Max Planck Institute for Human Development
Cognitive Neuroscience at the Center for Adaptive Rationality

Brandenburg Medical School Theodor Fontane
Faculty of Psychology

ON THE EVOLUTION OF MENTAL REPRESENTATIONS UNDERPINNING TRANSITIVE INFERENCE

A thesis submitted in partial fulfillment of the requirements of the Bachelor of Science in Psychology

Student: Clara Wicharz

Supervisors: Juan Linde-Domingo, Patrick Khader

September 2020

TABLE OF CONTENT

TABLE OF FIGURES

ABSTRACT

If being provided with the relational information A > B and B > C, humans and other non-human species are able to extract the implicit relationship between these bits and make a decision on the novel comparison A vs. C. This logical operation, known as transitive inference (TI), is a form of deductive reasoning that is assumed to underly more complex cognitive abilities like mathematical or scientific thinking. Current evidence suggests that at the moment of having overlearned premises (i.e., A > B or B > C), subjects have already elucidated implicit relationships between them and encoded elements according to their underlying order into a coherent semantic representation. What remains unknown is how these representations evolve over time. Furthermore, most studies have investigated TI and its underlying information organization on the basis of deterministic feedback. By redesigning the classic TI-paradigm, we assessed reaction time (RT) data shedding light on both the evolution of mental representations over time as well as the underlying organization of probabilistic information. We hypothesized a transitioning from using episodic strategies to solve TI towards applying semantic ones. Moreover, we adapted to the temporary closing of our laboratory due to the outbreak of COVID-19 by shifting to online-testing. Our findings suggest the reliability of online-testing even for demanding, RT-dependent tasks such as the underlying TI-paradigm. Furthermore, accuracy and RT results support the hypothesis that subjects are able to discover the underlying structure of elements and encode them in a semantic representation when being provided with probabilistic feedback. Inconsistent with our hypothesis about the dynamic of mental representations, RT did not support an episodic memory strategy to store information serving as premises in early learning phases. Instead, evidence suggests an early encoding of premises into a semantic representation that gains resolution over time.

INTRODUCTION

Reasoning is fundamental to human intelligence (Evans, Newstead & Byrne, 1993) and transitive inference (TI) is part of our ability to reason. Both abilities allow us to create an understanding of the world around us without having experienced every single aspect of it. If you are a professional speedskater and short track skating is on the agenda tomorrow, your trainer does not need to explicitly tell you to bring short track skates or a spandex skin suit, you can directly infer it. We connect the dots with different forms of inferences: inductive, deductive and statistical inferences. Inductive inferences add information, e.g. if it has been snowing for the past few weeks, we could infer that it will also snow today. Deductive inferences extract information that was already implicitly existing in the premises, e.g. we can deductively infer, that since it has been snowing for the past few weeks, it also snowed yesterday. Statistical inferences distinguish themselves from deductive ones through using probabilistic instead of deterministic premises (Evans, Newstead & Byrne, 1993).

Commonly TI is classified as one form of deductive inferring (Lazareva, 2012) that concerns entities that are placed upon a single scalar dimension (Evans, Newstead & Byrne, 1993). If A > B and B > C, humans and multiple animal species (Luo & Zhao, 2018), can put the pieces together and see that A > B > C and transitively infer that A > C. The “ > “ symbol can stand for “faster”, “prettier” or a relation in any other dimension that allows comparison (Acuna et al., 2002; Gazes et al., 2014; Jensen et al., 2015; Picklesimer et al., 2019). However, if premises become probabilistic, TI needs to be classified as a form of statical inference.

In real life, we don’t always acquire all the premises to transitively infer all at once. Instead we often sample relational information piecemeal to create our own broader understanding of the world. If we are speed skating enthusiasts, we might have witnessed, that Japanese contestant Nao Kodaira won against her Korean competitor Lee Sang-hwa in the Winter Olympics 2018, but was outpaced by Austrian speedskater Vanessa Herzog in the Single Distances Speed Skating Championships 2019. Ignoring all other factors influencing the performance in a speedskating contest, one could infer from these two bits of relational information, that Vanessa Herzog was a better speedskater than Lee Sang-hwa.

As simple as it is to solve the TI task like Nao Kodaira vs. Vanessa Herzog based on the previously given information, as powerful the underlying ability to combine relational information gathered from isolated episodes like speedskating contests might be. Transitive inference appears to be an underlying ability that supports more complex, abstract skills, such as mathematical, scientific and probabilistic thinking (Tversky & Kahneman, 1974). It not only enables us to understand the world around us, but also navigate through it (Evans, Newstead & Byrne, 1993).

Cognitive processes underlying TI: hypothetical information organization and extraction

In the venture of shedding light on the phenomenon of TI, one can differentiate its underpinning processes and subsequent research questions in two part:

1. Information organization and encoding: How is relational information (A > B, B > C) stored in memory?
2. Information extraction: How is implicit information retrieved from stored information to perform TI (A > C)?

Despite using different names for it, researchers proposed and tested two different representation models that explain how we organize and extract information from mental representations to perform TI:

Episodic-memory models. According to these models, we encode and retrieve information as single episodes and connect them episode by episode via "mental logic" to perform TI (Acuna et al., 2002). According to Braine and O'Brien (1998), mental logic is a set of inference schemas that are applied to concrete problems via a mental reasoning program. Kumaran et al. (2012) have described the phenomenon of storing episodes and connecting them on the fly in their theory REMERGE (Recurrence with Episodic Memory Results in Generalization), Brunamonti et al. (2016) described this scenario as "representing premises". According to the episodic-memory models, we perform TI by first retrieving all episodes that store information on adjacent items within a hierarchy, that implicitly connects the non-adjacent items. If we are asked, whether A > D, we need to first retrieve the episodes A > B, B > C and C > D. Secondly, we need to combine these episodes through their overlapping items. The bigger the distance within the hierarchy, the more episodes need to be retrieved and combined. The increasing number of computational steps to solve a TI problem should be reflected in increasing reaction times.

Semantic-memory models assume that we create a coherent representation in the sense of a mental model (e.g. unidimensional hierarchies can be mentally represented as a mental line). It is built according to the explicit relationships within and implicit relationships between the single bits of relational information. The representation can be manipulated on demand to derive implicit information in order to perform TI (Acuna et al., 2002; Johnson-Laird, 1983). Acuna et al. (2012) label this kind of representation "unified representation". Howard et al. are describing this scenario in their Temporal Context Model (TCM), which is presented by Kumaran (2012) as REMERGE's theoretical counterpart. Park et al. (2020) named a possibly coherent information representation "mental map", Brunamonti et al. (2016) spoke of a "mental schema" that is "scanned" for implicit information extraction. The assumed information deriving can be understood in the sense of Moyer and Landauer's (1967) specified information extraction in their study on judgements of numerical inequality. When being asked to compare pairs of numerals, subjects' RT decreased the bigger the difference between the numerals, revealing a so-called symbolic distance effect (SDE). It is suggested that subjects translate numerals into magnitudes and respective symbolic into analogue spaces. When being asked to judge on the inequality of numerals, subjects solve the comparison in the analogue space, in which high magnitude inequalities become more apparent, leading to less cognitive effort underlying the decision making. Semantic-memory models assume this mechanism also for the representation of symbolic spaces in general. Subsequently, when being asked to decide whether A > D, subjects just recall the analogue representations for A and D that are stored in the coherent semantic representation and compare them. The inequality of symbols should be more apparent and easier and faster to judge upon, the bigger the distance in symbolic space.

Previous Research: Evidence supporting semantic-memory models

TI is classically investigated with the following standard TI paradigm: A set of stimuli is employed in the experiment. Each stimulus gets assigned to a point within a one-dimensional space. The bespoken speedskaters could be mapped in a symbolic space representing speed according to the given relational information about the winners.

Acuna et al. (2002), pioneers in the study of mental chronometry of TI, chose geometrical figures as stimuli and randomly assigned values representing points in a physical space (further left - further right) to them. In Park's (2020) more recent and complex approach to TI, drawn portraits of people were employed and randomly assigned to characteristic attributes in two one-dimensional spaces: popularity and competence (unpopular-popular and incompetent-competent).

Being naïve to the arbitrary order of items in a classical TI paradigm, pairs of items were presented to participants. Subjects were instructed to relatively compare the items with regard to their position in the respective one-dimensional space ("Is a geometrical figure positioned further to the right on the underlying hierarchy?" or "Is a person more popular than the other?"). Participants only received feedback on their response accuracy on pairs of items, whose ranks in the one-dimensional space were adjacent like A > B or B > C. These items will further be referred to as "neighbours". The reinforced neighbour-comparisons would then serve as premises, when being asked to compare items, whose values are not adjacent ("non-neighbours") like A vs. C. Since participants have never received feedback on non-neighbours, they could only solve the non-neighbour comparisons through integrating the conditioned relational information and perform TI. Classically, training phases to condition neighbours and testing phases to perform TI on non-neighbours are separated in time (Acuna et al., 2002; Brunamonti et al., 2016; Jensen et al., 2019; Luo & Zhao, 2018; Park et al., 2020; Picklesimer et al., 2019). Commonly, participants first learned neighbours. Only if they met a certain accuracy criterion indicating the overlearning of neighbours, they were tested for TI in a segregated test-phase (Acuna et al., 2002).

In multiple studies employing the previously described design RT followed the predictions of semantic-memory models, immediately showing a symbolic distance effect when first testing for TI (Acuna et al., 2002; Brunamonti et al., 2016; Gazes et al., 2014; Jensen et al., 2019; Picklesimer et al., 2019). This indicates a coherent semantic representation had already begun to establish when just learning neighbours, consequently acquiring the premises for later performed TI.

Present study

We want to expand the acquired static perspective on information organization with a dynamic one: how do semantic representations evolve over time? Based on evidence, that semantic knowledge is progressively acquired via repeated encoding of episodes that share elements (Buzsáki & Moser, 2013) we hypothesize, that the coherent semantic representation supporting TI is evolving from first episodically encoded premises. According to this first and central hypothesis, RT in early learning phases should reveal a reversed SDE that transits into the classic SDE over time. If RT patterns should support this hypothesis, this would raise evidence for a time-dependent validity of two memory models, that until now have been assumed to be mutually exclusive. To test this hypothesis, the previously separated phases of learning premises and testing TI were intermixed. This alteration made it possible to keep track of the form of underlying representation(s) from the very beginning.

The second tested hypothesis is regarding the nature of informative feedback and its implications for the mental representations supporting TI. In the classic TI-paradigm informative feedback on neighbours is of deterministic nature. Namely, every bit of relational information is reflecting the true underlying ranks of items in the symbolic space. Illustrating this with the previous example, this would mean that since Vanessa Herzog had won against her competitor Nao Kodaira in 2019, Ms. Herzog would always remain the better speed skater. Looking at the results of the Speed Skating World Cup 2020 and seeing, that Kodaira was 0.43 seconds faster on the short skate than Herzog, the probabilistic nature of performance reveals itself. Many factors can influence performance, whether it regards speedskating, the economic success of a business venture or other areas in life, which makes it more difficult to gain an understanding of and navigate through our environment. Mapping this general challenge in the TI-paradigm, we provided probabilistic instead of deterministic feedback. In only 80 % of the trials the feedback was consistent with the underlying item hierarchy (e.g. A > B) and inversed in the remaining 20 % of the trials (e.g. A < B). Several studies employing reversal learning tasks showed the ability of humans and various non-human species to update previously acquired stimulus-response associations and generate adaptive behavior (Ghahremani et al., 2010; Tsuchida et al., 2010). On this basis we further hypothesize, subjects would still be able to learn the underlying order or items and encode it in a semantic representation despite receiving partially contradicting information across episodes. Consequently, participants should reach an above-chance accuracy over the experiment and still reveal a SDE in their RT in late learning phases.

Due to the outbreak of COVID-19 and the temporal closing of our laboratory, we needed to adapt the data collection. Given the technological progress in the field of online-based experiments regarding precision of stimuli-presentation and response-recording as well as in the feasibility to generate them (Bridges et al., 2020; Brooks, 2020; Grootswagers, 2020), we decided to collect data online. Therefore, we reprogrammed the experiment with PsychoPy 3, recruited participants via Prolific Academic and collected data with the hosting tool Pavlovia.

Subsequently, we did not only investigate the evolution of mental representations underlying TI in a more naturalistic environment, but also the methodological approach of collecting RT-data online by replicating previous evidence for semantic representations supporting TI. According to our third hypothesis, online testing allows the assessment of reliable RT data even in a demanding task such as the previously delineated TI-paradigm. Consequently, the SDE should show in late learning phases, thus replicating the previous findings of studies with laboratory-based data collection. A confirmation of the third hypothesis would further demonstrate the possibilities of online-testing, an alternative option for data collection, which has gained importance since the outbreak of COVID-19.

METHOD

Participants

A total of 80 volunteers (23 female; mean age 24.73 ± 5.40 years) took part in the behavioural experiment. The datasets from 8 participants could not be saved due to either technical problems on the participants' end or fraudulent submissions without having conducted the experiment, remaining 72 participants with associated data sets (22 females, mean age 24.85 ± 5.27 years). Moreover, 26 participants were excluded in the final analysis due to an accuracy lower than 60 % in the last block of the experiment.

Due to the custom pre-screening in Prolific Academic only participants were recruited, who met the criteria of being between 18 – 35 years old, native or highly fluent English speakers, having normal or corrected-to-normal vision, normal colour vision, and a Prolific Academic-internal approval rate of 95 – 100 %. We received informed consent from all participants via answering a survey hosted on Qualtrics. Other than the information of contributing to a research project investigating value-based decision-making, participants were naïve as to the goals of the experiments. Participants were financially compensated with a fixed amount of £4.87. Additionally, they could receive a bonus payment of £1.46, depending on their accuracy in the experiment.

The Ethics Committee of the Max Planck Institute for Human Development approved the project RETRIEVER T.I., of which this study is part of.

Stimuli

In total, 20 pictures of unique everyday objects and common animals were used in the main experiment. A further 8 were used in the practice task. All pictures were selected from the BOSS database (Brodeur et al., 2010). All original images were pictures in colour on a white background. Since this experiment is part of a broader study, that includes EEG-data collection to further examine neural signatures of episodic memory processes during value-based decision-making, the pictures chosen met certain EEG-relevant criteria regarding the categories they belonged to. These underlying categories were not intended to play a role in the behavioral studies.

Items were always presented in pairs in either the first position on the left side ((-250,0) pixel) or the second position on the right side ((250,0) pixel). Each item was presented with a rescaled size of 500 Å~ 500 pixels.

To counterbalance experimental conditions and avoid overly strong clustering in the presented stimuli, stimuli-value associations or sequences of presenting stimuli between participants, every participant received 8 items, that were uniquely and pseudo-randomly drawn from the pool of 20 items and pseudo-randomly assigned to values from 1 – 8.

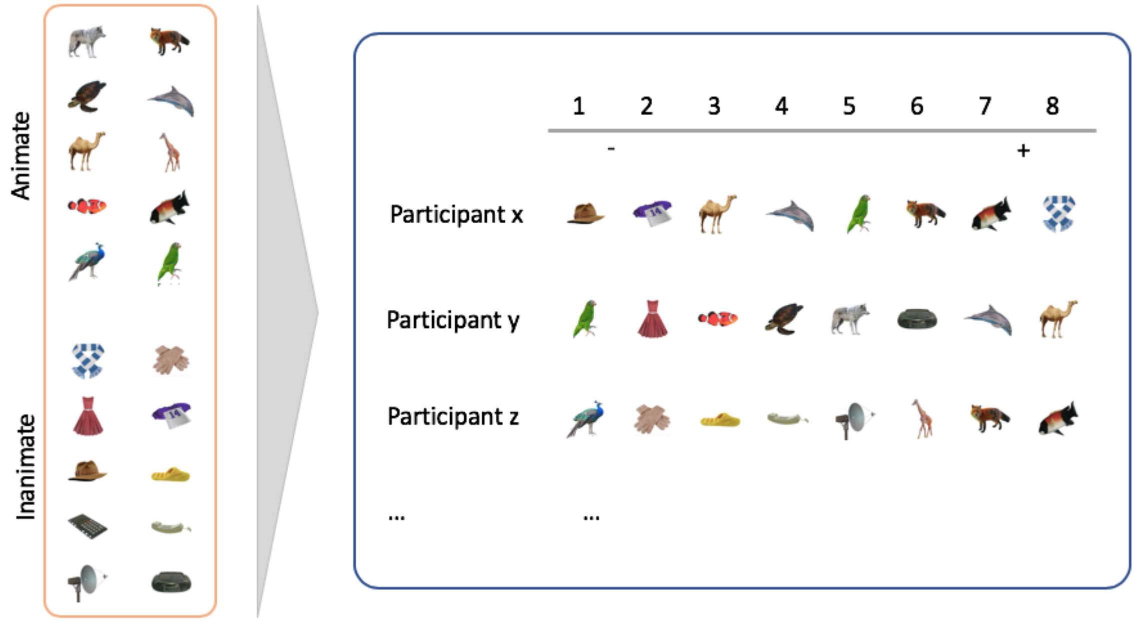

Figure 1: Stimuli and stimuli sequences of the behavioral and EEG-paradigm. Each participant received a unique sequence consisting of a set of 8 different stimuli. For each participant these stimuli were pseudorandomly drawn from a pool of 20 items, half of them representing animate, half of them inanimate objects. The drawn items were pseudorandomly associated with values from 1-8.

Moreover, stimuli were presented in a pseudo-random sequence, that was restricted to certain criteria: since participants were only provided with informative feedback on pairs of neighbours, neighour-comparisons were presented 2.5 as often as non-neighbour-comparisons to accelerate the learning and to ensure that enough participants would fully discover the underlying value hierarchy by the end of the experiment. This ratio of neigbour- and non-neighbour-pairs was maintained in every block.

For the intermitting attention checks, geometrical figures (circle, square, star and triangle) as well as their labels (letter height 0.05 pixel) were used as stimuli. Per attention check trial one geometrical figure was presented at the centre of the screen with a rescaled size of 200 Å~ 200 pixel and the labels of all four possible geometrical figures where written above (0, 0.25) , beyond (0, -0.25), to the left (-0.3, 0) and to the right (0.3, 0) of the figure. The geometric figures as well as the option to assign names to it varied between attention check trials.

Procedure

Instructions. First, participants were shown the instructions explaining the main task. They were told to see pairs of items on the screen, which would have specific values associated with them. It was emphasized, that these values had nothing to do with their price or value in real life. Participants were instructed to choose what they believe is the more valuable item in each trial and indicate their decision by pressing the respective arrow button as fast as possible while maintaining accuracy. Participants were not instructed how to press the response keys since comparisons containing the same items were presented equally often in both position arrangements (A-B and B-A) and eventually their results would be combined, disregarding elements' position on the screen and the assigned response keys. Any systematic difference in responses for items on the left and right side would be counterbalanced with respect to the central response question of whether and how fast an item with a certain associated value would be preferred over another one.

Participants were prepared for receiving both informative feedback on their decision ("WIN!"/"LOSS!") or uninformative feedback ("THANK YOU!"). Moreover, the probabilistic character of the feedback was explained in the sense, that an item's performance on a given trial could somewhat vary, but that the general value of each item would not change throughout the experiment.

Beyond explaining the main task, the instructions also informed about the attention checks and the importance of both reaction times and accuracy in the main task. It was explained how the latter influenced the chances to win the bonus. It was revealed that the bonus would be determined by the accuracy in 10 randomly chosen trials. If the accuracy was 60 % or higher in these trials, participants would receive the bonus. To compute the bonus a random trial drawing was chosen over calculating a general accuracy to create an incentive to perform well on every single trial.

Furthermore, the importance of passing the attention checks was stressed, since it was necessary to continue the experiment and receive the completion code at the end to get financially compensated for participating in the study.

Participants had the opportunity to ask questions to the experimenter via Prolific Academic's messaging service.

Practice. The practice part consisted of two blocks with a length of 8 trials each. Each block included the main task and an attention check. In the practice stimuli were presented that were not employed in the real task. Participants could familiarize themselves with the task type without gaining any insight or being biased regarding the underlying value hierarchy of the real task. Trials in the attention check were shortened to 4 instead of 8 in the real task. The early upcoming of attention checks served as to both familiarize the participant with the task type and check the attention. A participant could only continue the experiment with an average accuracy of over 50 % in the attention checks. Otherwise the experiment would get aborted, thus not providing the participant with a completion code to claim financial compensation for the participation.

The real task was separated in eight blocks of 77 trials each. All trials started with a jittered fixation cross (500–1500 ms) followed by the presentation of a pair of two items (2500 ms). Between the two stimuli the fixation cross remained until the participant responded. Subsequently, informative feedback on neighbours or respectively uninformative feedback on non-neighbours was given (500 ms).

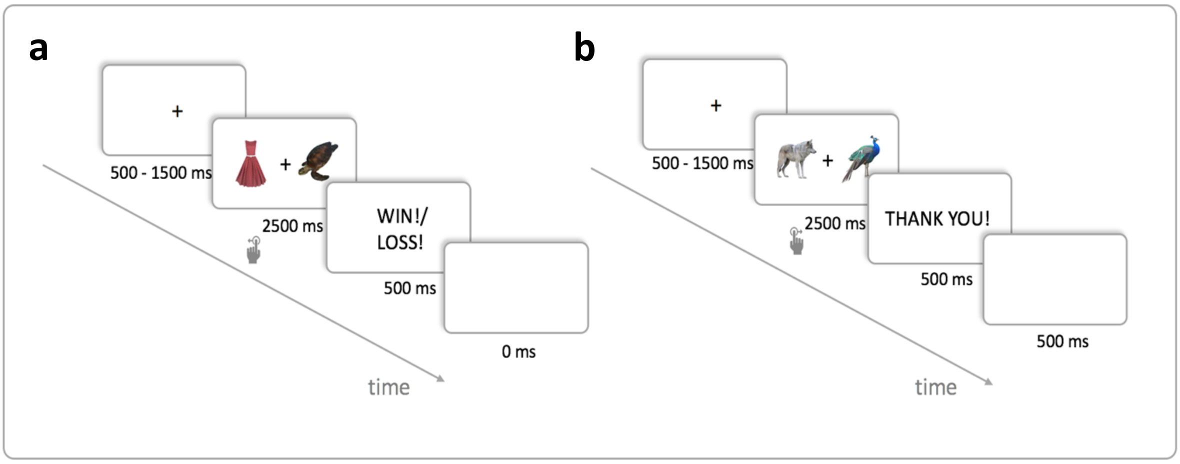

Figure 2: Experimental design. Participants were prompted to decide as fast and accurate as possible, which of the upcoming objects was more valuable. They were instructed to indicate their decision by pressing the arrow key pointing to the side of the chosen item. (a) Subsequently, they received informative feedback for neighbours and (b) uninformative feedback for non-neighbours.

After four blocks another attention check was being conducted, which served as a gateway to continue the experiment. At the end of the experiment the bonus was computed and the result communicated to the participant. Subsequently, the participant had the opportunity to give feedback. Finally, he or she would get redirected to Prolific Academic and receive the completion code.

The experiment took approximately 40 min, depending on the length of the self-paced breaks.

Data collection

COVID-19 globally broke out briefly before the planned start of the lab-based data collection. After having programmed the experiment using Psychophysics Toolbox Version 353 running under MATLAB 2014b (MathWorks), we needed to adapt to the new situation, that included a possibly longer lab-closing.

We searched for a feasible way to acquire reliable RT data online. A study comparing platforms, browsers, and participant's devices (Anwyl-Irvine, Dalmaijer & Hodges, 2020) tested for reasonable display duration and manual responses via modern web-platforms. Other laboratories published reports about their own positive experience with online testing as well as guides, giving high-level understanding regarding the general infrastructure with online data collection (Brooks, 2020; Grootswagers, 2020). Encouraged by the alleged feasibility of online testing, we took a closer look at specific experiment generators. Bridges et al.'s compared laboratory- and online-based ones. It was shown that the common generators for lab-based experiments Psychtoolbox, PsychoPy, Presentation and E-Prime delivered the highest level of precision with a mean of under 1 millisecond across the visual, audio and response measures. PsychoPy as one of the generators for online-based experiments however came close achieving a precision of under 3.5 ms on a number of browser configurations, thus being ranked as one of the top performers in online-based experiment generating (Bridges et al., 2020).

After reviewing further methodological publications as well as an institute-internal investigation we decided for the following infrastructure to support our online-data collection:

1. **Recruiting participants via Prolific Academic.** For three reasons Prolific Academic was chosen over other well-established players in the field of crowdworking platforms. First, its data quality is higher than on other popular crowdworking platforms such as CrowdFlower and in the study of Peer et al. (2017) even exceeds the quality of data collected from the classic university subject pool. Secondly, the subject pool is more heterogenous than the one of the generally dominating player Amazon's Mechanical Turk. Thirdly, in comparison to Amazon's Mechanical Turk Prolific Academic has established a fairer treatment of

crowd workers regarding a payment standard and strict rules for participant treatment (Palan & Schitter, 2018).

2. **Building a browser-compatible experiment via the Framework PsychoPy v3.0**. Beyond its high-precision timing (Bridges et al., 2020) the experiment generator was appraised to offer intuitively usable built-in functions over the graphical interface "Builder" and to allow for customizing with code in Python (Peirce, 2007) Retrospectively, we find that the built-in functions allow for an easy implementation of very basic experimental setups. The more complex the paradigm becomes, the less these built-in tools support this way of implementation. Additionally, customizing code becomes increasing difficult due to the limited translatability of Python libraries to JavaScript, the language the program is eventually compiled into to run online.
3. **Hosting the experiment on Pavlovia**. After programming the experiment with PsychoPy, it was reasonable to also use the inexpensive hosting platform Pavlovia (£0.2 per ran participant), since it was created and recommended by PsychoPy and users could directly upload their experiment from the builder view to debug, pilot and finally run it (*Using Pavlovia.org*, n.d.).
4. **Generating an anonymous participant's number via Qualtrics**. Even though PsychoPy offers easy implementation solutions for standard experimental aspects, keeping track of the participant's number is not one of them. Since a participant's number was crucial to counterbalance experimental conditions, we used the Qualtrics function "Quotas" (*Quotas*, n.d.), that incremented a number by one every time someone answered the last question of a small survey regarding demographic information and redirected the participant with the current number to the experiment, that was hosted on Pavlovia.org.

The experiment was run in three batches. The first batch (n_1=20) served as a pilot study and participants were informed about that. After not encountering any problems, two more batches (n_2=20, n_3=40) were collected with an unaltered experimental setup.

RESULTS

RT and accuracy for non-neighbour comparisons revealed three main effects. First, independent of symbolic distance accuracy increased while reaction times decreased over blocks. Secondly, a significant SDE was reflected in RT from the second paired block on, in accuracy from the first paired block on. Both measures revealed a SDE across blocks. Thirdly, the data revealed an early emergence of the SDE and a gradual increase in strength of the SDE over blocks.

Exclusion Criteria

For all of the following analyses, only data acquired from participants who had successfully learned the underlying value hierarchy was included (> 60 % performance criterion in the last block). Within these data sets trials with neighbour comparisons were excluded, since they could be solved via previous reinforcement learning. Moreover, trials without response or RT < 100 ms possibly reflecting accidental responses were excluded. After applying these exclusions, the data of 46 subjects was analysed.

Dynamic of RT and Accuracy

On average participants reached a general accuracy of 70.52 % (SD: 12,46 %) and responded with a mean reaction time of 976,3 ms (SD: 356,9 ms) on non-neighbour comparisons across blocks. To investigate the course of reaction times and accuracy over time, a combination of linear regression fitting for reaction times and logistic regression fitting for accuracy as well as subsequent t-tests were combined. On a participant-level the respective regression equation was fitted to predict reaction times or accuracy with the block-index (1-8). The second estimates for each regression equation were then tested against zero with a t-test. Results showed, that reaction times significantly decreased over blocks (T-test, $t = -6.0891$, $p \leq 0.01$), while accuracy significantly increased (T-test, $t = 5.4345$, $p \leq 0.01$), indicating that participants continuously learned throughout the experiment (Figure 3: Behavioral accuracy and RT results).

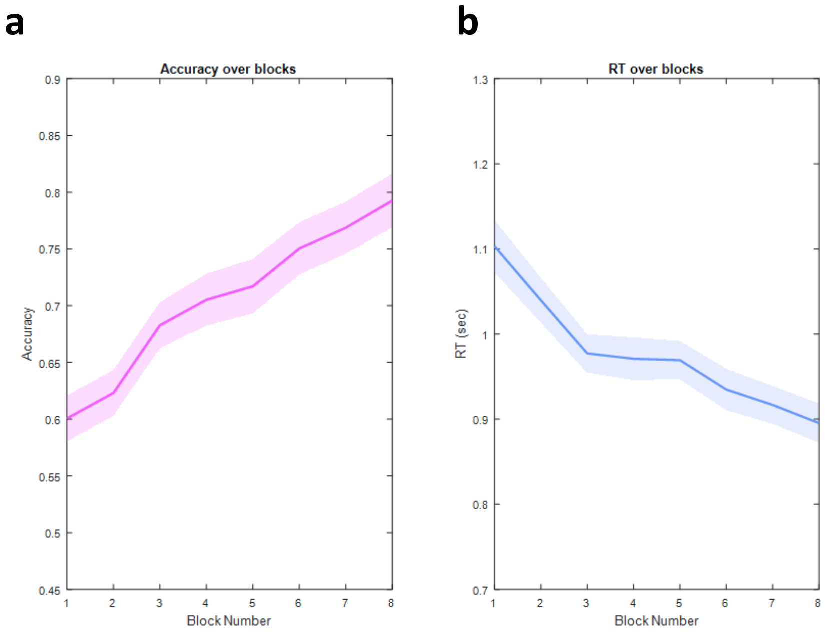

Figure 3: Behavioral accuracy and RT results. Line plots representing mean and standard error of (a) accuracy (y-achsis) and (b) RT (y-achsis) on non-neighbour comparisons across participants per block (x-achsis). We found that both measures were significantly predicted by the block-index (P < 0.01).

Symbolic Distance Effect: Static Perspective

Semantic-memory models assume a coherent representation, that elicits the implicit relationships between represented elements. According to this hypothesis, symbolic distances are mapped in the representational space, thus making items that are further apart more distinct and easier to compare. According to this hypothesis, the cognitive effort should decrease with increasing symbolic distance and should be reflected in increasing accuracy and decreasing RT.

To generally test an appearance of the SDE, further regression fittings were conducted on a participant-level and combined with two t-tests. First, for each participant a linear regression equation was fitted to reaction times and a logistic regression equation was fitted to accuracy data including all blocks, to predict the respective measure with symbolic distance. The second estimates of the generated regression equations for each participant and each measurement were tested against zero with a t-test. Both, the t-tests were significant revealing a significant SDE in both measures across blocks (RT: t-test, t= -5.7951, p > 0.01; accuracy: t-test, t= 8.2667, p < 0.01). The occurrence of the SDE and participants above-chance performance of non-neighbour comparisons confirms two of our working hypotheses: Firstly, the hypothesis, that participants can discover items' underlying ranks and encode them in a semantic representation when being provided with probabilistic instead of deterministic feedback. Secondly, with our results we could replicate the findings of Acuna et al. and thus provide evidence supporting evidence for our hypothesis, that online testing allows for the assessment of reliable RT data even in a demanding task such as this TI-paradigm.

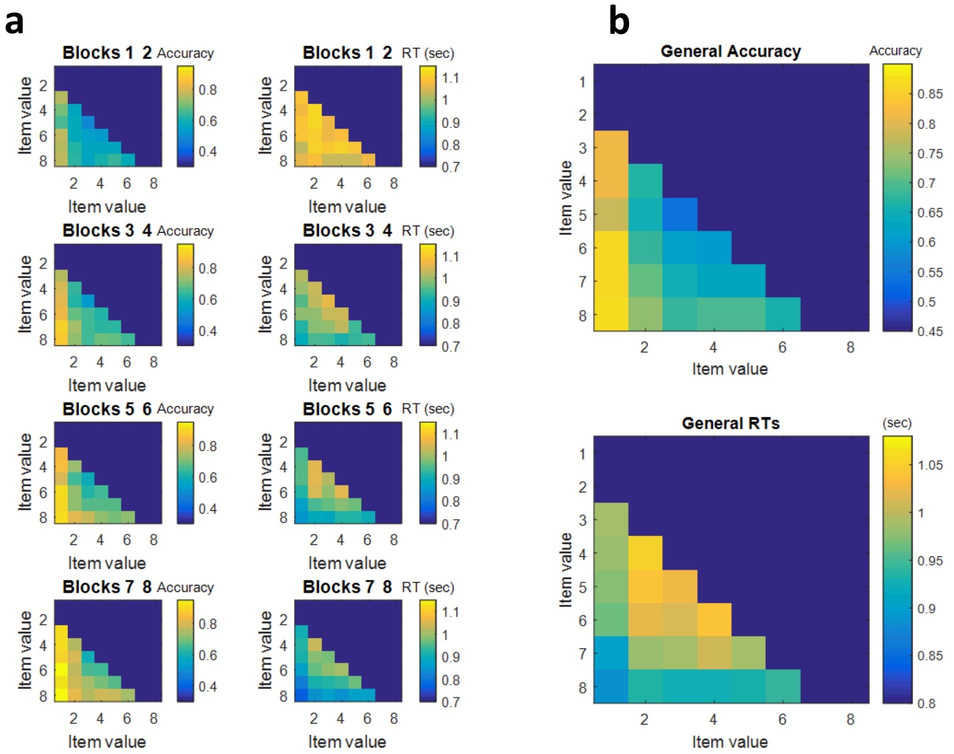

Figure 4: RT and accuracy matrices for each pair-combination. (a) Matrices show mean RT and accuracy on each pair-combination for each of four paired blocks and (b) across blocks. In each data matrix the x-achsis represents the value of one item in the respective pair, the y-achsis the value of the other item, independent of the items position in the pair (left or right). Accuracy in percent and RT in seconds are colour-coded with yellow representing high and blue representing low values in both measures. (a) With increasing index of the paired block an increasing accuracy and decreasing RT across pair-tpyes as well as an increasingly distinct SDE in both measures show. (b) Across blocks a dominating SDE shows in both measures.

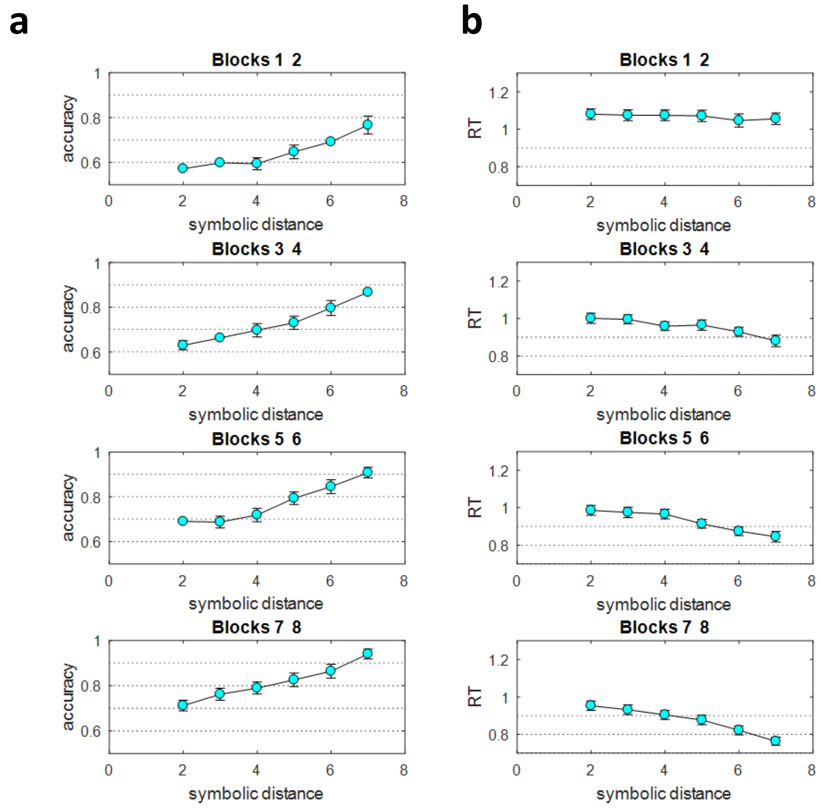

Figure 5: Behavioral results per symbolic distance over time. Boxplots show the mean and standard error of (a) accuracy and (b) RT for symbolic distances 2-7 per paired block. Participants' accuracy shows a significant SDE in each paired block (p < 0.05, while RT only reveal a significant SDE from the second paired block on (p < 0.05).

Symbolic Distance Effect: Dynamic Perspective

Due to the novel paradigm design with intermixed trials of learning the premises and testing for TI the data allowed to track the emergence and strength of the symbolic distance effect over time. To discover the first emergence of the SDE, the appearance of it was tested in paired blocks. The same analyses were conducted as for testing for the SDE across blocks, only on the basis of block-specific trials instead of all. Accuracy data showed a significant SDE in each paired block (paired block 1: t-test, t= 5.0303, p < 0.0125; paired block 2: t-test, t= 6.5331, p < 0.0125, paired block 3: t-test, t= 6.9941 p < 0.0125, paired block 4: t-test, t= 3.2105, p < 0.0125). RT however only revealed a significant SDE from the second block on (paired block 1: t-test, t= -1.2331, p > 0.0125; paired block 2: t-test, t= -3.9111, p < 0.0125; paired block 3: t-test, t= -5.0127, p < 0.0125; paired block 4: t-test, t= -6.8055, p < 0.0125).

Beyond testing for the existence of the SDE at different times, also the rate of change in its strength was tested. First, another regression fitting was conducted in the same manner as previously described, only differing by using single instead of paired blocks. Secondly, for each participant another regression equation was fitted, that predicted the previously generated second estimate reflecting the strength of the SDE in the respective block, with the block-index (1-8). The new second estimates, that reflected the rate of change in the SDE's strength were then tested against zero with a t-test. Both, RT and accuracy revealed a significant increase in strength of the SDE over blocks (RT: t-test, t = -4.3020, p ≤ 0.01; accuracy: t-test, t = 3.5448, p ≤ 0.01). These results are not consistent with our hypothesis, that humans first encode relational information episodically before merging it into a semantic representation in later learning phases.

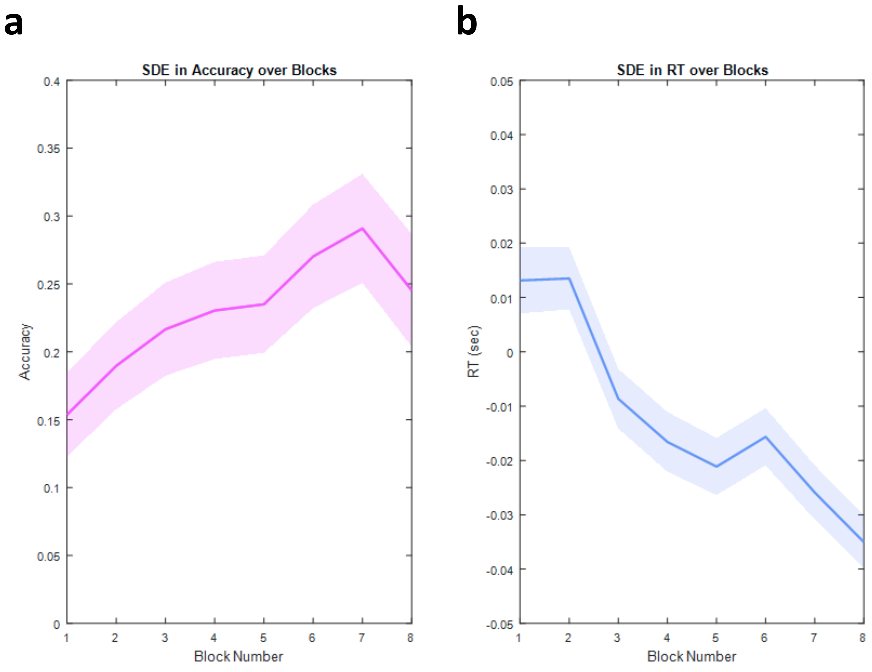

Figure 6: Evolution of the SDE. Line plots show the mean and standard error of the SDE reflected in (a) accuracy and (b) RT across participants per block. The strength of the SDE is increasing with the block-index (P < 0.01).

DISCUSSION

Three questions were guiding the underlying study: Firstly, how do mental representations underlying TI evolve over time? Secondly, are humans capable of performing TI even when being provided with premises of probabilistic nature and do they encode them in a semantic representation? Thirdly, does online testing allow for the assessment of reliable RT data even in a demanding task such as the delineated TI-paradigm? By redesigning the classic TI-paradigm and acquiring empirical data online, we could test for all three hypotheses. RT supported the reliability of data assessed online and the capability of humans to perform TI on the basis of probabilistic feedback and to integrate it into a semantic representation. However, the dynamic perspective on these mental representations revealed evidence that does not support our hypothesis of a transition from episodically to semantically encoded premises. Instead, RT suggest that relational information is soon embedded in its semantic context, namely the entire symbolic space, and semantic representations arise, that continuously gain resolution of symbolic space over time.

Lab-based vs. online-based data collection

Online-based data collection has become increasingly popular in social science (Palan & Schitter, 2018). It is less time consuming und less costly than in the laboratory-based data collection, participants have a more heterogeneous background than the classic university subject pool (Berinsky et al., 2012) and larger samples can be tested simultaneously (Anwyl-Irvine, Dalmaijer & Hodges, 2020). Particularly in the face of the current pandemic online-based data-collection has gained interest, since it temporarily became the only way of acquiring empirical data for a range of studies. However, the question arises how reliable data acquired online is. Several methodological studies tested for reasonable accuracy and precision of stimuli-presentation and response recording across different web-building platforms, browsers and operating systems (Anwyl-Irvine, Dalmaijer & Hodges, 2020; Bridges et al., 2020). These studies employed standard laboratory-based setups as well as robot actuators. What remains to be investigated is the quality of data acquired from diverging experimental setups and real participants. In our study we could replicate previous findings of classic RT-based TI-studies, that depend on precise stimulus presentation and response recording as well as a focused participant to perform well on the challenging TI-task. Thereby, we could confirm our hypothesis regarding the reliability of the assessed data and provide further evidence for the possibilities of online testing in praxis.

Deterministic vs. probabilistic information

Next to primarily investigating the evolution of TI's underlying mental representations we also wanted to test TI in a more naturalistic environment, namely when being provided with probabilistic instead of deterministic feedback. With our results we could replicated the findings of studies employing deterministic feedback: participants were able to perform TI and showed a SDE in RT across blocks and in the last paired block. These findings are consistent with results from reversal learning tasks. Ghahremani et al. (2010) measured performance and neurophysiological measures while participants were engaged in a feedback-driven discrimination task with altering contingencies. Behavioral results showed that participants could flexibly respond to the changing environment, a key-characteristic of adaptive behavior. fMRI data revealed, that other brain areas were additionally engaged when responding to newly altered contingencies, reflecting the additional computational effort to inhibit previously learned contingencies and consequent responses. The findings of Tsuchida et al. (2010) support these findings, providing further evidence on subjects with focal lesions affecting the respective regions supporting reversal learning. Despite this identified additional cognitive effort, participants' accuracy as well as the occurrence of the SDE in the underlying study indicate, that they had adaptively updated relational information and reintegrated it in the context of the semantic representation.

Representational Nature: Static Perspective

The general question of whether we organize relational information supporting TI in coherent semantic or in multiple isolated episodic representations leads way back in time. Current research investigating the nature of these representations is based on research regarded with mental representations underlying perceptual comparisons of physical entities. In the beginning of the last century scientists like Cattell (1902) or Henmon (1906) showed, that RT for comparing physical entities are inversely related to the difference of the entities regarding their physical attribute of interest like length or pitch.

Half a century later, Moyer and Landauer (1967) investigated, whether this phenomenon also occurred when comparing symbols instead of physical entities, and whether

RT would behave inversely to the distance between items in the symbolic instead of physical space. As symbolic space a number range from 1 to 9 was chosen. Numbers would always be presented in pairs and subjects were asked to choose as quickly as possible which number was higher. Moyer and Landauer hypothesized, subjects could solve comparisons either with a "direct memory look-up", that predicts a SDE in RT-patterns similar to the ones for perceptual comparisons or via counting symbolic spaces between symbols, which would lead to an inverse SDE. RT of the 10 tested participants revealed a SDE and could be fitted with equations similar to the classical psychophysical functions that delineated RT for comparisons in physical spaces. The result was consistent with the hypothesis of a "direct memory look-up", that numerals were translated to analogue magnitudes and numeral-comparisons would be solved within the reconstructed analogue space.

Similar to Moyer and Landauer's approach of investigating information organization underlying symbolic comparisons, Lovelace & Snodgrass (1971) assessed RT for comparing letters regarding their position in the alphabet. Analysing the relationship of RT and symbolic distance within the referential space of the alphabet again revealed the SDE. Potts (1972) started investigating symbolic comparisons in referential spaces that were not overlearned as the numerical space or the alphabet. For this he assigned different levels of friendliness to animals and delivered relational information about the animals regarding their property friendliness in written paragraphs, participants should read. Again, when asked to compare animals regarding their property friendliness, a SDE was reflected in RT.

Bryant and Trabasso (1971) merged comparisons in the physical and symbolic space. First, they associated differently long sticks with colours and then asked subjects to compare the colours regarding their symbolic meaning: the length of the sticks. The SDE also persisted when comparing symbols expressing a characteristic attribute of interest of a physical entity.

Subsequently, Moyer and Bayer (1976) picked up this approach of translating physical properties to symbolic attributes to directly compare comparisons in physical and symbolic space. For this, they employed four differently large circles as stimuli. To one group these circles were presented in pairs. The other group needed to learn an association for each circle with a CVC beforehand. Subsequently, instead of being asked to compare the circles the CVCs should be compared regarding their symbolic property: the size of the circle. In both task types, RT revealed a SDE leading to Moyer's notion, that memory-based comparisons are solved on the basis of a unified semantic representation, that can be "looked-up". This unified representation appears to map symbolic attributes similarly to physical attributes and allow "internal psychophysical judgements" to solve comparisons.

Based on the vast evidence for semantic representations underpinning comparisons in already overlearned symbolic spaces or spaces, whose structure was fully and explicitly revealed beforehand, Acuna et al. (2002) tapped into the form of mental representations of novel symbolic space, whose structure needed to be self-sufficiently reconstructed via trial and error learning of its premises. Almost a century after the first quoted study on mental representations supporting perceptual comparisons, Acuna also took into account further theoretical approaches. Since the structure of the symbolic space was not handed to participants directly, nor had they overlearned it like the numerical space or the alphabet before, subjects needed to discover it by gathering the single bits of relational information piecemeal in a learning phase and puzzle the pieces together to compare non-neighbour items in a separate test phase. Based on two theoretical streams Acuna formulated two hypotheses about the way to solve non-neighbour comparisons. Following the theoretical approach of Braine and O'Brien (1998) in which reasoning is assumed to be implemented via mentally running logical operations over premises, TI problems are solved through encoding single premises and connecting them along the underlying hierarchy on demand, leading to an inverse SDE in RT. Alternatively, based on Johnson-Lairdsons stream of thought, Acuna hypothesized a unified representation, in which the abstract structure, that implicitly exists between premises, is made explicit and serves as a mental model to flexibly and efficiently provide reference for non-neighbour comparisons. This hypothesis predicted a regular SDE, that subsequently also showed in RT. This result indicates that humas elucidate implicit relationships of episodically presented premises and reconstruct and encode them in a unified semantic representation reflecting the underlying symbolic space.

Acuna's TI paradigm established itself as the classic paradigm to first replicate evidence for the general representational nature underlying TI. The semantic nature of representations with its signature of the SDE was found in a multitude of studies in humans and non-human species (Bond et al., 2003; Lazareva, 2012; Paz-y-Miño C et al., 2004; Vasconcelos, 2008). Animal studies revealed, that successful TI performance does not need necessarily rely on explicit knowledge of the elements' order in symbolic space (Vasconcelos, 2008).

Park (2020) increased the level of complexity by adding another one-dimensional space to the classic referential space in a TI-paradigm. Again, premises were first trained in a learning phase. Here, items were not only mappable in one, but two one-dimenional spaces: popularity and competence. Separately the premises to infer the order of items in each dimension were overlearned. Subsequently, participants RT and were assessed when testing TI in both dimensions. Even though comparisons were always only regarding one dimension, RT inversely related to the Euclidian distance between stimuli, if they were mapped on a perfect cartesian coordinate-system, spanned by the dimensions competence and popularity. This form of a SDE indicates, that humans integrated bits of information into cross-dimensionally unified semantic representations.

These previous studies have all been regarded with a late form of information organization, since representation(s) underlying TI were only tracked after having overlearned the premises. In the underlying study we broaden this static perspective with a dynamic one. For this, we adopted the classic TI-paradigm employed by Acuna et al. or Brunamonti et al. and altered the first key-property: instead of separating the learning and testing-phase, we intermixed trials in which neighbours are learned through feedback and TI is tested on non-neighbours without being provided with feedback. This way it was possible to assess the form the mental representation right from the beginning, when participants were entire naïve as to underlying order of items in symbolic space. Consistent with previous findings, a SDE suggesting a semantic representation was found in RT of the last paired block, replicating the findings regarding the late form of information organization underlying TI.

Representational Nature: Dynamic Perspective

With the novel paradigm-property of intermixing trials in which neighbours are learned and non-neighbours are tested, for the first time, empirical evidence is provided to track the evolution leading up to the semantic representation from the beginning. With this tracking of representations insights into the formerly separated learning phase in classic TI-paradigms were gained. Based on evidence suggesting that semantic knowledge can be acquired via multiple episodes with common elements (Buzsáki & Moser, 2013) we hypothesized, that semantic representations evolve from episodic representations over time and challenged the assumption of a universal prevalence of semantic representations underlying TI.

Luo and Zhao (2018) had already addressed the topic of incidental, automatic learning of non-neighbour relationships when being exposed to neighbour-relationships. They employed a 1-back task, to present a sequence of coloured dots. Subsequently, a surprise two-alternative forced-choice test was conducted. Participants should decide between two sequences consisting of two dots which one would look more familiar. One pair consisted of dots in an order consistent with the order in the previously shown sequence underlying the 1-back task, one pair did not reflect the underlying order. The symbolic distance between dots in the sequence-consistent pair varied. For sequences reflecting the true order of items with a symbolic distance of up to 3 elements, participants tended to decide for the order-conform sequence, indicating that they had incidentally learned relationships beyond premises.

Jensen (2015b) was concerned with the algorithm, that could underly the integration of relational information into a coherent semantic representation. Therefore, he searched for a computational model matching behavioral RT data characterized by the SDE. While several reinforcement learning models relying on associative strength or reward prediction error failed to fit the data the algorithm "betasort" came closer. With this algorithm each item position in the symbolic space was represented along a unit span via a beta distribution, that determined the estimated position as well as the extent of uncertainty in this estimation. Each new bit of relational information updated not only the items in the respective rewarded comparison, but also all items, one can implicitly infer upon (2015b). In this model a unified semantic representation is assumed at every point in time. Before even receiving the first bit of relational information all item positions can be modeled with the same beta distribution: a flat line, that expresses an average position as well as full uncertainty in that estimate. With each bit of additional relational information, the peak and the slope of the distributions get updated resolving to a more and more differentiated representation of the items in symbolic space, that should lead to a stronger SDE in RT over time.

Complementary to Jensen's theoretical approach we collected empirical data to track mental representations over time. Inconsistent with our hypothesis of a transitioning from episodic to semantic encoding, RT in the first paired block did not show any significant positive or negative correlation with symbolic distance. First significant RT-patterns were found in the second paired block, that revealed a regular SDE, indicating an early emergence of a semantic representation. From then on, the regular SDE linearly increased in strength over blocks. The early emergence of the SDE as well as its increasing strength are consistent with the predictions of the betasort model: being naïve to the underlying positions of items in symbolic space, subjects assume an average position for each item and encode it in a coherent semantic representation. Each bit of additional information updates the estimates for the directly compared items as well as all items, that can be inferred upon. Thereby each bit of novel information gets embedded in the context of the entirely represented symbolic space. The more information a subject receives, the more item positions are differentiated, leading to an increasing resolution of the represented symbolic space.

Although current results are in line with the assumption that additional relational information gets embedded in the context of the entirely represented symbolic space, both episodic and semantic-memory models could account for the insignificant results in the first paired block. On the basis of the semantic-memory model a lacking SDE can be explained with item positions being mapped to close to each other in a semantic representation due to the sparse information updating item positions that are initially all assumed as average. However, it remains unrefuted, that initially relational information could have also been represented episodically, but too briefly to dominate RT patterns in the entire first paired block. To gain a higher resolution in time the same study could be conducted again on a larger sample. Complementary to this RT-based approach we are conducting a study employing the same task but representing elements of premises singly and record high-density electroencephalography (EEG) to examine neural signatures of episodic memory processes. This way theta-band oscillations are assessed, that are associated with episodic memory reactivation (Berens & Horner, 2017; Buzsáki & Moser, 2013). With this additional measure we hope to get further insights into the initial creation phase of the mental representations underlying TI.

How mental representations evolve over time and how we recreate the symbolic space surrounding us through them is an essentially epistemological question that ultimately leads to the beginning of thought. Beyond testing for the methodological approach of online testing, this study provides a dynamic perspective on the nature of TI's underlying representations that arise in more naturalistic environments. However, the initial moment of recreating symbolic space remains uncaptured, thus calling for further research on the beginning of mental representation creation.

BIBLIOGRAPHY

Acuna, B. D., Sanes, J. N., & Donoghue, J. P. (2002). Erratum: cognitive mechanism of transitive inference. *Experimental Brain Research*, *146*(1), 128. https://doi.org/10.1007/s00221-002-1226-2

Alex Anwyl-Irvine, Edwin S. Dalmaijer, Nick Hodges, J. K. E. (2020). *Online timing accuracy and precision: a comparison of platforms, browsers, and participant's devices*. https://doi.org/10.31234/osf.io/jfeca

Berens, S. C., & Horner, A. J. (2017). Theta rhythm: temporal glue for episodic memory. *Current Biology*, *27*(20), R1110–R1112. https://doi.org/10.1016/j.cub.2017.08.048

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Bond, A. B., Kamil, A. C., & Balda, R. P. (2003). Social complexity and transitive inference in corvids. *Animal Behaviour*, *65*(3), 479–487. https://doi.org/10.1006/anbe.2003.2101

Braine, M., & O'Brien, D. P. (1998). *Mental Logic*. Taylor & Francis. https://books.google.ne/books?id=IzB5AgAAQBAJ

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE*, *5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Brooks, H. (2020). *A brief guide to running an experiment using PsychoPy, Pavlovia, and Prolific*.

Brunamonti, E., Mione, V., di Bello, F., Pani, P., Genovesio, A., & Ferraina, S. (2016). Neuronal modulation in the prefrontal cortex in a transitive inference task: Evidence of neuronal correlates of mental schema management. *Journal of Neuroscience*, *36*(4), 1223–1236. https://doi.org/10.1523/JNEUROSCI.1473-15.2016

Bryant, P. E., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, *232*(5311), 456–458. https://doi.org/10.1038/232456a0

Buzsáki, G., & Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience*, *16*(2), 130–138.

https://doi.org/10.1038/nn.3304

Cattell, J. M. (1902). The time of perception as a measure of differences in intensity. *Festsch.*, *XIX*, 63–68.

Gazes, R. P., Lazareva, O. F., Bergene, C. N., & Hampton, R. R. (2014). Effects of spatial training on transitive inference performance in humans and rhesus monkeys. *Journal of Experimental Psychology. 40*(4), 477–489. https://doi.org/10.1037/xan0000038

Ghahremani, D. G., Monterosso, J., Jentsch, J. D., Bilder, R. M., & Poldrack, R. A. (2010). Neural components underlying behavioral flexibility in human reversal learning. *Cerebral Cortex*, *20*(8), 1843–1852. https://doi.org/10.1093/cercor/bhp247

Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods*, 1–9. https://doi.org/10.3758/s13428-020-01395-3

Henmon, V. A. C. (1906). The time of perception as a measure of differences in sensations. *The Journal of Philosophy, Psychology and Scientific Methods*, *Nr. 8*. https://books.google.de/books?id=D84uAAAAYAAJ

Jensen, G., Muñoz, F., Alkan, Y., Ferrera, V. P., & Terrace, H. S. (2015). Implicit value updating explains transitive inference performance: the betasort model. *PLoS Computational Biology*, *11*(9). https://doi.org/10.1371/journal.pcbi.1004523

Jensen, G., Terrace, H. S., & Ferrera, V. P. (2019). Discovering implied serial order through model-free and model-based learning. In *Frontiers in Neuroscience* (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fnins.2019.00878

Jonathan Evans, Stephen E. Newstead, R. M. B. (1993). *Human Reasoning: The Psychology of Deduction*. Lawrence Erlbaum Associates. https://books.google.de/books?id=iFMhZ4dl1KcC

Kumaran, D. (2012). What representations and computations underpin the contribution of the hippocampus to generalization and inference? In *Frontiers in Human Neuroscience* (Issue June 2012). Frontiers Media S. A. https://doi.org/10.3389/fnhum.2012.00157

Lazareva, O. F. (2012). Transitive Inference in Nonhuman Animals. *The Oxford Handbook of Comparative Cognition*. https://doi.org/10.1093/oxfordhb/9780195392661.013.0036

Lovelace, E. A., & Snodgrass, R. D. (1971). Decision times for alphabetic order of letter pairs. *Journal of Experimental Psychology*, *88*(2), 258–264. https://doi.org/10.1037/h0030922

Luo, Y., & Zhao, J. (2018). Statistical learning creates novel object associations via transitive relations. *Psychological Science*, *29*(8), 1207–1220. https://doi.org/10.1177/0956797618762400

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology*, *8*(2), 228–246. https://doi.org/10.1016/0010-0285(76)90025-6

Moyer, R. S., & Landauer, T. K. (1967). Time required for Judgements of Numerical Inequality. *Nature*, *215*(5109), 1519–1520. https://doi.org/10.1038/2151519a0

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). *Map making: Constructing, combining, and navigating abstract cognitive maps*. https://doi.org/10.1101/810051

Paz-y-Miño C, G., Bond, A. B., Kamil, A. C., & Balda, R. P. (2004). Pinyon jays use transitive inference to predict social dominance. *Nature*, *430*(7001), 778–781. https://doi.org/10.1038/nature02723

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Picklesimer, M. E., Buchin, Z. L., & Mulligan, N. W. (2019). The effect of retrieval practice on transitive inference. *Experimental Psychology*, *66*(6), 377–392. https://doi.org/10.1027/1618-3169/a000467

Potts, G. R. (1972). Information processing strategies used in the encoding of linear orderings. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 727–740. https://doi.org/10.1016/S0022-5371(72)80007-0

Quotas. (n.d.). https://www.qualtrics.com/support/survey-platform/survey-module/survey-tools/quotas/

Tsuchida, A., Doll, B. B., & Fellows, L. K. (2010). Beyond reversal: a critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback. *Journal of Neuroscience*, *30*(50), 16868–16875. https://doi.org/10.1523/JNEUROSCI.1958-10.2010

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Using Pavlovia.org. (n.d.). Psychopy.Org. https://www.psychopy.org/online/usingPavlovia.html

Vasconcelos, M. (2008). Transitive inference in non-human animals: an empirical and theoretical analysis. *Behavioural Processes*, *78*(3), 313–334. https://doi.org/10.1016/j.beproc.2008.02.017

CODE AVAILABILITY

The custom code used in this study is available on https://arc-git.mpib-berlin.mpg.de/thesis-retriever/myonlinestudy.

STATUTORY DECLARATION

Student: Clara Wicharz

Student Number: 20440183

I herewith formally declare that I have written the submitted thesis independently. I did not use any outside support except for the quoted literature and other sources mentioned in the paper. I clearly marked and separately listed all of the literature and all of the other sources which I employed when producing this academic work, either literally or in content. I am aware that the violation of this regulation will lead to failure of the thesis.

Clara Wicharz

Berlin, September 9, 2020