UNIVERSITÀ
DI PARMA

DIPARTIMENTO DI MEDICINA E CHIRURGIA

CORSO DI LAUREA MAGISTRALE IN

PSICOBIOLOGIA E NEUROSCIENZE COGNITIVE

MODULAZIONE DELLA MEMORIA A LUNGO TERMINE PER SCENE NATURALI: IL RUOLO DEL COMPITO E DELLA QUALITA' PERCETTIVA DEL CUE SULLA PERFORMANCE DI RICONOSCIMENTO UNA SETTIMANA DOPO

MODULATION OF THE LONG-TERM MEMORY FOR NATURAL SCENES: THE ROLE OF THE TASK AND THE PERCEPTUAL QUALITY OF THE CUE ON ONE-WEEK-LATER RECOGNITION PERFORMANCE

Relatore:
Chiar.ma Prof.ssa Vera Ferrari

Laureando:

Mariagrazia De Gioia

ANNO ACCADEMICO 2021 - 2022

INDEX

ABSTRACT

Visual memory of natural scenes has been described as a large-capacity system (Standing, 1973; Brady et al., 2008; Konkle et al., 2010; Evans & Baddeley, 2018). In this dissertation, we investigate whether this already efficient system can be further improved, using a modified testing effect paradigm. The testing effect is a broadly studied phenomenon in the literature: taking an intermediate test practising the retrieval of the new informations, after the learning phase, facilitates the subsequent retrieval of these informations in a final test. Therefore, an intermediate testing can consolidate and strengthen the memory trace in the long term. Nevertheless, most of the studies that have analysed the testing effect have used verbal material such as prose passages, scientific articles, lists of words and paired-associate materials like foreign language translations. Therefore, in our experiment we assessed the impact of the testing effect for natural scenes: after an initial encoding phase, some pictures went through an intermediate phase, while the remaining stimuli (30 out of 150 pictures) were only tested in the final memory test (two-alternative forced choice) one week later. Moreover, in the intermediate phase, we manipulated the type of task, to assess if the beneficial effects of testing are due to a simple re-exposure to the material (categorization task) or to an active retrieval process (old/new task). Furthermore, the role of the picture's perceptual quality in modulating the testing effect was investigated, manipulating the availability of picture's details during the testing phase: in both tasks, half of the pictures were presented as intact, while the other half were blurred. Results showed that memory performance in the final test was overall enhanced for pictures presented in the intermediate phase, compared to those directly tested in the final test. Furthermore, the type of task performed in the intermediate phase clearly affected the access to the memory trace of the stimuli in the final test: pictures assigned to the memory task prompted the highest recognition performance in the final test. Considering the role of the perceptual quality of the cue, pictures tested in the original version (intact) prompted an enhanced memory compared to the blurred version, and this effect was similar in the two tasks (memory and

categorization), suggesting that the richness of details during testing independently affected memory consolidation, regardless of the type of top-down process activated during the intermediate testing. All together, these results revealed a high potentiality to enhance long-term memory for complex visual scenes by reactivating the memory trace either by a degraded cue or by an intact cue, and this reconsolidation process is extremely effective in a context of active retrieval, compared to when the re-exposure to the stimulus does not directly contact the memory trace.

INTRODUCTION

Memory is a fundamental process in our everyday life: we rely on it to learn new informations and abilities, to solve problems, to remember past experiences and even to make predictions and simulations of possible future events (Schacter & Addis, 2007); it allows us to interact with the surrounding world and to have a stable and coherent representation of us. We can resume all its functions in one phrase: we use memory "to think about the past, live in the present, and plan for the future" (Madore & Wagner, 2022).

Given these premises, memory is a very complex phenomenon that researchers have tried to fathom, trying also to match the biological and neural mechanisms below: there are different stages of memory (encoding, storage, retrieval); distinct memory's structures, that can store different amount of informations for short or long periods; different kinds of memory according to the types of stored informations.

In this dissertation, we will focus on visual memory, especially the long term memory for natural scene pictures, in order to assess if, given the high amount of studies that have observed the massive and highly detailed capacity of this system, the visual memory can actually be modulated: thus, we proposed to use a paradigm largely used in the literature of the testing effect, with the purpose of investigating if an intermediate testing can consolidate and facilitate the subsequent retrieval of the picture on a later test, just like it happens for verbal material. Plus, we manipulated two factors in the intermediate phase: the type of task performed and the perceptual quality of the cue, because we wanted to understand if the consolidation in memory of the picture and the advantage for its later retrieval is due to a simple re-exposure to the material or to an active process of retrieval, and if the possibility to dispose of the details of the picture or not plays a further role.

1. VISUAL MEMORY

Before we go into details of the processes and the structure of the visual memory system, we should first give a definition of what memory is: according to Tulving (2000), the term “memory” can be defined as a “neurocognitive capacity to encode, store and retrieve information”, that can be a visual one, or an auditory one, or even an olfactory one. So, we can consider the type of information involved one of the many characteristics used to define the complex taxonomy of the memory systems.

A way to define the visual memory system is “any memory for which the stored information was acquired initially by the visual system” (Hollingworth & Luck, 2008), but as the same authors assert, this definition is too general and broad. The perfect example is the act of reading: despite the first and fundamental acquisition of the information takes place in the retina (the first step of the visual system that allows the information’s acquisition from the outside world), the text is then remembered not for the way it is written, but for his semantic contents, implicating a different type of memory, that is the semantic memory, where the stored memory trace has lost the visual properties acquired in the first place. Therefore, a more specific and proper definition of visual memory needs to take into account this important passage: “the memory must retain properties of the original perceptual states generated when the memory was encoded” (Hollingworth & Luck, 2008). So visual memories representations, according to this definition, should give rise to a “quasi-visual experience” (Hollingwoth & Luck, 2008), retaining the precise perceptual properties of the previously viewed stimulus. Many studies, by the way, have proved that memory is not that perfect and, most of all, it’s never a passive process like taking a simple photograph of an event: long-term memories are exposed to a constant process of recreation and manipulation, as time goes by. Even the memories of very intense, emotionally speaking, and unexpected events, the so called “flashbulb memories”, that according to Brown and Kulik (1977) take a picture that “indiscriminately preserves the scene”, are not so detailed, accurate and time resistant (Neisser &

Harsch, 1992; Talarico & Rubin, 2004).

This does not mean that visual memories can't retain the exact metric properties of early vision: such memories' type exists, still is very fleeting and is part of the visual sensory memory, also known as iconic memory, but not of the long-term memory. Thus, we can distinguish different levels of maintaining the information about the perceptual properties of a previously viewed stimulus (Hollingworth & Luck, 2008): a low-level representation that preserves all the metric properties of the stimulus, because the iconic memory is thought to possess a virtually unlimited capacity but for a very short period (only a few 100 ms) (Bradley and Pearson, 2012); a high-level representation that discards the precise metric properties, it still contains the details of the stimulus or of the scene but in a more abstract way and not focused on the precise spatial structure.

1.1 The Visual Memory's Structure

One of the most used model that describes the memory's structure is the one elaborated by Atkinson and Shiffrin (1968). This model, that we will use as a reference in this section for the description of the visual memory's structure, divides memory into three different components: the sensory register, where the incoming sensory information is first reported; the short-term store, also known as working memory (that gives the idea of a more active process), that receives part of the informations that were stored in the sensory register and besides reciprocally communicates with the third component of this model, the long-term store; this store is virtually permanent and unlimited, speaking of informations' capacity. These three components differ in the timing of information's maintenance, in the storage capacity and in the mechanisms and processes involved in the maintenance of the informations (Brady et al., 2011). Besides the structural division, the three memory components are constantly in communication one another.

We will now use the same subdivision proposed by Atkinson and Shiffrin to describe in detail the visual memory's components: the visual sensory memory, the visual short-term memory (vSTM) and the visual long-term memory (vLTM).

1.1.1 The Visual Sensory Memory

This component of the visual memory is responsible of the immediate registration and the maintenance, for a very brief period, of a visual stimulus. It can be actually described as a snapshot of the presented stimulus, that fades away soon after. A particular phenomenon that describes quite well this brief but at the same time precise retention of the visual image is the persistence of the vision (Goldstein, 2015, p. 124), that is a continuous perception of a visual stimulus even after it is no more present that lasts for less than a second. This persistence of the vision is very useful, for example, when watching an old movie, composed of a series of photograms: during the brief period of darkness between a frame and another, the persistence of the vision promote the perception of a dynamic scene, instead of separate static images presented in series.

A more systematic description of this phenomenon, already known in the psychological literature, was given by George Sperling (1960), that measured for the first time the duration and the capacity of the visual sensory memory, later on renamed Iconic memory by Neisser (1967). The subjects were briefly exposed to a complex visual stimulus, an array composed of, most of the times, 12 letters (3x4 array) and sometimes numbers, with an exposure time of 50 ms. Sperling noticed that, after this brief exposure, when the subjects were asked to make a total report of the letters presented, the performance was very low (a mean of 4.5 letters) and, interestingly enough, the subjects used to report the feeling that they had seen more than they remembered afterwards. Sperling decided to use a different testing method, a partial report, giving a specific instruction of which part of the stimulus (a specific row of the array of letters) the subject had to report: critically, this instruction was given after the termination of the stimulus. The purpose was to determine if the subject actually had more information available than he could indicate during the total report task. The performance, in a partial report task, was much better: the number of letters reported ranged from 8.1 to 11. After establishing the capacity of this visual sensory system, Sperling (1960) focused on the duration of this store, using a deferred partial report method: the same paradigm was used, but there was a manipulation of the delay's length between the termination of the stimulus and the auditory cue that indicated to the subject the row he had to report: the accuracy of the partial report can be represented as a decreasing function as the delay between the off-go of the stimulus and the instruction increased. In particular, when the delay was of the duration of one second, a substantial decline in accuracy was observed, near to the low performance of the immediate total report. As stated by Sperling (1960), when the visual image starts fading, also its information content decreases and this is revealed by a low performance in terms of accuracy.

After Sperling's study, other authors dedicated themselves to the phenomenon of the persistence of vision, especially Coltheart (1980a, 1980b) that described three different persistences of vision. For some time after the stimulus off-set, according to Coltheart, a subject experiences:

1. Visible or phenomenological persistence: a continuous experience of the visual stimulus, for 1/6 second on average (Sperling, 1960)
2. Neural persistence, of the neural activity evoked by the visual stimulus
3. Informational persistence: the information extracted from the stimulus continues to be registered in a visual form of memory. This type of persistence seems to be the one observed by Sperling (1960) in his experiment. It lasts a little bit longer than the visible persistence, approximately for 150-300 ms after the stimulus offset (Irwin & Yeomans, 1986)

The visible persistence can be described as a consequence of the residual neural activity, not only at a retinal level (Coltheart, 1980a; Di Lollo & Bischof, 1995). Furthermore, many authors provided three effects, related to the visible persistence: the inverse-duration effect, that is a longer duration of the stimulus determines a shorter visible persistence (Cork et al., 2020); the inverse-proximity effect, that is the spatial proximity of subsequent stimuli impairs the temporal integration (Cork et al., 2020); the inverse-intensity effect, even if in several studies this effect wasn't observed (Di Lollo & Bischof, 1995). Nevertheless, these three effects have not been observed in relation to the informational persistence (Coltheart, 1980a; Irwin & Yeomans, 1986), underlining once again the idea that the visible and the informational persistences are two different phenomena.

Even if this visual sensory memory has a really brief duration, some of the informations move to the second component of the visual memory: the visual short-term memory. Thanks to the crucial role played by attention, some selected informations survive and pass to the short-term memory.

1.1.2 The Visual Short-Term Memory

Unlike the visual sensory register, the vSTM has a limited capacity and lacks the metric precision (Hollingworth, 2008), but has a longer duration, even if it's still in the order of seconds (the information here is held for 15-20 seconds) (Goldstein, 2015, p. 127). The studies that focused on the STM capacity, instead, have led to contrasting results, through decades: Miller (1956) stated that the span of this store is 7 ± 2 items, but the use of strategies like chunking, can stretch this "informational bottleneck" (Miller, 1956). Later on, this storage capacity has been reconsidered, using a change detection paradigm. Luck and Vogel (1997) presented to the subjects two arrays, a sample array and a test array separated by a brief delay; each subject, on each trial, was asked to indicate whether the two stimuli differed or not in terms of single features: for example, the two arrays could be different for colour or orientation of one of the items that composed them. The manipulated variable was the number of items in the array, that the authors called set size, to assess the quantity of items that are accurately retained in the vSTM. The set size could go from 1 to 12 squares, that differed in colour, and the sample array was presented for 100 ms. After a 900 ms interval, the test array was presented for 2 seconds and the subject had to perform an identical-different judgement. The performance was very high when the sample array had a set size of 1-3 items, but sistematically declined with the increasing of the set size. It was estimated that the vSTM capacity has a limit set to 4 items. The same result was observed manipulating the feature orientation. The next step was to assess if the vSTM capacity regarded individual features of an object or the whole object as a complex perceptual unit. Using the same task, the authors created another condition: the items of the array (2, 4 or 6 bars) could differ in colour, orientation or both.

In this last condition, the subjects would have retained eight features (four colours and four orientations). The results showed that the visual information in the vSTM is stored not as individual features but as integrated objects: the performance was the same when the difference concerned both one feature and a conjunction of features. Even with a further condition, a conjunction of four features (colour, orientation, size presence/absence of a gap), the performance was still very high. Just like the chunking proposed by Miller (1956), this type of informations' storage makes possible an increase of the vSTM capacity.

Successively, Alvarez and Cavanagh (2004) suggested that the capacity of vSTM depends not only on the number of items, but also by the visual information load. We could explain this idea with a metaphor from the computer's world (Goldstein, 2015, p. 131): when we want to load some images on a pendrive, the amount of images that can be stored depends, first, on the capacity of the pendrive, but second also on the images' size. A smaller quantity of richly detailed images can be loaded on a pendrive, because they take up more space. So, these two authors tested whether the amount of features of an object could affect the capacity of the vSTM: the visual information load they talk about refers to the object's total visual details stored. They used five different stimuli (letters, colored squares, chinese characters, random polygons and shaded cubes) within a change detection task with an array of a variable number of objects from 1 to 15; on 50% of the trials, the sample array and the test array were identical. The results showed that the vSTM capacity is not constant across the different visual stimuli, because "the greater the information load of each item […], the fewer objects from that class one can hold in memory" (Alvarez & Cavanagh, 2004): taking constant the number of items, the performance was lower when the arrays presented on the screen were composed of very detailed objects (for example, the cubes). By the way, this experiment also repeated what was first observed by Luck and Vogel (1997): the maximum capacity of the vSTM is 4. From these results, the authors stated that both the visual information load and the number of objects determine the vSTM capacity limits.

1.1.3 The Visual Long-Term Memory

Differently from the previous visual memory stores, the vLTM has a larger capacity and can store informations for long periods of time, from a few minutes to a lifetime. One of the first studies about the huge and detailed capacity of vLTM was led by Standing (1973), that observed also a pictorial material superiority to the verbal material, using a forced-choice recognition task. Each stimulus was presented for 5 seconds, with an ISI of 600 ms; three types of stimuli were used: 11.000 normal pictures, 1200 vivid pictures and 25.000 most common English words. The set size was variable (20, 40, 100, 200, 400 or 1000 stimuli), but for Normal pictures, some subjects observed 4000 or 10.000 items. Two days later, the subjects performed a recognition test (20, 40, 80 trials, using a Two-alternative forced choice task: this task provides that two stimuli are presented side by side, one of the two is a stimulus that the subject has already seen during the encoding phase, the other one is a new stimulus; the subject has to indicate the position (right or left) of the old stimulus. With the two biggest set size (4000 and 10.000), the subjects observed 2000 images per day; in the case of the 10.000 images, the fifth and last day of encoding phase, the recognition test was immediately performed. In general, visual stimuli were better remembered than verbal stimuli and, in particular, vivid pictures were better retained compared with normal pictures.

This study can be considered as a pioneer of a long series of studies, till nowadays, that replicated this massive storage capacity (Brady et al., 2008; Evans & Baddeley, 2018). In particular, Brady and colleagues (2008) tried to estimate not only the amount of items that can be remembered but also the level of detail (not only the gist) with which each item is remembered. The subjects observed 2500 objects, for 3 seconds each, not embedded in a scene, to control the role of contextual cues that could explain the high performance in previous experiments. Then, 10 minutes after, the subjects performed a 2-afc task, using three types of pairing: a new object from a different category (novel condition); a new object from the same category (exemplar condition); the same object presented in a different state or pose (state condition). The performance, for each type of pairings (also the exemplar and state conditions that required memory for details to perform the task well), was very high, respectively 92%, 88% and 87%.

Konkle and colleagues (2010) decided to use more complex stimuli (scenes) in order to assess the fidelity of the stored scene representations, as was previously done for object images (Brady et al., 2008). In this study, the subjects observed 2912 images from 128 different scene categories, with a variable number (1, 4, 16 or 64) of exemplars presented per category. Twenty minutes after, the subjects performed a two-alternative forced choice task, created using two pairing’s types: an old image paired with a new one from a novel scene category or with an exemplar from the same scene category. The performance was better (96%) when the new item belonged to a different category from the old one, replicating previous results that described the vLTM as a massive storage system. But what about fidelity? The performance obtained with pairings composed of items from the same scene category can give us an answer: the performance was significantly lower, despite it was still remarkable. Plus, the number of items from the same category, observed during the encoding phase, influenced the performance too: the more the exemplars observed, the lower the performance obtained during the test (84% with 4 exemplars, 80% with 16 exemplars, 76% with 64 exemplars). By the way, these results demonstrate that the vLTM can store a significant and sufficient amount of details to distinguish a scene, not only objects, from many other exemplars of the same category.

A step forward was made by Evans and Baddeley (2018), that tested two important aspects: the awareness or not that the memory will be tested after the encoding phase, so that the subject could activate some strategies during the encoding to better perform in the subsequent task; the amount of attentional capacity available during the encoding. The stimuli used in this study were complex scenes, from complex manmade scenes (cityscapes, landscapes, waterscapes and interiors) to door scenes where some distinctive features could be removed. During the encoding phase, to ensure the processing of the stimuli, the subjects were asked to make a judgement of pleasantness or a judgement of presence/absence of a dot. Given the high level performance observed in many studies about the vLTM, the hypothesis of the authors is that the encoding of visual material, unlike the verbal material, requires less resources and an effortful elaborative processing, less intent and may be more resilient to load: according to this view, the visual features should be rapidly and richly encoded. In Experiment 1, 400 complex scenes stimuli were used, while in the other 304 door scenes were presented; the subjects were assigned to intentional or incidental memory group and subsequently tested in an old/new task. As expected, no enahncement in performance was observed in the intentional memory group compared to the incidental one. To prevent the possibility that the judgement task performed during the encoding phase could have induced an effective learning strategy because of a deep processing's level, a different task was used, detection of a dot, that also required a scanning of the scene but a more superficial processing. This second experiment was realized, starting from the results obtained by Baddeley and Hitch (2017), that manipulated the level of processing for visual stimuli and compared this LOP effects between visual and verbal stimuli: a deep processing consisted of judgement of pleasantness, while a shallow processing consisted of indicating the stimulus color. The results showed, for visual stimuli, a small positive effect of LOP, even if the magnitude of this effect was significantly less than the effect obtained with verbal stimuli. This pattern led the authors to suggest that, unlike verbal stimuli, visual stimuli have less potential for further semantic elaboration, because the relevant diagnostic cues may be more dependent on perceptual features. However, coming back to Evans and Baddeley's study, once again, the intention to learn didn't represent an advantage for the performance on the test. However, when the encoding condition was made more difficult, which means that door scenes stimuli with peculiar features removed were used, intention to remember did produce an advantage in the test's performance. When the attentional capacity available during the encoding phase was manipulated using verbal tasks, it was observed that visual memory is quite sensitive to distraction, with greater impairment when more demanding tasks were used (backward counting by threes) for all types of material used. In general, the authors suggested that the little effect of intention reflects the fact that visual stimuli may encode a high amount of visual features rapidly and in parallel: this does not mean that our vLTM can encode and store all the features of the scene, but certainly a sufficient amount of discriminative features to successfully perform during the old/new task; it is when these necessary diagnostic features are removed, that the awareness of being tested soon after encourages the subjects to rely on a memory strategy. This parallel and rapid encoding of the relevant features of a scene, however, is not entirely automatic, since it is influenced by the attentional capacity available. To conclude, the authors proposed a two-level processing, influenced by the encoding conditions: a first stage, based on the gist of the stimulus, that is rapid and not influenced by the intention to remember; a more detailed stage that is used when the diagnostic features are not present, in order to retain more details that could be necessary to later remember the stimulus.

One important aspect to take into account is what happens to the vLTM over time: through the remember/know procedure it's been possible to observe that the recognition process can be divided in two different processes. In the first case, called recollection, the old stimulus is remembered as familiar and the subject can also retrieve the circumstances of when he first saw the stimulus, it means that the subject can recollect the encoding experience and that the episodic memory is involved; in the second case, called familiarity, the old stimulus generate in the subject a sense of looking familiar, but the subject can't get access to the episodic memory of the event. The loss of episodic details is called semantization of past memories (Goldstein, 2015, p. 166) and is a peculiar process that involves the oldest memories.

1.2 Visual memory for Natural scenes

As previously said, the studies that have focused on the characteristics of the vLTM, have used different type of stimuli, from objects context-free (Brady et al., 2008) to more complex stimuli, like natural scenes (Konkle et al., 2010). Thus, the focus in this section will be the specific material of natural scenes, since it is the material we used in the experiment that we will soon describe in this dissertation.

Many recent studies have focused on the memory for visual scenes, certainly because we spend most of the time perceiving, moving and behaving within complex scenes. Experiments like the ones described above, such as those led by Konkle and colleagues (2010) and Evans and Baddeley (2018) have demonstrated that vLTM for scenes is actually a massive system, specific enough to successfully discriminate between a high amount of natural scenes' stimuli, but they do not clarify the nature of the stored information (Hollingworth, 2008). So, one important question regards if the high capacity of vLTM is due to either the retention of scene gist or actually the retention of visual details of the scene. Already Standing and colleagues (1970) obtained an evidence that representations of natural scenes in vLTM can preserve specific visual details: the participants observed 120 pictures for 2 seconds and then performed a final test, after either 30 minutes or 24 hours: the subjects were tested both for the recognition memory using a Two-alternative forced choice and further for the orientation of the picture, that is they had to specify if the orientation was the same as during the encoding phase. If it's true that natural scene's representation is limited to the scene gist, then the information about the picture's orientation should not be encoded, thus the subjects' perfomance would be at a chance level. Yet, orientation judgements, after 30 minutes, were very high, but became poorer, even though well above chance, after 24 hours.

More recently, Hollingworth (2005) investigated the relationship between online visual representation of natural scenes and their long-term memory, from less than one minute to 24 hours. Already the previously described studies demonstrated that the LTM representation of a natural scene is quite detailed, but in this experiment subjects were asked to remember information more specific than the "simple" scene gist and, as a result, was observed a good ability to recollect specific object details of a scene and also the robustness of this representation over time. Using a change detection task, the subjects observed an initial natural scene picture for 20 seconds; then, they performed either an immediate test or a delayed test (after all scenes had been viewed), where in each picture an object either remained the same or changed because of a rotation in depth or a replacement by another object from the same category. In the test phase, it was specified which target object the subjects had to make judgement about, to ensure a retrieval and comparison processes aimed at the target. The results showed that performance in the delayed final test was only modestly reduced compared with the performance obtained in the immediate test; plus, 24 hours later, even though there was a significant forgetting, performance remained well above chance. Taken together, these results demonstrated that vLTM stores quite detailed representations of a visual scene and that they also are relatively stable and robust over time.

The next question involving the vLTM is how the representation of scenes is formed, since it is the result of a sequential and local acquisition of visual informations that compose the environment and that are bound together to form the final, comprehensive and more complex scene representation. This is caused by a structural limit of the human eye that cannot detect all the visual details of a scene in a single glance, since our eye disposes of a relatively small area (fovea) where the visual acuity is maximum: therefore, visual scene perception is extendend over time and space (Hollingworth, 2005). Hollingworth (2006) tried to verify if actually single objects of a scene are represented as a part of a lager representation or as independent from the context in which they are embedded, so the purpose was to demonstrate if there is an object-to-scene binding. Natural scenes pictures were presented for 20 seconds each; then, with a 2-AFC task, the subjects were asked to identify the original target object, presented with a distractor stimulus composed of a different token or the same object with a different orientation. To verify the possibility of an object-to-scene binding, some pairings were assigned to the background present condition and some others were instead assigned to the background absent condition, which means that the target objects were presented isolated from the context. The results demonstrated the presence of object-to-scene binding, which means that the “object representations are episodically bound to scene context”, since the performance in the final test was better for objects assigned to the background present condition.

Is there also a spatial, object-position binding? In the same work, Hollingworth (2006) verified this hypothesis presenting, in the 2-AFC task, the target object either in the original position or in a different one; the performance was better when the target object was presented in its original position, revealing a same-position advantage. “Thus, visual object representations are not only bound to scene context but are bound to particular locations within a spatially organized representation of the scene” (Hollingworth, 2008).

1.2.1 The role of global and local information in Natural scene identification

When we observe a picture, like a natural scene, we start to analyse the visual informations that compose the image, in order to make a final correct identification of the stimulus. But, through the years, an important question arose: which is the role of the two information’s levels (global and local informations) of a scene representation in its correct identification? In other words, which of the two informations is first processed in the perception and in the following identification of natural scenes? We can consider the situation from two points of view (Castelhano et al., 2019): the global contents precede the processing of local details and provide an organizing principle for a subsequent integration of the local details; the local details precede the processing of global contents, which means that only in a second moment the local features are collected together in a more global contextual representation. This is an important step in our visual search process, since it is governed by memory and the way the environment in which we move is organized (Brooks et al., 2010). Thus, knowledge of the typical location of objects in a specific natural scene and its structure’s memory bring the subjects to guide the visual search in regions where it is more probable to find the target object. Using a contextual cuing paradigm, subjects are asked to search for a specific target in a natural scene picture; some pictures are presented more than once, and this allows the subjects to learn the target location in relation to a particular context. But exactly what is this context? Indeed, the target position could be learned in relation to a local context (for example, the closest and immediately adjacent objects) or in relation to a global configuration.

Starting from works that studied this relationship on a perceptual level, many authors have suggested the presence of a “global precedence hypothesis”, like the study realized by De Cesarei and Loftus (2011), since it was demonstrated that subjects can accurately identify visual scenes also in degraded conditions, when fewer incoming visual detailed information are available. The stimuli presented were manipulated in their spatial frequencies, using a low-pass filter and a high-pass filter. Plus, the time of exposure was manipulated. The results showed that low-spatial frequency information is available earlier in time and, according to the authors, this information is then integrated with high-spatial frequency information that eventually maximaze the identification efficiency.

A clearer explanation of how this encoding strategy (that gives priority to global contents) works, in relation to the memory fo natural scenes, is given by Brooks and colleagues (2010): “when a consistent environmental structure is available, spatial representations supporting visual search are organized hierarchically, with memory for functional sub-regions of an environment nested within a representation of the larger scene”. Despite previous studies (Olson & Chun, 2002) had recorded that the spatial contextual cueing was driven by local context informations, Brockmole and colleagues (2006), using the same paradigm but a different type of stimuli (natural scenes) than arrays composed of simple and artificial nonscene stimuli, showed that the subjects are biased to associate the location of a target object with the global context. In their experiment, the subjects had to search for a target letter contained in the scene picture, for example the target was placed on a coffee table in a library and this picture was presented multiple times during the learning phase; the use of a letter as a target ensured that its location could not be predicted by previous knowledge of object placement in a specific scene context, the so called “scene schema” (Mandler & Ritchey, 1977), consisting in an abstract representation of a specific scene type that contains an inventory of objects typically found in that scene and their typical spatial positions. Then, during the transfer phase, the repeated stimulus was replaced by a new one, in which either the local aspects (a different table) or the global context (a different room) was changed: only the change of the global context eliminated the contextual cueing. Even when, in a second experiment, only the local context was repeated, a little facilitation was observed; whereas only when the global context was repeated, a significantly better contextual cueing was observed. According to the author, thus, the local context is used only in a second moment, when the global information is not predictive, is absent or is difficult to derive. Starting from these conflicting results, Brooks, Rasmussen and Hollingworth (2010) tried to shed light, using different terms like sub-region of a scene, since it is difficult to define precisely the global-local context in natural scenes. They proposed, observing their results, that the informations about specific subregions of a real-world environment bound to locations within a representation of the scene at a larger spatial scale; thus, the subject must primarily get access to the representation of the broader scene, in order to, in a second moment, get access to the informations about the structure of the scene's subregions. "For example, although the position of the milk is unlikely to be remembered relative to the large-scale spatial structure of the kitchen, retrieval of one's knowledge of the structure of the refrigerator depends on first recognizing the kitchen" (Brooks, Rasmussen & Hollingworth, 2010).

A different explanation of how a natural scene is represented in our memory, is offered by Castelhano and colleagues (2019) with the "parallel memory model": rather than a hierarchy in a scene structure, they hypothesized a parallel model that stores flexibly the various levels of informations, that vary in strength, and, critically, can be accessed independently, while in the hierarchical model the global information must be retrieved as first. The authors compared the two models with the purpose to find which one best captures how a scene is represented in our memory. They used as usual a contextual cueing paradigm, and created, together with normal natural scene, the so called "chimera scenes", where the background and the foreground were not semantically related. The experiment was composed of three phases: learning phase where the subject had to search for a target letter which could appear either in the background (global information) or in the foreground (local information) of a normal or a chimera scene and, for the repeated scenes, the target appeared always in the same position; transfer phase, composed of a local change condition (the foreground was changed, while the background remained the same), a global change condition and a no change condition; memory phase for the natural scene pictures without the presence of the target, once the subjects defined a picture as “old”, they had to indicate the target location. According to the parallel model hypothesis, opposed to the hierarchical model, the local scene information should be retrieved regardless of a change in the background; plus, both with a local and a global change, the search times should be faster than novel scenes. Analysing the search time in the transfer phase, the authors found mixed results: as predicted by the hierarchical model, search times in the local change condition, with constant global information, were faster than novel scenes, for both background and foreground targets, and instead they did not differ from no change condition; but in the global change condition, the situation differed in part from the predictions advanced by the hierarchical model. When the target was presented in the background, the results were actually consistent with the hierarchical model, with global change scenes’ search times significantly slower than no change condition and did not differ from novel scenes; in contrast, when the target was presented in the foreground, in the global change condition, search times were significantly faster than novel scenes and did not differ from no change condition. These results perfectly explain the parallel model hypothesis, that states that global and local informations varies in strength, on the grounds of the search task prioritizing the former or the latter. Thus, to summarize the results, when the target was presented in the background of the natural scene, the global context was emphasized and the performance obtained may appear more consistent with a hierarchical model’s prediction, therefore global changes eliminated the contextual cueing; however, for foreground targets, only a subregion of the scene was emphasized, so global changes had only a small effect on the contextual cueing, as predicted by the parallel model. So, the representation of a natural scene can actually be described as flexible, retrieving independent various levels’ information depending on the task’s demand.

1.3 Summary

In this chapter, we have explored the vast and complex field of the visual memory, depicting the principal characteristics of this system, in particular the three visual stores and their processes, and how the representations of visual stimuli, with a regard for natural scenes, are created, stored and then retrieved. The first component of the visual system route that ends with the perception of the visual stimulus and the creation of its representation in the memory system, is the Visual Sensory Register which, as the name indicates, is responsible of the immediate registration of the visual stimulus, preserving its metric features just like a snapshot (Sperling, 1960); the passage from this store to a more lasting one (in the order of seconds) but with a limited capacity is assured by the attentional system, which selects some informations that survive and pass to the Visual Short-Term Memory, otherwise they decay very quickly (Sperling, 1960; Irwin & Yeomans, 1986). This representation differs from the one created in the previous store, since it lacks the metric precision (Hollingworth, 2008); then, we have presented some studies, that focused on the STM capacity, starting from Miller's "the magical number seven, plus or minus two" (1956) till more recent studies, that have downsized this capacity (Luck & Vogel, 1997; Alvarez & Cavanagh, 2004). The last step is the passage from this temporary store to the Visual Long-term Memory. The studies for this store, have focused on different aspects and, summarized, they show us that the vLTM is a massive system (Standing, 1973), quite detailed not only for objects (Brady et al., 2008), but also for more complex stimuli, like natural scenes (Konkle et al., 2010; Evans & Baddeley, 2018); this high capacity is also connected to the retention of specific visual details in a scene (Hollingworth, 2005), not only the gist, plus they are maintained for a long period of time.

At the end, we have focused on the relationship between the global and local informations in a natural scene, that have been studied using the spatial contextual cueing paradigm: despite some studies (Olson & Chun, 2002) have observed that the local details precede the processing of global contents, which means that the contextual cueing was driven by local context informations, these studies haven't been considered generalizable to the visual scene, since they used a completely different kind of visual material (arrays composed of simple and artificial nonscene stimuli).

Authors like Brockmole and colleagues (2006) and later Brooks and colleagues (2010) obtained results that led to the formulation of the hierarchical model, where the priority is given to the global information, first processed than the local one. In particular, they proposed a nested representation of a natural scene where local informations that they call "functional sub-regions" (Brooks et al., 2010) are nested in a representation that is based on the larger scene. Differently, Castelhano and colleagues (2019) have proposed a parallel model, where instead the local and global informations of a natural scene are stored and can be retrieved flexibly and independently, depending on the task's demand, that can prioritize the former or the latter type of information.

2. THE TESTING EFFECT

Before starting to describe the experiment we conducted, in this section we are going to review another paradigm related to the LTM and the creation and consolidation of the mnestic trace, a paradigm that has been used a lot with verbal stimuli and has obtained solid basis in the literature: the so called “testing effect”.

Two authors that have taken an interest in this phenomenon are Roediger and Karpicke; they defined the testing effect as a “phenomenon of improved performance from taking a test” (Roediger & Karpicke, 2006a). In particular, taking an intermediate test and practising the retrieval of new informations, shortly after the learning phase, facilitates the retrieval of these informations in a subsequent moment, because of a slowing down of the forgetting process and a consolidation of the learned informations in the long term (McDermott, 2021). The testing effect is based on the process of retrieval practice, which is, according to McDermott (2021) “one of the most effective ways of solidifying new knowledge”.

How is that possible? We can refer to two important processes (McDermott, 2021): an iterative process, which means that when an information is retrieved, its memory trace in the LTM is consolidated and this makes the trace more accessible over time; an interactive process between encoding and retrieval, in fact retrieval can be described as a re-encoding of the event. As stated by Bjork (1975; Bjork & Bjork, 1992), retrieval is “a memory modifier”: it is actually a mechanism that leads to an updating of the memories (Goldstein, 2015, p. 199), so that when a memory is retrieved, it returns to a fragile condition as it was when first created, that allows the memory to be modified or, in case, even eliminated; then, the memory is consolidated again, so that the “the act of retrieval is itself a potent learning event” (Bjork & Bjork, 1992).

In the present chapter we will review some of the studies that have focused on this phenomenon and the possible underlying explanations and mechanisms, before moving to a specific type of stimuli: the testing effect for natural scenes.

2.1 Evolution of the testing effect and its classic paradigm

Even before in the literature researchers started to talk about the testing effect (also called retrieval-based learning), Gates (1917) demonstrated the advantage for the memory of recitation, a self-testing approach that consists in alternating reading, attempts to recite and consulting the text to get a feedback when the information's retrieval fails. He gathered a group of students from the first to the eigth grade. His purpose was to "answer a practical question of the school-room", so all the study was realized in a more ecological context, in real schools and not in a laboratory. The materials used were nonsense syllables and biographies; Gates gave 9 minutes to learn the new material, manipulating the time spent in recitation (from 0% to 80% of the 9 minutes). The instruction for recitation was the following "Now you are to try to say to yourselves as many of the syllables as you can without looking at the card. When you cannot remember the next word look down at your card and then go on saying as many of them as possible without looking. Glance at the card again whenever you cannot remember" (Gates, 1917). At the end, students performed a final free-recall test; for the nonsense syllables, the results were quite clear (while for the biographies, the results were mixed): the more time spent during the encoding phase in recitation, the better the final test's performance.

The first evidence of the testing effect was obtained by Spitzer, that also defined the paradigm that is still used nowadays, with some modification. Spitzer (1939) recruited elementary school students that had to read an article and then they could be subjected to a multiple-choice test without feedback. Spitzer created many conditions, dividing the students into ten groups: some were not tested at all after reading the article; some instead took an initial test, manipulating the delay between the reading phase and the testing phase. The results showed that taking a test, especially with a short delay from the study phase, attenuates the forgetting curve, so a more delayed intial test led to a lower performance in a final test.

It's with the works of Roediger and Karpicke that the testing effect has obtained very strong evidence of its efficacy, introducing a control group that, differently from previous experiments, restudy the new material, which means that the subjects could re-experience the whole previously learned material. An important step forward in the two experiments led by these two authors (2006b) is that they investigated the testing effect under educationally relevant conditions and using a different type of testing (free recall test) without feedback. In experiment 1, the college students studied two short prose passages; one passage was then restudied, while the other one was tested through a free recall task. The final test was performed 5 minutes, 2 days or 1 week later: after 5 minutes, the students that had the opportunity to study twice the material, recalled more informations than the test group, but this tendency was reversed when the final test was performed 2 days and 1 week later. Another interesting finding is that the performance in the final test 1 week later, for the test group, was even slightly better (56%) than the performance of the restudy group that had the final test 2 days after (54%), indicating that "taking an initial recall test prevented forgetting of information for an additional 5 days relative to repeated study" (Roediger & Karpicke, 2006b). In the second experiment, the authors investigated the effects of repeated intermediate testing, demonstrating once again that repeated testing is a powerful tool for preventing forgetting and for long-term retention. These two experiments supported, more than previous studies that had no test group as a control condition, that the advantage obtained from testing new material is not purely due to an additional exposure to the material, since the performance was better than the restudy group. Using a metaphor from the world of physics, they described like this the phenomenon "Just as measuring the position of an electron changes that position, so the act of retrieving information from memory changes the mnemonic representation underlying retrieval - and enhances later retention of the tested information." (Roediger & Karpicke, 2006a).

Despite these strong results, constantly replicated by lots of studies, the retrieval-based learning as a powerful and effective study method is still underestimated and not much used by students: Roediger and Karpicke (2006b), when asking how well students thought they would have remembered the learned material in a final test one week later, the ones who had performed an initial test, instead of a restudy opportunity, predicted that they would do worse, while students from restudy group were more confident that they would remember the passage one week later even though they actually obtained a lower performance on the final test. This suggests that students "lack metacognitive awareness of the mnemonic benefits of testing" (Karpicke et al., 2009), preferring rereading their notes or textbooks instead of self-testing and practising retrieval: the short-term benefits produced by repeated studying (Roediger & Karpicke, 2006b), called "illusions of competence" (Karpicke et al., 2009) may be the reason why students believe in a better performance in the future.

During the years, the testing effect has been studied manipulating many variables, in order to determine the generalizability of the effect. In the review written by McDermott (2021) are cited many works that demonstrate this statement: in relation to the material used, a testing effect has been observed with both verbal (articles, prose passages, word lists ecc.) and nonverbal / visuospatial (faces, maps, ideogrames ecc) stimuli, that will be the focus in the next section of this chapter. An advantage from testing has been observed also among different age groups, involving not only different grade's students which are certainly the most studied population in the testing effect (since it has an incredible potential for educational learning), but older adults too: Rogalski and colleagues (2014) tested subjects from 60 to 75 years with a not only immediate, but also delayed improvement for retrieval of the material that was previously tested; in a study of Meyer and Logan (2013) older adults' performance was also compared with both students and non-students young adult groups, observing a similar benefit from testing, in a delayed performance too.

2.2 Testing effect with visual material

As already mentioned above, most of the studies that have analysed the testing effect are characterized by the use of verbal material such as prose passages, scientific articles, lists of words and paired-associate materials like foreign language translations (Dunlosky et al., 2013). Even though this type of material is the most representative in the testing effect's literature, other studies have been planned using visual material, that will be the focus of the present dissertation. In this section we will examine the studies that have used visual material.

Some interesting studies have paired visual and verbal materials, for example lots of studies have used face-name pairs (Landauer & Bjork, 1978; Morris & Fritz, 2002; Carpenter & DeLosh, 2005) observing that the names associated to a face and tested during an intermediate phase were actually better remembered than restudied pairs, after a retention interval from 5 minutes (Carpenter & DeLosh, 2005) to 30 minutes (Landauer & Bjork, 1978) and also a good performance two weeks later (Morris & Fritz, 2002); other studies used different visual-verbal materials pairings like pictures and foreign language translations, objects and names, obtaining a support of this effect for visual and spatial information too (see Dunlosky et al., 2013 for a review of studies that observed the testing effect for visual materials).

Jacoby and colleagues (2010) investigated the testing effect in relation to the classification learning of natural concepts, in this particular case exemplars of bird families. In the repeated study condition, birds' images were presented with their family name; in the repeated testing condition, birds' images were presented and the subjects were asked to retrieve the correct category label, followed by feedback. For the final test, the subjects performed both a recognition test (old/new task) of birds' pictures that they had previously seen during the study phase or not, and a classification decision task, where eight family names appeared below the bird's picture (the task was performed both for old and new images) and the subjects had to choose the correct family name to which the exemplar belonged. The results demonstrated that both of the final test tasks were facilitated by repeated testing.

Also the visuospatial learning seems to benefit from retrieval practice. Carpenter and Pashler (2007) tried to replicate the testing effect for map learning. Each subject studied two maps, containing 12 features each; one map was assigned to restudy condition, while the other to intermediate testing. For the last condition, the subjects observed the map but with one feature deleted; the subject was told to figure out what was missing and trie to covertly recall the feature with its exact location; then, as a feedback, the complete map was showed. Differently from the few studies that have used visual stimuli that still required verbal responses, for example to remember the name of the objects or faces, in this study a visuospatial output was required; thus, 30 minutes later, the subjects drew the two maps; the performance was significantly better for testing condition's map. Another experiment designed by Rohrer and colleagues (2010) tested the benefits of retrieval practice for spatial informations: students had to learn the location of regions and cities on a map. In this study too, the previously tested items were better remembered than items assigned to the study-only condition.

However, as Kang (2010) observed, in studies like the ones previously described, it is still possible that some form of verbal elaboration mediated the effect observed by the two authors, since location of features could be verbally coded. As a consequence, Kang investigated the testing effect using abstract visuospatial information, Chinese characters. In particular, the author used a paired-associate learning, where English words were paired with the Chinese equivalent (of course, the subjects had no experience of Chinese language). Here, like the previously described study, a visuospatial output was required: drawing the Chinese characters, that are relatively difficult to verbalize. Subjects were assigned to a restudy group where the pairings were presented three times, or to the retrieval practice group, where the pairs were presented one for the study and twice for testing (an English word was presented and the subject should trie to covertly retrieve the Chinese character) followed by a feedback. The final test was performed 10 minutes or 24 hours later. As a result, a testing effect also for visuospatial stimuli was observed, in comparison with a restudy condition, both after a 10 minutes and a 24 hours delay. Plus, just like it happens for verbal stimuli (Karpicke et al., 2009), the subjects lack metacognitive awareness of the advantage gained from the testing. Another interesting observation is the advantage obtained from the retrieval practice, which is still present even when no overt response is emitted during the testing phase: the same beneficial effect, even without an overt response, had been already observed with verbal material (Putnam & Roediger, 2013). All these results suggest that the testing effect does not entirely and exclusively depend on verbal strategy. Therefore, as stated by Kang (2010), this evidence of testing effect for visual material, once the verbal strategy is excluded, does not weaken the elaborative retrieval hypothesis (Carpenter, 2009), relevant to verbal stimuli, but suggest that another mechanism could be involved when visual material is used. With this theory, Carpenter (2009) sustains that, in a paired-associate condition, when the subjects try to recall the target word, they activate a network of semantically related items; these multiple pathways should help to get access to the target item on a later test, while in a restudy condition the information is currently presented. By the way, we decided to not present and describe this theory in the next section, where we are going to list the models and theories that have been proposed in order to explain the testing effect and that could be used to explain the same effect in the field of visual material, even if some of them have been disconfirmed for verbal material. First, because the elaborative retrieval hypothesis fit only with verbal stimuli, since it is difficult to imagine that nonverbal materials like symbols, that are non verbalizable, could activate semantic mediators (Karpicke, 2017); second, because many recent studies (Karpicke & Smith, 2012; Lehman & Karpicke, 2016) have demonstrated that the processes and the effect on a long-term retention prompted by the retrieval practice are different and independent from an elaborative processing of the material. As Karpicke and Smith (2012) have stated, the results "cast doubt on the idea that the mnemonic effects of retrieval practice stem from elaborative encoding", since it was demonstrated that not only the generation of the semantic mediators wasn't more likely during retrieval practice than during restudy condition, but also that these mediators didn't actually facilitate later recall of the targets (Lehman & Karpicke, 2016).

Another study, like Kang (2010), used symbols as visual material in order to assess the testing effect. In particular, Coppens and colleagues (2011) used forty Adinkra visual symbols, developed by Asante tribe from Ghana. Following a procedure similar to Kang's study, they found that the testing effect is beneficial for learning symbol meanings, after a long retention interval (seven days).

2.3 Possible explanations for the testing effect: theories and frameworks

During the decades, different theories have been elaborated, trying to explain this effect that have obtained strong evidence also with a variety of materials, ages and different test formats. We will now briefly present the main theories that have hypothesized the underlying mechanisms that could explain the advantage observed for tested material over other study methods. Despite the fact that the testing effect is well documented in the literature, the factors and mechanisms behind it are not well established, yet. So, some of the theories we will describe could be considered better candidates than others, since the literature on the argument still has obtained many variable results, that sometimes seem to be in favour of a specific theory but, at the same time, a specific theory can't explain the testing effect observed in different conditions: for example, most of them have been proposed after studying the testing effect in relation to the verbal stimuli and, despite the fact that they have obtained a lot of support from empirical data, they can't explain the same effect observed with visual stimuli.

2.3.1 Re-exposure to the material

In experiments like the one led by Spitzer (1939), the advantage observed for the tested material could be advocated not by the retrieval practice that contributed to consolidate the information, but by a simple re-exposure to the material, resulting in an enhanced performance: thus, the effect of a memory test, if the re-exposure to the material hypothesis is true, would not be due to the memory retrieval itself (Kang et al., 2007). This alternative hypothesis was quite legitimate, since in the first studies on testing effect the variable of the different exposure to the material was not controlled: in fact, the control group consisted in not being tested during an intermediate phase after the encoding condition, but only in a final test. In order to assess this hypothesis, more recent studies have used a different control condition (Wenger et al., 1980; Carrier

& Pashler, 1992; Roediger & Karpicke, 2006b). These studies proposed a different control group (the group that wasn’t exposed to the testing condition) that consisted in a restudy group where, not only the subjects assigned to this condition can re-experience the studied material, but they are even exposed to the 100% of it. The results of both Carrier and Pashler study (1992) and Roediger and Karpicke study (2006b) exclude that the possible explanation could be the re-exposure to the material because the restudy group had a poor and lower performance in a final test, than the testing group. This is even more impressive if we consider that there are differences in re-exposure to the material in favor of re-study condition and still the testing group outperforms, in a final test. These results showed that actually retrieving an item from memory gives more beneficial effects and facilitates its later memory trace access, than merely studying the item.

Finally, another problem with the additional exposure theory is that it does not explain another effect of testing, known as “retrieval-induced facilitation” (Chan et al., 2006), consisting in a memory enhancement for related but not initially tested material; facilitation that is not observed, instead, for the restudy group. All these empirical data contribute to make reference to other explanations and underlying mechanisms, different from those proposed by the additional exposure theory.

2.3.2 The “desirable difficulty”: Bjork and the retrieval effort

Another theory sustains that “manipulations that appear to introduce difficulties for the learner during training can enhance posttraining performance” (Bjork, 1994): simply re-reading a material can be seen as an easy and less challenging activity for the learner, while practicing retrieval requires an effort on the part of the learner and the result is that the retrieved material, for example during an intermediate test, is strengthened (Karpicke, 2017), so that it could be easily accessed again in the future. Bjork (1975) analyzes the retrieval process as acting in a similar framework of the depth of processing during the encoding, proposed by Craik and Lockhart (1972): the two authors described a framework for the human memory, where the deeper the level of processing, the stronger and longer lasting the memory trace; in particular, in this "hierarchy of processing stages" (Craik & Lockhart, 1972), they showed that the deeper level consists of a semantic or cognitive analysis, while the more shallow level consists of a perceptual features analysis of the stimulus. In the same way, a deeper and more effortful retrieval should enhance the testing effect. This "desirable difficulty" (Bjork, 1994), by the way, should not be too excessive: more retrieval effort is helpful and gives a later advantage, but if the test is so hard that retrieval attempts often fail, such increased effort is less beneficial for later memory (McDermott, 2021). In other words, the retrieval effort hypothesis states that a difficult but still successful retrieval gives more advantage for later retrieval than an easier successful retrieval.

The idea that retrieval effort is associated with a larger testing effect has been supported by some studies: Gardiner and colleagues (1973) tested the hypothesis that a difficult initial retrieval may have facilitated the later retrieval, using a tip-of-the-tongue paradigm. The subject were presented with uncommon words' definitions, read by the experimenter with the initial letter of the corresponding words: the subjects had to say the word as rapidly as possible, within 60 seconds. Soon after this phase, the subjects had to unexpectedly recall the target words, writing down as many they could remember, in any order. The retrieval difficulty was assessed as the latency of retrieval in the definition session because, according to the author, a longer response's latency should reflect low accessibility of the information and a greater effort trying to retrieve it. The authors, analysing the performances in the final recall task, observed that the words that were retrieved with difficulty were then better remembered than words that instead were retrieved easily, during the definition session. Even if this study didn't actually focused on the testing effect, it demonstrates, with other similar studies, the positive effects of a retrieval effort on later retention, and the testing effect, according to Bjork, should reflect an example of retrieval effort. However, someone could claim that the times required to retrieve the information is not a good operationalization of the construct "retrieval effort". So, Pyc and Rwason (2009), studied directly the testing effect using a different representation of "difficulty": they manipulated the variables ISI (interstimulus interval) and criterion level (how many times the subject had to correctly recall the items in order to end the practice phase). The prediction was that the performance in the final test should be greater for items correctly retrieved after a longer ISI and should instead decrease as the number of times the items are correctly retrieved increased. The materials used were Swahili-English word pairs: after the study phase, they were presented in a restudy or test phase where the subject was given 8 seconds to recall the target word; the correct retrieved words were represented until they reached the assigned criterion level of performance, from 1 to 10 correct retrievals. Analysing the results, both of the predictions were confirmed, supporting the retrieval effort hypothesis for the testing effect: as the retrieval difficulty for some items during the practice phase increased, so the final performance of those items increased.

Another study, led by Glover (1989) gave support to this hypothesis, sustaining that "more complete intervening retrieval events having a more beneficial effect": here, the effect of three different formats in an intermediate test was compared, in terms of performance in a final test. The intermediate test consisted in a free recall, a cued recall or a recognition task and the results showed that, in a final test four day later, regardless of the matching or not between the intermediate and the final tests' format, the subjects who performed a free recall intermediate test performed much better, followed by the cued recall and, in the last position, the recognition task. In other words, the more demanding the retrieval and the more effort required during the intermediate test, the better the perfomance in a final test. But, in Glover's study, no restudy group was present; Kang and colleagues (2007) added a control restudy group and observed that the performance for items tested with an intermediate multiple-choice task was not significantly better that those assigned to the restudy condition.

The main problem with the effortful retrieval hypothesis is that it doesn't offer a mechanism that could explain why and how the effort strengthens memory traces: stuck in a tautology, "the proposal that retrieved items become strengthened is essentially a redescription of the retrieval practice phenomenon instead of an explanation of it" (Karpicke, 2017).

2.3.3 Transfer-Appropriate Processing

Differently from Craik and Lockhart (1972), who sustained the depth of levels of processing with deeper levels (specifically semantic) associated with stronger, longer lasting memory traces, Morris and colleagues (1977) proposed to reconsider the assumptions made by the levels of processing theory and suggesting, instead, a transfer appropriate processing. According to this idea, there is no reason to think that shallow and non-semantic levels of processing should be considered inferior than a deep semantic processing, they should instead be considered in the context of an eventual retrieval test: a "shallow" encoding strategy could have beneficial effect on a final test, if the encoding operations matches with the process required to perform well in this final test. So, taking tests requires retrieval operations, allowing the subject to practice with the same skills that, later, he will need to perform in a final test. Therefore, in the framework of the transfer-appropriate processing, the testing effect would be explained as a result of the similar operations and strategies involved during both the initial test and the final test. The main prediction is that the performance on a final test should be better when the format of both of the tests is the same. A support to the transfer-appropriate processing prediction comes from a study led by Duchastel and Nungester (1982): the subjects of this study, high school students, took either a short-answer or a multiple-choice intermediate test; the results showed that the performance in a final test was better for those items that both in the intermediate and final phase were tested with the same format. This beneficial effect on the performance for items tested in the same tests' format was described by the authors as "test practice effect" (Duchastel & Nungester, 1982).

Many studies, by the way, have observed a persistent testing effect even when there is no match between the format of the initial test and the format of the final test. For example, Carpenter and DeLosh (2006) tested the role of the transfer-appropriate processing, using three different type of test's format (recognition, cued recall, free recall); contrary to the prediction proposed by the transfer-appropriate processing model, the final test performance was not better in a matching condition where the format of both the intermediate and final tests was the same, than a mismatching condition. The same result has been obtained by many other studies (Wenger et al., 1980; Glover, 1989; Kang et al., 2007; Rohrer et al., 2010), that demonstrated that taking an intermediate test with a particular format still led to a positive performance in a final test with a different format and, especially, it was not lower than the condition of format-matching between intermediate and final test. Thus, we could sum up these results with a phrase of Wenger and colleagues (1980): "the facilitative effect of initial recall trials on subsequent recognition is not an artifact reflecting the fortuitous matching of the modality of presentation on training and test trials."

2.3.4 The bifurcation model

This model has been proposed by Kornell and colleagues (2011); they stated that an intermediate test produce a bifurcated distribution of the item: retrieved items become stronger than non-retrieved items, which instead remain weak since they are not boosted compared with the successfully retrieved items; on the other hand, restudied items are stengthened too, but to a lesser degree and for a shorter time than retrieved items. This means that, while in the restudy condition all the items gain an equal memory strength, the items in the testing condition are instead bifurcated. "As time passes and all items become less accessible, restudied items pass below the final-test threshold before the stronger tested items do so" (Kornell et al., 2011). A point of reference for this model is the theory of disuse proposed by Bjork and Bjork (1992). They proposed two separate strengths for representations in memory: the storage strength, which refers to the permanent strength of a representation; the retrieval strength, which refers instead to the ease of access to the representation in memory at a particular time. According to the authors, these two strengths are independent (items with a low storage strength can have a high retrieval strength and viceversa) and the probability that an item is recalled, when a cue is given, only depends on its retrieval strength, while the storage strength has no direct effects on the subject's perfomance, in fact it's defined as a "latent variable" (Bjork & Bjork, 1992). In their theory of disuse, the authors proposed a series of assumption, and one of them is directly connected with the testing effect compared with the restudy condition and, especially, it represents a common point with the bifurcation model proposed by Kornell and colleagues; this assumption states that "The act of retrieving an item from memory, and the act of studying an item, both result in increments to that item's retrieval strength as well as to its storage strength, but retrieval is the more potent event. That is, the act of retrieving an item (successfully) results in larger increments to storage strength and retrieval strength than does studying an item" (Bjork & Bjork, 1992).

The hypothesized bifurcated distribution was tested and obtained in three experiment designed by Kornell and colleagues (2011). One interesting prediction that was tested by these authors is that a bifurcated distribution can prevent forgetting even if items are not tested, but are assigned to a testing-like condition. So, the authors created an experiment with only a restudy condition: in the Study-40 conditions, 40 word pairs were studied just once, so that all their memory strengths were boosted; in the Study-20-Twice condition, instead, 20 items were boosted since they were studied twice, while the other half not at all, creating the typical situation of a test condition. The authors expected to observe, in this last condition, a bifurcated distribution and that's exactely what they observed: after a short retention interval, more items from Study-40 condition were recalled, but with a longer retention interval the forgetting rate appears to be larger for these items than those from Study-20-Twice condition.

This model, thus, explains two phenomena widely observed in the testing effect studies: first, an intermediate testing can immunize against forgetting, since successfully retrieved item are boosted more and for a longer period than restudied items; second, it also explains the reason why, with a short retention interval, restudied items gain a higher beneficial effect, since they are all boosted, thus obtaining an equal memory strength, which is though to a lesser degree and for a shorter time than retrieved items. The consequence is that, after a long retention interval, more items, previously assigned to restudy condition, have already passed below the recallability threshold.

2.4 Summary

In this chapter we have focused on the testing effect and its paradigm, that we have used, with some changes, in our experiment too. As we have already examined, most of the works that have studied the testing effect, have used verbal material, but recently a higher amount of authors are testing this phenomenon with various visual material (from objects, to faces, to maps and videos too), obtaining promising results also in the visual field.

After a brief description of the main theories, models and hypotheses proposed, we have to underline that they all have been created in a verbal material frame, so in this experimental work we are going to evaluate if they also apply to visual material, as for natural scenes that we process in our daily life.

3. THE PRESENT STUDY

The aim of the present study is to investigate the factors that can modulate the long-term memory visual system, that has been described in the first chapter of this dissertation, a system characterized by a massive capacity and that can store a high amount of details from a complex stimulus like a natural scene. Given the high memorability of this type of material (Standing, 1973; Brady et al., 2008; Konkle et al., 2010; Evans & Baddeley, 2018) that we experience in our daily life, we are interested in understanding the cognitive mechanisms underlying this efficient consolidation process. Besides, despite the extraordinary memory for natural scenes (Standing, 1973; Konkle et al., 2010), several factors may further modulate this type of long-term memory: many studies have focused on the emotional valence, demonstrating a better memory for emotional pictures than neutral pictures (Dolcos et al., 2005; Finn & Roediger, 2011). Thus, in our study, since the effect of emotional valence on visual stimuli has been largely demonstrated, we decided to use neutral natural scene pictures and to manipulate other factors. So, we have designed this experiment manipulating two variables, that could play a significant role in the consolidation and recollection of pictures of natural scenes: the nature of the task performed during picture processing and the amount of perceptual details provided. So, our main questions that have guided the creation of the experiment are: what type of information and how is this information is encoded and then used to retrieve visual scenes? Does the task performed during an intermediate testing phase, and as a consequence the type of activation of the memory trace in this phase, modulate the performance in a final memory test?

We tried to answer these questions using a modified paradigm that has been widely described in the literature as the paradigm created to study the testing effect (Roediger & Karpicke, 2006b): this effect has been studied extensively using verbal material, so our study could shed light on the testing effect for stimuli that have been little used in the literature of this phenomenon, that is complex visual stimuli like natural scenes. The testing paradigm is composed of three phases, an encoding phase, followed by an intermediate testing phase and, after a week, a final test phase.

The manipulation of the two variables task and perceptual quality of the cue (intact vs blurred) has been made in the intermediate testing phase, in order to let us understand which condition could promote a better memory for the stimuli, by strengthening the memory trace and promoting a subsequent successful recognition in a final test, which consisted in a Two-alternative forced choice task. Thus, with this experiment, we could assess whether and to what extent the re-exposure to natural scenes previously seen prompts memory enhancement as a function of the type of task on the picture and the perceptual quality of the stimulus, or instead it promotes memory enhancement independently of the ongoing context, only by virtue of reactivating the memory trace. The hypothesis that led our study are different. First, we would expect to observe a testing effect for natural scene pictures that have been presented during the intermediate testing phase, compared with the pictures that have been presented only once during the encoding phase and then have undergone a test in the final phase, one week later.

Focusing on the type of task, we wanted to verify the mechanism underlying the testing effect, in other words we tried to understand if the testing effect can be modulated, manipulating the way a picture is retrieved during the intermediate testing phase. Therefore, we divided the pictures into two groups (memory task and categorization task), in order to analyze this hypothesis. We could define two different scenarios (see Fig. 1).

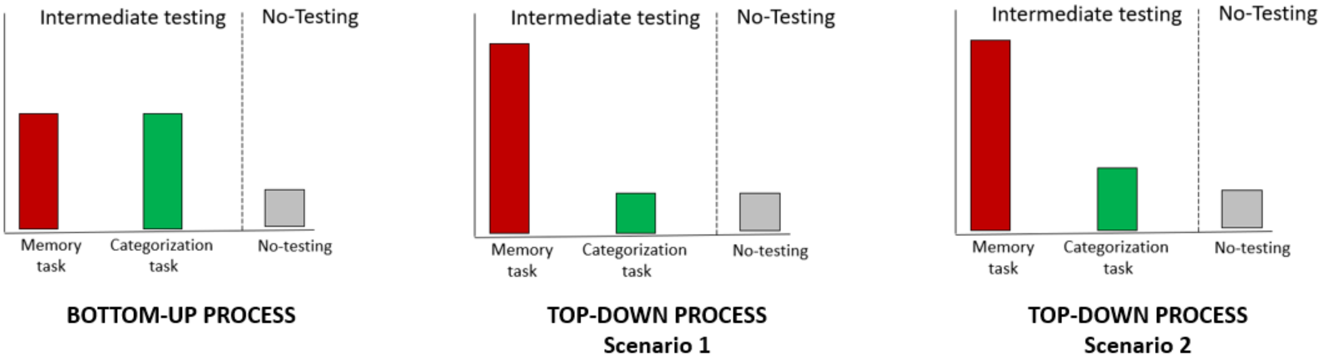

Figure 1: Predictions about the role played by the type of task during the intermediate testing phase, in the final recognition memory test: 3 possible scenarios

If a bottom-up process is involved, it should be sufficient a re-exposure to the material, regardless of the type of task, to reconsolidate the memory trace of the pictures and facilitate their later retrieval in the final test: thus, we would expect no difference in the final recognition memory test between pictures presented in the intermediate memory and categorization tasks; instead, if a testing effect is present, we will just observe a difference between the previously tested pictures and the non-tested ones, with a better performance in the final test for the former.

Differently, the memory task (old/new) requires an active retrieval of the picture's memory trace with an involvement of the episodic memory; the categorization task, instead, only requires to identify the semantic category of the pictures and does not explicitly need an activation and retrieval of the corresponding episodic memory trace to perform the task, since it involves only the semantic memory. We could then expect an advantage for those pictures whose memory trace has been actively retrieved during the intermediate phase, since this active process should strengthen the memory trace facilitating its reactivation and access in a subsequent memory test. As Carrier and Pashler (1992) sustained, "retrieving an item from memory has beneficial effects for later retention above and beyond the effects due to merely studying the item": thus, we could state the same observation in the context of visual material, where the "restudy" condition is the categorization task, characterized by the simple re-exposure to the material. When there is a retrieval attempt, the subject also attempt "to reinstate the former temporal context to help the retrieval search process" (McDermott, 2021). According to the episodic context account (Karpicke et al., 2014), a retrieval-based learning is characterized by a reinstatement of the prior learning context, followed by a context updating to include the features of the new context and, as a consequence, thanks to this "composite context representation" (McDermott, 2021), there would be a restriction of the search set, which facilitates the later access to the memory trace. Thus, "future retrieval is enhanced because updated context representations can be used to restrict the search set and hone in on a desired target" (Karpicke et al., 2014). According to this top-down process hypothesis, we would observe a better recognition performance for the pictures tested with an intermediate memory test; thus, the advantage would be only for the pictures presented in the memory task and there would be no difference between the non-tested pictures and the pictures presented in the categorization task. At the same time we could figure a second possible scenario: we could hypothesize that, parallel the top-down process, also a bottom-up process, but to a smaller extent, is involved. It means that we would still observe a strong advantage for the pictures actively retrieved, but also a little advantage for the pictures presented in the categorization task, compared with the non-tested pictures. If a similar pattern would emerge in the results, we could thus hypothesize that, even though the categorization task mainly involves the semantic memory, it could not be completely separate from the episodic memory of the same picture; therefore, some retrieval of the prior episodic context could probably occur also during the execution of this task: by the way, the context updating would occur at a lesser degree when an incidental retrieval occurs compared with an active retrieval. Moreover, we used a recognition task (Two-alternative forced choice) as a final test: in a review written by Diana an colleagues (2006), the authors argue that, for a recognition memory task, a dual-process model is involved: a first familiarity process, which is though not sufficient for a successful recognition; thus it is followed by a recollection process which consists in the retrieval of episodic informations, that are divided into two different nodes, the "general experimental context node" and the "specific context node" which are then bound together in an "episode node" (Diana et al., 2006): its activation produces a recollection-based judgement, which is certainly more accurate than the familiarity-based judgement. As a consequence, the activation of the episodic memory of a picture would be associated to a better performance, since a recognition task is involved.

Focusing on the perceptual aspect of the cue, in both the tasks, we presented two types of stimuli: intact pictures (which means that they had a normal definition), and blurred pictures, where there is a lack of details. Thus, in the blurred condition, the subject can get access only to the gist of the picture. We can expect three scenarios (see Fig. 2), based on different theories found in the literature.

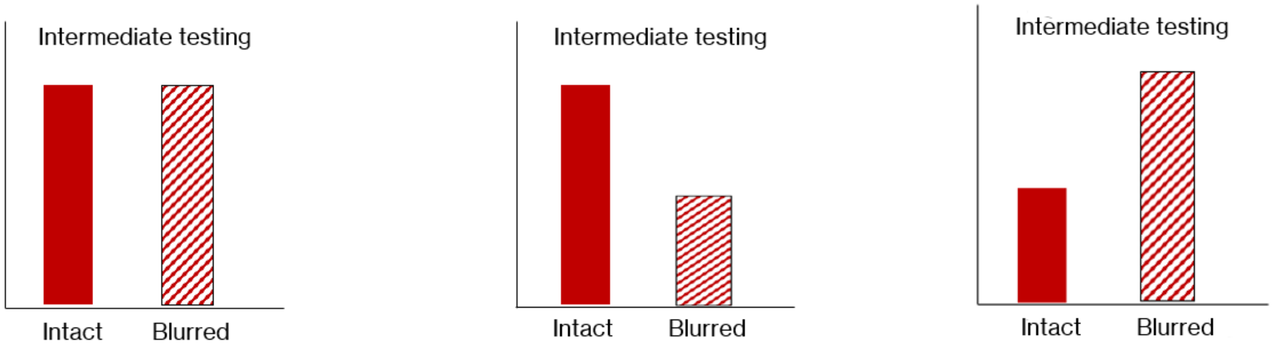

Figure 2: Predictions about the role played by the degradation of the cue, in the final recognition memory test: 3 possible scenarios

According to many studies, like the one described in the first chapter of this dissertation (De Cesarei & Loftus, 2011), subjects can accurately identify visual scenes also when fewer incoming visual detailed informations are available. This "global precedence hypothesis" therefore sustains that we should expect no difference between the intact and the blurred conditions, since the gist of the scene enough to allow the subject to dispose of sufficient informations for a subsequent successful retrieval. Seemingly, Brooks and colleagues (2010) have offered an explanation of how there is a priority for global contents in a natural scene, proposing a hierarchical model where the first necessary and sufficient access is to a gist level: taking into account this hypothesis, even if poor of details, also the blurred image can provide informations about the gist of the scene and it should not invalidate the recognition performance, compared with the intact condition.

The parallel model (Castelhano & Fernandes, 2019), instead, sustains that local and global informations of a natural scene are stored and can be retrieved flexibly and independently, depending on the task's demand, that can prioritize the former or the latter type of information. In this case, an intact picture can provide more "diagnostic details" (Evans & Baddeley, 2018) that could facilitate the process of recognition in the final test. As the authors stated, two levels could be involved in the process of recognition: the first one, familiarity, is based on the gist of the picture, where very general features are encoded; if we would stop at this level, though, the general features of a picture could be "easily confusable with the features of the lures of the same semantic category" (Evans & Baddeley, 2018). That's why we need the second level, known as "checking stage", a more detailed process. So, while "familiarity is usually sufficient to discriminate between old and unrelated new items" (Hintzman & Curran, 1994), in cases like our experiment, where the chosen pictures belonged to only two different semantic categories, a subsequent "recall-like process" should be involved, based on "retention of more detailed, distinctive and diagnostic features" (Evans & Baddeley, 2018). Thus, since the subjects, according to this hypothesis, can't rely uniquely on the gist of the image, in order to successfully remember the target picture when presented with a distractor one, we would expect a higher recognition performance for intact pictures.

The third and last scenario refers to the "retrieval effort hypothesis" proposed by Bjork (1975), which states that more retrieval effort is helpful and gives a later advantage, so introducing difficulties during the intermediate testing phase should enhance the final recognition performance. Therefore, the prediction of this theory is that memory for blurred pictures should be better than intact pictures, since the subject is involved in a more challenging retrieval practice, improving the memory trace consolidation given by this additional cognitive effort.

3.1 Methods and materials

3.1.1 Participants

Thirty subjects were recruited by e-mail, after signing an informed consent. The group of participants was composed by 17 women and 13 men, between 20 and 41 years old (M = 25, SD = 3.9). Before the experiment, each participant was asked to fill in a form, to obtain anagraphical informations. The questions were about the possible presence of neurological, psychological or psychiatric pathologies, allergies, sight's problem, consumption of medicines, hours of sleep the night before the experiment, age, education's level and job, gender, dominant hand. All the participants, except one, completed this preliminary passage. Since the full experiment was entirely performed online, the participants received detailed instructions about the optimal conditions for both of the experiment's sessions, for example the distance from the screen, a silent room, the absence of possible distractions.

3.1.2 Materials and design

Stimuli were 540 neutral pictures selected from the Web; 270 of them belonged to the category "animals" (dogs, cats, rabbits, birds, cows, sheeps, horses, wild animals like bears, foxes and lions) and the other 270 to the category "means of transport" (cars, bicycles, motorcycles and trucks). For both of the semantic categories, some pictures were set outdoors and some indoors; besides, some pictures consisted of a single element (for example, one cat or one car) and some instead consisted of a couple of the same element (for example, two dogs or two trucks) (see Fig. 3).

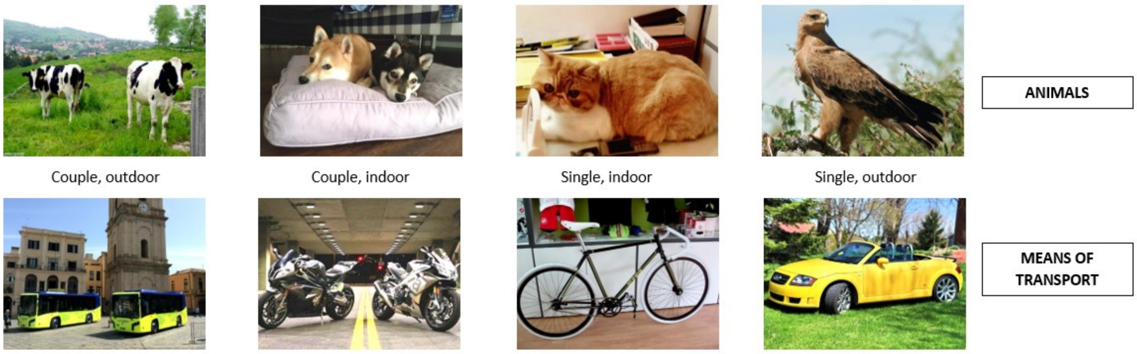

Figure 3: Samples of stimuli used in this study

From the 540 pictures, 2 sets were created, each one made up of 270 pictures: 135 animals and 135 means of transport. In addition to the balancing of the semantic category (animals vs means of transport), both sets were balanced taking into account three more conditions: indoor/outdoor pictures; single/couple pictures; the animals' species and the type of transports, so that, for example, between the two sets there was a similar number of birds, horses or cars (see Fig. 4).

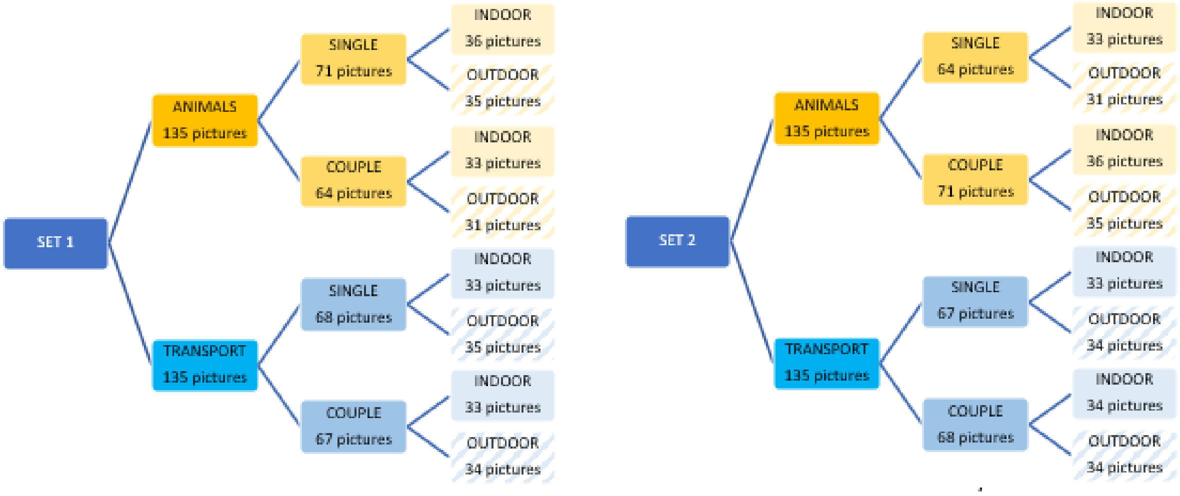

Figure 4: A schematic representation of the two balanced sets

All the chosen pictures were made homogeneous, as regards to specific visual parameters: horizontal orientation, so all the pictures with a vertical orientation were excluded; the same dimension (800x600 pixels), using the programme IrfanView.

Of the 270 pictures that composed each set, 150 (75 animals and 75 means of transport) were presented as dynamic pictures, during the first phase called "Encoding condition". 120 of them were then presented during a second phase, the "intermediate testing condition", 10 minutes after the encoding phase, with other 120 new pictures (the remaining ones from the 270 images of the set). In this second phase, the pictures were divided equally among two tasks (memory and categorization) and were presented either intact or blurred. One week later, during a "final test", 300 pictures were presented in a two-alternative forced choice task: 150 of them were pictures that the participant had previously seen, exclusively during the encoding condition or during both the encoding and the intermediate testing conditions (120 out of 150 images); the other 150 pictures were completely new images (see Fig. 5).

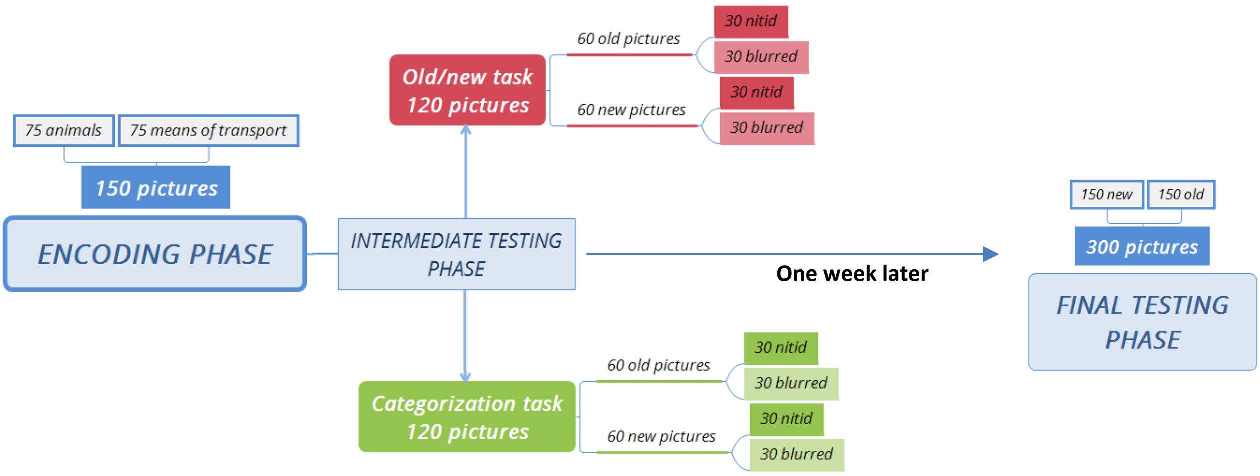

Figure 5: A representation of the experimental paradigm.

Starting from the two sets, four sequences were built (two sequences for every set), changing the order of the stimuli's presentation and the conditions assigned (intact or blurred, tested during the intermediate phase or not), both for the old pictures and the new ones. For the set 1, sequences A and B were built, while for the set 2, sequences C and D were built. The order and the conditions given to the stimuli, in each sequence, were always the same and were settled before the beginning of the experiment. This pseudo-randomization was chosen not to create too much variability between the participants. During this preliminary phase, an important rule was followed for the creation of the sequences: in the pictures' list, there were not more than three stimuli assigned to the same condition (semantic category, new/old, intact/blurred, right/left for the 2-afc final test).

3.1.3 Procedure

Encoding phase: after signing an informed consent (where the general purpose of the experiment was explained), the participants received an e-mail, containing the instructions both for the optimal environment to create during the experiment (since it was entirely performed online and not at the laboratory) and for the data to put before the beginning of the experiment. First of all, the participant was asked to place himself at a distance of 50 cm from the screen, in a silent room, without the presence of possible distracting stimuli, making sure not to be disturbed during the course of the experiment; more important, the subject had to dispose of a stable Internet connection. After the subject clicked the provided link, an online page opened, with a table the subject had to fill with a specific numeric code (so that the identity of the participant was protected) and, under the entry "group", with a letter from A to D, that indicated one of the four sequences created for the experiment. Then, the participant was told to wait until the bar "downloading" was completely grey; after 2-3 minutes, this phrase appeared "All resources downloaded" and the subject clicked "OK" to proceed. After that, the participant had to fill some additional form, reporting some personal data: the possible presence of neurological, psychological or psychiatric pathologies, allergies, sight's problem, consumption of medicines, hours of sleep the night before the experiment, age, education's level and job, gender, dominant hand. Once completed this preliminary part, the experiment actually began.

Each subject was instructed that a series of pictures would have been presented at the center of the screen and that he had just to observe them for the entire time of exposition. The exact sequence was the following: a fixation cross lasting 150 ms; a dynamic pictures (gifs), lasting 4 s, that went gradually from blurred to intact and then from intact to blurred again (see Fig. 6); a green cross the subject had to click to proceed. This first phase was divided in two blocks: block 1 consisted of 75 pictures of animals and means of transport; then there was a pause lasting two minutes (but after one minute, the subject could decide to proceed with the second block); block 2, consisting of the last 75 pictures. In order to avoid possible primacy or recency effects, at the beginning and at the end of each block, were put, respectively, 4 and 3 filler pictures. No more than three consecutive pictures of the same semantic category were presented. The four groups differed for the order of the two blocks, in order to avoid a possible effect due to the order of presentation: while in the group A was firstly presented the block 1 and then the block 2, in the group B was firstly presented the block 2 and then the block 1. The same was done for the groups C and D.

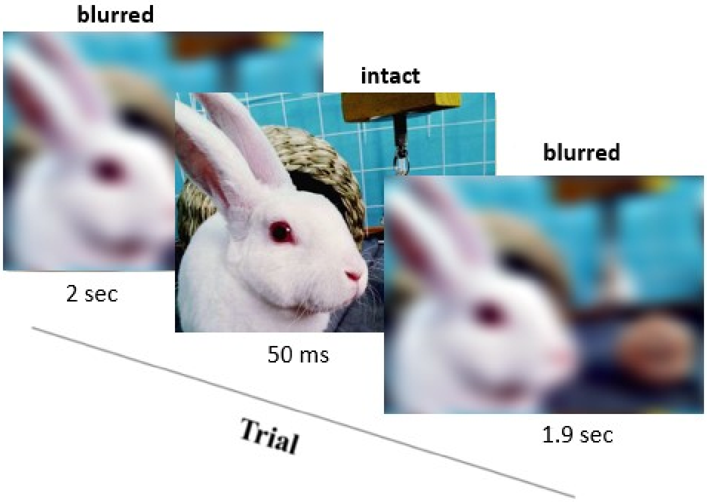

Figure 6: The figure represents the dynamic stimulus (gif) used during the encoding phase: the same image changed gradually from blurred to intact, and then from intact to blurred again.

The 150 dynamic pictures were built in *.gif* format, using a MATLAB script, but then were changed in a *.mp4* format, using CloudConvert, because the platform Pavlovia, where the experiment was uploaded, didn't support the previous .gif format. Each stimulus was presented at the center of the screen, for a total of 3.95 seconds: every dynamic picture was composed of 40 frame, starting from a degradation's level of 4.0 for the frame 1, to a degradation's level of 9.0 for the frame 21 (that corresponded with the intact image), and then going back to a degradation's level of 4.0 for the frame 40. This change in the image's degradation was gradual. Every frame lasted 100 ms, except for frame 21 (the intact image) that lasted instead 50 ms. Of these 150 pictures, 75 were animals and 75 were means of transport; 120 pictures (60 animals and 60 means of transport) were presented again during the intermediate testing phase, while the remaining 30 pictures (15 animals and 15 means of transport) were represented only during the final test, one week later. At the end of the encoding phase, the subject had at his disposal a 10 minutes pause. During the pause, the subject was allowed to temporarily go away from the computer and do something else. When the interval was about to finish (one minute left), a warning red message appeared on the screen, so that the subject had the time to go back in position.

Intermediate testing phase: during this second phase, each participant had to perform two different tasks, an old/new judgement as a memory task and a categorization task: they were not mixed, so they were performed separately, after an interval of 5 minutes (only 3 minutes were obligatory) to prevent the subject from getting tired. In order to avoid a possible influence of the task the subjects performed first, the participants assigned to group A or C performed the memory task first, while the participants assigned to group B or D performed the categorization task first. As for the encoding phase, at the beginning of each task there were 5 filler pictures. For both of the tasks, the participant was instructed to observe the picture presented at the center of the screen: first, a fixation cross appeared for 1 second; then, the stimulus, lasting 3 seconds, that could be intact or blurred (see Fig. 7); in the end, the question “Did you see these image before?” for the memory task and the question “Which category does the image belong?” for the categorization task. The participant answered by pressing one of two keys on the computer keyboard, these keys where the same for both of the tasks, to avoid possible response’s errors due to confusion in pressing too many different keys: a “Z” press button when the participant had already seen the picture or when the picture belonged to the animals’ category; a “M” press button when a new picture was presented or when the picture belonged to the transport’s category. As suggested by Brady in many articles (Brady et al., 2008; Konkle et al., 2010), “participants proceeded at their own pace and were told to emphasize accuracy, not speed, in making their judgements”. No more than three consecutive old or new images and no more than three consecutive intact or blurred images were sequentially presented.

Intact condition

Blurred condition

Figure 7: An example of the intact and blurred trial of the intermediate testing phase, each presented for 3 seconds

For each task, 120 stimuli were presented (for a total of 240 images): 60 images (30 animals and 30 means of transport) have been observed during the encoding phase; half of them were presented in the intact condition (15 animals and 15 means of transport), while the other 30 pictures were presented in the blurred condition (15 animals and 15 means of transport). The same procedure was followed for the 60 new images. The pictures that, in one of the sequences, were assigned to the intact condition, were then assigned to the blurred one in the other sequence, so that all the stimuli were balanced across the blurred/intact condition. Before the end of this first session, the participant had to answer to few questions: if there have been technical problems during the experiment and, if yes, which ones; from 1 to 5, the level of tiredness and how much boring and laborious the experiment was, from 1 to 5.

At the end of this second phase, the participants received by email a reminder that they would have been contacted again the following week to perform the last part of the experiment.

Final testing phase: for this final phase, the participants received another mail, containing the link to access to the second session of the experiment and the same instructions of the first mail, about the optimal environment to create during the experiment and the data to put in the table that appeared on the screen, before the beginning of the experiment. The participants were instructed that a series of pairs of pictures would have been presented and to perform a Two-alternative forced choice task: the subject was asked to make a judgement about which of the two images he had already seen during the first session of the experiment. The participant answered by pressing one of two keys on the computer keyboard, these keys where the same of the intermediate phase's tasks: a "Z" press button if the old image was presented on the left; a "M" press button if it was presented on the right. Critically, the response was made during the presentation of the two images. Like the previous tasks, the participant was told again to emphasize accuracy, not speed, in making the judgements. Immediately after the participant's answer, another screen appeared where he was asked to point out how sure he was of the previous response, through a confidence rating scale from 1 (Not very confident) to 6 (Extremely confident) (see Fig. 8).

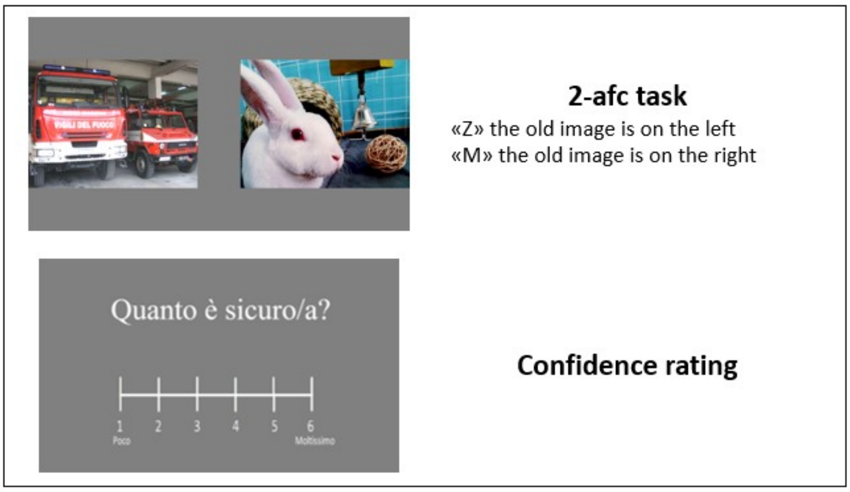

Figure 8: An example of the trial presented during the first part of the final test

During this phase, all the images presented were intact, for a total 0f 300 images, 150 of them were old images and the other 150 were completely new images. For this final test, the pairings of images, composed of one old image and one new image, were realized in advance, randomizing their presentation's order, so that there was the same settled sequence for the subjects assigned to a specific group. For the groups A and B, that had the set 1 as a reference, the new images were taken from the set 2; viceversa for the groups C and D. Four types of pairing were created (see Fig. 9): 38 pairings “old animal image – new animal image”; 38 pairings “old transport image – new transport image”; 37 pairings “old transport image – new animal image”; 37 pairings “old animal image – new transport image”. The right or left position of the old image was then balanced between the four pairings’ types.

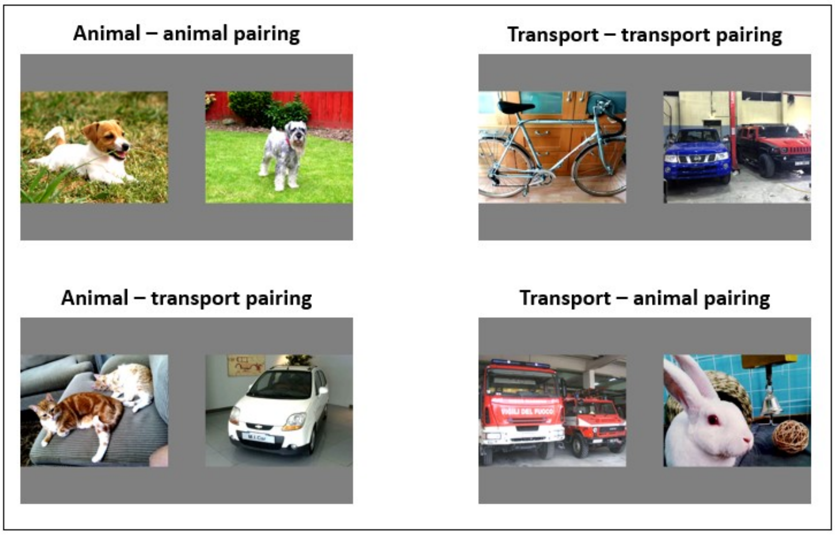

Figure 9: An example of the four types of pairing, used during the 2-afc task

The sequences were created considering the rule that no more than three consecutive old or new images were presented on the same position and no more than three consecutive pairings belonged to the same type. Another consideration regarded the conditions task and degradation of the images during the intermediate testing phase. We can group all the pictures presented in the intermediate phase in five different conditions: memory task and intact image, memory task and blurred image, categorization task and intact image, categorization task and blurred image, not tested image. These five conditions were equally distributed to the four pairing’s types and equally balanced for the right and left position.

For this last part of the experiment, a different memory task was chosen: first of all because we wanted to exclude that the supposed advantage, in the final test, for the images presented in the memory task during the intermediate phase, was due to the fact that the subject had performed the same type of task (old/new), rather than a deeper processing and an active retrieval involved in comparison with the categorization task. Second, we chose a Two-alternative forced choice task because, according to Brady and colleagues (in press), it should be the preferred task when interested in mnestic traces' performance, rather than a binary old/new task. This type of task requires a discriminating judgement, getting round a very important problem in an old/new task: the response criterion. According to Brady, the 2-afc task is more reliable than others when the purpose is to evaluate the performance in a memory task, because it allows to compare the different experiment's condition, false alarm rate being equal, in order to find which condition actually possesses the highest hit rate and results in stronger memories. This is because, unlike an old/new task that gives us just one pair of hit and false alarm rates, a 2-afc task allows to capture the full latent distribution of memory strengths, "because in forced-choice, the old and new items on each trial are experimenter-controlled random samples from anywhere in the entire memory strength distribution" (Brady et al., in press), which means that with this method we can calculate an accurate index of memory, estimating the proportion of genuinely old items that have a stronger memory strength in comparison with the genuinely new ones. Another way to assess the entire latent memory strength distribution is to construct a ROC (receiver operating characteristic) curve, by asking participants to make a confidence rating, in an old/new task; from such a curve, it is possible to find out which conditions or stimuli would be associated with a better memory performance, because at a given false alarm rate, the hit rate is higher than the hit rate of other conditions or stimuli. The main problem with the measures used in an old/new task, such as d', is that they try to infer the shape of the ROC curve, starting from one single hit and false alarm rate: the d' measure incorrectly predicts a normal distribution, while the ROC curves are generally curvilinear and asymmetric. Confidence ratings seem to be a reliable measurement of the memory strength, it has been demonstrated (Mickes et al. 2007) that people are generally very accurate when assessing the strength of their memories.

At the end of the 2-alternative forced choice task, after an interval of five minutes, the participant was asked to perform one last task: a confidence rating of the pictures assigned to the category "new" in the intermediate testing phase. We created this further condition to state the "destiny" in terms of memory trace of images that are encoded in such different conditions (memory/categorization task, intact/blurred). The participant was just told to observe the pictures presented on the screen and to judge from 0 (never seen before) to 6 (extremely confident) how confident he was that he had already seen the pictures. The task's sequence was the following: a fixation cross, lasting 1 second; then the picture, lasting 2 seconds; last, the question "How sure are you that you have already seen this picture?". Once again, like all the previous tasks, the subject was told to be accurate in his judgement, not to be fast. This task was composed of 150 pictures, all of them presented as intact: 120 of them were the "new" images of the intermediate test (60 from the memory task and 60 from the categorization task); the other 30 were instead completely new images.

Before the end of the experiment, the participant had to few questions: if there have been technical problems during the experiment and, if yes, which ones; from 1 to 5, the level of tiredness and how much boring and laborious the experiment was.

3.2 Statistical Analysis of the behavioral data

For this experiment, we only recorded the behavioral data, that were subsequently analysed using the softwares JMP and SPSS. Thanks to an option available on PsychoPy (used for the experiment's building), the emitted responses of the participants in both the intermediate tasks and the final test (except for the confidence ratings) were stored and then compared with a list, created on Excel, composed of all the correct answers associated to the corresponding trials, automatically generating a further column in the same Excel document with the Hits and False alarms for each participant. From this column, we obtained the level of accuracy (mean of correct response) in the intermediate tasks (categorization and memory), taking into account the two conditions of intact and blurred images for each task; plus, we calculated the d prime for old/new task (intermediate phase), in order to obtain an objective measure of sensitivity. Finally, we analyzed the mean accuracy in the final test using an ANOVA, including as factors the degradation levels (intact vs blurred) and the tasks performed in the intermediate phase (memory. categorization); furthermore, we made a comparison with the non-tested condition. We also analyzed the mean of the responses in the confidence rating task for the stimuli that were new in the intermediate testing, in order to observe if the different encoding conditions play a role in the following retrieval of the stimuli.

4. RESULTS

4.1 Intermediate testing phase

The intermediate testing phase was performed 10 minutes after the encoding phase; the pictures presented were 240: 120 presented in the old/new task, half of them were intact, while the remaining ones were blurred; 120 were presented in the categorization task, intact for one half and blurred for the other half. Of the 240 pictures observed during the intermediate testing phase, 120 were old (already seen in the encoding phase) and 120 were new. We calculated the accuracy for both of the tasks, evaluating the performance obtained in the intact and blurred pictures condition for each task (see Fig. 10).

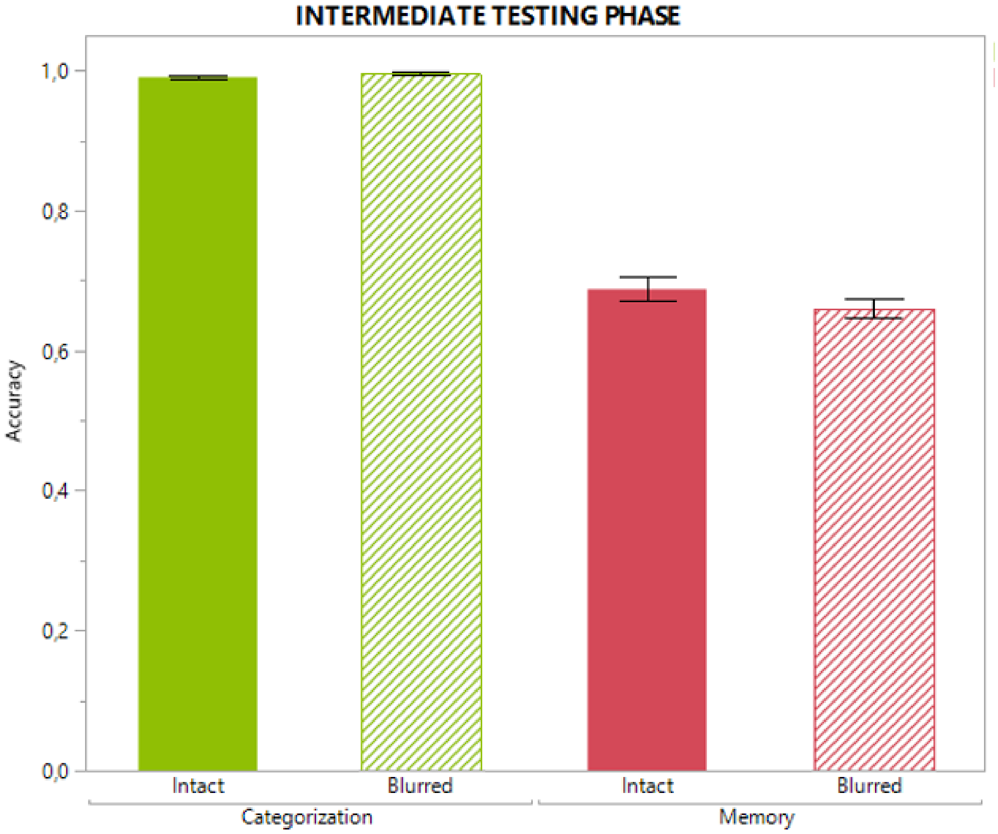

Figure 10: Accuracy of responses, in the intermediate testing phase, for intact and blurred pictures, in the categorization and memory tasks

Considering one task at a time (see Table 1), it looks like the categorization task is a very easy task to perform, since the accuracy is almost 100% not only for intact pictures, but also for the blurred ones: the performance obtained by the subjects, in this task, is not significantly different between the intact and blurred conditions [F(1, 29) =1.6 , p = .211, η_p^2 =.053]. Differently, the old/new task can be considered a more difficult task, since the performance drastically drops, with an average accuracy of 67%: again, no difference was found between the intact and the blurred picture condition [F(1, 29) =3.6 , p = .066, η_p^2 =.112].

In addition to the descriptive statistics, we also calculated the d prime for the memory task (see Table 1), subtracting the standardized Hits from the standardized False Alarms for each condition, in order to obtain an objective measure of sensitivity.

CONDITION	**Mean(D')**	**SD(D')**
Intact	-0,196	0,637
Blurred	-0,482	0,549

Table 1: Descriptive statistics relative to d', for the intermediate memory task

The resulting d' is negative for both of the two conditions, which means that the performance of the subjects is under che chance level: the Hit rate is lower than the False Alarme rate, reflecting a tendency of a yes answer when a new picture is presented. Also, the performance is significantly worst for the blurred than intact condition [F(1, 29) =8.3 , p = .007, η_p^2 =.222].

4.2 Two-alternative Forced Choice Final Test

In presenting the results of the final test, we will analyze the two manipulated variables separately and then, performing an ANOVA 2x2, with the two factors "task" (memory vs categorization) and "degradation" (intact vs blurred), we will see if there is an interaction between them.

4.2.1 The role of the task

We have analysed the level of accuracy for the three conditions (memory task, categorization task, no-testing) (see Fig. 11). Performing a three levels ANOVA, a main effect for the testing condition is showed [F(2, 58) =80.7 , p <.001, η_p^2 =.736]. Also, focusing on the histogram, it looks like the top-down process scenario 2 is involved: the pictures presented during the intermediate memory task have a big advantage over the pictures assigned to the categorization task and this difference, that consists in a better recgnition performance in the final test, is significant [F(1, 29) =146.7 , p <.001, η_p^2 =.835]. Moreover, an advantage for the categorization task pictures is observed, compared with the non-tested pictures [F(1, 29) =19.5 , p <.001, η_p^2 =.402]: so, also the re-esposure to the material, even if to a less extent than an active retrieval process, plays a role and gives a beneficial effect in the final recognition test performance. This result would suggest that even a local process, like the one involved in this particular semantic task, can weakly reactivate the memory trace of the stimulus: making an attempt to explain this phenomenon, we can define it a “spontaneous retrieval”, since the memory trace is unintentionally reactivated.

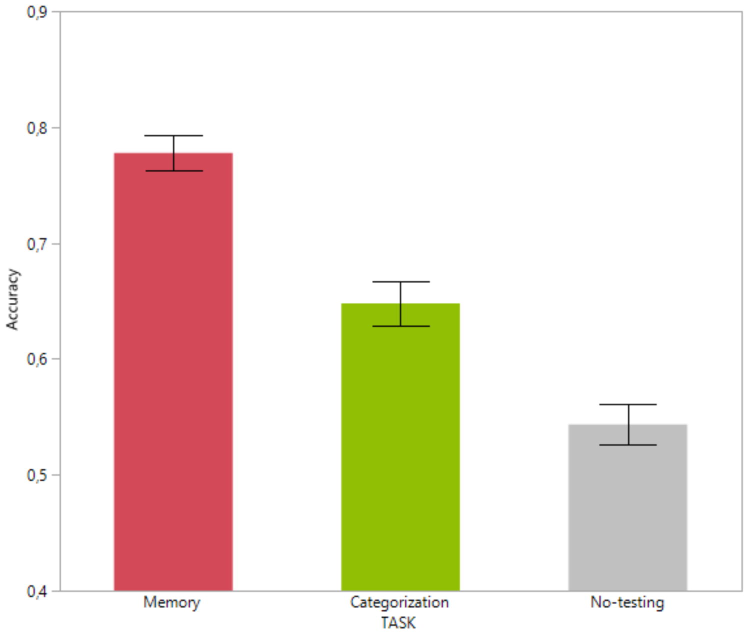

Figure 11: Accuracy of responses, in the final test, considering the task performed in the intermediate phase, compared with the condition no-testing

4.2.2 The role played by the perceptual quality of the cue

After we have demonstrated that the type of task performed during an intermediate phase actually affects the access to the memory trace of the stimuli in a final recognition test, we will now consider if and to what extent the degradation of the cue (intact vs blurred), presented during the intermediate testing phase, can modulate the recognition performance in the final recognition test. We performed an ANOVA 2X2, with the two level factors task (memory vs categorization) and degradation of the cue (intact vs blurred). We found a significant main effect for the task [F(1, 29) =146.7 , p <.001, η_p^2 =.835], with the pictures presented in the memory task more benefited; plus, we found a significant main effect for the degradation of the cue condition [F(1, 29) =16.7 , p <.001, η_p^2 =.365], with the intact pictures more benefited in the final test. There was no significant interaction between the two factors [F(1, 29) = .363 , p = .552, η_p^2 =.012]. (see Table 2)

	Mean accuracy	**SD accuracy**
Memory Intact	0,811	0,104
Memory Blurred	0,744	0,106
Categorization Intact	0,691	0,119
Categorization Blurred	0,604	0,134

Table 2: Descriptive statistics relative to accuracy, for the final memory task

We have also analyzed if the mnestic advantage due to the re-exposure of the material in the condition associated with the lower recognition performance (categorization blurred), is still present and significant compared with the condition no-testing, that is those pictures that were not presented again during the intermediate phase, but instead were observed only during the encoding phase and then went directly to the final phase. The difference is actually significant [F(1, 29) = 4.5 , p = .043, η_p^2 =.134], which means that even if the stimulus is perceptualy degraded and is presented in a task that involves a local process, it is stil and more favoured than a stimulus that is not tested at all in an intermediate phase (see Fig. 12).

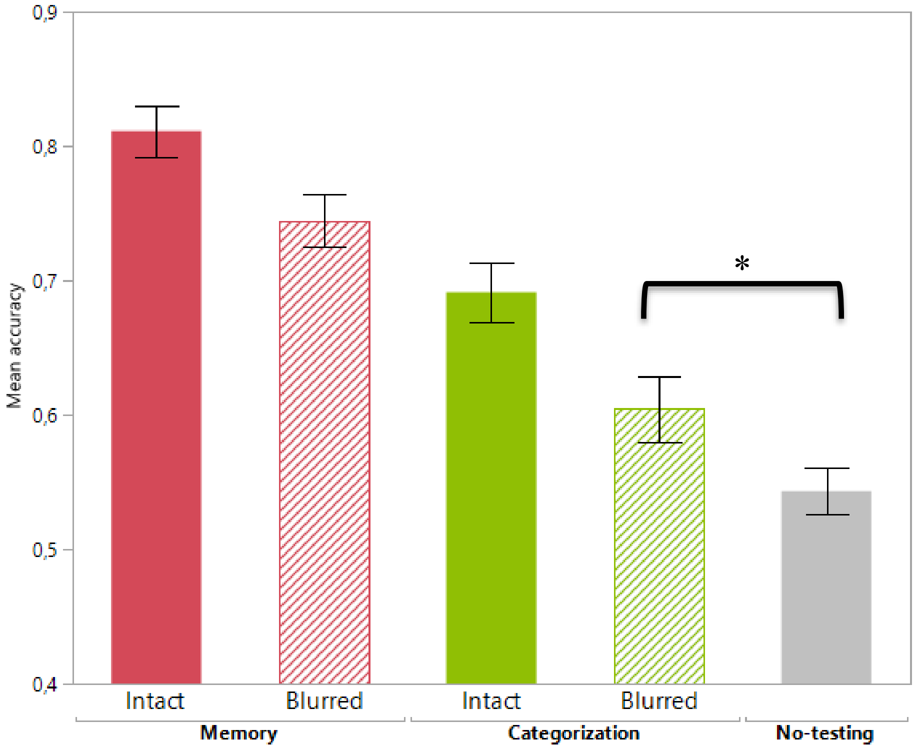

Figure 12: Accuracy of responses, in the final test, considering all the 5 conditions (memory intact, memory blurred, categorization intact, categorization blurred, non-tested pictures)

During the Two-alternative forced choice task, after every couple of pictures, the participant was asked to make a judgement of how sure he was of the previous response, through a confidence rating scale from 1 (Not very confident) to 6 (Extremely confident). As stated by Mickes and colleagues (2007), confidence ratings seem to be a reliable measurement of the memory strength, since people are generally very accurate when assessing the strength of their memories: in fact, in our experiment, we can observe the same trend (see Fig. 13) as in the Figure 12, where the mean accuracy of the recognition performance is represented. All the differences in the five conditions were significant.

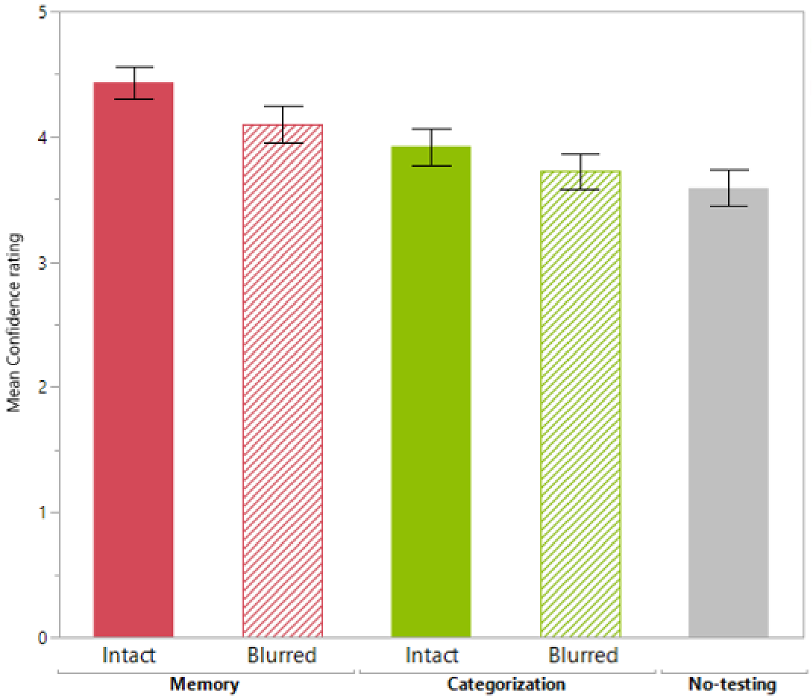

Figure 13: Judgement, in the final test, of how sure the participant was of the previous response (2-afc), through a confidence rating scale from 1 (Not very confident) to 6 (Extremely confident).

4.3 Confidence rating for the new stimuli of the intermediate phase

In the last part of our experiment, soon after the Two-alternative forced choice task, the subjects were asked to perform one last task, a confidence rating of the pictures assigned to the category "new" in the intermediate testing phase. They had to indicate from 0 (never seen before) to 6 (extremely confident) how confident they were that they had already seen the pictures, that were encoded during the intermediate phase, under four different conditions (memory intact, memory blurred, categorization intact, categorization blurred). We created this further task to verify if the advantage observed for the pictures presented in the memory task could be exclusively explained by the fact that this type of task requires more effort, therefore a higher engagement of attention, than the categorization task. Thus, we would observe an advantage for the memory task, independently from the blurred and intact condition, compared with the categorization task. So we performed a 5 levels ANOVA, to examine the different conditions and we observed that the better memory for the pictures presented in the memory task can't be explained exclusively by a higher amount of attention involved to perform this task, since there is no significant difference between the condition memory blurred and the condition categorization intact [F(1, 28) =.471 , p = .498, η_p^2 =.017]. (see Fig. 13)

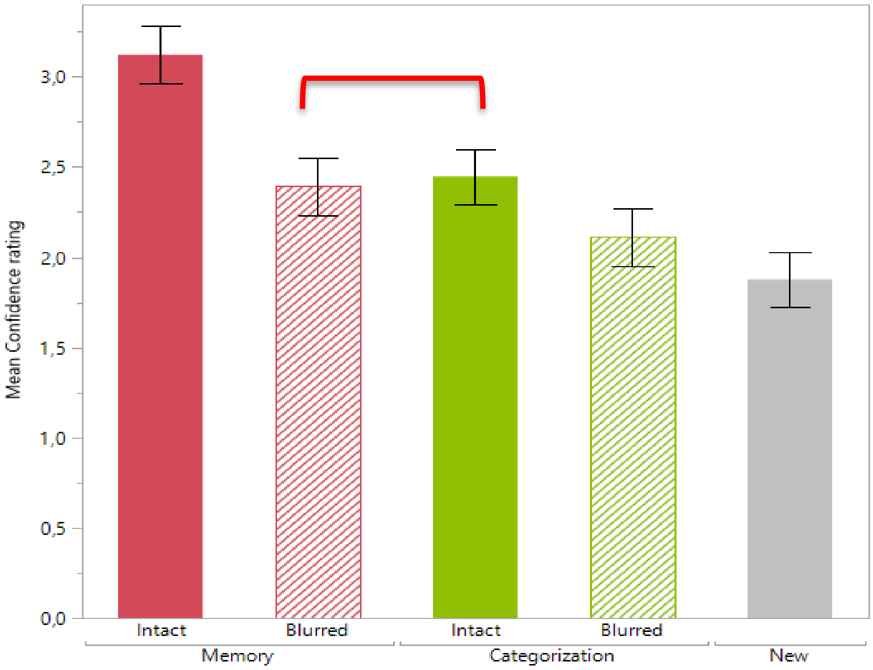

Figure 13: Mean confidence rating for the pictures "new" of the intermediate testing phase

It is not just a matter of attentional effort and difficulty of the task, since when the bottom-up information is reduced in the blurred condition, we also observe a disadvantage. In the end, we can assert that the available amount of details actually enriches the memory trace, set aside the nature of the task. So, the pictures presented as intact in the memory task are significantly better remembred then the ones presented blurred in the same task [F(1, 28) = 34.9 , p < .001, η_p^2 =.555], but, as we stated above, there is no signicant difference between the condition memory blurred and categorization intact; instead, the picture presented as intact in the categorization task are significantly better remembered than the blurred ones in the same task [F(1, 28) = 10.6, p = .003, η_p^2 =.275]; then, there is also a significant difference between the blurred pictures of the categorization task and the new ones [$F(1, 28) = 13.7$, $p < .001$, $\eta_p^2 = .329$].

5. DISCUSSION

The present study aimed to investigate whether the visual memory system, especially for natural scenes, can be modulated, since a lot of study have observed that it is a system of massive capacity that richly and rapidly encodes a high amount of visual information: in particular, we manipulated the task (memory vs categorization) and the perceptual quality of the picture (intact vs blurred). To test how and to what extent the visual long-term memory can be modulated, we used a modified version of the paradigm for the testing effect, consisting of three phases. The first part was an initial encoding phase where the pictures dynamically changed from a blurred condition to an intact condition and returned to be blurred for a total duration of 4 seconds: our purpose, in creating this dynamic gif, was to make possible an association between the blurred version with the intact one of the same picture, so that the subject could encode both the cues that in the following phase would have been tested. Then, the subjects were tested in an intermediate testing phase where only a part of the pictures presented in the encoding phase were presented again, while the others went directly to the final phase (non-tested pictures); here we manipulated the two factors of task (memory task, in particular an old/new task vs categorization task) and perceptual quality of the cue (intact vs blurred) with the purpose to determine a possible role played by the two of them on the memory trace of the natural pictures in terms of consolidation and future recall, in order to test our main hypothesis, which is the possible modulation of the visual memory. One week later, the subject were asked to perform a final memory task, different from the old/new task: we opted for a Two-alternative forced choice task, first because authors like Brady and colleagues (in press) have suggested that it is a better and more reliable task for testing memory, second because one of the theories that have been proposed to explain the testing effect is the ”transfer-appropriate processing” (Morris et al., 1977) that states that the performance on a final test should be better when the format of both of the tests is the same. Even though many studies have disconfermed this theory as an explanation of the testing effect, they focused mainly on verbal stimuli, so in order to prevent the possibility that the possible advantage observed for the memory condition could be explained by this correspondence of the task's format, we used a different memory task, testing recognition memory in the final test.

Considering now the pictures presented in the intermediate phase (both the ones tested with the memory and the categorization tasks) opposed to the pictures that instead were tested only once, in the final testing phase, the results showed us that there was a general testing effect: independently from the task performed, the pictures that were presented again during the intermediate phase, show an advantage in the final performance, compared with those pictures that haven't been submitted to an intermediate task. Plus, we found a main effect for the type of task: the performance in a final recognition memory test is better for the pictures assigned to the memory task condition than the ones assigned to the categorization task. So an active retrieval operated by the subject, makes the memory trace stronger and that facilitates its reactivation in a subsequent test: as stated by the episodic context account (Karpicke et al., 2014), a retrieval-based learning is characterized by a reinstatement of the prior learning context, followed by a context updating to include the features of the new context, since the act of retrieval has been described by Bjork (1975) as a "memory modifier". Thus, the resulting "composite context representation" (McDermott, 2021) favours a restriction of the search set, which not only consolidates but also facilitates the later access to the memory trace: as McDermott sustained (2021), the retrieval practice is "one of the most effective ways of solidifying new knowledge". Nevertheless, the re-exposure to the material (categorization task condition) offers a little, yet significant, advantage than the condition no-testing: we explained this beneficial effect obtained from this local process as a "spontaneous retrieval", an unintentional reactivation of the memory trace in this condition.

We will now analyze the results concerning the perceptual quality of the cue, where we proposed three different scenarios. The first one is consistent with the "global precedence hypothesis" (De Cesarei & Loftus, 2011) and the hierarchical model (Konkle et al., 2010) that sustain that the information provided by the gist of the natural scene is sufficient to correctly recognize the picture, thus we should not expect any difference between the intact and blurred pictures in the final memory test. The second one, instead, consistent with a parallel model (Castelhano et al., 2019), states that the memory representation of a natural scene doesn't depend completely on the global content of the scene, but can actually store flexibly the various levels of informations (global and local), that vary in strength, and, critically, can be accessed independently, according to the task's demand. In this case, considering the fact that we presented pictures belonging to only two different semantic categories, we should expect a better performance for the intact pictures, since the details offer a higher diagnosticity value (Evans & Baddeley, 2018), that facilitates their later recognition in the final test. The third scenario instead aligns with the effort retrieval hypothesis by Bjork (1975) which states that more retrieval effort is helpful and gives a later advantage, so the retrieval practice for the blurred pictures should be more challenging and therefore we would expect a better final performance for these ones, than the intact pictures. The results of our study support the second scenario: intact pictures are better remembered than the blurred ones, both in the memory and the categorization tasks, suggesting that details provide a stronger memory trace, that is simpler to get access to in a final test.

Moreover, even in the condition associated to the lower perfomance, that is the categorization task and blurred condition, the pictures still have an advantage on the non-tested pictures. This means that the simple re-exposure to the material in a poor-of-details condition, is enough to make the memory trace easier to retrieve in a final test, than pictures that weren't submitted to an intermediate testing phase, demonstrating once again the power of the testing effect and extending this effect to the domain of the visual material, too.

To conclude, we created a second task to perform in the last session one week after; this task has been performed right after the Two-alternative forced choice task, and provided that the subjects should make a confidence rating for the pictures that were assigned to the "new" condition in the intermediate testing phase, both for the memory and the categorization tasks; this means that these pictures have not been observed during the previous encoding phase. So, they have been encoded for the first time in the intermediate testing phase, in different conditions, considering the task and the intact or blurred condition. Of the 150 pictures presented during the confidence rating task, 120 were pictures that the subjects had already seen in the intermediate testing phase (60 in the categorization task, 60 in the memory task), while the other 30 pictures were completely new. We wanted to assess the consequences on the memory traces of an encoding realized in the various conditions. In particular, this task was created for a particular reason: we wanted to exclude the possibility that the advantage for the pictures (both intact and blurred) presented in the memory task, could be explained uniquely by a higher engagement of attention, since the task is more difficult and requires a bigger effort than the categorization task. The results show us that this is not the case: it is not just a matter of effort, otherwise we would observe an advantage for intact and blurred pictures in the memory task; we observe, instead, a better memory for the intact pictures of the memory task, but a lower and not significant difference for the blurred pictures of the same task and the intact pictures of the categorization task. So, we can state that the effort required to perform a task is not the only factor involved, since the intact condition represents a further advantage over the blurred one.

Considering these results in relation to the testing effect's models and theories we presented in the second chapter, the one that could fit quite well is that proposed by Kornell and colleagues. In fact, taking the bifurcation model (Kornell et al., 2011) as a point of reference and making the due differences, since the paradigm we adopted is a modified one from the original testing effect paradigm and since we used a different type of stimuli (visual material), we could state a similar result: positing the memory task as the traditional testing condition and the categorization task as the "restudy" condition and considering the third condition of non-tested stimuli, we could affirm that pictures actively retrieved are more boosted than pictures presented in the categorization task; plus, the pictures of the categorization task, though to a lesser degree than actively retrieved pictures, are still boosted, compared to the non-tested pictures. Moreover, taking into account the level of details of the pictures (intact vs blurred), we could explain our results, considering the experiment conducted by the same authors (Kornell et al., 2011): remaining in the bifurcation model framework, the authors created a test-condition-like situation, but using only restudied material at a different level. They still observed a bifurcated ditribution: thus, we could expect, and that's what we actually observed, that intact pictures, not only from the memory task but also from the categorization task yet at a lesser degree, would be more boosted than blurred ones, just like items that were studied twice had a higher advantage on items both restudied only once and not restudied.

In summary, our data suggest that the testing effect can be generalized beyond the verbal material, since we observed it also for the visual material, in particular visual scenes. Plus, we assessed that, despite the rapid and richly detailed encoding of this type of material and the massive capacity of the visual system, there are conditions that modulate the visual long-term memory: an active retrieval process and the richness of details strengthen and consolidate the memory traces and facilitate the recognition of the natural scene pictures in a final memory test.

BIBLIOGRAPHY

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science*, *15*(2), 106–111.

Atkinson, R.C., & Shiffrin, R.M. (1968). Human Memory: A Proposed System and its Control Processes. *Psychology of Learning and Motivation*, *2*, 89-195.

Baddeley, A. D., & Hitch, G. J. (2017). Is the Levels of Processing effect language-limited?. *Journal of Memory and Language, 92,* 1–13.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). The MIT Press.

Bjork, R.A. (1975). Retrieval as a memory modifier: an interpretation of negative recency and related phenomena. In Solso, R.L. (Ed.), *Information Processing and Cognition: The Loyola Symposium* (pp. 123–144). Erlbaum, Hillsdale, NJ.

Bjork, R.A., & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In Healy, A.F., Kosslyn, S.M., & Shiffrin, R.M. (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes*, vol. 2 (pp. 35–67). Erlbaum, Hillsdale, NJ.

Bradley, C., & Pearson, J. (2012). The sensory components of high-capacity iconic memory and visual working memory. *Frontiers in psychology*, *3*, Article 355.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of vision*, *11*(5), Article 4.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329.

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J.T. (in press). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. *Psychonomic Bulletin & Review*.

Brockmole, J. R., Castelhano, M. S., & Henderson, J. M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of experimental psychology. Learning, memory, and cognition*, *32*(4), 699–706.

Brooks, D. I., Rasmussen, I. P., & Hollingworth, A. (2010). The nesting of search contexts within natural scenes: evidence from contextual cuing. *Journal of experimental psychology. Human perception and performance*, *36*(6), 1406–1418.

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition*, *5*(1), 73–99

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563–1569.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the Testing and Spacing Effects to Name Learning. *Applied Cognitive Psychology, 19*(5), 619–636.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. *Memory & cognition*, *34*(2), 268–276.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: using tests to enhance visuospatial map learning. *Psychonomic bulletin & review*, *14*(3), 474–478.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & cognition*, *20*(6), 633–642.

Castelhano, M. S., Fernandes, S., & Theriault, J. (2019). Examining the hierarchical nature of scene representations in memory. *Journal of experimental psychology. Learning, memory, and cognition*, *45*(9), 1619–1633.

Chan, J. C., McDermott, K. B., & Roediger, H. L., 3rd (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of experimental psychology. General*, *135*(4), 553–571.

Coltheart M. (1980a). The persistences of vision. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *290*(1038), 57–69.

Coltheart M. (1980b). Iconic memory and visible persistence. *Perception & psychophysics*, *27*(3), 183–228.

Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology, 23*(3), 351–357.

Cork, T. M., Jankovic, N., Di Lollo, V., & Spalek, T. M. (2020). Are the inverse-duration and inverse-proximity effects in temporal integration subserved by independent mechanisms?. *Vision research*, *167*, 24–30.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11*(6), 671–684.

De Cesarei, A., & Loftus, G. R. (2011). Global and local vision in natural scene identification. *Psychonomic bulletin & review*, *18*(5), 840–847.

Di Lollo, V., & Bischof, W. F. (1995). Inverse-intensity effect in duration of visible persistence. *Psychological bulletin*, *118*(2), 223–237.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: a review of arguments in favor of a dual-process account. *Psychonomic bulletin & review*, *13*(1), 1–21.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2005). Remembering one year later: role of the amygdala and the medial temporal lobe memory system in retrieving emotional memories. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2626–2631.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *The Journal of Educational Research, 75*(5), 309–313.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological science in the public interest : a journal of the American Psychological Society*, *14*(1), 4–58.

Evans, K. K., & Baddeley, A. (2018). Intention, attention and long-term memory for visual scenes: It all depends on the scenes. *Cognition*, *180*, 24–37.

Finn, B., & Roediger, H. L., 3rd (2011). Enhancing retention through reconsolidation: negative emotional arousal following retrieval enhances later recall. *Psychological science*, *22*(6), 781–786.

Gardiner, F. M., Craik, F. I., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & cognition*, *1*(3), 213–216.

Gates, A.I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 40,* Pp. 104.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392–399.

Goldstein, E. B. (2015). Long-Term Memory: Encoding, Retrieval, and Consolidation. In

Goldstein, E. B. (Ed.), *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience* (4th ed.) (pp. 178-205). Cengage learning, NY.

Goldstein, E. B. (2015). Long-Term Memory: Structure. In Goldstein, E. B. (Ed.), *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience* (4th ed.) (pp. 150-177). Cengage learning, NY.

Goldstein, E. B. (2015). Short-Term and Working Memory. In Goldstein, E. B. (Ed.), *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience* (4th ed.) (pp. 118-149). Cengage learning, NY.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33*(1), 1–18.

Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of experimental psychology. Learning, memory, and cognition*, *31*(3), 396–411.

Hollingworth A. (2006). Scene and position specificity in visual memory for objects. *Journal of experimental psychology. Learning, memory, and cognition*, *32*(1), 58–69.

Hollingworth, A., & Luck, S. J. (2008). Visual Memory Systems. In Luck, S. J., & Hollingwoth, A. (Eds.), *Visual Memory* (pp. 3-8). Oxford University Press, Inc, NY.

Hollingworth, A (2008). Visual Memory for Natural Scenes. In Luck, S. J., & Hollingwoth, A. (Eds.), *Visual Memory* (pp. 123-161). Oxford University Press, Inc, NY.

Irwin, D. E., & Yeomans, J. M. (1986). Sensory registration and informational persistence. *Journal of experimental psychology. Human perception and performance*, *12*(3), 343–360.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: effects on recognition memory, classification, and metacognition. *Journal of experimental psychology. Learning, memory, and cognition*, *36*(6), 1441–1451.

Kang, S. H. (2010). Enhancing visuospatial learning: the benefit of retrieval practice. *Memory & cognition*, *38*(8), 1009–1017.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4-5), 528–558.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*(1), 17–29.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., 3rd (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own?. *Memory (Hove, England)*, *17*(4), 471–479.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In Ross, B. H. (Ed.), *The psychology of learning and motivation* (pp. 237–284). Elsevier Academic Press.

Karpicke, J.D. (2017). Retrieval-based learning: A decade of progress. In Wixted, J. T. (Ed.), *Cognitive Psychology of Memory, Learning and Memory: A Comprehensive Reference* (Volume 2, pp. 487–514). Academic Press: Cambridge, MA, USA.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychological science*, *21*(11), 1551–1556.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85–97.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In Gruneberg, M. M., Morris, P. E., & Sykes, R. N. (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory?. *Journal of experimental psychology. Learning, memory, and cognition*, *42*(10), 1573–1591.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Madore, K. P., & Wagner, A. D. (2022). Readiness to remember: predicting variability in episodic memory. *Trends in cognitive sciences*, *26*(8), 707–723.

Mandler, J. M., & Ritchey, G. H. (1977). Long-term memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 3*(4), 386–396.

McDermott, K. B. (2021). Practicing Retrieval Facilitates Learning. *Annual review of psychology*, *72*, 609–633.

Meyer, A. N., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: benefits for lifelong learning. *Psychology and aging*, *28*(1), 142–147.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*(5), 858–865.

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81–97.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16*(5), 519–533.

Morris, P. E., & Fritz, C. O. (2002). The improved name game: better use of expanding retrieval practice. *Memory (Hove, England)*, *10*(4), 259–266.

Neisser, U. (1967). *Cognitive psychology.* New York, NY :Appleton-Century-Crofts.

Neisser, U., & Harsch, N. (1992). Phantom flashbulbs: False recollections of hearing the news about Challenger. In E. Winograd & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb" memories* (pp. 9–31). Cambridge University Press.

Olson, I. R., & Chun, M. M. (2002). Perceptual constraints on implicit learning of spatial context. *Visual Cognition, 9*(3), 273–302.

Putnam, A. L., & Roediger, H. L., 3rd (2013). Does response mode affect amount recalled or the magnitude of the testing effect?. *Memory & cognition*, *41*(1), 36–48.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?. *Journal of Memory and Language, 60*(4), 437–447.

Roediger, H. L., 3rd, & Karpicke, J. D. (2006a). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on psychological science: a journal of the Association for Psychological Science*, *1*(3), 181–210.

Roediger, H. L., 3rd, & Karpicke, J. D. (2006b). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249–255.

Rogalski, Y., Altmann, L. J. P., & Rosenbek, J. C. (2014). Retrieval practice and testing improve memory in older adults. *Aphasiology, 28*(4), 381–400.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of experimental psychology. Learning, memory, and cognition*, *36*(1), 233–239.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*(1481), 773–786.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*(11), 1–29.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*(9), 641–656.

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly journal of experimental psychology*, *25*(2), 207–222.

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science, 19*(2), 73–74.

Talarico, J. M., & Rubin, D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. *Psychological science*, *14*(5), 455–461.

Tulving, E. (2000). Concepts of memory. In Tulving, E., Craik, F. I. M. (.Eds), *The Oxford Handbook of Memory* (pp. 33–43). New York, NY: Oxford University Press, Inc.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 135–144.