The Effect of Relative Size Consistency on Object Recognition

Alexandra Theodorou[1]

Supervisors: Genevieve L. Quek[1], Marius V. Peelen[1]

[1]*Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands*

Abstract

Objects that frequently co-occur in the real-world appear to have canonical size relations. For example, a milk carton should be twice the size of the glass next to it. Similarly, a chair next to a table needs to be appropriately sized to form a functional group. Although the real-world size of objects has been identified as an integral property of object processing and search, little is known about how canonical size relations between objects influence recognition for semantically congruent object groups. Here, we used electroencephalography (EEG) and behavioural measures to test whether real-world size consistency between familiar object pairs facilitates grouping and recognition. We constructed silhouette object pairs containing semantically associated objects drawn from two possible real-world scale categories (eg., large objects – desk and chair, small objects – a bottle and a glass). We perturbed size consistency by rescaling one of the two items by 1:2 ratio. In Experiment 1 we validated the stimulus set through behavioural testing. In Experiment 2 brain activity was recorded using EEG while participants performed a one-back task during which size consistent and inconsistent targets were viewed. We tested whether event related potential (ERP) magnitudes differed as a function of size consistency. We found differences between mean amplitudes for size consistent and inconsistent trials specific to scale category. That is, for targets that appear large in the real-world responses significantly differ as a function of consistency earlier (P200) as opposed to small targets (P300, P600). We further hypothesized that scene understanding, measured by scale decodability, should be better for consistent as opposed to inconsistent pairs. However, multivariate classification analysis found no evidence that those underlying representations differ significantly.

Keywords: relative-size, size constancy, object recognition, visual cognition, EEG decoding

During our everyday experience with the visual world we do not encounter objects in isolation but embedded within the context of broader complex scenes and object groups. Relationships between individual objects can be defined by a variety of parameters such as their absolute locations in space, positional associations, and size relations (Biederman, Mezzanotte & Rabinowitz 1982). For objects that frequently tend to occur in everyday vision, as observers we can easily establish canonical associations for those parameters. For example, when bringing to mind the familiar pair of a bathroom mirror and sink we can clearly picture their canonical positional associations both in relation to location in space and each other, but we also have an inherent understanding and expectations about their size relations.

Thus far, existing literature has identified neural and cognitive mechanisms which utilize positional properties to optimize object perception and recognition in the real world (Kaiser, Quek, Cichy & Peelen 2019; Stein, Kaiser & Peelen 2015; Kaiser & Peelen 2018). However, despite the intuitive nature of size relations between objects that appear on the same depth plane, whether similar mechanisms apply to facilitate grouping and recognition of objects based on their relative size remains unknown. Here, considering relative size as an integral physical property defining relations between visual objects, we aim to investigate its role in object and scene understanding.

Scene to object contextual congruency on recognition and processing

Much of the classic research on the effect of context on object processing has centered on object-scene congruency – that is, the 'match' between an object and the scene surrounding it. Scene to object congruency has been known to influence perceptual processing. Early computational models have identified low-level features defined by scene statistical properties to prime object recognition within scenes (Torralba 2003; Oliva & Torralba 2007). In addition to low-level properties, early work has shown that semantic consistency between objects and scenes facilitates recognition of relevant targets. In an early study by Palmer (1975), when showing participants pictures of scenes that could be semantically congruent, incongruent or neutral to an object following scene presentation, participants were more likely to correctly name an object for the congruent scene condition. However, scene to object contextual facilitation is not one directional. In the task described by Davenport and Potter (2004), participants were asked to identify backgrounds and foreground objects for consistent and inconsistent targets (e.g., a priest in a church as opposed to a priest on a football field). Beyond increased accuracy when reporting either the background or foreground object in consistent conditions, when asked to report both, participants underperformed in identifying the backgrounds with foreground objects in comparison to naming isolated backgrounds. This serves as evidence of integrative processing of scene and object context for optimal perceptual facilitation for scene to object context.

Given the behavioural evidence of the effects of scene context on object recognition, focus has shifted to uncovering the underlying neural mechanisms of contextual facilitation effects. In an EEG study where participants had to identify an object appearing within a contextually congruent or incongruent scene, neural responses differed as a function of whether the semantic context of the scene matches the object (Ganis & Kutas, 2003). This difference was reflected in the N390 ERP component emerging around 300ms post stimulus presentation

(peaking at 390ms). This modulation of the late N390 component points to contextual facilitation resulting from high level semantic associations rather than low-level visual features. Later behavioural evidence showed that contextual facilitation occurs by extracting the overall scene gist and not based on matching of low-level feature similarities between the object and the scene or as a result of the modulation of spatial attention (Munneke, Brentari & Peelen, 2013). In support of scene-enabled object processing, evidence from decoding of neural signals from functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) data, identified enhanced decoding of object category when presented in the appropriate scene context in the object selective cortex (Brandman & Peelen, 2017). Decoding of MEG signals revealed scene-facilitated classification of objects to arise 100ms after peak decoding for single objects, providing evidence for the backpropagation of contextual scene information updating object representations.

Typical absolute location of objects in scenes facilitates processing

Semantic congruency between scenes and objects, might inform observers about what objects to expect in a given scene. However, objects also tend to occupy canonical absolute locations which help observers form expectations about where certain objects should be located. For example, certain objects in indoor scenes reliably occupy different parts of the visual field - we rarely encounter chairs in the upper visual field, or lamps in the lower visual field. The visual system appears to exploit these regularities in object positions, and has developed mechanisms that facilitate processing for objects appearing in their expected location (Kaiser et al., 2019). For example, behavioural evidence indicates that objects appearing in their typical absolute positions in space enter perceptual awareness faster than atypically positioned objects during continuous flash suppression tasks (CFS) (Kaiser & Cichy, 2018a). On a neural processing level, object identity representations in object selective regions are more robust for objects appearing in their typical locations (Kaiser & Cichy, 2018b). Based on their temporal profiles, those representational differences seem to emerge at the early stages of visual processing indicated by reliable decoding for the regular positioned items emerging at 140ms post stimulus presentation (Kaiser, Moeskops & Cichy, 2018).

Relations between individual objects matter for perceptual processing

Beyond associations between an object and the surrounding scene, the positional and semantic relationships between individual object units affect how viewers interact with and process the visual world. Larger objects are often indicative of the overall semantics, have been described to function as ‘anchors’ guiding attention and search based on their canonical spatial and semantic relations to other smaller objects (Boettcher, Draschkow, Dienhart, & Võ 2018; Boettcher, Dienhart, & Võ, 2017; Draschkow & Võ, 2017).

Furthermore, those kinds of top-down prediction guiding object-to-object associations also seem to impact processing on a more fundamental level. Evidence from behavioural work has indicated that semantically congruent objects interacting in meaningful manner form coherent groups that gain an advantage in perceptual processing. In an object matching task, participants were presented with a target object and a distractor and were asked to match the target object to a label. Naming accuracy for target objects was higher when the target and

distractor were positioned in a way to imply the objects could meaningfully interact, suggesting that those objects can be perceptually grouped (Green & Hummel, 2006)

In addition, familiar objects that frequently co-occur (e.g., A table and a chair), have been shown to enter conscious awareness sooner (e.g., break through continuous flash suppression sooner) when appearing in their regular relative positions (Stein et al., 2015). These perceptual facilitations have been proposed to arise from a reduction in visual competition mediated by the *grouping* of objects in typical meaningful configurations. Instead of processing each object individually it has been proposed that the visual system integrates objects into meaningful groups taking advantage of previous knowledge about scene structure. This proposed mechanism serves to efficiently cope with the complexity of real-world scenes and limit attentional competition (Kaiser et al., 2019).

Contextual integration signatures have also been observed on the neural processing level. In a study by Kaiser and Peelen (2018) a neural contextual integration index was identified in the lateral occipital cortex. When modeling responses from individual objects from a semantically congruent object pairs, the reflected mean activity in the region was more similar to objects that were irregularly positioned, signifying that those objects could not be perceptually grouped effectively. On the temporal domain a contextual integration response supporting the extraction of relevant features from meaningful sets, has been shown to emerge approximately 300ms post target presentation over right temporal channels (Quek & Peelen, 2020). Taken together those findings provide evidence for object grouping based on positional regularities between familiar objects as a key factor aiding object processing in naturalistic conditions.

The size of objects matters for perceptual processing

In addition to positional regularities what other factors might determine whether frequently co-occurring objects are grouped by the visual system? On possibility is size relations between related objects. When two objects appear on the same depth plane, they display highly predictable size relations. For example, we can think of an egg and an eggcup. The strict size relation between the two objects in this case is not just apparent but also has significant functional implications. A disproportionately large egg in relation to the cup would disrupt the functional relationship between the objects indicating that the two do not exist on the same depth plane and therefore do not form a meaningful group. Grouping objects based on real-world size can have real life implications. Think of a car approaching a traffic sign from afar. When the car is further away in depth there is no reason to group the two objects. This way observers can assume there is enough time for them to get across safely. Failing to group the two objects when the car approaches the traffic sign could have disastrous real-life consequences, yet as human observers we are not prone to such perceptual miscalculations.

Although intuitive, to date no existing work has examined whether real world size consistency mediates the degree to which the visual system groups objects to facilitate recognition. This is somewhat surprising given the substantial literature that has investigated how real-world size of *individual* objects matters to the visual system. Real world size selectivity has been demonstrated across the visual processing hierarchy in the ventral stream,

irrespective of the retinal image the objects appear on the screen (Konkle & Oliva, 2012a). Moreover, real-world size appears to be an automatic and essential property of object processing, such that in a size judgement task familiar objects appearing consistent with their real-world size during experimental presentation are processed and categorized more efficiently (Konkle & Oliva, 2012b).

The relative size of objects has also been shown to impact visual search strategies (Wolf, 2017; Collegio, Nah, Scotti, & Shomstein, 2019). In the context of visual search, in comparison to artificial deep neural networks (DNNs), human observers tend to employ attentional strategies utilizing templates tuned to the real-world size of target objects when searching in naturalistic displays. Template-based strategies might cause observers to miss size-inconsistent targets in relation to scene context, even when those manipulations are as extreme as a giant toothbrush on a sink (Eckstein, Koehler, Welbourne & Akbas, 2017). In the study by Eckstein et al. (2017), objects were presented in canonical locations, indicating that spatial regularities between objects alone cannot be held accountable for template-based search efficiency in naturalistic settings. Therefore, in addition to attentional guidance based on semantic associations or positional regularities, the expected relative size of objects seems to influence search.

Present study

Where the above findings suggest that observers are sensitive to the relative size of co-occurring objects (Collegio et al., 2019; Eckstein et al., 2017; Konkle & Oliva, 2012a; Konkle & Oliva, 2012b), to the best of our knowledge, there is no existing paradigm investigating whether the relative sizing of semantically-related objects affects object recognition itself. Specifically, it could be the case that when two related objects are appropriately sized (e.g., a proportional egg + eggcup), the visual system is able to 'group' the units in a way that facilitates recognition. Here we examined how the neural response to pairs of related objects (eg., desk + chair) varies as a function of their relative sizing. For the purposes of this study, relative object size refers to the proportional size of associated real-world objects appearing on the same depth plane. Since associated objects typically have a canonical size relationship (e.g., a milk carton would normally be 2-3 times taller than a glass), scaling either object up or down serves to disrupt the *size consistency* of the pair. We hypothesize that disrupting size relations between semantically congruent object pairs will inhibit efficient grouping and as a result affect recognition of the pair. We tested our hypothesis by investigating differences in the underlying neural representations and emerging temporal dynamics of objects pairs that display consistent or inconsistent size relations. To achieve this, we constructed a novel stimulus set of assorted pairs of two distinct real-world size categories in order to probe ensemble recognition of a semantic template when presented in their size-consistent configuration (see figure 1B for stimulus examples).

As our stimulus set is novel, in Experiment 1, we attempted to validate it by behavioural testing to ultimately select the exemplars that drive a size-consistency effect. We characterize this effect of interest as degraded recognition for size inconsistent pairs compared to facilitated perceptual processing for their size-consistent counterparts. Such an observation for individual

exemplars was used as an inclusion criterion for stimuli to be included in the following EEG paradigm for Experiment 2.

We used ERP and multivariate decoding methods of EEG signals to compare responses to size consistent as opposed to inconsistent pairs. Target pairs were of two distinct real-world scale categories, objects that appear large in the real world (e.g., a chair and a desk) or small (e.g., a spoon and a bowl). We trained a linear discriminate classifier (LDA) to distinguish between real-world scale categories of targets for size consistent and inconsistent pairs. If relative size is a factor which enables recognition and grouping, we expect lower classification accuracy for size inconsistent as opposed to consistent targets. By observing differences between conditions through ERPs we expect differences for size consistent and inconsistent pairs in later components associated with indexing semantic congruency around 300ms post stimulus presentation (Gannis & Kutas, 2003). Differences are also expected later in the time course for components previously linked to structural inconsistencies in scenes after 600ms (Vo & Wolf, 2013).

Experiment 1

Methods

Participants

Data from 24 (20 female) healthy individuals with an age range of 19 – 28 (M_{age} = 23.2 ± 3.09) were collected and included in the final analysis. Participants were recruited on a voluntary basis using the online research recruitment system at Radboud University in return for monetary compensation. All participants reported no neurological history and had normal or corrected-to-normal vision. Prior to their participation, all participants were briefed, read, and signed written informed consent forms. The study received approval from the Radboud University Faculty of Social Sciences Ethics Committee. Collected data was stored using pseudorandomized codes according to the European General Data Protection Regulation (EU GDPR).

Stimuli

The stimuli used for the behavioural paradigm consisted of 48 silhouettes of familiar object pairs drawn from two distinct real-world scale categories. Half the object pairs (24 pairs) contained ‘small’ scale objects, i.e., objects that are manipulable, easily grasped, found on indoor surfaces and are typically no larger than a computer monitor (e.g., A teapot and a cup). The remaining half (24 pairs) contained ‘large’ scale objects, i.e., objects that cannot be easily grasped or manipulated and typically convey the gist of a larger indoor scene observers operate within (e.g., A sofa and a floor lamp). We chose these specific real-world scale categories since there is both behavioural and neural evidence to suggest they are processed differently (Konkle & Oliva, 2012b; Josephs & Konkle, 2019).

Since our aim was to examine whether size-consistent object pairs would be recognised/processed more efficiently than size-inconsistent pairs, we presented objects in a Silhouette form that maintained high-level semantic and relative size information while losing other low and mid-level visual features that could trigger recognition (e.g., texture, within-

object contour variance and colour). By isolating objects pairs from the overall scene, we prohibit recognition based on scene-driven contextual cues and attentional guidance, constituting the size and semantic relations between the items as the only form of contextual information.

To introduce the central manipulation of relative size consistency, we created four versions of each object pair (e.g., plate + spoon). We began by creating a size-consistent version in which the objects were scaled proportionally (i.e., as you would encounter them at the same depth plane, see Figure 1A). We then created an identical, larger version of this pair (1:2 size increase). For the size-inconsistent exemplars, the objects from the two different size-consistent versions were recombined to yield the two different size-inconsistent pairs (Fig. 1B).

We attempted to limit the interference of mid and low-level confounds interacting with size-consistency induced effects, by taking low-level differences into consideration when constructing size inconsistent pairs. Namely, we ensured the distance between the two objects in the size-inconsistent pairs was be equal to the distance to between their size consistent counterparts. To achieve this, we randomly labelled the objects in each pair as object 1 and object 2. We then utilized the smaller distance whenever object 1 was the small object in the pair and the large distance when it was the larger object. For the small real-world scale category, the small mean distance in pixels between objects in the pair was $M_{smallDist} = 59.17$ (SD = 25.75) and the large average distance $M_{largeDist} = 117.14$ (SD = 51.51). Similarly, for the large real-world scale category, $M_{smallDist} = 61.94$ (SD = 38.19) and $M_{largeDist} = 117.02$ (SD = 74.77).

An additional possible low-level confound is alignment, in that large-scale objects tend to be aligned when viewed from the same viewpoint as opposed to smaller objects appearing on surfaces. For example, a bookcase and desk are aligned horizontally along the floor; an oven is vertically aligned under a rangehood, etc. To control for such alignment-related confounds, we tried to maintain any alignments between objects across conditions (see Fig. 1B for examples of stimulus subsets).

Fig 1. Stimulus construction overview and example subsets. A. Stimulus construction overview. Size inconsistent object pairs were constructed by recombining the objects from the two versions of size consistent pairs. The two distances between objects in the size-inconsistent versions were the same as in the size consistent condition. For each version for the size consistent and inconsistent pairs the distance between the two objects in the pair are the same. **B.** Examples of stimulus subsets. All displayed exemplars were both included in experiments 1 and 2.

Procedure & Design

The task was performed on a 1920x1080 pixel desktop computer monitor with a 120Hz refresh rate. All participants were seated approximately 65 cm away from the screen. All targets were presented on a uniform grey background on as 14x14cm squares. The experimental sequence presentation was programmed and controlled using MATLAB 2013b (MathWorks, Natick, MA) and the Psychophysics Toolbox (Brainard, 1997).

We used a 2 x 2 x 2 design with the factors *Size Consistency* (consistent/inconsistent), *Real World Scale* (Desk/Room), and *Rotation* (upright, inverted).

Participants were asked to perform a two-part computerized task consisting of ten experimental blocks, of 170 trials each and a 20-trial practice block. Trials were presented pseudo randomly as we implemented a counterbalancing procedure to ensure each stimulus was presented in one condition only for each participant. That is, for a given participant, a given object pair (e.g., plate + spoon) could appear only in one condition, and not in any of the remaining 7 conditions This way, we attempted to prevent carry-over learning of exemplar identity from one condition to another (i.e., seeing an object in the consistent condition first might make it more recognisable in the inconsistent condition later)

Each experimental trial started with a 75ms central fixation cross followed by a phase-scrambled mask presented for a variable interval between 300 to 800ms. The target was then centrally presented within the phase-scrambled mask for 150ms at 40% opacity. Following the target presentation, a second (different) phase-scrambled mask appeared prompting participants to provide a speeded response about the real-world spatial-scale identity of the stimulus pair by pressing the 'M' or 'Z' key for 'large' and 'small' scales. Response keys were counterbalanced across participants. After categorising the objects' real-world scale, participants were prompted to type the individual object identities they saw (see Figure 2). Order of entry for each object was emphasized as irrelevant since object configurations varied and items were not labelled upon presentation. We explicitly asked them to indicate the identities of individual objects and not a general scene category. For example, given the target stimulus was a bathtub and a shower curtain, accepted responses would be, 'object 1: bathtub, shower, tub' and 'object 2: shower curtain, curtain, towel'. An example of an undesirable response for the task would be 'object 1: shower, bathroom' while leaving the second object prompt blank. In addition, we encouraged observers to type in each object prompt with their best guess, only leaving one or both prompts blank if they were completely unaware of object identities. Everyone was encouraged to provide their answers in English. Exceptions were made when participants knew how to identify objects only in their native language, especially for Dutch natives. Due to the self-paced nature of the second part of each trial sequence, experimental presentation time varied with an average run of one and a half to two hours.

Fig 2. Behavioural task overview. Typical trial procedure. Participants for each trial performed a scale categorization and a following object naming task during which they were requested to type in the blank spaces the individual object names.

Analysis

We had three metrics of interest in Experiment 1: categorization accuracy and response time on the real-world scale categorization task, and naming accuracy on the object naming task. For the latter, object identification performance scores were measured based on the number of items named correctly (0 for no responses or no correct responses, 1 for one correct response and 2 for two correct responses). We evaluated the typed responses against a pre-generated list of possible correct labels for both objects within the pair. To minimise bias, the experimenter evaluating responses was blind to the rotation and size-consistency of the presented exemplar but was aware of exemplar number and spatial scale category. Responses

that were not on the pre-generated list of object names, but semantically related to the object in the pair and coincided with the real-world size were also considered correct (e.g., a kitchen cupboard being referred to as 'piece of furniture').

For each metric, we performed a three-way repeated measure analysis of variance (ANOVA) to access effects of real-world scale category, size consistency and rotation. In addition, we also calculated a standardised score (i.e., z-score) by subtracting the grand average of all items from the mean average of each item, and dividing by the standard deviation of the grand average $\frac{m_{item} - m_{grandav}}{SD_{grandav}}$. We then used these standardised z-scores to compute difference scores between size consistent and inconsistent pairs of the same exemplar based on z-score values. It was important to use standardised scores for this difference calculation since each subject saw each item in only one consistency condition, the difference scores are between participants. For all three metrics, a positive difference value indicates a size consistency effect in the expected direction (e.g., higher classification accuracy, faster RTs, and more accuracy naming for consistent vs. inconsistent pairs).

Results

Group analysis

Scale categorization task. Scale categorization accuracy showed main effects of real-world scale, $F(1,23) = 22.62$, $p < 0.01$, and image rotation, $F(1,23) = 60.10$, $p < 0.001$. As can be seen in Figure 3A, observers classified the scale of object pairs more accurately overall when the stimuli appeared on their upright as opposed to inverted condition. Overall small scale targets were more often accurately classified. There was a marginal interaction between scale and size consistency ($F(1,23) = 3.86$, $p = 0.061$). No other main effects or interactions reached significance (Fig. 3A).

Concerning reaction times, main effect of all three factors of scale $F(1,23) = 8.69$, $p = 0.007$, rotation $F(1,23) = 16.15$, $p < 0.001$ and size consistency ($F(1,23) = 4.62$, $p = 0.042$) were observed for the response time measure. With overall upright and consistent targets being categorized faster. In terms of scale, small targets demonstrate faster reaction times as seen on Figure 3B. However no significant interactions between any factors were noted.

Object-naming task. We observed a main effect of scale on the number of objects named correctly $F(1,23) = 120.59$, $p < 0.001$, such that observers were better at naming objects from the small scale. There was also a significant main effect of rotation $F(1,23) = 248.54$, $p = 0.000$, with upright objects producing higher naming scores than inverted ones (see Fig. 3C). Size consistency effects were not statistically significant ($F(1,23) = 1.80$, $p = 0.191$) as was also the case for interactions.

Fig 3. Group level results from behavioural Experiment 1. A. Scale classification accuracy scores. **B.** Scale classification reaction times **C.** Object naming performance.

Individual item analysis

Since a primary goal of behavioural Experiment 1 was to identify a subset of stimuli capable of driving strong size-consistency effects to be used in the subsequent EEG paradigm

of Experiment 2, we also inspected the data at the individual stimulus level. To be selected, a given object pair had to express consistency effects for at least two or more metrics. Based on this criterion, ten items for each spatial scale category qualified to be included in the EEG paradigm of Experiment 2.

Figure 4 shows the group performance on each metric of interest for each individual item. Here we ranked the object pairs by the magnitude of the consistency effect in the upright condition. On all measures of interest, a positive difference score for a given target indicates the expected effect of size consistency for this item. Based on this ranking, we identified ten targets of each spatial scale to serve as stimuli in the subsequent EEG Experiment 2 (Fig. 4D). Figure 1B shows examples of selected targets in all eight different versions.

Fig 4. Individual item analysis. The difference scores based on size consistency for Experiment 1 for **A.** *Scale classification accuracy.* **B**. *Scale classification reaction times*. **C.** *Object naming performance*. Positive values denote a size consistency effect in the expected direction (better performance for size-consistent pairs). Each label corresponds to an individual stimulus number. The items are ranked by effect magnitude in the Upright condition, with objects displaying the highest effect magnitude for the upright condition in favour of a consistency effect. Selected items are noted by a green number label. **D.** *Size consistent versions of the final stimuli selected for inclusion in Experiment 2*. Ten for each distinct scale category

Interim discussion

In Experiment 1, we used a behavioural paradigm to test whether relative size consistency affects object recognition as operationalized through scale categorization accuracy, scale categorization reaction time, and object naming performance. We then performed an individual item analysis to select the stimulus exemplars driving a size consistency effect, by ranking difference scores for the upright versions of the pairs. In terms of a group level analysis for all targets, we identified a significant main effect of the size consistency manipulation only for reaction times in the scale categorization task. Inconsistently sized targets displayed slower reaction times than their size consistent counterparts. Observing an effect of size consistency for only one out of three metrics of interest, provides little evidence that size consistency affects recognition on the behavioural level for the entire stimulus set used. Overall, a main effect of scale was found for all three tasks. Small scale targets seem to display increased performance in all three measures. This indicates that targets which belong in the small-scale category are better recognized than larger objects. However, the primary goal of this experiment was to validate the stimulus set and select targets for which the size inconsistent versions produce lower categorization accuracy, slower reaction times and lower naming performance. We therefore further performed an individual item analysis to select the stimulus exemplars driving an effect in the expected direction by ranking differences score for the size consistent and inconsistent pairs based on standardized z-scores for each target subset. This approach selected a subset of ten stimuli for each real-world scale category to include for EEG testing which display an effect of consistency for two or three metrics of interest (Fig. 4).

Experiment 2

Methods

Participants

Participants were recruited on a voluntary basis using the online research recruitment system at Radboud University in return for monetary compensation. Data from 24 (15 female) healthy individuals with an age range of 19 – 31 (M = 23.3 ± 3.03) were collected and included in the final analysis. All participants reported no neurological history and had normal or corrected-to-normal vision. Prior to their participation everyone was briefed, read and signed written informed consent forms. The study received approval from the Radboud University Faculty of Social Sciences Ethics Committee. Data was stored using pseudorandomized codes according to the European General Data Protection Regulation (EU GDPR).

Stimuli

Based on the behavioural data obtained in Experiment 1, we selected a subset of ten stimuli from each real-world scale category (Fig. 4D). The exemplars were chosen for their demonstrated capacity to elicit size consistency effects on two or more measures of interest (see Experiment 1). Unlike for the behavioural Experiment 1, here we further included the objects from every selected pair as singleton exemplars during experimental blocks (see Figure 1A for single object examples). Singletons were located between their original position as a part of the consistent and inconsistent version. The single objects were presented in both their large and small versions, segmented from the consistent silhouette pairs of each size respectively (see Fig. 1A).

Procedure & Design

The beginning of each trial was indicated by a central fixation cross appearing for a 200-500ms randomly asynchronous intertrial interval (ITI), followed by phase-scrambled noise mask, for a 400ms duration. Within the mask one of the chosen exemplars was shown briefly for 100ms at 40% opacity followed by a second mask with a duration of 200ms (see Figure 5 for the experimental presentation sequence).

Each participant performed 10 experimental blocks of 80 trials preceded by one practice block of 20 trials with a slower presentation sequence speed. None of the exemplars in the practice block appeared during the experimental presentation blocks. Unlike the behavioural Experiment 1, here all participants saw all targets in all conditions across the course of the experiment. Counterbalancing of targets was done by presenting each target in only one version within each block to minimise learning of object identities. We asked participants to perform a simple one-back task (press the button when a target repeated across trials). This task was orthogonal to the size consistency manipulation as size differences within object pairs were not task relevant. The main function of the task was to enforce participants to sustain their attention to the presentation sequence. Accuracy scores were provided as feedback at the end of each block; calculated by considering both correct rejections of non-repeat targets and hits for repeat targets. Each experimental block lasted for approximately five minutes with an optional short break half-way through each block and a longer break before initiating the next block.

Fig 5. Experimental presentation sequence for Experiment 2. Participants were asked to perform a one-ack task whenever they detect an exact repeat target. This is an example of two discrete experimental trials including a one-back target.

Experimental setup

The task was performed on 1920x1080 pixel desktop computer monitor with a 120Hz refresh rate. All participants were seated approximately 65 cm away from the screen, targets were presented on a uniform grey background on as 14x14cm squares. The experimental sequence presentation was programmed and controlled using Presentation software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

EEG acquisition

Scalp EEG was recorded using a 64-channel active electrode actiCap system at a sampling rate of 500Hz. Eye movement artifacts were measured using external passive electrodes situated at the outer canthus of each eye (to monitor horizontal eye-movements) and above and under the right eye (to monitor blinks). The ground electrode for these passive ocular channels was placed on the tip of the nose. Before initiating recording, impedances for all individual scalp channels were confirmed to be below 40 kOhm. The left mastoid acted as an online reference in BrainVision recorder. The experimenter monitored the EEG trace throughout recording. Each experimental block was initiated manually after following visual inspection of the EEG trace ensuring that there were no high amplitude deflections resulting from ocular or other muscle artefacts.

Analysis

ERP Analysis

Pre-processing

EEG pre-processing was carried out using the FieldTrip toolbox (Oostenveld, Fries, Maris & Schoffelen, 2011; http://fieldtriptoolbox.org) in MATLAB 2018a. Data were re-referenced to the left and right linked mastoid channels. A fourth order Butterworth bandpass filter (0.05-120 Hz) was applied to the raw data for 64 scalp channels. We also used a multi-notch filter to remove electrical line noise at 50, 100 and 150Hz. Following filtering, we down sampled the continuous EEG trace to 250Hz for easier processing and storage. We then proceeded to perform blink-related artifact rejection using independent component analysis (ICA) with a square mixing matrix, by visually inspecting the emergent components. By visually inspecting the results of the ICA we replaced artefact ridden channels with the weighted average of their neighbouring electrodes. The maximum number of channels replaced per participant was three.

We used a custom MATLAB function in conjunction with existing Fieldtrip functions to segment individual epochs around each target presentation (-200 to 1000ms). The targets were initially separated by two trial types, single and paired item targets, resulting to 800 trials of each type. One-back targets were identified and excluded from the final analysis. By visually inspecting the resulting trials, we identified 1.5% of total trials as artifact-ridden and excluded them from the final analysis. Segmented trials were split into consistent and inconsistent pairs according to the consistency condition of the presented target. As we were also interested in possible scale-specific consistency effects, further segmentation categorized trials based on consistency for each of the two real world scale categories.

ERP component selection

Since we did not have a priori assumptions about relevant time windows for our effects of interest, time windows for ERP analysis were identifyied by averaging trials across all conditions and observing the resulting peaks in the grand-averaged waveform independent of condition. This procedure identified four component windows of interest, namely an N170 (105-195ms), P200 (160-240ms), P300 (250-400ms) and a late P600 (560-750ms) (See Figure 6A).

ROI selection

Similar to time-window selection, regions of interest (ROIs) were identified by observing the emerging scalp topography of the grand-averaged waveform across the duration of the trial blind to condition. We identified three ROIs, including frontal (Fp1, Fp2, AF7, AF3, AF4, F7, F5, F3, F1, Fz, F2, F4, F6, F8) and midcentral (FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, CP4) and posterior electrodes (P7, P5, PO9, PO7, PO3, O1, P6, P8, PO4, PO8, PO10, O2).

Fig 6. **Averaged responses across conditions**. **A.** *Averaged waveform*. The averaged waveform independent of condition was plotted for ERP time window selection. **B.** *Average scalp topography per component*. The average scalp topography per component was used for ROI selection. **C**. *Selected ROIs.*

Statistical analysis of ERPs

Cluster based permutation statistics. To evaluate whether the evoked responses differ significantly for the two conditions of interest, size consistent and size inconsistent pairs for all pairs and for trials segmented by scale category (large and small), we performed a non-parametric clustered based permutation test using the FieldTrip toolbox in MATLAB 2018b (Maris & Oostenveld, 2007). We tested for differences between the two conditions both across the entire trial time window (-200 to 1000ms) and within the individual pre-defined component windows. To evaluate whether our two conditions differ significantly using the cluster-based permutation test, the larger clusters in the data are sampled by performing an independent samples t-test at each individual sample. The permutation distribution for the selected t-test statistic was created by drawing 10000 random permutations from the observed data. Finally, by using a Monte Carlo estimate we computed the p-value under the permutation distribution for an alpha level of 0.05.

Amplitude per component window. In addition to the timepoint-wise analysis above, we also computed conditional subject means within specific component windows (indicated in Figure 6). For each component, we subjected the means to a three-way analysis of variance (ANOVA) using the factors ROI (posterior, midcentral, frontal), scale category (large, small) and consistency (size consistent, size inconsistent). Furthermore, we also performed a two-way repeated measures ANOVA using factors consistency and scale for a priori selected

combinations of ROI and component. We focused on the posterior ROI for the N170 component to investigate possible differences at the early stages of visual processing. Statistical testing for the P200 component was restricted to midcentral regions due to increased positivity in those regions for the averaged responses (Fig. 1B). Finally, for later components (P600, P300) we focused on the midcentral and frontal regions, given prior evidence for the differentiation of responses over such channels for semantic and structural violation in visual scenes (Vo & Wolf, 2013). When we observed significant interactions between consistency and scale, we performed follow-up paired sampled t-tests between the means of the consistent and inconsistent conditions for each real-world scale category, using Bonferroni to correct for multiple comparisons.

Multivariate pattern analysis of EEG waveforms

In addition to the above univariate quantifications, we further employed a neural decoding approach using Multivariate Pattern Analysis (MVPA) to assess whether our size consistency manipulation affected perceptual processing and recognition at the representational level. MVPA is understood to have higher sensitivity for distinguishing differences between experimental conditions than univariate analysis methods such as ERPs (Carlson, Grootswagers & Robinson, 2019).

We performed MVPA decoding for the electrophysiological data using segmented trials that were subject to the same pre-processing pipeline described above, except that here we did not perform ICA for artifact rejection and channel correction, nor were any trials excluded (As suggested by Carlson et al. 2019). We performed two different types of analysis:

First, we were interested in identifying representational differences between targets based on their real-world scale. To achieve this, we trained a classifier using linear discriminant analysis (LDA) to distinguish between trials containing small-scale vs. large-scale object pairs. By using a decoding parameter orthogonal to our central size consistency manipulation, we hoped to use this analysis as an indirect measure of scene understanding/recognition. We hypothesize that in the case size consistency is a factor affecting recognition and grouping of familiar object pairs, by disrupting relative size relations the objects will no longer form a coherent recognizable group. Therefore, information about the semantic context of real-world scale category of the pair should not be as accessible for size inconsistent targets. This way higher and earlier classification accuracy for scale for size consistent as opposed to inconsistent pairs would indicate facilitated recognition of target category for the former and degraded recognition for the later.

We applied a 10-fold cross validation procedure in which we trained and tested the classifier on independent trials, by training on 90% of the total trials and testing on the remaining 10%. This procedure was performed ten times for each participants data, ensuring that all trials served in the training and testing set. To statistically assess whether classifier performance was higher than chance for specific timepoints while controlling for false discovery rates (FDR), we used a threshold-free cluster enhancement approach (TFCE) (Smith & Nichols, 2009). For TFCE, the grand average of the real classification accuracy scores are

tested against a null distribution of the data based on 100 permutations of the decoding analysis where targets are randomly re-assigned.

In a subsequent separate analysis, we employed the same decoding approach to examine size consistency representations directly. Here we trained the classifier to distinguish size-consistent or inconsistent object pairs, again training and testing on independent data following the procedures outlined above.

Results

ERP results

Cluster based statistics.

Size consistency effect averaged across Scale. Cluster based permutation statistics did not reveal any significant clusters between the size-consistent and inconsistent pairs when testing over the entire trial window (-200 to 1000ms around stimulus presentation). For the specific component windows, we identified a marginally significant cluster for size consistent and inconsistent pairs at the N170 latency between 140 to 175ms (p = 0.065) that was most prominent over frontal channels. No other component windows showed significant clusters.

Next, we split consistent and inconsistent trials by real world scale category and examined the effect of size consistency within each scale separately.

Size consistency effect for large scale pairs. No significant differences between size-consistent and inconsistent pairs were revealed by the permutation test for the entire epoch time window for large scale object pairs. When restricting statistical analysis to our components of interest a significant cluster was observed for the P200 time window (p = 0.020). This difference between consistency conditions was more prominent over frontal channels spreading towards posterior channels throughout the component window.

Size consistency effect for small scale pairs. The permutation test did not identify any statistically significant differences between consistency conditions for the entire trial window. For the small real-world scale category, the permutation test within the specific component time windows revealed a significant difference between size consistent and inconsistent pairs during the P300 window over posterior channels (p = 0. 007) and a marginal significant difference for the P600 latency over frontal and midcentral channels (p = 0. 055)

Amplitude differences per component.

Overall, three-way ANOVAs did not yield any significant interaction effects between ROI, size consistency and real world-scale category for any of the four pre-defined components. However, given a priori assumptions about the site and latency of the effect, we nevertheless performed two-way ANOVAs containing the factors scale category (large scale, small scale) and consistency (size consistent, size inconsistent) for the posterior ROI for the N170 window, midcentral ROI for the P200 and midcentral and frontal ROIs both for the P300 and P600 window.

Early N170 time window. For the N170 component, we were specifically interested in the posterior ROI over visual cortex, where we expected early differences between conditions should manifest. Here we identified a significant main effect of scale $F(1,23) = 58.01$, $p < 0.001$ and an interaction between size consistency and scale $F(1,23) = 5.75$, $p = 0.024$. For the main effect of scale, large targets displayed a stronger negative mean amplitude that small targets. In terms of the interaction effect, the large real-world scale category displayed stronger negative mean amplitudes for the consistent as opposed to the inconsistent condition while small real-world scale targets showed the reverse pattern (see Figure 7A). However, follow up paired sample t-tests of the consistency effect at each scale failed to reach significance in either case (small scale $t(23) = -1.57$, $p = 0.1595$, $t(23) = 1.24$, $p = 0.1279$).

P200 time window. For the P200 time window, we focused on the midcentral ROI. Here the two-away ANOVA revealed a significant interaction between scale and size consistency $F(1,23) = 4.38$, $p = 0.047$. We followed up this interaction by performing a paired sample t-test between the consistent and inconsistent conditions at each level of scale individually. For large targets, there was a marginally statistically significant difference between the consistent and inconsistent conditions, $t(23) = 2.77$, $p = 0.064$, with inconsistent pairs displaying a higher mean amplitude (see Figure 7B). The size consistency effect for small targets did not reach significance ($t(23) = 5.51$, $p = 1.000$).

P300 time window. Here a two way ANOVA for the midcentral channels revealed a significant main effect of scale, $F(1,23) = 20.15$, $p = 0.000$, with large scale targets yielding overall larger mean amplitudes than small scale targets. There was also a marginally significant interaction between scale and consistency $F(1,23) = 3.18$, $p = 0.087$, which we followed up by conducting a paired samples t-test of the consistency effect in each scale category individually. For small scale targets, we observed stronger positive mean amplitudes for the size consistent vs. inconsistent condition, $t(24) = -3.273$, $p = 0.006$ (see Fig. 7C). For large targets the paired samples t-test did not yield statistically significant results ($t(23) = 0.22$, $p = 0.720$). In contrast, large targets displayed the opposite effect. For the frontal ROI the two way ANOVA showed only a significant main effect of scale, $F(1,23) = 14.439$, $p < 0.001$, with large scale pairs overall displaying larger positive amplitudes than small pairs.

Late P600 time window. For the late P600 time window, a two-way ANOVA over midcentral channels displayed only a marginally significant main effect of size consistency $F(1,23) = 3.69$, $p = 0.067$. Over frontal channels, we identified a significant main effect of size consistency $F(1,23) = 4.267$, $p = 0.050$ where small scale targets show an overall larger mean amplitude and a marginally significant interaction between scale and consistency $F(1,23) = 4.06$, $p = 0.055$. A grater positive mean amplitude is reported for the consistent pairs for the small scale category while the opposite effect is present for large scale targets with size consistent pairs yielding a slightly greater positive mean amplitude. Following up this interaction using paired samples t-test we identified significant differences in mean amplitudes between small scale targets $t(24) = -2.39$, $p = 0.050$, where inconsistent targets display higher positive amplitudes.

Fig 7. Evoked responses within the three ROIs according toas a function of size consistency for large (left column) and small (right column) scale categories. The zero timepoint marks stimulus onset. Each column reflects a scale category (left – large scale, right – small scale) and each row one of the three ROIs.

Fig 8. Mean amplitudes per component/ROI as a function of scale and size consistency. Mean amplitudes for all three ROIs for components N170 **(A),** P200 **(B),** P300 **(C)**, and P600 **(D)**. See Figure 6 for component windows.

MVPA results

Decoding real-world scale category

Here we began by training a classifier to distinguish between trials containing small scale and large scale object pairs. We used decoding of scale category as a proxy for recognition. If consistent size relations play a significant role for recognition and grouping of the object pairs, then this should be reflected by decreased decoding of the semantic content of scale/scene category for size inconsistent as opposed to consistent pairs. Collapsing across size-consistent and inconsistent targets, we observed that real-world scale was decodable from 135 to 450. This above chance performance seems to be sustained over that time-window and peaks at 210ms. This confirms that information about the real-world scale of the targets is contained in the signal. As a sanity check analysis, we also conducted the same type of decoding of real-world scale category for single targets assess the validity of the pair decoding performance. No above chance accuracy was found for single targets at any time point, suggesting that the representation of real-world scale elicited by the object pairs resulted from the two objects forming a sparse or basic scene (Fig. 8A).

Since we were most interested in whether scale decoding, reflecting the semantic content of the pair, would vary with the objects' size relations, we next separated the data by the two conditions and trained and tested on scale category. This analysis revealed above chance accuracies for both types of targets (see Fig. 8B). Here we observed above-chance decoding of scale for both size consistent and inconsistent pairs, with decoding emerging sooner in the former condition (160ms vs. 190ms). Decoding also persisted for slightly longer in the size-consistent condition (250ms vs. 210ms). Scale decoding in both conditions peaked around the 200ms mark. To access whether the two scores significantly differ from one another, we performed a two-tailed t-test at every timepoint from -100ms to 700ms post stimulus onset. However, there was no significant difference in decoding accuracy between the consistent and inconsistent pairs at any timepoint. A secondary analysis by averaging accuracy means across participants and consistency conditions was performed using a one-tailed t-test within the 170 to 230 ms time-window. This window of interest was chosen since both size consistent and inconsistent targets display above chance classification accuracies in order to test for effects of consistency for the average of the two. Analysis of the averaged classifier performance within the 170-230 time window did generate statistically significant results ($t(23) = 1.406$, $p = 0.173$, CI [-0.004 0.021]).

Fig 9. Decoding real world scale category for independent trials. Decoding scale category implementing 10-fold cross validation for independent trials. The 50 percent line denotes chance classification performance and the

zero-time point notes the stimulus onset. Coloured stars below the 50 percent line indicate timepoints where classification accuracy for the given condition is significantly above chance. Black stars above the 50 percent line indicate the results of a two-tailed t-test, testing differences between classification accuracies for two conditions. **A.** *Decoding real-world scale category independently for all pairs and single targets.* Significant above chance classification performance is reported from 140 to 450ms only for pair targets. Classification performance differs significantly for pair and single targets within the 170 to 240ms time window **B.** *Decoding real-world scale category independently for consistent and inconsistent targets*. Above chance classification accuracy is reported for both size consistent and inconsistent pairs peaking at 200ms. Difference between performance for the two consistency conditions did not reach statistical significance.

Decoding size consistency

In a secondary analysis, we trained a new classifier to distinguish between size consistent and inconsistent pairs, separately for large and small scale object pairs. Here we observed no above chance decoding of size consistency at any timepoint, for either scale category (Fig. 9). That is, we found no evidence that the neural response to object pairs contained information about their relative size appropriateness.

Fig 10. Decoding size consistency by scale category for independent trials for small and large real-world scale targets. No above chance classification accuracy for decoding of size-consistency was reported at any time point.

Discussion

In this study, we sought to investigate the question of whether size consistency facilitates neural processing of object pairs. We suspected that semantically associated object pairs may be recognized better when they display canonical size relations (eg. a carton of milk being twice the size of a glass). We tested our hypothesis recording brain activity using EEG while observers saw object pairs of either consistent or inconsistent size proportions. Object pairs could be drawn from either a small or large real-world scale (ex. large objects – desk and chair, small objects – a bottle and a glass). The selected stimulus set used for EEG testing was pre-validated through behavioural testing (see Experiment 1). Through EEG testing we explored the underlying temporal dynamics using ERP and MVPA classification methods. We were primarily interested in whether the neural response to object pairs would differ as a function of the objects' size-consistency and if those effects are specific to scale category.

Using ERP methods, we were interested in identifying differences within pre-selected components between evoked responses for size consistent and inconsistent pairs. Based on the mean amplitude of responses over the entire trial window blind to condition we pre-defined four component windows (N170, P200, P300, P600) and three ROIs (posterior, midcentral, frontal). For the early N170 component window we considered to focus statistical testing on the posterior ROI to identify differences for early visually evoked potentials. For the P200 component we focused testing on midcentral channels based on observations of the averaged evoked responses independent of condition (see Fig. 5B). Finally, for later components P300 and P600, we focused statistical testing on the frontal and midcentral ROIs based on previous accounts in the literature concerning semantic and structural congruency in visual scenes (Gannis & Kutas, 2004, Vo & Wolf 2013).

At the P200 latency over midcentral channels large scale targets displayed higher positive mean amplitudes for the inconsistent condition (see Fig. 7B). For the P300 latency we tested mean amplitude differences over the midcentral ROI. For small scale targets, we report that consistent pairs display higher positive amplitudes (see Fig. 7C). Finally, for the P600 latency we investigated differences between scale and consistency relations within for frontal and midcentral channels. A significant difference in mean amplitude for size consistent and inconsistent pairs was identified only for small targets in frontal channels, with consistent pairs demonstrating higher positive amplitudes (Fig. 7D).

Following the analysis of ERP data, we performed MVPA decoding to identify representational differences between size consistency conditions not evident in ERP differences. We first trained an LDA classifier to distinguish between real-world scale category for size consistent and inconsistent pairs of objects. The two distinct scale categories (large and small) were chosen as there is evidence for differences in cortical processing (Konkle & Oliva, 2012b; Josephs & Konkle, 2019). We therefore treated scale category classification as an indirect measure of recognition. Following this rational, higher classification accuracy for real-world scale would indicate that high level semantic information about the pair is present in the neural signal. In terms of our size consistency manipulation, lower classification performance for inconsistent pairs, would indicate that the object pair is less recognizable and therefore semantic information about the scale of the objects in the real-world less accessible. In this study, we observed above chance sustained classification accuracy for all pairs independent of size-consistency from 135 to 450ms peaking at 200ms post stimulus presentation. When splitting the data into consistent and inconsistent pairs, the neural response contained decodable information about the real-world scale of the object pairs regardless of the size consistency of the target. That is, we observed comparable scale-decoding for size-consistent and size inconsistent object pairs peaking at 200ms.

On a second type of analysis to further understand whether representations differ between size consistent and inconsistent pairs we trained the classifier to distinguish between consistency conditions for all pairs and on a consecutive analysis, pairs separated by real-world scale category. This analysis did not yield significant results at any time point.

Size inconsistencies modulate early P200 and P300 components specific to scale category

In this experiment we report a higher positive mean amplitude for size inconsistent as opposed to consistent pairs within the P200 latency (160-240ms). This effect was evident only for targets belonging in the large real-world scale category. Previous literature does not provide evidence for a similar effect to the one we report in the present study. A study investigating hemispheric differences for meaning processing Federmeier and Kutas (2002) presented observers sentences followed by pictures that could either be expected or unexpected in terms of semantic content. They observed amplitude differences at a 150-250ms latency were larger as a response to expected pictures when targets were presented on the right visual field. Earlier accounts report higher P200 amplitudes during pop-out detection tasks when pop-out search targets display distinct low-level visual features such as colour and orientation (Luck & Hillyard, 1994). If we consider the size-inconsistent targets as unexpected/surprising and harder to interpret, then given those previous findings we would expect an effect in the opposite

direction where size-inconsistent targets display higher mean amplitudes. However, as large real-world scale targets display more rectilinearity and alignment between the two objects, modulation of the P200 component could possibly result from an enhancement of those low-level features during the size inconsistent condition as opposed to differences in semantic interpretations.

We further identified modulations in the P300 time window, however here the influence of size consistency was constrained to small-scale object pairs. Specifically, we found that over midcentral channels consistently sized small scale objects drove higher positive amplitudes than inconsistently sized objects. Importantly, these effects were somewhat weaker than those identified for the P200 latency (i.e., the interaction between real world scale and size consistency was only marginally significant, $p = 0.087$). Nevertheless, previous work has identified modulations of the P300 component as a function of semantic inconsistencies. In their 2013 paper Vo & Wolf during a one back task where observers viewed real world scenes with objects that had undergone semantic (ex. A bar of soap next to a laptop computer) or syntactic (ex. A mouse on top of a computer) violations, they observed a negative peak associated with structural and semantic violations within the 300ms latency. Negativity peaking around 300ms for anterior regions driven by semantic incongruencies between objects presented sequentially has been previously established in the literature (Eddy, Schmid & Holcomb 2006, McPherson & Holcomb 1999). Based on those previous accounts we can infer that the dissociation of responses driven by size-consistency conditions could result from a disruption of semantic content elicited by a disruption of relative size. This way constituting size-inconsistent targets as lacking semantic content. However, given our marginal statistical finds and the qualitative difference between the stimuli and tasks used in previous studies we must make interpretations with caution.

P600 modulated by size consistency only for small scale targets

Analyses within the P600 time range identified that the late positivity over frontal channels was stronger for consistent than inconsistent pairs, but only when the objects belonged to the small real-world scale category. Late ongoing positive potentials from 600 to 1000ms over midcentral have been attributed structural violations in scenes (Vo & Wolf, 2013). Similar findings where the P600 reflects mild syntactic inconsistencies was first reported in the language domain (Osterhout & Holcomb, 1992). Here we identified significant differences between size-consistent and inconsistent potentials within the P600 latency. Our effect differs to the one reported by Vo & Wolf (2013) for mild structural violations in real world scenes displaying a larger positive amplitude than for structurally consistent scenes in the sense that we observe grater positivity for size-consistent pairs. Despite those differences we could argue that a disruption between size-relations for a semantically congruent pair of objects resembles a structural inconsistency. If that is the case, then as it is suggested for structural irregularities participants might need to employ additional cognitive resources to re-evaluate this disruption between size relations after initial recognition later in the time course.

Consistency effects specific to scale category per component

Based on the results from our analysis of ERP data we identified significant differences between size consistency conditions specific to scale category. In addition, those differences seem to be relevant for large scale targets early on in the time course and for targets that are small in terms of real-world size later, but effects do not seem to overlap. A possible explanation could stem from differences in terms of low-level and mid-level confounds between the two scale categories (see section on low level differences). Large scale targets are often vertically and horizontally aligned and appear to have more edges that their small-scale counterparts (Fig. 1B). Distinguishing between those low-level differences between the size consistent and inconsistent pairs for the large scale category could emerge earlier in the time course. On the contrary, small scale objects that are not defined by such differences might contain more semantic information about the pair. Therefore, viewers identifying a disruption of semantic content might need additional time and resources to re-evaluate the given target later in the time course.

As it is evident from the behavioural data of Experiment 1 (Fig. 3), small scale targets were easier for observers to recognise (faster/more accurate categorisation and displayed higher naming scores). As a result, it might be the case that if observers are indeed better at recognizing small, handheld objects they can better identify inconsistencies within such pairs and need additional time and resources to re-evaluate them. Conversely, if observers struggled to recognise the large scale objects, they might not have perceived their size inconsistency.

Representation of real-world scale is not modulated by size-consistency

We initially hypothesized that if relative size is a vital aspect of object recognition for semantically associated pairs, then by disrupting size relations we would also degrade the semantic content. By choosing two distinct real-world scale categories and training and testing an LDA classifier to distinguish between them, disruption of semantic content would be reflected by lower classification performance for inconsistent pairs. Our results do not support our initial hypothesis as scale category was equally decodable both for size consistent and inconsistent targets. That is, observers appeared to represent the real-world scale of the objects to an equal extent regardless of the objects' size consistency. Those results imply that there is no evidence that the neural representation of real-world scale is depended on whether the objects' relative size was appropriate or not. What might account for this lack of an effect? One possibility is that our manipulation of relative size was not sufficiently extreme to elicit differential neural responses. Further investigation of the stimulus set and potentially increasing the ratio between the two objects from 1:2 to 1:3 could possibly uncover representational differences.

In addition to pairs we also performed decoding of scale category for single targets as a validation procedure. We expected that isolated silhouettes of objects would be less recognizable for participants in the absence of contextual information of the second object. The results from this analysis confirms this hypothesis that inferences on the real-world size identity of the object are not available without a semantically associated object to inform scale identity.

Does size consistency affect recognition after all?

We set out to understand whether size consistency plays a significant role on object recognition and grouping. There is sufficient evidence to think of the size of objects in relation to one another as an essential property of object processing in the real world (Konkle & Oliva, 2012a, Konkle & Oliva, 2012b). As human observers we can extract and learn positional properties between arbitrary shapes on artificial displays within a small amount of time (Fiser & Aslin, 2001, 2005, Yu & Zhao, 2018). It is therefore reasonable to assume that through a lifetime of exposure to the regularities of the visual world around us we can learn the canonical size relations between objects we frequently encounter and interact with. However, despite size constancy being an inherent property defining object relations, our manipulation did not elicit significant differences in the neural responses to consistent and inconsistent objects pairs. Decoding analysis found no evidence that the neural response to object pairs contains information about their relative size consistency. Nevertheless, we did find some more local differences in ERP magnitudes between size consistent and inconsistent pairs. This could imply that even if size consistency might not be an inherent aspect of recognition within the context of this study, it still affects the underlying neural responses. Nonetheless, as this is the first study to our knowledge investigating size consistency in the context of recognition further assessment of the current paradigm is necessary.

The absence of significant findings in favour of our main hypothesis through the decoding analysis could be a results of an insufficient size manipulation to elicit the effects of interest. By implementing a more radical disruption of size consistency we could gain a better understanding of the effect of the manipulation.

Low-level differences between scale categories for target pairs

Although the size consistency of the object pairs was not represented in the neural response, we did manage to observe strong decoding of scale category – that is, whether the object pairs were drawn from a large or small real world scale. It is interesting to consider whether this reflects a high-level representation of real world size, or more low level visual differences between categories. For example, large scale objects thend to display overall more alignment, rectilinearity and edges than smaller hand-held objects. This results from the fact that those larger objects consistently appear on the same plane and viewpoints aligned on the floor or vertically (ex. An oven and a rangehood). It has been previously reported that large and small objects in the real world differ in terms of mid-level features such as shape information and junctions, both of which are maintained throughout our manipulation (Long, Yu & Konkle, 2018, Long, Konkle, Cohen & Alvarez, 2016).

To disentangle whether low-level features affect processing between the two categories we could perform a complementary decoding analysis to consider would be to train and test on the parameters of interest leaving one stimulus set out per run. For example, when decoding for real-world scale category train on exemplar sets 1-9 for each scale and then test classifier performance on the remaining one for each scale. Forcing the classifier to generalise to novel object pairs like this will reduce (but not eliminate) the degree to which above chance classifier performance can be driven by low-level visual differences between categories. If it is the case that decoding performance is driven overall target visual features as opposed to semantic content, then through this analysis classification accuracy should decrease.

It is worth noting that so far, to our knowledge, only one study has investigated the effects of size constancy between parts within individual objects. Electrophysiological evidence from neuronal populations in the monkey inferior temporal cortex (IT), revealed a subpopulation that codes for relative size consistency within single objects, by demonstrating selective tuning for proportionally scaled parts, providing a potential neural mechanism for relative size constancy (Vighneshvel & Sripati Arun, 2015). However, even though those findings suggest a possible substrate size constancy, given the study population and the nature of single stimuli with proportionally and disproportionally scaled parts, no inferences about higher level semantics and associations can be drawn.

Future directions

In addition to identifying the contribution of low-level confounds, to identify representational differences we could perform a pair approximation analysis akin to the analysis described in the paper Kaiser and Peelen (2018). As we presented single targets during the EEG paradigm of Experiment 2 that appeared on the intermediate locations between their position is size consistent and inconsistent pairs, we could perform linear addition of those signals. This type of analysis can be implemented by averaging the signals across time points of the single targets that belong to a specific pair both for the size consistent and size inconsistent conditions. This will result in eight different types of targets, the size consistent and inconsistent pairs in their two versions and their pair approximations formed by linear addition of single targets. By comparing the pair approximated targets with single and pair temporal responses we could establish perceptual grouping effects. If pair approximations are more similar to consistent than inconsistent pairs we can infer that the size inconsistency manipulation inhibits grouping between objects in the pair and therefore relative size of objects constitutes a factor that plays a significant role for object grouping and perceptual facilitation.

Conclusion

In this study we sought to determine the pending question of whether size relations between contextually associated objects facilitates recognition by efficiently grouping objects proportionate in size into a meaningful pair. Our results from decoding of EEG waveforms suggest that relative size consistency is not represented strongly in the neural response to object pairs. However, we identified differences in the underlying neural responses though the analysis of evoked potentials. Those differences seem to be specific to real-world scale category with objects that appear large in the real-world displaying differences early on as opposed to small objects where we identify differences later in processing.

References

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143–177.

Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *Journal of Neuroscience*, *37*(32), 7700-7710.

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision 10*:433-436.

Boettcher, S., Dienhart, E., & Vo, M. (2017). Anchoring spatial predictions: Evidence for the critical role of anchor objects for visual search in scenes. *Journal of Vision*, *17*(10), 304-304.

Boettcher, S. E., Draschkow, D., Dienhart, E., & Võ, M. L. H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of vision*, *18*(13), 11-11.

Carlson, T. A., Grootswagers, T., & Robinson, A. K. (2019). An introduction to time-resolved decoding analysis for M/EEG. *arXiv preprint,* arXiv:1905.04820.

Collegio, A. J., Nah, J. C., Scotti, P. S., & Shomstein, S. (2019). Attention scales according to inferred real-world object size. *Nature Human Behaviour*, *3*(1), 40-47.

Davenport, J. L., & Mary C. P. (2004) Scene consistency in object and background perception. *Psychological science, 15*(8), 559-564.

Draschkow, D., & Võ, M. L. H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific reports*, *7*(1), 1-12.

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, *27*(18), 2827-2832.

Eddy, M., Schmid, A., & Holcomb, P. J. (2006). Masked repetition priming and event-related brain potentials: A new approach for tracking the time-course of object perception. Psychophysiology, 43(6), 564-568.

Federmeier, K. D., & Kutas, M. (2002). Picture the difference: Electrophysiological investigations of picture processing in the two cerebral hemispheres. Neuropsychologia, 40(7), 730-747.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychological science, 12(6), 499-504.

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: statistical learning of visual feature hierarchies. Journal of Experimental Psychology: General, 134(4), 521.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, *16*(2), 123-144.

Green, C., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1107–1119.

Josephs, E. L., & Konkle, T. (2019). Perceptual dissociations among views of objects, scenes, and reachable spaces. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(6), 715.

Kaiser, D., Cichy R. (2018a). Typical visual-field locations facilitate access to awareness for everyday objects. *Cognition, 180*, 118-122.

Kaiser, D., & Cichy, R. M. (2018b). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology*, *120*(2), 848-853.

Kaiser, D., Moeskops, M. M., & Cichy, R. M. (2018). Typical retinotopic locations impact the time course of object coding. *NeuroImage*, *176*, 372-379.

Kaiser, D., & Peelen, M. V. (2018). Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *NeuroImage*, *169*, 334-341.

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in cognitive sciences*, *23*(8), 672-685.

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, *74*(6), 1114-1124.

Konkle, T., & Oliva, A. (2012b). A familiar-size Stroop effect: real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(3), 561.

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General, 145*(1), 95.

Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences, 115*(38)

Luck, S. J., & Hillyard, S. A. (1994). Electrophysiological correlates of feature analysis during visual search. Psychophysiology, 31(3), 291-308.

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. Psychophysiology, 36(1), 53-65.

Maris, E., Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods.*

Munneke, J., Brentari, V., & Peelen, M. (2013). The influence of scene context on object recognition is independent of attentional focus. *Frontiers in psychology*, *4*, 552.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J. M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience, 2011*.

Oosterhof, N. N., Connolly, A. C., and Haxby, J. V. (2016). CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab / GNU Octave. Frontiers in Neuroinformatics.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. Journal of memory and language, 31(6), 785-806.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in cognitive sciences*, *11*(12), 520-527.

Palmer, T. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*, 519-526.

Quek, G. L. & Peelen, M.V. (2020) Contextual and spatial associations between objects interactively modulate visual processing. *Cerebral Cortex*. *197*.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage, 44*(1), 83-98.

Stein, T., Kaiser, D., & Peelen, M. V. (2015). Interobject grouping facilitates visual awareness. *Journal of Vision*, *15*(8), 10-10.

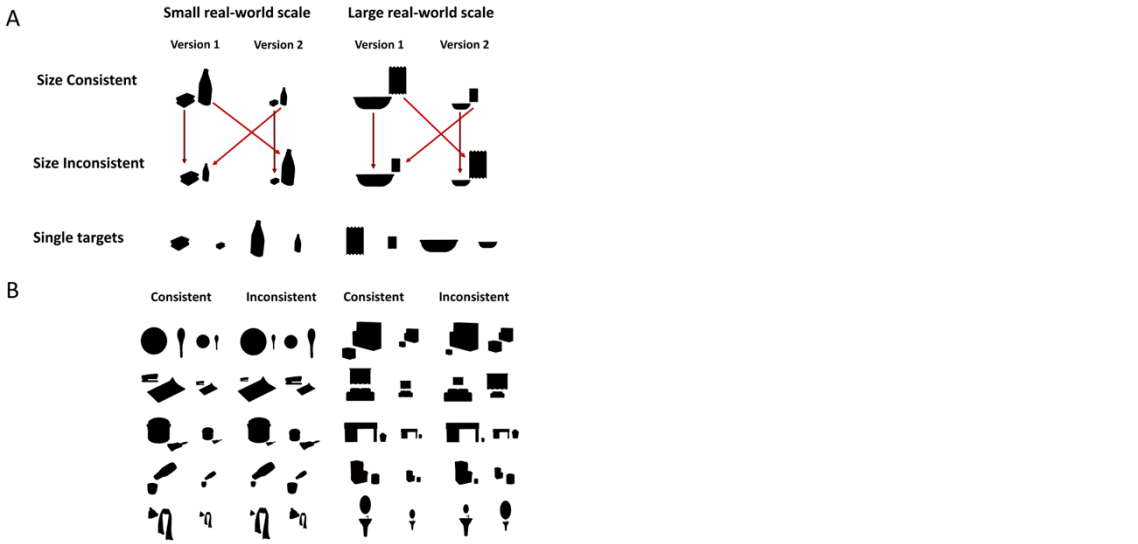
Torralba, A. (2003). Contextual priming for object detection. *International journal of computer vision*, *53*(2), 169-191.

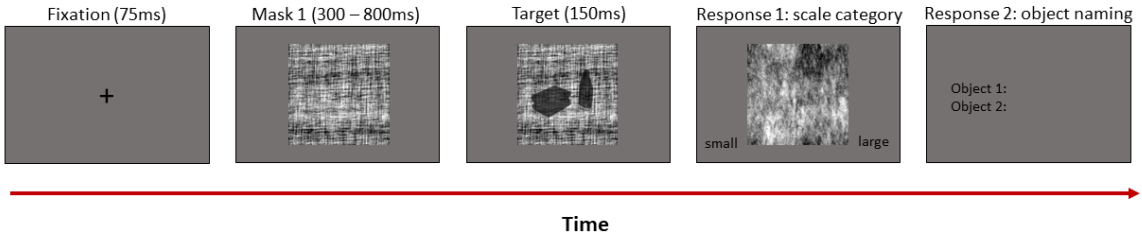
Vighneshvel, T., & Arun, S. P. (2015). Coding of relative size in monkey inferotemporal cortex. *Journal of neurophysiology*, *113*(7), 2173-2179.

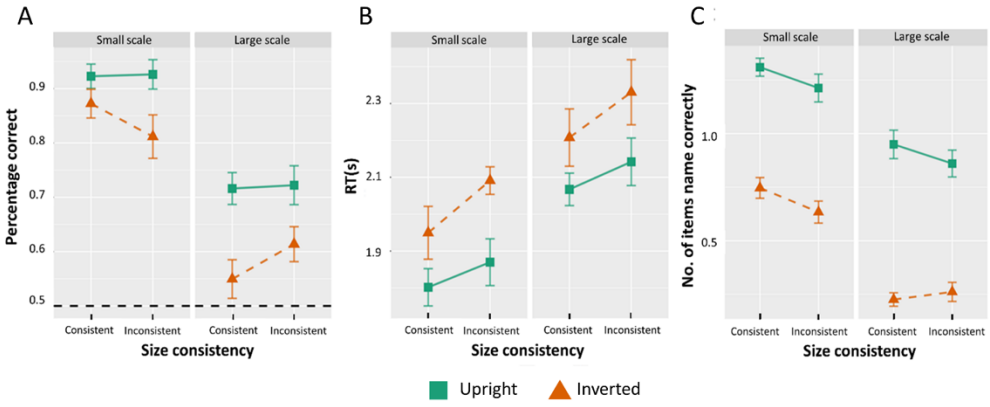
Võ, M. L. H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. Psychological science, 24(9), 1816-1823.

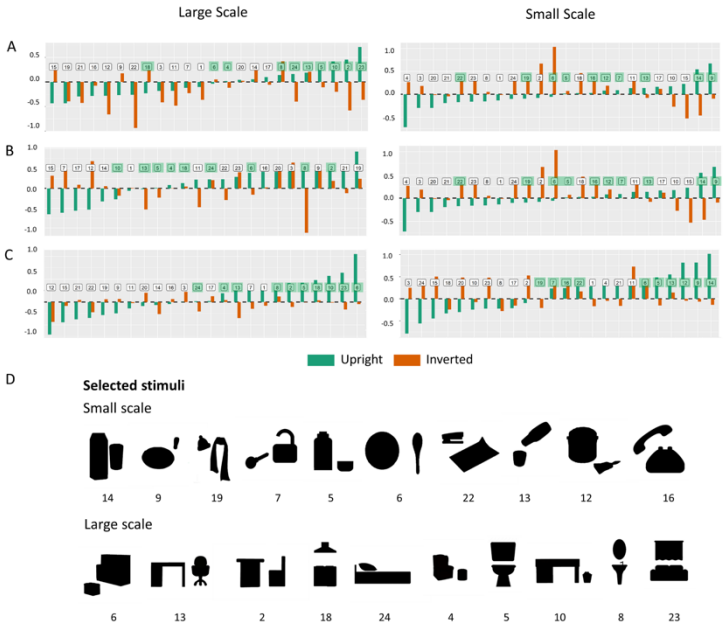
Wolfe, J. M. (2017). Visual attention: size matters. Current Biology, 27(18), 1002-1003.

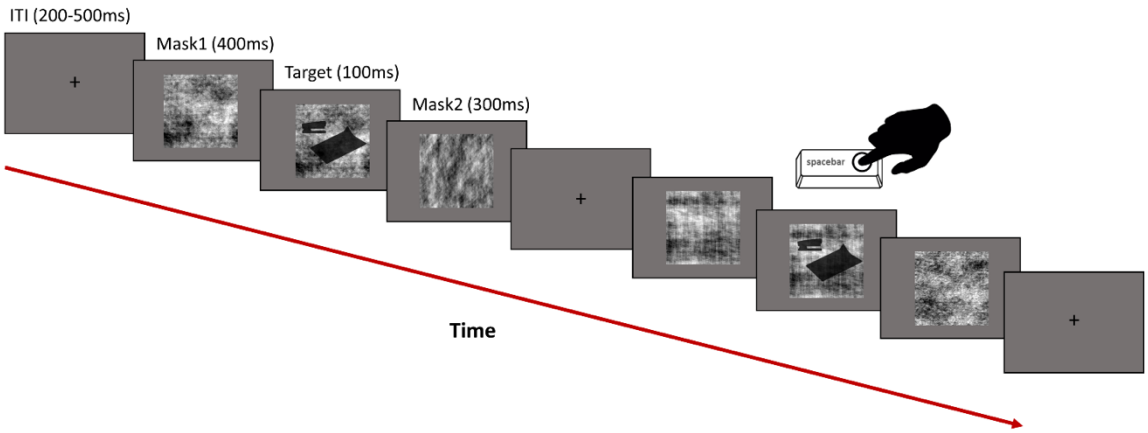
Yu, R. Q., & Zhao, J. (2018). Object representations are biased toward each other through statistical learning. Visual Cognition, 26(4), 253-267.

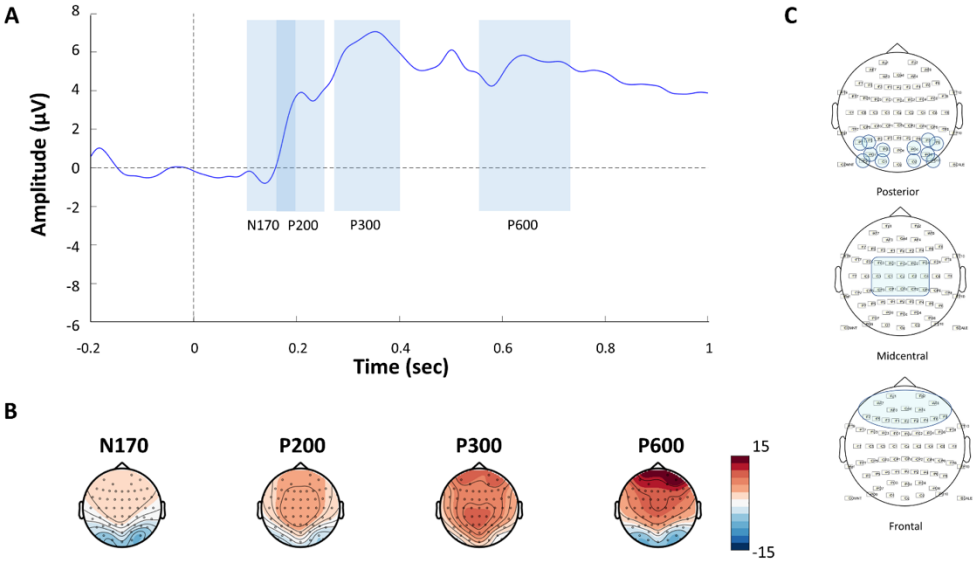

A
Small real-world scale
Large real-world scale
Version 1
Version 2
Version 1
Version 2
Size Consistent
Size Inconsistent
Single targets
B
Consistent
Inconsistent
Consistent
Inconsistent

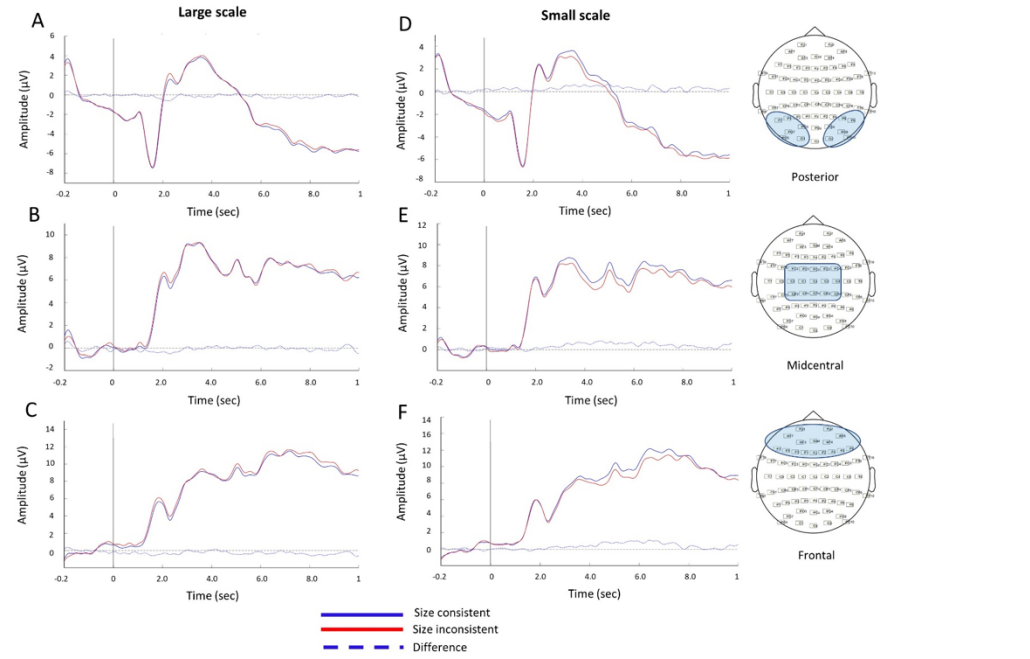

Fixation (75ms)
Mask 1 (300 – 800ms)
Target (150ms)
Response 1: scale category
Response 2: object naming
+
small
large
Object 1:
Object 2:
Time

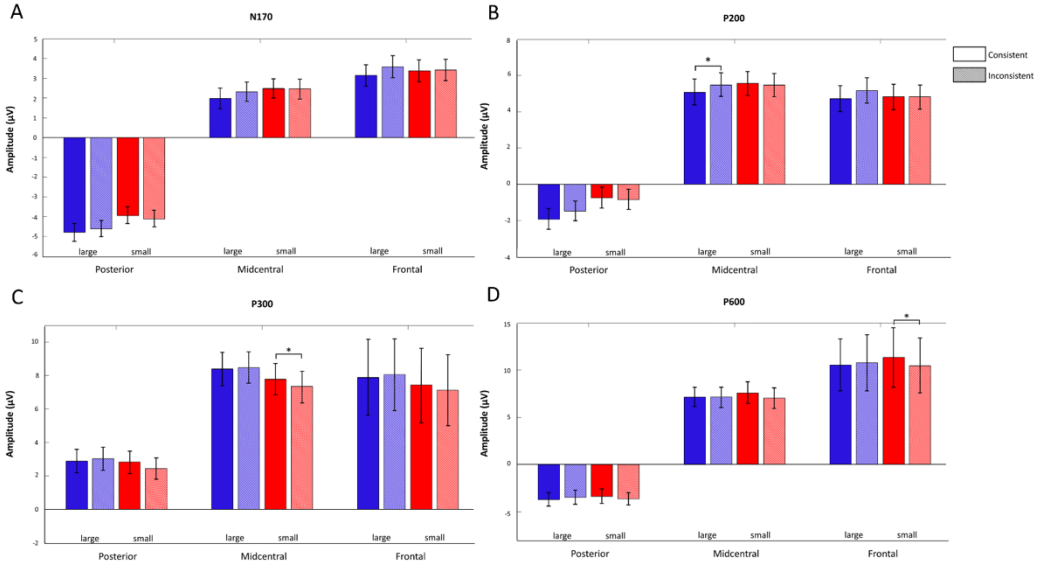

A
Small scale
Large scale
Percentage correct
0.9
0.8
0.7
0.6
0.5
Consistent
Inconsistent
Consistent
Inconsistent
Size consistency
B
Small scale
Large scale
RT(s)
2.3
2.1
1.9
Consistent
Inconsistent
Consistent
Inconsistent
Size consistency
C
Small scale
Large scale
No. of items name correctly
1.0
0.5
Consistent
Inconsistent
Consistent
Inconsistent
Size consistency
Upright
Inverted

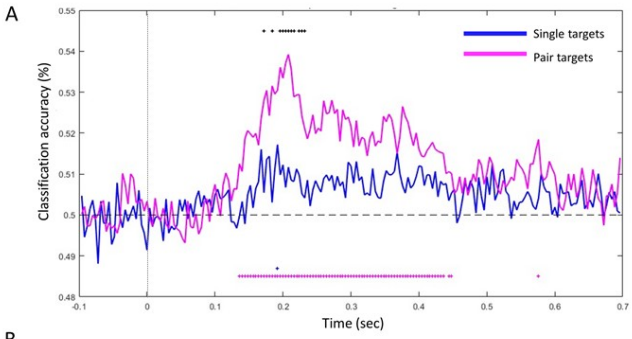

Large Scale
Small Scale
A
B
C
Upright
Inverted
D
Selected stimuli
Small scale
14
9
19
7
5
6
22
13
12
16
Large scale
6
13
2
18
24
4
5
10
8
23

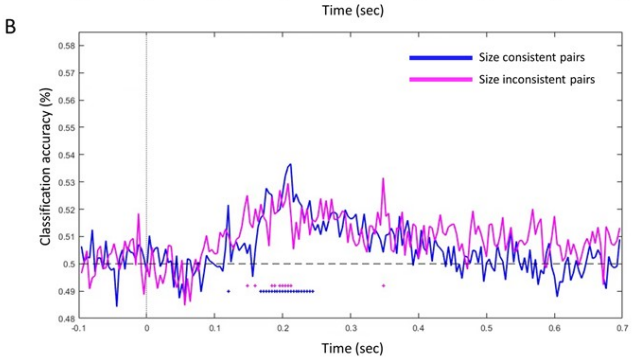

ITI (200-500ms)
Mask1 (400ms)
Target (100ms)
Mask2 (300ms)
spacebar
Time

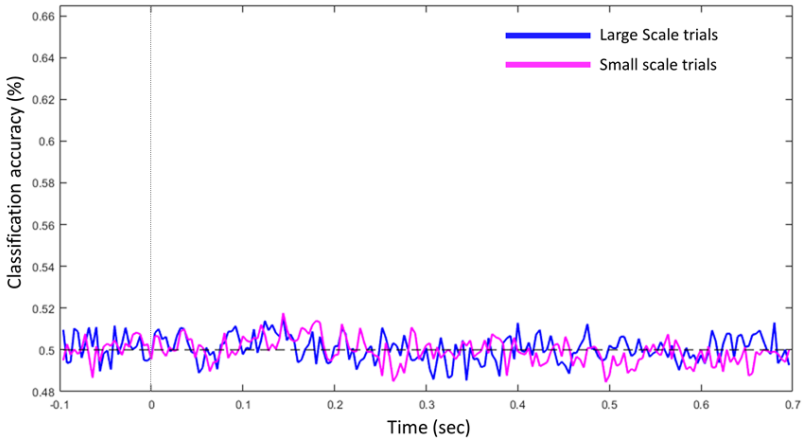

A
Amplitude (μV)
8
6
4
2
0
-2
-4
-8
-6
-0.2
0
0.2
0.4
0.6
0.8
1
Time (sec)
N170
P200
P300
P600
B
N170
P200
P300
P600
15
-15
C
Posterior
Midcentral
Frontal


A
Large scale
D
Small scale
Amplitude (μV)
Time (sec)
B
E
C
F
Posterior
Midcentral
Frontal
Size consistent
Size inconsistent
Difference


A
N170
B
P200
C
P300
D
P600
Amplitude (μV)
large
small
Posterior
Midcentral
Frontal
Consistent
Inconsistent
*


A
Single targets
Pair targets
Classification accuracy (%)
Time (sec)


B
Size consistent pairs
Size inconsistent pairs
Classification accuracy (%)
Time (sec)


Large Scale trials
Small scale trials
Classification accuracy (%)
Time (sec)