective median. Thereby, we obtain two splits, one based on the syllogistic reasoning performance, and one based on spatial reasoning performance. Comparing such splits can serve as a first impression, shedding light on the question if the other domains relate to different cognitive traits, or if there is essentially something akin to a general reasoning capability.

Figure 3 shows the response patterns in the syllogistic reasoning tasks provided by participants together with the most-frequently selected response sets (denoted by the red and purple circles) for both splits. First, we compare the two patterns originating from splitting by correctness in syllogistic reasoning. Thereby, the most striking differences are with respect to *NVC*, as well as to the direction. The *Low* pattern has rarely any *NVC* responses, and the most frequent sets of responses are always bidirectional (i.e., featuring the *ac*- and the *ca*-direction), while the figure effect is far more prominent in the *High* pattern. From an interpretational standpoint, it is corroborating the commonly reported findings that attribute most of the inter-individual differences in the domain to *NVC* (Riesterer, Brand, Dames, & Ragni, 2020; Ragni, Dames, Brand, & Riesterer, 2019; Brand, Riesterer, & Ragni, 2022) and the respective dominance of an *NVC-pattern* and a *non-NVC-pattern* (e.g., Brand et al., 2023). Commonly, the *Low* pattern would be considered to reflect heuristic reasoning, having resemblance with the *Matching hypothesis* or the *atmosphere effect* (Wetherick & Gilhooly, 1995). Splits based on the other domains and tests show similar results than what we observed with spatial reasoning, highlighting that, from a purely qualitative perspective, the splits resemble roughly the patterns obtained by splitting by syllogistic correctness itself. This hints at a potential use of the respective domains and tests as predictors for individual reasoning behavior in cross-domain modeling.

Recommender Analysis

To estimate the predictive potential of the other domains for syllogistic reasoning, we used a recommender system using user-based collaborative filtering (for an introduction, see Aggarwal, 2016). As a nearest-neighbour based approach, it determines the similarity between participants and uses a voting system (weighted by similarity) to predict the patterns for each user. For the evaluation, we used a leave-one-out crossvalidation, predicting patterns for each user using all other participants as training data. To quantify the prediction quality, we again used the Jaccard coefficient.

For the use in a recommender, we are interpreting the information from other domains as features of a participant's reasoning profile. To this end, we derived additional *features*: The NFC was included as well as the percentages of *NVC* for spatials, conditionals, and the cardinal directions, which is combined into the *NVC metrics*. To have reference values, we included several baselines: First, a randomly chosen participant instead of the recommender system. Second, the most frequently pattern with the most commonly selected set of responses. Third, a run of the recommender system using the correctness of the syllogisms directly. Despite being a post-hoc metric, and therefore not being suited for a true predictive setting, the latter can still serve as a baseline for the other metrics. Finally, we included the optimal combination of features, selected from a full search through the powerset of features. The optimal selection of features were the the conditionals, the spatials without memory ("visible") and the verbal substitution task. This is a reasonable combination given the correlations and the interpretations: Conditionals and spatials are intuitively related domains, and the verbal substitution task was created as a predictor for the ability to manipulate mental verbal representations. For comparison, a version including all features is included.

The results are shown in Figure 4. Most features, as expected, offer at least some improvement over the most frequent pattern, indicating that they cover some kind of individual trait important for the task. Additionally, the other reasoning domains are in fact even better predictors than the correctness on syllogisms alone, with conditionals and spatials surpassing all baselines. However, even the optimal combination is barely able to surpass the best single metrics, with all differences being only marginal. Overall, however, it seems that an upper bound is approached, where the inter-individual differences may not be explained by related domains.

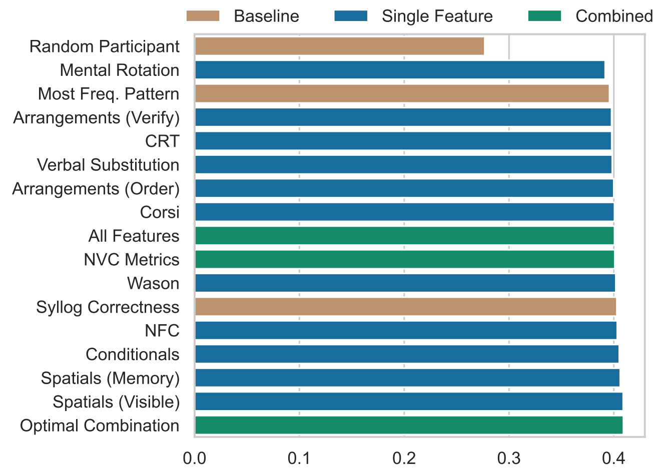

Figure 4: Predictive performance of the recommender system using different features as well as baseline performances. Performance is measured using the average Jaccard coefficient between participants' responses and predictions.

Discussion

In this paper, we present a comprehensive dataset covering some of the most important domains of human reasoning research. The dataset contains the responses of 95 participants to an extensive set of reasoning tasks across domains as well as additional tasks to assess cognitive capabilities of the participants. We believe that such a data foundation is essential to tackle the problem of cross-domain modeling that seems overdue in reasoning research: Despite theories built on concepts that work across domains exist (e.g., the Mental Model Theory; see Johnson-Laird, 1983), there are no model implementations yet that actually work across domains, even less in a predictive setting.

As a first step towards this, we investigated the data with a focus on getting initial insights into the relations between the different domains, as well as making a first step towards predictive modeling leveraging information from other domains, using syllogistic reasoning as an exemplary domain. We found that all three major reasoning domains were in fact significantly correlated with respect to correctness. This indicates that there seems to be a general and transferrable way of reasoning, which allows participants to achieve comparable performance across domains. However, this does not mean that the different domains do not have their own peculiarities and respective processes, since for example the spatial-relational reasoning tasks were the only ones profiting from spatial working memory. Interestingly, the verbal substitution task we introduced with this paper correlates well with both spatial and syllogistic reasoning, making it a potentially useful test for complex premise-based tasks.

The relation between domains also showed when assessing patterns of reasoners achieving above or below the median correctness in syllogistic and spatial reasoning tasks. The splits showed slight and subtle differences, even though a general reasoning capability seems prevalent. Finally, we used a recommender system as a proxy for predictive modeling. It showed that the additional information from other domains can be used effectively to inform predictions on an individual level, surpassing even the correctness in syllogisms themselves slightly. However, it also indicated a clear limit that seems to be approached, since not even the optimal combination of features could achieve any substantial performance gain. It is important to note that more complicated features could be investigated, including creating elaborate *reasoning profiles* for a participant.

To this end, analyses similar to the ones presented in this paper can shed light on the other domains in the dataset, with spatial reasoning being the most promising candidate. The challenge for cross-domain modeling will thereby be to capture the subtle differences between domains efficiently without treating them as separated entities.

Acknowledgments

This project has been funded by grants to MR in the DFG projects 529624975 and 283135041.

References

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

Beißert, H., Köhler, M., Rempel, M., & Beierlein, C. (2015). Deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition NFC-K [German short scale for measuring the construct Need for Cognition NFC-K]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*.

Brand, D., & Ragni, M. (2023). Effect of response format on syllogistic reasoning. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (pp. 2408–2414).

Brand, D., Riesterer, N., & Ragni, M. (2022). Model-based explanation of feedback effects in syllogistic reasoning. *Topics in Cognitive Science*, *14*(4), 828-844. doi: 10.1111/tops.12624

Brand, D., Riesterer, N., & Ragni, M. (2023). Uncovering iconic patterns of syllogistic reasoning: A clustering analysis. In C. Sibert (Ed.), *Proceedings of the 21th International Conference on Cognitive Modeling* (pp. 57–63). University Park, PA: Applied Cognitive Science Lab, Penn State.

Brand, D., Todorovikj, S., & Ragni, M. (2024). Necessity, possibility and likelihood in syllogistic reasoning. In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society* (p. 2776–2782).

Brunetti, R., Del Gatto, C., & Delogu, F. (2014). ecorsi: implementation and testing of the corsi block-tapping task for digital tablets. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00939

Byrne, R. M., & Johnson-Laird, P. N. (2020). If and or: Real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 760.

Byrne, R. M., & Tasso, A. (2019). Counterfactual reasoning: Inferences from hypothetical conditionals. In *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 124–130).

Castañeda, L. E. G., & Knauff, M. (2021). Everyday reasoning with unfamiliar conditionals. *Thinking & Reasoning*, *27*(3), 389–416. doi: 10.1080/13546783.2020.1823478

Ceraso, J., & Provitera, A. (1971). Sources of error in syllogistic reasoning. *Cognitive Psychology*, *2*(4), 400-410. doi: 10.1016/0010-0285(71)90023-5

Espino, O., & Byrne, R. M. J. (2020). The suppression of inferences from counterfactual conditionals. *Cognitive Science*, *44*(4), e12827. doi: https://doi.org/10.1111/cogs.12827

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. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA, USA: Harvard University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

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.

Lewis, D. K. (1973). *Counterfactuals*. Malden, Mass.: Blackwell.

Manktelow, K. I. (1999). *Reasoning and thinking*. Hove, East Sussex, UK: Psychology Press.

Neys, W. D., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory & Cognition*, *30*, 908-920.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Ragni, M., Brand, D., & Riesterer, N. (2021). The predictive power of spatial relational reasoning models: A new evaluation approach. *Frontiers in Psychology*, *12*, 626292.

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.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588.

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.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703. doi: 10.1126/science.171.3972.701

Singmann, H., & Klauer, K. C. (2011). Deductive and inductive conditional inferences: Two modes of reasoning. *Thinking & Reasoning*, *17*(3), 247–281. doi: 10.1080/13546783.2011.572718

Singmann, H., Klauer, K. C., & Beller, S. (2016). Probabilistic conditional reasoning: Disentangling form and content with the dual-source model. *Cognitive Psychology*, *88*, 61-87. doi: https://doi.org/10.1016/j.cogpsych.2016.06.005

Spearman, C. (1904). 'general intelligence,' objectively determined and measured. *The American Journal of Psychol-*

ogy, *15*(2), 201–292. doi: 10.2307/1412107

Spiegel, L.-P., Kern-Isberner, G., & Ragni, M. (2019). Rational inference patterns. In A. C. Nayak & A. Sharma (Eds.), *Pricai 2019: Trends in artificial intelligence* (pp. 405–417). Cham: Springer International Publishing.

Sternberg, R. J., & Grigorenko, E. L. (Eds.). (2002). *The general factor of intelligence: How general is it?* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Tooby, J., & Cosmides, L. (2015). The theoretical foundations of evolutionary psychology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 3–87). John Wiley & Sons. Retrieved from `https://www.cep.ucsb.edu/wp-content/uploads/2023/05/2015ToobyCosmides-BussEPHandbook.pdf`

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147-168. doi: 10.1080/13546783.2013.844729

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. doi: 10.1080/14640746808400161

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