Reaction Time as an Indicator of Instance Typicality in Conceptual Spaces

Elektra Kypridemou[1][0000-0003-1575-9311] and Loizos Michael

Open University of Cyprus, Nicosia, Cyprus
[1] elektra.kypridemou@st.ouc.ac.cy

Abstract. In typical categorization tasks, humans are presented with a sequence of instances and report whether each instance is a member of a given category or not. In the current study, we examine the relationship between the reaction times (RTs) of human participants and the position of the instance in the conceptual space. Our main hypothesis is that instances closer to the boundary of the two categories, which are harder to be categorized, will require longer cognitive processing, resulting in longer RTs. Human subjects categorized images of novel objects to one of two given categories (represented by images of their prototypes); the selected category, RT and confidence rating for each trial were recorded. For trials with longer RTs people responded with less confidence and were more prone to making errors than for trials with shorter RTs. Moreover, people responded faster to stimuli with high similarity to at least one of the prototypes of the given categories than to stimuli that were distant from both prototypes, and hence closer to the boundary of the two categories, confirming our main hypothesis.

Keywords: Conceptual Spaces, Categorical Perception, Exemplars, Prototypes, Classification, Categorization, Reaction Time, Confidence Rating.

1 Introduction

In typical supervised and semi-supervised learning settings of machine learning, human teachers are presented with a series of elements and are asked to report for each element whether it is a member of a given category or not, usually by assigning a positive or a negative label to the element. Similarly, in psychophysics experiments, participants are given a series of stimuli and are asked to decide for each stimulus in which of the given categories it belongs. In such experimental designs, even if the task does not explicitly require a positive/negative label, the given categories are usually two well defined and complementary concepts, making the task analogous to the supervised learning training setting of machine learning. For example, Graf and Wichmann [5] implemented a gender categorization task with visual images of human faces, where participants had to categorize images to males and females. If we assume that the two categories of males and females are complementary (i.e., each face is either male or female), then the task would be logically equivalent to assigning posi-

CEUR-WS.org/Vol-2090/paper1.pdf

tive and negative labels w.r.t. one of the categories (even if, psychometrically speaking, changing the instructions of the task could possibly alter the results).

Certain previous approaches combined empirical psychophysics results and machine learning, aiming at a better understanding of human categorization processes. On the contrary, the main purpose of the present work is towards the opposite direction. Instead of using input from machine learning techniques to explain the experimental psychophysical results, our aim is to examine how the use of additional input coming from human teachers could presumably improve the existing machine learning techniques.

Existing predictive models of supervised and semi-supervised classification use labels produced by human teachers as a training set to classify future observations. We posit that considering reaction times (RTs) as an indicator of instance typicality in conceptual spaces, and incorporating RTs in the training material of the machine learning procedures, could possibly lead to better classification algorithms. As a first step towards this direction, in the current work, we examine the relationship between the RTs and the position of the instance to be categorized (target) in the conceptual space. Given that *(i)* RTs are found to provide a good approximation of distance between the element to be classified and the SH [5–7], and that *(ii)* in experimental settings it is easier to measure RTs than distances, which are internal representations of human minds, we argue that "*considering RTs in addition to the labels given by human teachers in supervised and semi-supervised settings, could potentially provide valuable input for more efficient learning algorithms*". Specifically, based on previous experimental results, we suggest that targets closer to the boundary of two categories are harder to be categorized, in the sense that they require longer cognitive processing, which is manifested by longer RTs.

Although we are unaware of any previous work trying to examine the above hypothesis, there has been work that examined an analogous hypothesis using the self-reported confidence of the users (confidence rating; CR). Ji and Lu [11] developed SVMAC, a novel support vector machine with automatic confidence, which is found to be significantly more accurate for gender classification than other traditional algorithms. Conceivably, one could also consider additional information from the teachers, including for example an explanation regarding their judgments, beyond their CR, to gain even more quantitative input about their decisions. If the improvements of such additional requirements are significant, then it might be worthwhile sacrificing some of the teacher's time for better machine learning performance.

Unlike Ji and Lu's [11] suggestion towards more efficient classification algorithms, the approach we suggest does not require any extra effort or time by the teachers, since the value of the RT is automatically recorded along with the teacher's response. As an extension of our approach, we could also consider other types of passive sources of information, acting as valuable input for our algorithms. For example, using some eye-tracking techniques during an image categorization task, we could track the visual processing of the stimuli. Examining the parts of the image where the eye is focused for longer time periods we could gain some valuable insight about the features, or the parts of the images, that guided the teacher's decision. Combining quantitative results coming from RTs with qualitative information coming

from eye-tracking techniques could give us some valuable insight into the cognitive processing and the factors that guided the decision making for each label.

The experiment of this paper is part of a longer research path towards our goal. The next step is to practically test whether the use of such additional input in the implementation of learning algorithms accelerates the learning process and improves the efficiency of the algorithms. The use of additional input could be implemented in several ways. One way is to filter the responses based on certain criteria and exclude the responses that do not meet these criteria from the training data. For example, excluding the responses for which the RTs are shorter than a minimum value (to avoid instances selected without any processing of information), or ignoring the responses for which the RTs are longer than a maximum value (implying less typical instances of a category) could be some types of filtering. Another way is to implement some already established techniques for using additional information such as the LUHI [18] and the LUPI [17] paradigms. In such techniques, the additional information is only provided during the training phase and is not available during the testing phase.

In the following sections, we demonstrate current empirical work on categorization, followed by a detailed description of the experimental design and the stimuli we used. We provide a detailed explanation of why we chose such a setting, and point out which methodological gaps of previous studies we are trying to fill. We then provide a more detailed description of the method we used regarding the participants, the materials used, the experimental design, and the procedure we followed. After presenting and discussing the results, we conclude our findings and we suggest the next steps to be taken.

1.1 Current Empirical Work on Categorization

Previous studies [5, 6] have reported lower RTs for correct responses than for incorrect ones, indicating that people respond faster when their response is correct. There is also experimental evidence [5, 6] that for higher metacognitive judgments of confidence the RTs are lower, indicating that people respond faster when they are more confident about their response. Moreover, participants of classification tasks were found to have metacognitive abilities, since their self-reported CR is negatively correlated with the classification error (CE; Eq. (1)); i.e., people are more confident about their choice when their response is correct [5]. Altogether, the above findings imply that longer RTs indicate classification cases in which people respond with less confidence and are more prone to making errors. In other words, cases that are more 'difficult' to be classified by humans require longer processing of information by the human brain. But which are these 'difficult' cases to be classified?

Taking it a step further, Graf and his colleagues [5–7], in order to better understand human classification processes, compared psychophysics results to machine learning techniques. They asked human participants to classify images of human faces to males and females, and correlated the human responses to the distance between the stimuli and the separating hyperplane (SH), as provided by several learning algorithms. What they found is that people are more accurate, respond faster, and report

higher confidence for their judgments when they classify human faces that are farther from the SH, than for those closer to the SH. Hence, one could argue that *the 'difficult' cases to be classified come from the stimuli closer to the SH, while stimuli that are farther from the SH are classified easier.*

However, using the above experimental designs, human responses about categories (as well as the corresponding RTs and CRs) might be affected by *(i)* participants' prior knowledge and personal interpretation of the given categories, and *(ii)* previously presented stimuli from the same category acting as exemplars of the category, a phenomenon known as the "old-items advantage effect".

In the present paper we are going to explore the relations between input from humans performing a categorization task and the similarity between the stimulus to be categorized (target) and the prototypes of the candidate categories. At the same time, we will try to limit potential effects arising from the nature of previous experimental designs, and check whether results are replicated. In the following paragraphs, we describe a new experimental design that addresses the above effects.

1.2 Introducing our approach

In the experimental design we used, participants were presented with three images of novel objects and were asked to categorize the image on the left part of the screen (the target *t*) in one of the two given categories, represented by two images *a* and *b*, on the right part of the screen (**Fig. 1**). After their selection, they were asked to report their confidence about their decision, on a scale from 1 (unsure) to 3 (sure). For each trial, we recorded three values: *(i)* the selected category, *(ii)* the reaction time (RT), and *(iii)* a self-reported confidence rating (CR) about the response. The experiment comprised eighty trials, which were presented sequentially to the subjects in random order.

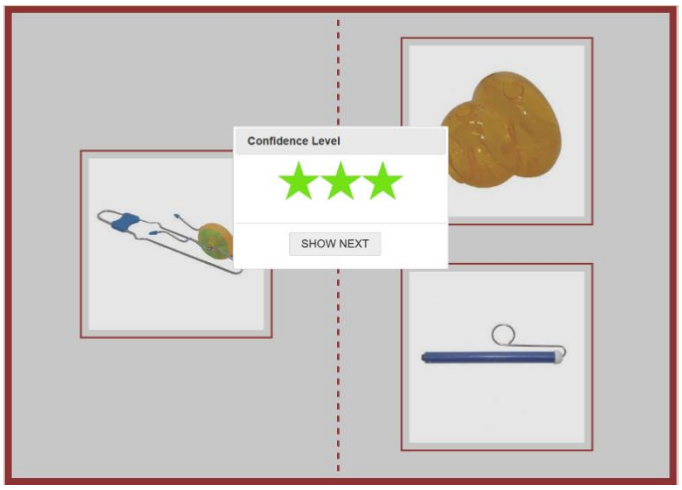

Fig. 1. A screenshot of a trial of the experiment with images of (*i)* the target *t* to be categorized (left), (*ii)* the prototype *a* of the category *A* (top right), and (*iii)* the prototype *b* of the category *B* (bottom right) and the confidence level rating window, which was presented after a category was selected.

The nature of the stimuli (images of novel objects) as well as our experimental design (presenting randomly-created triplets of images in each trial) resulted in a less straightforward categorization task. In some cases, the item to be categorized (target) could not easily fit to any of the given categories, while in some other cases the target could almost equally fit to both given categories. The purpose of such a setting was twofold. First, to test our hypothesis, we needed a range of possible arrangements of the target and the prototypes of the two categories in the conceptual space. Second, we argue that using a setting where the items to be categorized are not always clearly members of one and only one category better simulates more realistic situations. For example, everyday objects could be members of more than one category (e.g., a smartphone is also a camera), images might depict more than one object or concept (e.g., a picture of a beach view depicts the sea, the sun, the sky and maybe more concepts all at once), excerpts of text do not always have a unique style to be characterized (e.g., a text might be characterized as scientific and educational at the same time), and users' reviews might involve more than one emotion (e.g. a buyer might be angry and also disappointed by a product). In such cases, where there is not only one unique category where the instance fits, considering additional input such as the RTs could give us some more insight about the most dominant or representative category among all the candidate categories. In the following two sections we introduce some more technical benefits arising from our experimental design and we explain how our design limits potential effects that might be present in the standard classification tasks.

Limiting possible prior knowledge effects for concept representation

Human categorization of instances in commonly known categories such as males and females inevitably triggers effects arising from individual differences based on participants' prior knowledge related to the given categories. Such differences might arise either by individual experiences or by other social or geographical factors (e.g., Asian male faces significantly differ from Caucasian male faces). Human information processing and decision-making depend on personal pre-existing mental representations of the category, whether the category is represented by a prototype, a set of exemplars, or even by a set of rules of necessary and sufficient conditions. Even if experimenters explicitly ask participants to ignore any prior knowledge about the category and base their judgments only on some given prototypes or rules, it is not guaranteed that such effects of the prior knowledge will be successfully inhibited.

To avoid any pre-conceived categories, in our experiment we use a categorization task of unfamiliar objects coming from the NOUN database [9, 10], a collection of 64 images of novel objects specially created for experimental research studies. Since participants are not familiar with the visual stimuli of the task, and hence they have no

a priori knowledge of the target images and the categories represented[1], we argue that the prior-experiences that might influence participants' behavior are being limited.

Moreover, experiments in previous studies make space for individual representations of the categories based on prior knowledge, allowing participants to use their own prototypes of the category. In our experiment, instead of naming the given categories, we represent categories with images coming from the NOUN database. This way, we explicitly define the prototype of each category by an image, preventing any possible subjective interpretations of the categories. Participants, having no other clue to base their decision, are somehow 'forced' to use the given image as the category's prototype.

Controlling the use of exemplars

Considering the interaction between the prototype-based and the exemplars-based categorization processes [1, 3, 4, 12, 13, 19], shorter RTs do not necessarily indicate a lower distance between the stimulus and the prototype. Experimental psychology results [14–16] have shown that stimuli that are found to be similar to previously encountered exemplars of the category are categorized more easily (i.e., faster and more accurately) than non-familiar stimuli that are equally typical (or even more typical) members of the category. Moreover, when there is a pre-encountered exemplar of the category corresponding to the stimulus to be categorized, the categorization process is based on the similarity between the stimulus and the known exemplar rather than between the stimulus and the prototype. This privilege of the exemplar w.r.t. the prototype is known as the "old-items advantage effect". To highlight even more this effect, Hahn et al. [8] reported that exemplar similarity was dominant even in cases where basing categorization on a given rule would lead to perfect performance.

Even if in our experiment we use a categorization task of novel objects, according to the "old-items advantage effect", previously-encountered targets of a category could favor the categorization of new targets to the same category. This is why in our experimental design the candidate categories change between trials, instead of being fixed throughout the experiment. This way, we ensure that the only representation of the categories that will be used by the participants will be the prototype, as it is defined by the experimenters for each trial.

However, since the available images from the NOUN database were limited, some of the images would inevitably be presented more than once throughout the experiment (either in the form of a target or in the form of a prototype of a category). To control any sequential effects caused by previously presented images, we randomized the order of the trials for each participant.

Overcoming the obstacles caused by using unspecified concepts

[1] Please note that even if the novelty of the objects implies that the categories are not well-defined a priori, this does not imply that the categories of such objects are not pre-defined in the sense of Barsalou's 'ad hoc' categories [2], which are categories of known familiar objects.

Given that in our experiment the two categories of each trial are represented only by an image, and that the two categories differ from trial to trial, we only have two points in the conceptual space for each trial (acting as a prototype of the category). Therefore, the boundary between the two categories cannot be computed. In other words, the use of unspecified concepts implies the absence of an explicit boundary separating the two given categories.

To overcome this obstacle, we approximate the notion of distance between the target to be categorized and the boundary of the two categories, by using the notion of distance between pairs of images (i.e., the target and the prototype of each category). For images of the NOUN database, the empirically derived distance between all pairs of images is provided (see Materials section). This is one more reason why we decided to use the NOUN database.

Using the above approximation, we make the following assumptions. First, comparing the two distances (i.e., the distance between the target and the prototype of each category), we can prescribe the expected categorization of the target (which we are going to consider as the 'correct' label). Second, looking at the value of the two distances, we could get an idea about the position of the target in the conceptual space w.r.t the boundary separating the two categories. *When the target is distant from both categories, we suppose that it is close to the boundary of the two categories. On the contrary, in cases where the target is close to one prototype and distant from the other one, we suppose that the target is distant from the boundary, lying on the side of the closest prototype.* Under these assumptions, we are going to examine whether our experimental results are consistent with previous work, despite the methodological differences between the two experimental settings.

2 Empirical Method

Participants

Our data derived from a human sample of 40 adults (25 males, 15 females), aged 22 to 66, who participated voluntarily to the study by completing an online experimental task. Participants were naïve to the purpose of the experiment and received no financial or other compensation for their participation. They reported to have a normal or corrected-to-normal vision and provided informed consent. All participants completed all trials of our experiment.

For the descriptive analysis, we used the entire data set, while for the remaining part of our analysis we excluded two participants who were identified as outliers based on their RTs (see Results section).

Materials

In our experiment, we use the NOUN database [9, 10], a collection of 64 images, specifically designed for experimental research, especially for categorization studies. The objects depicted in the images are naturalistic, complex, multipart and multicolored, three-dimensional real objects [9], which in some respects resemble everyday familiar objects but at the same time are distinct and novel.

Sixty of the images were used in our experiment due to their higher quality, while the remaining 4 images were used only in the practice session. All images that we used were resized to 300×300 pixels, to ensure fast loading during each trial of the experiment.

Additionally, the NOUN database comes with a similarity matrix, providing a similarity rating for each pair of images. To obtain these ratings, Horst & Hout [9] performed an experiment, based on the spatial model of similarity. In their experiment participants completed a task of spatial arrangement, comprising 13 trials. In each trial, participants were given 20 images of the NOUN database and they were asked to arrange the images in the two-dimensional space, based on their perceived similarity (i.e., more similar items placed closer). Following the participants' ratings, the experimenters calculated all pairwise similarity ratings using multidimensional scaling (MDS) on the Euclidean distance for each pair of images. Lastly, Horst & Hout rank-ordered all pairs of images into four quartiles, based on the distances between their elements. Pairs belonging to the first quartile were the most similar pairs, while pairs of the fourth quartile were the most dissimilar ones. In our experimental design, we group the pairs of images to similar and dissimilar, based on the given quartiles of Horst & Hout.

Experimental design

For our experiment, we created ordered triplets *(t, a, b)* of images, one for each trial, where *(i) t* is the target to be categorized, *(ii) a* is the prototype of category *A*, and *(iii) b* is the prototype of category *B*. Using the 60 of 64 images of the NOUN database, we created 205,320 = 60·59·58 ordered triplets of different images ($t \neq a \neq b$), by creating all possible permutations of 60 without repetition.

We then characterized the above triplets based on the similarity ratings of each pair *(t, a)*, *(t, b)*, *(a, b)* of images, as provided by the creators of the database using a multidimensional scaling analysis [9]. To limit the number of our experimental conditions, we created two groups of pairs; pairs of similar items (by merging the first and second quartiles), and pairs of dissimilar items (by merging the third and fourth quartiles). Subsequently, we named the families of triplets w.r.t. the similarity between the elements of each pair *(t, a)*, *(t, b)*, *(a, b)*. Pairs *(t, a)* and *(t, b)* were characterized as High (H) when their elements were similar, and as Low (L) when their elements were dissimilar. Similarly, pairs *(a, b)* were characterized as Similar (Sim) or Dissimilar (Dis) when the prototypes *a*, *b* of categories *A*, *B* were similar or dissimilar, respectively. Based on the above terminology, we ended up with the following families of triplets: LL-Sim, LL-Dis, LH-Sim, LH-Dis, HL-Sim, HL-Dis, HH-Sim, HH-Dis, which consisted the eight conditions of our experimental design (**Table 1**, **Fig. 2**).

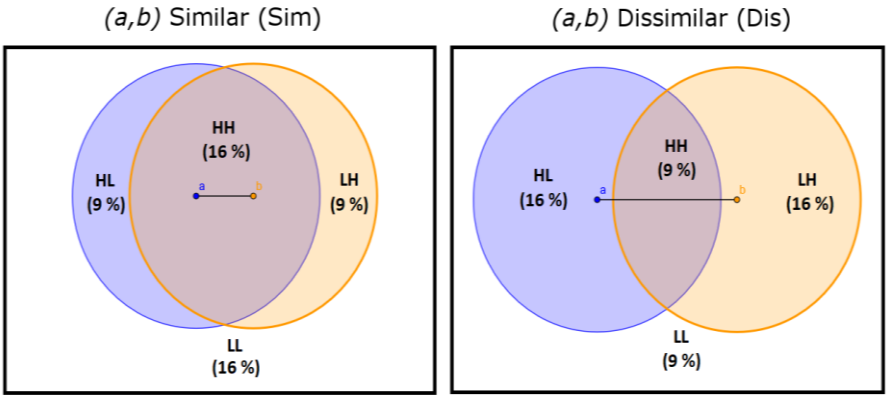

Fig. 2. Percentages of triplets produced for cases of similar (left) and dissimilar (right) pairs of prototypes *(a, b)*. The blue and orange circles represent all similar pairs *(t, a)* and *(t, b)*, respectively.

In examining the families of triplets produced, we made some surprising observations. Presumably, one would not expect to find any HH-Dis pairs, since this would imply that the target *t* is highly similar to both the prototypes *a* and *b*, while *a* and *b* are not similar to each other. Similarly, LH-Sim and HL-Sim families of triplets were also unexpected, since in such scenarios the target *t* would be similar to only one of the two prototypes *a* and *b*, while *a* and *b* are similar to each other. However, such families of triplets, that we considered as less possible, were also created, in smaller proportions (**Table 1**, **Fig. 2**).

Regardless the size of the produced families of triplets, we balanced our final experiment across conditions. Hence, the final experiment consisted of 80 trials in total, 10 trials for each condition, which were selected uniformly at random. The pool of the final selected triples was fixed for all participants, but trials were presented in random order for each participant, to avoid any sequential and order effects.

Table 1. Multitude of triplets created for each condition.

	Sim	Dis	Total
LL	32,150 (16%)	19,458 (9%)	51,608 (25%)
LH	18,902 (9%)	32,150 (16 %)	51,052 (25%)
HL	18,902 (9%)	32,150 (16 %)	51,052 (25%)
HH	32,706 (16%)	18,902 (9%)	51,608 (25%)
Total	102,660 (50%)	102,660 (50%)	205,320 (100%)

Procedure

Participants were personally invited to participate in our study. Before initializing the procedure, we had a personal session with each participant, to make sure all experimental criteria were met. First, we made clear that a desktop or laptop is needed

for participation (no mobiles, tablets, or other smart devices were allowed). In case they reported the use of a laptop, we imposed the use of a mouse (instead of the laptop's trackpad) for submitting their answers. Even if the web interface was light enough to ensure flawless loading between trials, we also made clear that an average Internet connection speed is necessary during the experimental task. Finally, we strongly recommended that participants were in a quiet environment with no distractors, while completing the experiment.

After making sure that all above criteria were met, we sent to the participants a first link to one of the experiment's trials (from the practice phase) to calibrate their browsers. We guided them to zoom in / zoom out their browsers so that the frame surrounding all three images of the trial would cover most of the surface of their monitor. After everything was set, we sent them a second link directing them to the web interface of the experiment and invited them to start.

On the first screen of the experiment, participants were informed about the study and completed an electronic consent form. A screen with detailed instructions followed, where participants were informed about the task, the timing and the self-reporting rating about their confidence for each response. Regarding timing, participants were advised to answer as fast as possible without sacrificing accuracy, so that we ensure that their decisions involved not only perceptual but also conscious cognitive processing. At the end of the instructions, participants were informed that a practice phase will follow, to ensure that the procedure is clear.

The practice phase consisted of four trials, identical to the trials of the actual experiment, during which no responses were recorded. Images that were presented in the practice phase were excluded from the actual experiment. After the practice was completed, participants were informed that the experiment begins.

In each trial of the experiment, a triplet *(t,a,b)* was randomly selected from the pool of the pre-selected triplets of the experiment. To record the RTs, time started counting by the time all three images *t*, *a*, and *b* were presented on the screen and stopped as soon as the participant clicked on one of the two images *a*, and *b*. After their selection, a smaller window appeared and participants had to evaluate their confidence about their previous response. Participants selected one, two, or three stars, to report their confidence level and then they had to click on the "Show next" button to proceed to the next trial. To control the distance between the position of the mouse when initializing a trial and each category image *a*, *b*, we placed the "Show next" button in a position equidistant from both category images. Eighty trials (ten trials from each condition) sequentially appeared in random order for each participant. After completing all trials of the experiment, we thanked participants and redirected them to the webpage of our lab.

3 Results

3.1 Descriptive statistics

Since our experimental design and the nature of the stimuli did not allow for an "objective truth", 'correct' responses were considered only for the families of triplets for

which the target *t* was similar with one of the prototypes and dissimilar with the other one (i.e., for the LH-Sim, LH-Dis, HL-Sim, and HL-Dis families). For these less 'ambiguous' families of triplets, we considered as 'correct' response the prototype *a* or *b* that was similar to the target *t*. For example, for trials coming from the family of triplets LH-Sim, the 'correct' response was the image *b* (i.e., the one positioned on the bottom right of the screen), while for trials from the family HL-Sim, the 'correct' response was the image *a* (i.e., the one on the top right of the screen).

Based on the number of 'correct' and 'wrong' responses given by participants for each trial, we calculated the variable classification error (CE) by dividing the number of wrong responses to the number of the valid responses given for each trial (1). The CE value could only be calculated for the families of triplets where the 'correct' response could be defined (i.e., for the less 'ambiguous' families).

$$CE = \frac{number\ of\ wrong\ responses}{number\ of\ correct\ responses + number\ of\ wrong\ responses} \quad (1)$$

Descriptive statistics for CE, RT, and CR are shown in Table 2, Table 3, and Table 4, respectively.

Table 2. Descriptive statistics for CE.

Triplets family	# of trials	n/a (%)	'Correct' responses	'Wrong' responses	CE
LL-Dis	400	400 (100%)	0	0	-
LL-Sim	400	400 (100%)	0	0	-
LH-Dis	400	0	181	219	54,75%
LH-Sim	400	0	199	201	50,25%
HL-Dis	400	0	244	156	39,00%
HL-Sim	400	0	289	111	27,75%
HH-Dis	400	400 (100%)	0	0	-
HH-Sim	400	400 (100%)	0	0	-
Overall	3200	1600 (50%)	951	649	40,56%

Table 3. Descriptive statistics for RTs (milliseconds).

Triplets family	# of trials	Mean	SD	SE	95% CI	
					Lower	Upper
LL-Dis	400	4907	3737	187	4539	5274
LL-Sim	400	5366	4186	209	4955	5778
LH-Dis	400	4689	4198	210	4276	5102
LH-Sim	400	5104	3894	195	4722	5487
HL-Dis	400	4788	3960	198	4399	5177

HL-Sim	400	4460	3562	178	4110	4810
HH-Dis	400	4247	3301	165	3923	4572
HH-Sim	400	4719	3730	187	4353	5086
Overall	3200	4785	3842	68	4652	4918

SD = standard deviation; SE = standard error; CI = confidence interval.

Table 4. Descriptive statistics for CR (stars).

Triplets family	# of trials	1 star (%)	2 stars (%)	3 stars (%)	Average
LL-Dis	400	**173 (43,25%)**	156 (39,00%)	71 (17,75%)	1,745
LL-Sim	400	**203 (50,75%)**	136 (34,00%)	61 (15,25%)	1,645
LH-Dis	400	135 (33,75%)	**162 (40,50%)**	103 (25,75%)	1,920
LH-Sim	400	**161 (40,25%)**	159 (39,75%)	80 (20,00%)	1,798
HL-Dis	400	150 (37,50%)	**167 (41,75%)**	83 (20,75%)	1,833
HL-Sim	400	144 (36,00%)	**165 (41,25%)**	91 (22,75%)	1,868
HH-Dis	400	119 (29,75%)	**179 (44,75%)**	102 (25,50%)	1,958
HH-Sim	400	123 (30,75%)	**174 (43,50%)**	103 (25,75%)	1,950
Overall	3200	1208 (37,75%)	1298 (40,56%)	694 (21,69%)	1,839

3.2 Correlations between RT, CR, and CE

Bivariate correlations between (a) RT and CE, (b) CR and CE, and (c) RT and CR were also calculated. Correlations (a) and (b) were calculated only for triplets where the CE could be calculated (i.e., only for the non-'ambiguous' families of triplets; N=1600), while correlation (c) was calculated for the entire dataset (N=3200).

According to the results, *(a)* there was a significant correlation between the RT and the CE, $r = .077$, p (one-tailed) $< .01$, indicating that people spent more time for trials for which they selected the wrong category, *(b)* there was a significant correlation between the CR and the CE, $r = -.097$, p (one-tailed) $< .01$, indicating that people were less confident for trials for which they selected the wrong category, and *(c)* there was a significant correlation between the RT and the CR, $r = -.143$, p (one-tailed) $< .01$, indicating that people spent more time for trials for which they were less confident.

3.3 Screening data and testing assumptions

All participants fully completed the experiment, and hence there were no missing values in our dataset. For each condition of the experiment, we tested our data for normality. Since normality assumption was violated, we checked for cases identified as outliers (i.e. participants with high RTs compared to the sample's mean RT). Two participants were identified as outliers in most of the experiments' conditions (6 of 8 and 7 of 8 conditions, respectively), and a third one only in 3 of 8 conditions (HH-

Dis, HH-Sim, and LL-Sim). The first two were excluded from the sample, whereas for the third one we used winsorization to limit extreme values. Hence, for the rest of our analyses, our final sample consisted of 38 participants (n=38). After the above corrections, the assumption of normality for RT was met.

3.4 Examination of the RTs

To examine the RTs among the eight families of triplets, we considered the variable Target Position, with four levels (HH, HL, LH, HH), and the variable Categories Similarity, with two levels (Sim, Dis), and we conducted a two-way repeated measures analysis of variance for these two within-subjects factors (**Fig. 3**).

Mauchly's test indicated that the assumption of sphericity was not violated for both the Target Position factor ($\chi^2(5) = 1.90, p > .05$), and for the interaction of the two factors ($\chi^2(5) = 6.10, p > .05$). The results show that there was a significant main effect for both the Target Position ($F(3,111) = 9.53, p < .01$), and the Categories Similarity ($F(1,37) = 8.47, p < .01$), as well as for their interaction ($F(3,111) = 2.82, p < .05$).

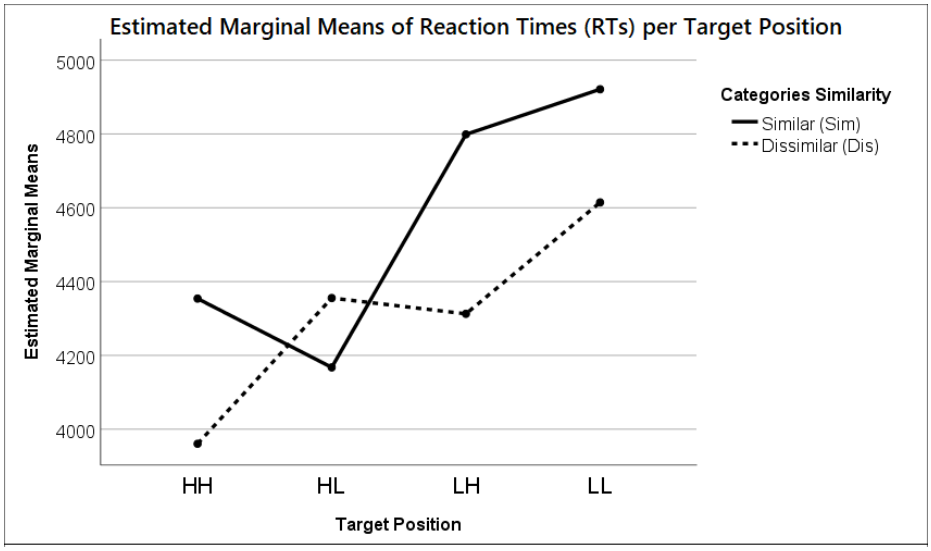

Fig. 3. Two-way repeated measures ANOVA for Reaction Times (RTs).

Further analysis of pairwise comparisons revealed that there was a significant difference of the average RTs only between the (HH,LH), (HH,LL), and the (HL,LL) Target Positions ($p < .01$). There was also significant different between the Similar and Dissimilar triplets ($p < .01$).

Even if the mean RTs between the HL and LH Target Position were not found to be significantly different, we observed that participants did not behave the same in these two cases, which was unexpected. To further explore this trait, we had to consider some additional factors. One possible interpretation could be that the ten triplets

selected for each of the families of triplets were not balanced. Another interpretation could be that the position of the two prototypes *a* and *b* also influences the RTs, and hence the decision-making process.

To check our first assumption, we examined whether the pairs consisting the triplets of the HL-Sim, HL-Dis, LH-Sim, and LH-Dis families were biased w.r.t. their similarity ratings. The mean similarity rating for the low similarity pairs was 540,15 and 406,05, for the HL and LH cases respectively. The mean similarity rating for the high similarity pairs was 1440,05 and 1481,80, for the HL and LH cases respectively. This is an indicator that triplets were balanced between HL and LH cases.

To examine the second assumption, we ignored the analysis of the participants' responses w.r.t. the 'correct' response and we only examined the responses w.r.t. the position of the selected image (i.e., top right of the screen or bottom right of the screen). Results show that for trials where the 'correct' response was at the bottom (LH-Dis, LH-Sim), participants' accuracy was not better than a random selection, whereas for trials where the 'correct' response was at the top, people tended to select the 'correct' response, regardless its position (Table 5).

Table 5. Statistics of the dependent variables, based on the position of the selected category for each family of triplets.

Triplets family	Expected selection	Users who selected top prototype			Users who selected bottom prototype		
		%	Average RT	Average CR	%	Average RT	Average CR
LL-Dis	n/a	42	4957	1.71	58	4870	1.77
LL-Sim	n/a	36	5688	1.62	64	5183	1.66
LH-Dis	bottom	45	5018	1.80	55	4417	2.02
LH-Sim	bottom	50	5517	1.77	50	4688	1.83
HL-Dis	top	61	4703	1.85	39	4921	1.81
HL-Sim	top	73	4305	1.95	27	4864	1.66
HH-Dis	n/a	62	4168	1.95	38	4373	1.97
HH-Sim	n/a	46	4975	1.81	54	4499	2.07
Overall	n/a	52	4830	1.83	48	4737	1.85

* Underlined values indicate statistically significant differences ($p < .05$) between the means of RTs and CRs of the two independent groups.

4 Discussion

Our results replicate previous findings exploring the meaning of RTs in categorization tasks while limiting potential effects arising from the nature of previous experimental designs. For trials with longer RTs people responded with less confidence and were more prone to making errors than for trials with shorter RTs, which is consistent with

previous work. Moreover, people responded faster for targets with high similarity to at least one of the prototypes of the given categories (HL and LH conditions) than for targets that were distant from both prototypes (LL), and hence closer to the boundary of the two categories, confirming our main hypothesis.

The shortest RTs were found in the HH-Dis family of triplets, where the target *t* was similar to both prototypes *a* and *b*, but the two prototypes were dissimilar. Although we expected that trials from this family would require longer processing in order to choose the best option, the experimental results showed that this was the case where participants responded faster. Additionally, we also found the highest average CR for this family of triplets, with most people reporting they were almost confident about their selection (self-rated their confidence with 2 stars out of 3). One interpretation of this phenomenon could be that participants, as soon as they identified one fitting category for the target, did not spend any extra time for checking whether there is a second fitting category or trying to decide which is the most appropriate one among the two. Hence, lower RTs do not always indicate instances typical for one category and not typical for the other, as we initially assumed. This could be a very useful finding for cases where targets could be members of more than one category, since lower RTs do not always imply excluding the categories which were not selected by the participant.

Finally, the fact that the HL and LH conditions were not symmetrical, highlights the need for a further examination of other factors, such as the position that appear the candidate categories.

5 Conclusion

The above results, though preliminary, are very promising. First, they replicate previous findings exploring the meaning of RTs in categorization tasks, while limiting potential effects arising from the nature of previous experimental designs. We consider that replicating previous results even with the use of novel images that form unspecified concepts, indicates that our basic hypothesis is primitive w.r.t the basic processes of human categorization. Second, the experimental design we used, combined with the findings of the present study, uncover many hidden aspects of previous studies, opening the way to future work towards multiple directions.

We are currently investigating possible bias effects arising from the position of the prototypes (top / bottom) or by any other presentation effects. Eye-tracking techniques can also be used to better interpret findings from RTs, as a quantitative method of the cognitive processes involved in the task, as well as a tool for exploring other possibe effects and revealing biases. Future work could also involve experimentation with more familiar stimuli, such as *(i)* images of familiar objects, *(ii)* images depicting more than one objects, or *(iii)* excerpts of text, which could be characterized by multiple labels, etc.

Acknowledgements.

We thank Christos Rodosthenous for assistance with creating the web interface of the experiment and for comments that greatly improved the manuscript.

References

1. Anderson JR, Betz J (2001) A Hybrid Model of Categorization. Psychon Bull Rev 8:629–647.
2. Barsalou LW (1983) Ad hoc categories. Mem Cognit 11:211–227. doi: 10.3758/BF03196968
3. Frixione M, Lieto A (2012) Prototypes vs. Exemplars in Concept Representation. Procedings of KEOD 2012, Int Conf Knowl Eng and Ontol. Dev, 226–232.
4. Frixione M, Lieto A (2012) Representing concepts in formal ontologies. Compositionality vs. typicality effects. Log Log Philos 21:391–414. doi: 10.12775/LLP.2012.018
5. Graf A, Wichmann F (2004) Insights from Machine Learning Applied to Human Visual Classification. In: Thrun S, Scholkopf B (eds) Adv. Neural Inf. Process. Syst. 16, Nips-16. MIT Press, Cambridge, MA, pp 905–912
6. Graf A, Wichmann F, Bülthoff H, Schölkopf B (2003) Study of Human Classification using Psychophysics and Machine Learning. 6:149.
7. Graf A, Wichmann F, Bülthoff H, Schölkopf B (2006) Classification of faces in man and machine. Neural Comput 18:143–65. doi: 10.1162/089976606774841611
8. Hahn U, Chater N (1998) Similarity and rules: distinct? Exhaustive? Empirically distinguishable? Cognition 65:197–230. doi: 10.1016/S0010-0277(97)00044-9
9. Horst J (2009) Novel Object & Unusual Name (NOUN) Database [PDF document]. 1–17.
10. Horst J, Hout M (2015) The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. Behav Res Methods 1–17. doi: 10.3758/s13428-015-0647-3
11. Ji Z, Lu B (2009) Gender Classification Based on Support Vector Machine with Automatic Confidence. Neural Comput 685–692.
12. Lieto A, Radicioni D.P, Rho V, (2017) Dual-PECCS: A Cognitive System for Conceptual Representation and Categorization, Journal of Experimental and Theoretical Artificial Intelligence, Vol. 29 (2), pp. 433-452. https://doi.org/10.1080/0952813X.2016.1198934.
13. Lieto A, Lebiere C, Oltramari A (2018) The Knowledge Level in Cognitive Architectures : Current Limitations and Possible Developments. Cognitive Systems Research, Vol. 48, pp. 39–55. doi: https://doi.org/10.1016/j.cogsys.2017.05.001.
14. Medin DL, Schaffer MM (1978) Context Theory of Classification Learning. Psychol Rev 85:207–238. doi: 10.1037/0033-295X.85.3.207
15. Smith JD, Minda JP (1998) Prototypes in the Mist : The Early Epochs of Category Learning. 24:1411–1436.

16. Smith JD, Minda JP (2000) Thirty Categorization Results in Search of a Model. 3–27. doi: 10.1037//0278-7393.26.1.3
17. Vapnik V, Vashist A (2009) A new learning paradigm: Learning using privileged information. Neural Networks 22:544–557. doi: 10.1016/j.neunet.2009.06.042
18. Vapnik V, Vashist A, Pavlovitch N (2009) Learning Using Hidden Information (Learning with Teacher). Int Jt Conf Neural Networks 3188–3195. doi: 10.1109/IJCNN.2009.5178760
19. Zaki SR, Nosofsky RM, Stanton RD, Cohen AL (2003) Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. J Exp Psychol Mem Cogn 29:1160–1173. doi: 10.1037/0278-7393.29.6.1160