Response Variability and Stability in Human Reasoning

Clemens Bombach (clemens.bombach@hsw.tu-chemnitz.de)[1], Rajmadan Lakshmanan[2] & Marco Ragni[3]
[1,2,3]Predictive Analytics, TU Chemnitz

Abstract

Understanding how humans reason – and how reasoning responses vary across tasks and individuals – remains a core challenge for modeling and explanation in cognitive science. We investigate the stability of response patterns within reasoners and whether variation in these patterns can be used to predict learning effects. We introduce a formal, geometry-based method to quantify distances between individual reasoning patterns and their internal variability, grounded in heuristic theories. The proposed framework is tested against experimental data via generalized linear mixed-effects models and clustering, where we find that our proposed variation measure interacts with correctness to predict performance gains. Moreover, we find that reasoning patterns are stable over time within the same reasoner. The method is general enough to be applied to other reasoning domains.

Keywords: Reasoning; Similarity Analysis; Cognitive Modeling; Syllogisms; GLM; Cluster analysis

Introduction

Many (classical) cognitive models implicitly assume an almost deterministic mapping from stimulus to response: holding the task and context constant, repeated presentations of the same input are expected to yield almost the same or at least similar response. This assumption simplifies model identification and parameter estimation, because variability can be treated as negligible. But, substantial trial-to-trial and inter-individual variability has been shown in reasoning and decision making, arising from factors such as lack of focus, attention, and context or sequence effects—often conceptualized as unwanted noise in decision making (e.g., Kahneman et al., 2022).

Consider the domain of human reasoning, more precisely the domain of syllogistic reasoning. A syllogism consists of two quantified premises over three terms, for example

All Architects are Bakers,
All Chemists are Bakers,

and a conclusion such as "All Architects are Chemists" or "Some Chemists are not Architects". Normative and observed human responses in syllogistic reasoning differ substantially. For example, in a meta-analysis using data from 6 studies of all 64 categorical syllogisms, it was reported that 47% of all participants responded "All Architects are Chemists" (Khemlani & Johnson-Laird, 2012). However, the logically correct response is that no valid conclusion can be derived from the premises, chosen by only 31% of participants. None of the 9 reported theories in Khemlani and Johnson-Laird (2012) adequately explain all gaps between normative and empirically observed responses, so the development of a general psychological theory remains an open problem. Interestingly, most of the 9 theories reported assume a static cognitive process, i.e., for the same input of the task they predict in most cases the same response.

Previous studies indicate a shift of the responses towards more logically correct responses if human reasoners are tested a week apart (Dames et al., 2022; Ragni et al., 2018). Responses in reasoning tasks do not vary at random, but instead reveal systematic patterns. These patterns for individuals hold across different tasks (Stanovich & West, 1998). Consistency judgments have been shown to be systematically influenced by underlying representations and processing constraints (Ragni et al., 2014). However, with the exception of one theory (mReasoner) that change has not been explained by other theories and it has rarely been investigated on the individual level. A related question that remains open is, to which degree do small variations in the task change the response? For instance, the syllogism above is obtained from the syllogism

All Architects are Bakers,
All Bakers are Chemists,

by swapping the terms in the second premise. Superficially, this is a small change, but it changes the normative response from "All Architects are Chemists" to "No valid conclusion". However, the result of Khemlani and Johnson-Laird (2012) indicates that most reasoners are insensitive to this change.

A major barrier in this regard is that there is no generally accepted formalism for determining whether a response pattern is more or less variable, or for defining the distance between two response patterns. Hence, this paper is concerned with developing such a formally sound framework. The principal idea is to identify individual reasoning patterns with *functions* acting on spaces of binary vectors, thereby transforming questions about reasoning patterns into questions about functions on discrete spaces. Such functions are well studied in the mathematical literature, allowing us to draw on suitable notions of distance and variability. This method of identification also incorporates findings from heuristic theories of syllogistic reasoning — theories that predict responses from surface features of the premises alone, without evaluating logical validity — including the atmosphere theory (Woodworth & Sells, 1935), Probability Heuristics Model (PHM) (Chater & Oaksford, 1999), and TransSet (Brand et al., 2019). More

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

precisely, the atmosphere theory predicts conclusions from the universal/particular and positive/negative character of the premises. The PHM ranks conclusions by informativeness derived from a probabilistic reading of the quantifiers, predicting that reasoners prefer more informative conclusions. TransSet models the reasoner's attempt to construct a transitive inference path across the figure of the syllogism. These three theories together motivate our encoding of both tasks and responses into binary feature vectors, as detailed in the upcoming discussion.

We show the value of this framework by re-analyzing a dataset collected in a study on the stability of syllogistic reasoning over time (Dames et al., 2022). Participants solved twice all 64 syllogisms, the second time a week later. Performance often improved in the retest, however the degree of improvement was highly dependent on the particular task and reasoner. Based on this result and the discussion above, we formulate the following research questions:

- **Q0** What is a useful distance measure on syllogistic reasoning patterns? How can we measure response variability?
- **Q1** To what extent does response variability explain individual differences in performance improvement? In other words, are highly variable response patterns a predictor of retest performance?
- **Q2** Related to the result of Stanovich and West (1998): Are reasoning patterns stable within individuals? That is, is the distance between a reasoner's initial response pattern and their retest pattern smaller than the distance between their initial pattern and another reasoner's retest pattern?

Materials and Methods

We begin by briefly describing the dataset used for our analysis (Dames et al., 2022). This dataset was chosen because of its large sample size and because it includes repeated measures for all 64 syllogisms. Session 0 assessed personality traits and general cognitive abilities. In Session 1, participants selected one response (out of the 9 canonical responses) for each of the 64 syllogistic problems; no feedback was provided, and task order was randomized. This procedure was repeated in Session 2, which took place one week after Session 1. The final sample consists of $N = 100$ individuals.

For the data analysis, we employ generalized linear mixed models (GLMMs) which were fit in R using the `lme4` package (Bates et al., 2015). GLMMs a are well-known tool for hierarchical modeling. For model diagnostics, we employed the `DHARMa` package (Hartig, 2024). For the calculation of p-values, we employed the `lmerTest` package (Kuznetsova et al., 2017). All models were fit using the BOBYQA optimizer. Goodness of fit was assessed with the `performance` package using pseudo-R^2 (Lüdecke et al., 2021; Nakagawa & Schielzeth, 2013). The data and the scripts used to run the analyses are openly available on GitHub[1].

[1] https://github.com/c-bombach/cogsci-response-variability

The new ingredient of our approach is the systematic use of response variability and distance features, which allow a finer characterization of individual responses than correctness alone and are defined in the next section.

Q0: Framework for Feature Sensitivity

Feature sensitivity is operationalized here as local response variability with respect to a psychologically motivated task distance. A syllogism, like above, consists of two quantified premises about three terms a,b,c. Quantifiers are $\mathtt{A}$ = All, $\mathtt{E}$ = No, $\mathtt{I}$ = Some and $\mathtt{O}$ = Some . . . not.

There are 4 possible orderings of the terms and these are called figures:

Figure 1	Figure 2	Figure 3	Figure 4
a-b	b-a	a-b	b-a
b-c	c-b	c-b	b-c

A syllogistic problem is uniquely determined by the quantifiers of the two premises and the figure (e.g., the syllogism from the introduction is AA3), which we refer to as its *surface features*. Analogously, responses can be encoded in terms of surface features, with NVC (No Valid Conclusion) as the logically correct response in 37 of the 64 syllogisms. The remaining 27 syllogisms each license a valid conclusion, which must be one of 8 possible response types consisting of a quantifier from $\{\mathtt{A}, \mathtt{E}, \mathtt{I}, \mathtt{O}\}$ combined with a direction $\{\mathtt{ac}, \mathtt{ca}\}$ (e.g., Eac for "No a are c").

In the following, we denote by $X = \{(q1, q2, r) : q_1, q_2 \in \{\mathtt{A}, \mathtt{E}, \mathtt{I}, \mathtt{O}\}, r \in \{1, 2, 3, 4\}\}$ the set of syllogistic tasks. The set of responses is defined as $Y = \{(q, d) : q \in \{\mathtt{A}, \mathtt{E}, \mathtt{I}, \mathtt{O}\}, d \in \{\mathtt{ac}, \mathtt{ca}\}\} \cup \{\mathtt{NVC}\}$. As outlined in the introduction, we identify reasoning patterns with functions $f : X \to Y$ that take a task $x \in X$ as an input and produce an output $f(x) \in Y$.

To obtain notions of distance and variability from X to Y, we begin by defining distance functions (also called *metrics*) between elements of the underlying spaces X and Y respectively. Formally, a metric on a set Z is a function $d_Z : Z \times Z \to [0, \infty)$ such that $d(z, w) = 0$ if and only if $z = w$, and for all $z, w, u \in Z$ it holds that $d_Z(z, w) = d_Z(w, z)$ and $d_Z(z, u) \leq d_Z(z, w) + d_Z(w, u)$. The pair (Z, d_Z) is called a metric space.

Metrics on task and response sets

A simple metric can be obtained by one-hot encoding. This means that each syllogism is encoded as a unit vector in $\mathbb{R}^{64}$. Then, we could take the Euclidean distance on $\mathbb{R}^{64}$. This, however, would make all syllogisms equidistant to each other. Moreover, in applications such as clustering, such an embedding suffers from the curse of dimensionality. By contrast, our goal is to obtain a metric to express response similarities and which lives on a low-dimensional space. If we wish to stick with a binary encoding, we need at least 6 bits since $2^6 = 64$.

To motivate our definition of a metric on X, we recall that the syllogistic quantifiers can be divided into *universal* quantifiers $\mathtt{A}, \mathtt{E}$ versus *particular* quantifiers $\mathtt{I}, \mathtt{O}$ and *positive* quantifiers $\mathtt{A}, \mathtt{I}$ versus *negative* quantifiers $\mathtt{E}, \mathtt{O}$. This classification is the basis of the atmosphere theory of syllogistic reasoning (Woodworth & Sells, 1935), which postulates that reasoners select conclusions according to the *atmosphere* created by the premises. Specifically, the theory predicts that the universality and positivity of the premises jointly determine the predicted conclusion: a single particular premise (`I` or `O`) is sufficient to bias reasoners toward a particular conclusion, and a single negative premise (`E` or `O`) toward a negative one. Because conclusion selection is thus driven by these two binary features of each premise quantifier, encoding each quantifier along precisely the universality and positivity dimensions embeds the theoretically relevant structure directly into our metric.

We identify each quantifier q with a tuple (u_q, p_q) of bits u_q, p_q that encode the universality or the positivity of q, respectively. More precisely, we have $u_q = 1$ if q is universal and 0 if it is particular. Analogously, we set $p_q = 1$ if q is positive and 0 otherwise. In a similar fashion, we encode the figure r using two bits t_r, s_r. In particular, we set $t_r = 1$ when $r \in \{1, 2\}$ The definition of t_r is motivated by the following considerations: In figure 1 syllogisms, the ordering of subject and terms allow for a simple transitive inference from a to c (even when this inference may not be logically warranted). A figure 2 syllogism can be reduced to a figure 1 syllogism by swapping the ordering of the premises. For the figures $3, 4$, it is not immediately possible to obtain such a transitive path. We therefore see figure 1 and 2 syllogisms as *transitive*. This distinction between transitive and intransitive figures plays an important role in TransSet (Brand et al., 2019, 2020), which is an algorithmic model of syllogistic reasoning. It is based on the empirically supported observation that reasoners attempt to create a transitive path between inferences. If such a path is not immediately available, they resort to a (possibly illicit) reordering of the terms.

If we wish to obtain a 6-dimensional embedding, we are now left with one free bit s_r that is essentially uniquely determined. Thus, we set $s_r = 1$ if $r \in \{1, 3\}$. We can therefore encode a syllogism $x = (q_1, q_2, r)$ as the 6-dimensional binary vector

$$j(x) = (u_{q_1}, p_{q_1}, u_{q_2}, p_{q_2}, t_r, s_r) \in \{0, 1\}^6 . \tag{1}$$

We obtain a one-to-one correspondence $j : X \to H_6$ between X and the 6−dimensional Hamming cube $H_6 = \{0, 1\}^6$. On H_6, we have a natural metric given by the Hamming distance

$$d_{H_6}(v, w) = \sum_{j=1}^{6} |v_j - w_j| .$$

Essentially, this counts the number of bits on which v and w disagree. We pull back this distance to a metric on X using j by setting $d_X(x, y) = d_{H_6}(j(x), j(y))$.

To obtain a metric on Y, we proceed in a similar fashion: Given $y \in Y$ we set $k(y) = (0, 0, 0, 0)$ if $y = \texttt{NVC}$. Otherwise, if $y = (q, d)$ where $q \in \texttt{A}, \texttt{E}, \texttt{I}, \texttt{O}$ and $d \in \{\texttt{ac}, \texttt{ca}\}$, we set $k(y) = (u_q, p_q, 1, 0)$ if $d = \texttt{ac}$ and $k(y) = (u_q, p_q, 0, 1)$ if $d = \texttt{ca}$. Finally, we set

$$d_Y(y, \tilde{y}) = \sum_{j=1}^{4} |k(y)_j - k(\tilde{y})_j| .$$

Geometrically, we can think of k as mapping the valid responses of Y into a proper subset of the four-dimensional Hamming cube H_4.

Note that the `NVC`-response is mapped to the origin. Our choice of k implies that `O`-responses lie closest to `NVC`. This choice is inspired by the `O`-heuristic of the PHM theory of syllogistic reasoning (Chater & Oaksford, 1999). PHM postulates that reasoners use heuristics that are based on the informativeness of the premises. This informativeness is itself derived from a probabilistic interpretation of the quantifiers: `A` is maximally informative as it rules out the most possibilities, while `O` is least informative. The `O`-heuristic states that reasoners generally avoid drawing `O`-conclusions due to their low informativity, suggesting a psychological similarity between `NVC` and `O`-conclusions. By contrast, the most informative conclusions $\texttt{A}ac$ and $\texttt{A}ca$ have high distance to `NVC` in our embedding.

Distance and variability of response patterns

Having defined metrics d_X, d_Y on X and Y, we may now view reasoning patterns as maps $f : (X, d_X) \to (Y, d_Y)$ between metric spaces. On the set of these maps we consider, for $p \geq 1$, the $\ell^p(X; Y)$-metrics,

$$d_{\ell^p(X;Y)}(f, g) = \left(\sum_{x \in X} d_Y(f(x), g(x))^p \right)^{1/p} .$$

This is well-defined since X is finite and generalizes the familiar Euclidean distance for $p = 2$.

Since the Hamming cubes carry a natural graph structure, we may also see reasoning patterns as maps between graphs. This means we view X, Y as graphs where the vertices x, y are elements of X, Y and two vertices are connected with an edge (written as $x \sim y$) if and only if the distance between them is 1. Thus, the degree of a vertex in X is given by $\deg(x) = |\{y : y \sim x\}| = |\{y : d(y, x) = 1\}|$.

We use this graph structure to define the desired similarity measure $f : X \to Y$. It is a generalisation of the graph energy for real-valued functions on graphs (see, e.g., Keller et al., 2021) to the setting where the target space Y is a metric space:

Definition 1 *Let $p \geq 1$, $(X, d_X), (Y, d_Y)$ be metric spaces such that X is a finite graph. We define the local p-energy of $f : X \to Y$ at $x \in X$ as*

$$\mathcal{E}_p(f; x) = \frac{1}{\deg(x)} \sum_{y \sim x} d_Y(f(x), f(y))^p .$$

The total energy of f is defined as

$$\mathcal{E}_p(f) = \sum_{x \in X} \mathcal{E}_p(f; x) .$$

The respective normalizations are defined as

$$E_p(f; x) = \mathcal{E}_p(f; x)^{1/p} , \quad E_p(f) = \mathcal{E}_p(f)^{1/p} .$$

Note that the local energy of f at x is 0 if f is constant on a neighborhood of x. Thus, the energy measures the degree to which f is not (locally) constant and can therefore be seen as a measure of variability. Our definition is itself a special case of the notion of p-energy of maps between metric measure spaces as defined in the work by Jost (1994). Such energies have been used extensively in machine learning for the purpose of clustering, for example in the Laplace eigenmaps algorithm (Belkin & Niyogi, 2003).

Meta-data inference

In preparation for the predictive modeling, we perform a meta-data inference using a k-means clustering. The dataset consists of repeated task–response trials collected from a fixed set of participants. Let $I = \{1, \dots, M\}$ denote the set of participants, where each participant $i \in I$ completes multiple trials indexed by $t = 1, \dots, T$. Each trial involves a syllogistic task $x_{i,t} \in X$, a response $f_i(x_{i,t}) \in Y$, and an associated local energy value derived from the response pattern.

The present analysis focuses exclusively on participant-level information. This aggregation allows us to infer general response tendencies and stability in reasoning behavior.

For each trial t of participant i, we define a feature vector $\mathbf{v}_{i,t} = (\mathbf{t}_{i,t}, \mathbf{r}_{i,t}, E_2(f_i; x_{i,t}))$ where $\mathbf{t}_{i,t} \in \{0,1\}^6$ denotes the binary task-feature encoding introduced in (1), $\mathbf{r}_{i,t} \in \{0,1\}^4$ denotes the binary response-feature encoding. $E_2(f_i; x_{i,t})$ denotes the normalized local 2-energy introduced in Definition 1, evaluated at task $x_{i,t}$ for participant i.

To obtain a participant-level representation, we compute the empirical mean $\bar{\mathbf{v}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{v}_{i,t}$. This vector summarizes the average task exposure, response tendencies, and energetic profile of participant i. Before clustering, we standardize each feature dimension across participants to zero mean and unit variance. This step ensures that task features, response features, and energy contribute comparably to the distance computations.

We perform participant-level clustering using the k-means algorithm. Let $\tilde{\mathbf{v}}_i$ denote the standardized participant-level feature vector. For a fixed number of clusters k, the algorithm identifies cluster centroids $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k \in \mathbb{R}^p$ by minimizing the objective function

$$\min_{\{c_i\},\{\mu_j\}} \sum_{i=1}^{M} \left\| \tilde{\mathbf{v}}_i - \boldsymbol{\mu}_{c_i} \right\|^2,$$

where $c_i \in \{1, \dots, k\}$ denotes the cluster assignment of participant i. Each centroid $\boldsymbol{\mu}_j$ represents a prototypical participant profile. This profile captures a characteristic combination of task-feature exposure, response-feature usage, and energy measure. To find the optimal number of clusters, we employ the elbow criterion and silhouette score. The elbow criterion measures how much the inertia, which measures within-cluster variability, decreases as the number of clusters increases. On the other hand, the silhouette score quantifies how well-separated and coherent the resulting clusters are. From Fig. 1, both criteria indicate a clear balance at $k = 3$, which we therefore select as the optimal number of participant clusters.

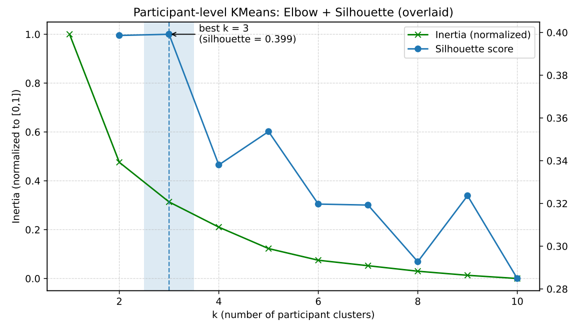

Figure 1: Participant-level parameter (k) selection using the elbow and silhouette criteria. The shaded region marks the selected number of clusters k.

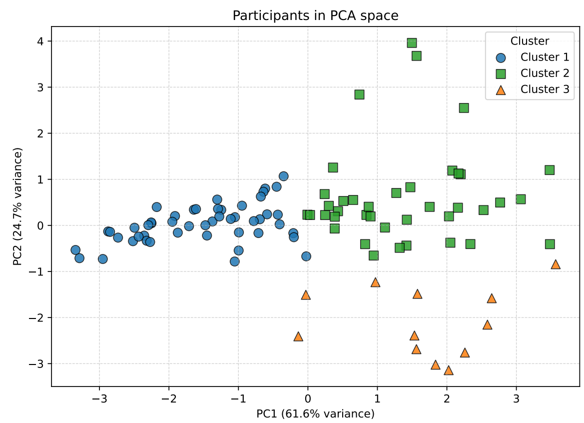

Figure 2: Principal component analysis of participant-level feature vectors. Points represent participants projected onto the first two principal components; colors and marker shapes indicate k-means cluster membership. Axis labels show the variance explained by each component.

Participants in Cluster 1 were found to have substantially higher mean correctness (67.2%) than participants in Clusters 2 (37.9%) and 3 (30.3%). By contrast mean energy was highest for Cluster 2 (0.471), followed by Cluster 3 (0.389) and Cluster 1 (0.307). We apply Principal Component Analysis (PCA) after clustering to obtain a low-dimensional projection of participant-level feature vectors. Each participant is represented by averaged task features, response features, and local energy, so principal components reflect joint variation in strategy choice and response smoothness. Fig. 2 shows the participant-level clustering projected onto the first two principal components.

The PCA projection reveals clear structure across the three participant clusters. Principal components are influenced by both response tendency and energy, with the PC1 showing a stronger positive influence of energy, and PC2 being more influenced by response tendency, in particular with respect to direction (`ac` vs. `ca`). Participants in Cluster 1 occupy low values along the first principal component and form a relatively compact group along the second, indicating a *coherent, correct response strategy* with low flexibility. Participants in Cluster 2 show high values along the first principal component and a broader spread along the second. Participants in Cluster 3 align with Cluster 2 along the first principal component but separate strongly along the second, occupying markedly lower values. This separation indicates that participants in Cluster 3 exhibit substantially lower response flexibility.

Mixed-effects modeling

Q1: Prediction of retest performance

For this research question, we studied the effect of the energy in the inital test on correctness in the retest. This can be reformulated as follows: When reasoners give a response that is exceptional in the sense that it has a large distance from responses to neighboring syllogisms (in the sense of the metric defined above), does this indicate a performance improvement when they are tested again? We fit a sequence of GLMMs with logit link function with the aim of predicting retest correctness C_{retest} from the following variables:

- initial correctness C, encoded such that $C = 1$ for correct responses and $C = 0$ for incorrect responses
- task validity V, encoded such that $V = 1$ for valid tasks and $V = -1$ for invalid tasks
- participant-centered and scaled, normalized local 2-energy, E_2^{pc}
- (standardized) correctness per task averaged over all participants $\overline{C}$

In particular, we were interested in the effect of the energy variable and its interactions.

Continuous variables were standardized to aid convergence and to make results easily interpretable. We decided to use the normalized 2-energy since it corresponds to the Dirichlet energy. The normalization reduces skewness and ensures that it is in the same "length unit" as the underlying spaces. We opted for participant-centering since we are interested in predicting the effect of local deviations in energy on learning. Mean correctness $\overline{C}$ was included as a predictor since models that didn't include it showed substantial underdispersion during model diagnostics. We presume that this is caused by the large variation in task success rate since some syllogisms (such as `AA1`) are solved correctly by almost all participants.

To find a suitable model, our aim was to find a convergent model with a maximally complex random effects structure and minimal AIC. The best model we were able to fit under these criteria is given by the following formula in Wilkinson-Rogers notation (Wilkinson & Rogers, 1973):

$$\begin{aligned} C_{\text{retest}} \sim & C * V * E_2^{\text{pc}} - C : V : E_2^{pc} + \overline{C} \\ & + (C + V|\,\text{participant}) + (C + E_2^{\text{pc}}|\,\text{task}) \qquad (M_1) \end{aligned}$$

In order to have a baseline for assessing the effect of energy, we fitted the reduced model

$$C_{\text{retest}} \sim C * V + \overline{C} + (C + V|\,\text{participant}) + (C|\,\text{task})\,. \quad (M_0)$$

Table 1: Fixed effects for models M_1 and M_0. $N = 6400$ observations; 100 participants; 64 tasks.

	M_1	M_0
Intercept	-0.467^{**} (0.152)	-0.360^{*} (0.141)
Correctness C	1.560^{***} (0.139)	1.446^{***} (0.133)
Validity V	0.242^{*} (0.121)	0.227^{*} (0.109)
Energy E_2^{pc}	0.225^{**} (0.076)	
Av. Corr. $\overline{C}$	1.307^{***} (0.074)	1.281^{***} (0.073)
$C : V$	0.037 (0.121)	−0.018 (0.111)
$C : E_2$	-0.438^{***} (0.100)	
$V : E_2$	−0.122 (0.064)	
AIC	5442.222	5481.051
BIC	5577.503	5575.748
Log Likelihood	−2701.111	−2726.526

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

A model comparison showed a significant improvement in fit of (M_1) over (M_0) ($\Delta AIC = 39, G^2(6) = 50.82, p = 3.2 \cdot 10^{-9}$). This strongly indicates that energy has substantial predictive value for the problem under consideration. Overall, model fit was acceptable with $R^2_{\text{conditial}} = 0.644, R^2_{\text{marginal}} = 0.382$. The results for the fixed effects and the variance components are listed in Table 1. As expected, Correctness and mean correctness have the largest effects on retest correctness with log-odds $\beta_C = 1.560$ and $\beta_{\overline{C}} = 1.307$ respectively. The effects of validity and energy are smaller ($\beta_V = 0.242, \beta_{E_2} = 0.225$) but the low standard deviation for energy in particular ($\sigma_{E_2} = 0.076$) suggests a robust finding. That improvement in correctness is higher for valid syllogisms is consistent with the findings of Dames et al. (2022). We note the significant correctness-energy interaction $\beta_{C:E_2} = -0.438$.

To get a better sense of the effect of energy on performance improvement, it is interesting to compute effective odds ratios dependent on the value of the correctness and validity variables for energy in (M_1):

$OR(E_2)$	valid	invalid
correct	0.72	0.91
incorrect	1.11	1.41

In particular, we have $OR < 1$ for correct and $OR > 1$ for incorrect responses. One possible interpretation for these effects is follows: When a correct response has high energy, it is different from responses to related syllogisms and thus it is less likely that it can be associated to a consistent strategy of the participant. On the other hand, low energy can be taken as a sign of a systematic response pattern. Thus, syllogisms with correct but unsystematic responses are less likely to be answered correctly on retest. By contrast, for incorrect responses, high energy raises the probability of improvement. In particular, this effect is stronger for incorrect responses to invalid syllogisms.

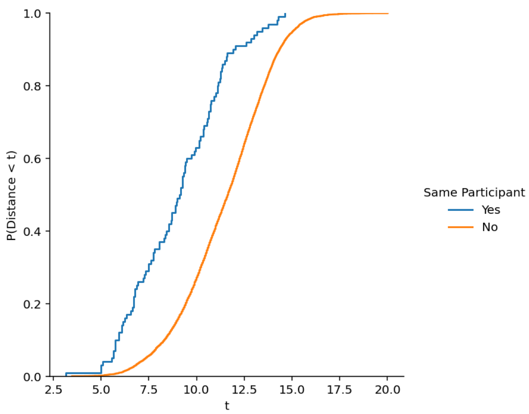

Figure 3: Empirical cumulative distribution functions (cdfs) for test-retest distance

As a sanity check, we replaced the theory-driven encoding based on task features by a simple one-hot encoding. As expected, this led to a substantial reduction in fit over the best model with theory-driven encoding ($\Delta AIC = -25$) and made the main effect of energy insignificant ($\beta = 0.070$). However, we note the observation of a very significant validity-energy interaction ($\beta = -0.025$, $p < 0.0001$). A possible explanation is that a preference towards `NVC`-responses leads to a large subset on which the reasoning pattern is constant. This lowers the energy in an encoding-independent fashion. Such an effect can therefore be detected by a coarse encoding.

Q2: Stability of reasoning patterns

Formally, the question is: Suppose that i, j are reasoners, f_0^i, f_0^j their response patterns on the initial test and f_1^i, f_1^j their response patterns on the retest. Moreover, let us write $d_{ij} = d_{\ell^2(X;Y)}(f_0^i, f_1^j)$ in the following. If response patterns are stable within reasoners, then we should have

$$\mathbb{E}(d_{ij} | i = j) < \mathbb{E}(d_{ij} | i \neq j) \tag{2}$$

with $\mathbb{E}$ denoting the (conditional) expectation. In other words, the expected distance of a reasoner between a test and retest pattern is smaller conditional on the fact that they come from the same reasoner. A plot of the empirical conditional distribution functions suggests that this is indeed the case. In Figure 3, the blue curve (same participant) is shifted to the left of the orange curve (different participants), indicating that within-reasoner distances are stochastically smaller than between-reasoner distances. Formally, our sample fulfills the stronger condition $\mathbb{P}(d_{ij} < t | i = j) > \mathbb{P}(d_{ij} < t | i \neq j)$. In order to estimate the effect size, we fit a LMM of the form

$$d_{ij} \sim \mathbf{1}_{i=j} + (1|i) + (1|j)\,,$$

where $\mathbf{1}_{i=j}$ denotes a binary variable that is 1 if and only if $i = j$ and 0 otherwise. The terms $(1|i)$, $(1|j)$ are random intercepts for the participants i, j. We obtained a (fixed) intercept of $\alpha = 11.48$ and a fixed effect of $\beta = -2.44$ distance units with a standard deviation $\sigma = 0.13$, resulting in the 95% confidence interval $[-2.69, -2.19]$, indicating a very significant and strong effect. We note that $R^2_{\text{conditional}} = 0.709$ whereas $R^2_{\text{marginal}} = 0.011$, suggesting that while the effect itself is strong, it only explains a small part of the total variation in distance. This is to be expected since reasoning patterns differ greatly between individuals.

Discussion

We developed a method to define psychological distances between individual reasoning patterns, incorporating findings from heuristic theories of reasoning. We employed a notion from metric geometry, the so-called p-energy (Definition 1) as a measure of response variability. The framework supports fine-grained analyses of response variability at the level of individual reasoner–task combinations. The clustering analysis suggests that reasoners differ systematically both in response strategies and variability.

Our findings from the mixed-effects modeling are twofold: First, we showed that high energy predicts an increase in correctness on retest, but only for initially incorrect syllogisms. For initially correct syllogisms, a decrease in energy is predicted. Energy can thus be seen as a measure of psychological consistency in response patterns: If reasoners are stably incorrect, an improvement is less likely than when they are unstably incorrect. Second, we demonstrated that individual differences in reasoning patterns are stable: reasoners were closer to their own retest patterns than other reasoner's retest patterns. This effect seems to be particularily robust.

We note that our framework abstracts away content effects such as belief bias (Chater & Oaksford, 1999; Khemlani & Johnson-Laird, 2012). However, the task formulations in the dataset under consideration here are specifically designed to minimize content effects. Thus, it limits the impact on the present results. More broadly, extra-logical factors such as content effects and belief bias represent additional sources of response variability that the present framework does not capture; extending the task encoding to incorporate them would be a natural refinement.

Regarding generalizability, some aspects of our construction depend on the domain of syllogistic reasoning and the psychological theories adopted; alternative distance functions could be derived from mental model or logic-based approaches. However, the mathematical framework itself is domain- and theory-independent: defining distances on task and response spaces directly induces a distance and energy functional on response patterns. Recent Bayesian approaches to syllogistic reasoning predict probability distributions over responses (Tessler et al., 2022). Extending our framework from deterministic response functions to non-deterministic response distributions would therefore be a natural next step and would accommodate experimental designs allowing multiple responses, which are common in valid syllogisms.

Acknowledgements

The authors thank the German Research Foundation (DFG) for financial support via grants awarded to MR within the projects 283135041 and 427257555. CB was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) –- Project number 283135041. RL was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) –- Project numbers 283135041; 427257555.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373–1396. https://doi.org/10.1162/089976603321780317

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. *Proceedings of the 17th International Conference on Cognitive Modeling*, 5–10.

Brand, D., Riesterer, N., & Ragni, M. (2020). Extending TransSet: An individualized model for human syllogistic reasoning. *Proceedings of the 18th International Conference on Cognitive Modeling*, 17–22.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*(2), 191–258. https://doi.org/https://doi.org/10.1006/cogp.1998.0696

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

Hartig, F. (2024). *Dharma: Residual diagnostics for hierarchical (multi-level / mixed) regression models* [R package version 0.4.7]. https://CRAN.R-project.org/package=DHARMa

Jost, J. (1994). Equilibrium maps between metric spaces. *Calculus of Variations and Partial Differential Equations*, *2*(2), 173–204. https://doi.org/10.1007/BF01191341

Kahneman, D., Sibony, O., & Sunstein, C. R. (2022). *Noise: A flaw in human judgment*. Little, Brown Spark.

Keller, M., Lenz, D., & Wojciechowski, R. K. (2021). *Graphs and discrete Dirichlet spaces* (Vol. 358). Springer. https://doi.org/10.1007/978-3-030-81459-5

Khemlani, S. S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457. https://doi.org/10.1037/a0026841

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of statistical software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Ragni, M., Khemlani, S., & Johnson-Laird, P. N. (2014). The evaluation of the consistency of quantified assertions. *Memory & Cognition*, *42*(1), 53–66. https://doi.org/10.3758/s13421-013-0349-y

Ragni, M., Riesterer, N., Khemlani, S., & Johnson-Laird, P. (2018). Individuals become more logical without feedback. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *40*. https://escholarship.org/uc/item/3kx0p15x

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. https://doi.org/10.1037/0096-3445.127.2.161

Tessler, M. H., Tenenbaum, J. B., & Goodman, N. D. (2022). Logic, probability, and pragmatics in syllogistic reasoning. *Topics in Cognitive Science*, *14*(3), 574–601. https://doi.org/10.1111/tops.12593

Wilkinson, G., & Rogers, C. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *22*(3), 392–399. https://doi.org/10.2307/2346786

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451–460. https://doi.org/10.1037/h0060520