UNIVERSITY OF PAVIA

DEPARTMENT OF BRAIN AND BEHAVIOURAL SCIENCES

PHD COURSE IN PSYCHOLOGY, NEUROSCIENCE AND DATA SCIENCE

THE BINDING OF FALSE MEMORY:

BEHAVIORAL AND BRAIN STIMULATION EVIDENCE

SUPERVISOR:
Prof. Tomaso Vecchi

PhD Candidate
Daniele Gatti

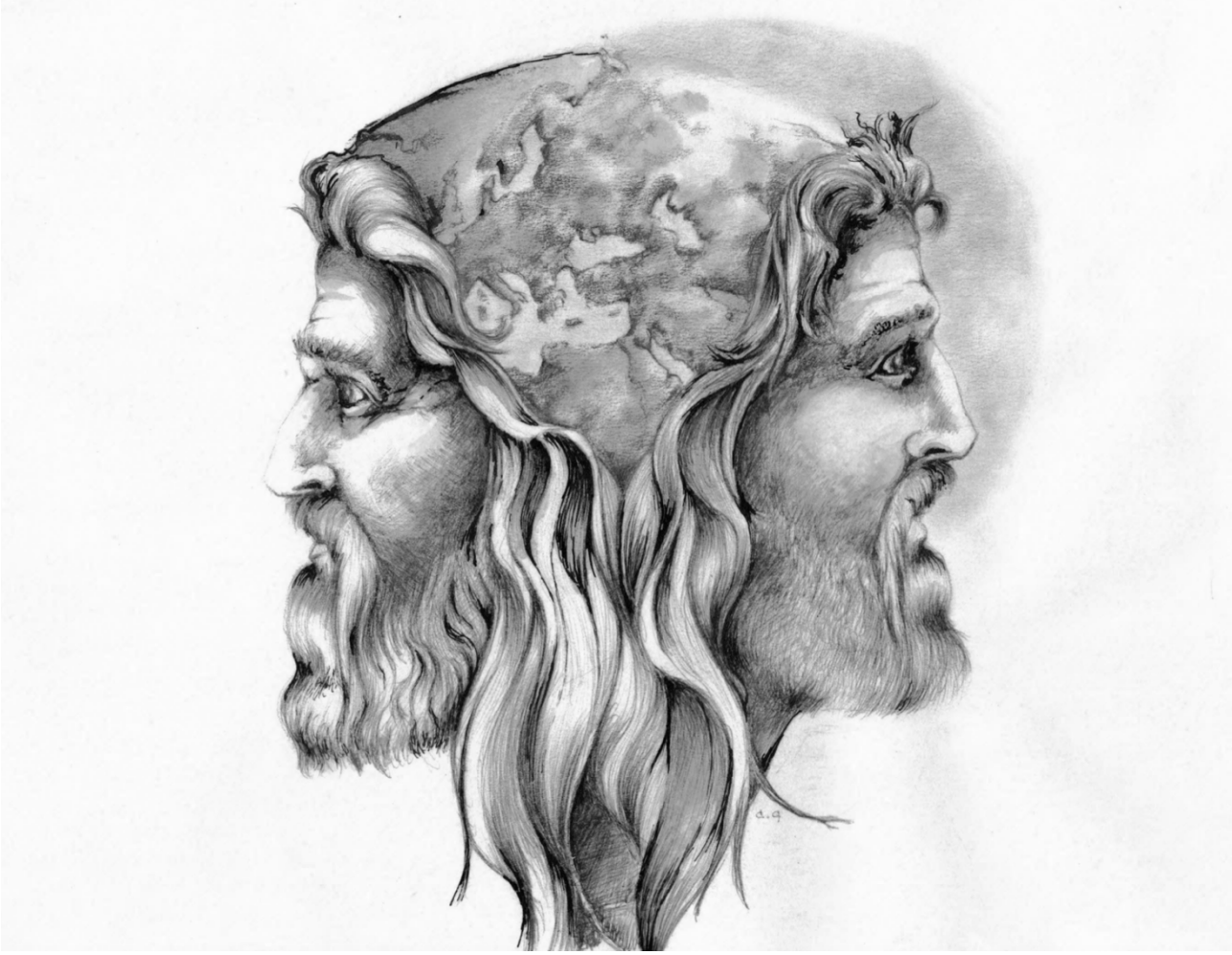
XXXIV Doctorate cycle

Alle due persone che più stimo al mondo:
A mia madre e a mio padre.

Al llegar al aeropuerto de Neuquén se me acercó un tipo que dijo haberme marcado muy fuerte en un partido de fútbol que escribí pero que nunca existió. [...] ¿Quién soy en aquel que fui a orillas del Limay? ¿Seré los ojos de mi madre y la desazón de mi padre? Poco importa: el árbol sigue dando peras y por la ventana de mi pieza todavía entra el sol. Mi padre solía contarme de una curtiembre en Campana y tal vez ahí, en ese lugar al que nunca fui, esté el Rosebud del que no me habló. Un trompo olvidado en un sótano o unas pocas bolitas todas cascadas.

Osvaldo Soriano

INDEX



Janus Bifrons, Roman god of beginnings, gates, transitions, time, passages, and endings. Like human memory, he is looking towards the past and towards the future. Drawn by my father, Carlo Gatti.

Acknowledgements

It is time to say thank you.

Without useless complexity, I can say that during these three years I took more than I gave, and I can only be grateful for this.

I am grateful to my mentor, Tomaso Vecchi, for his advice and all the help provided across professional, experimental, and human perspectives.

To Giuliana Mazzoni and Luca Rinaldi for their supervision and their time.

To my parents, for their priceless teachings and their humanity.

To my lifetime friends, Luca, Edoardo and Davide, for being there without a real reason, for remind me where I belong, where I will always be.

And then again, as if we were in a book written by Henry Miller, simply, to her.

It has been a nice trip.

1. Abstract

Memory is one of the most studied topics in cognitive science. Recent perspectives proposed that human memory is not actually a memory system, but rather a predictive system adaptively shaped. However, the role of false memory in such frameworks is not clear. Here, across five studies, we directly investigated the adaptive bases of false memory using the Deese–Roediger–McDermott (DRM) task. Participants were required to study lists of associated words and then to perform a recognition task. In Study 1 we show that participants' memory performance follows a continuous semantic gradient, while in Study 2 we show that semantic memory plays a role in participants' performance even when correctly rejecting semantically related new non-studied items. Then, in Study 3 and Study 4 we adopt an individual differences approach and show that participants' episodic and semantic memory scores differently predict false memory, as well as that participants' reliance on semantic memory when falsely recognizing new words is predicted by theory of mind indexes. Finally, in Study 5 we show that cerebellar perturbation through TMS can over-activate semantic memory traces and thus increase the number of false recognitions. Overall, the studies presented in this Thesis point to the need to build a more global view of memory and of memory ultimate function itself.

2. Introduction

2.1

Is memory a memory system?

Memory is one of the most studied topics in cognitive science. Attention to memory has branching out to various domains such as neurobiology, artificial intelligence, and, more recently, neuroscience. Several models have been developed to explain how memory works, some of which, such as those based upon the distinction between short-term and long-term memory have found a place in the collective imagination and have provided an extremely effective terminology within both the scientific and the popular literatures. Generally speaking, we can define memory as the ability to encode and retrieve information, but this definition does not account either for the fact that this information is continuously changed, modified, or for the lack of a precise correspondence between what is originally encoded and what is later retrieved. At this point, we could ask ourselves: is memory a memory system? Is the purpose of memory to remember?

Indeed, the fact that by using memory we can remember information does not necessarily imply that the final purpose of memory is to remember (Klein, 2013). Surprisingly, this idea has several theoretical issues, first of all the fact that human memory, under normal conditions, makes a great number of *errors*: it does not store information as it was perceived or processed, but it transforms it in order to maintain reduced, but more useful and updated information. Indeed, several studies have shown that the retrieval of long-term memories involves active rather than passive reproduction (for a review: Schacter, 2021). That is, humans use their semantic knowledge to encode, store and remember information, generally adapting it to their own expectations, with systematic errors that may

occur when the same individuals have to retrieve the presented material (e.g., Bartlett, 1932; Brewer & Treyens, 1981; Sulin & Dooling, 1974). We tend to forget the precise features of the information memorized in favor of an extraction of its meaning. Memory makes errors, a good memory makes errors, and it would be extremely difficult to explain, either phylogenetically or ontogenetically, the development of a memory system making so many errors.

With particular focus on memory accuracy, although we subjectively perceive our own memory, at least episodic one, to be highly accurate, numerous studies have shown that in the majority of cases what we remember undergo transformations and distortions and, also, that it is even possible to maintain "memories" of events that never actually occurred (for an in-depth discussion, see, e.g., Laney & Loftus, 2010; Schacter & Loftus, 2013). The presence of these distortions constitutes a challenge for the classic paradigm that considers memory as a system of retention and, indeed, motivated many authors to search for an alternative theoretical framework. In particular, the discussion has focused on – and still today focuses on – the possibility of considering memory as the result of an adaptive process and/or to what extent it may contain maladaptive components (Newman & Lindsay, 2009; Schacter et al., 2012).

At this point it is worth asking: well, if remembering is not the purpose of memory, then *why do we remember?*

Recently, we tried to answer this question by arguing that human memory is a predictive system (Vecchi & Gatti, 2020). Indeed, several studies have shown, both on a cerebral and a cognitive level, that memory traces undergo transformations that render the memory inaccurate and therefore make a testimony that is based on it unreliable (Schacter & Loftus, 2013). On the other hand, there are conditions in which memory is not only durable but also extraordinarily accurate (Hirst & Phelps, 2016). Interestingly, these latter conditions happen as a consequence of abnormal events and in more extreme cases constitute

maladaptive phenomena in the daily life of individuals. Thus, if the answer is not in the past, it should be in the future.

Predictive memory processes are consistent with a large number of characteristics of human memory, from the ease of forgetting, passing for the update of memory, to false memory in general.

For example, reconsolidation processes (for a review of the experimental evidence, see Lee, Nader, & Schiller, 2017), which cause the alteration of a reactivated memory as a result of a pharmacological or behavioral intervention, are "ideally placed to enable memories to be updated with new information" (Lee et al., 2017, p. 532; but see also: Agren, 2014; Exton-McGuinness, Lee, & Reichelt, 2015; Nader & Hardt, 2009). Indeed, following the reactivation of a stored memory, the original trace can be modified, even radically, leading to incorporation of new material, updating the original memory (Lee et al., 2017).

The purpose of this kind of transformations that memory continuously undergoes would therefore be to enable the updating of memory traces at the cost of their accuracy, thereby ensuring that they maintain a certain relevance in an environment that constantly changes (Dudai, 2004, 2006; Sara, 2000; Schiller & Phelps, 2011; Tronson & Taylor, 2007). The relevance of the memory trace is obviously established by how adaptive it is, that is, the relevance is based on its future usefulness in a predictive phase (e.g., Klein, 2013), and therefore the memory would transform itself so that it can yield a more precise and up-to-date prediction.

This framework is particularly desirable within the topic of memory distortion and false memory. It seems, indeed, that the tendency to cause memory distortions to arise through misattribution, suggestibility, and bias would be the result of an essentially normal mental functioning related to the updating of memory traces. Individuals with mediotemporal lesions and amnesia have a lesser tendency to develop these memory distortions than healthy controls (Koutstaal, Verfaellie, & Schacter, 2001; Verfaellie, Schacter, & Cook, 2002). In parallel, it has been shown that this tendency is linked, in the encoding phase, to the activity of the prefrontal ventromedial area (Garoff, Slotnick, & Schacter, 2005; Kim & Cabeza, 2006; Kubota et al., 2006), which in turn is linked to the

semantic coding of novel materials. This led to the hypothesis that semantic coding, which has great adaptive value, could contribute to the distortions of memories (Schacter et al., 2011). The existence of a shared system for the encoding of real memories and for the errors in the incorporation of information (e.g., Baym & Gonsalves, 2010; Okado & Stark, 2005) could lead to the conclusion that false memory is the ultimate consequence of the tendency to update memory in order to guarantee better adaptability (Schacter et al., 2011).

Within this context, a very relevant point has been made by Klein (2013). Klein argued that "from an evolutionary perspective, memory's function is to enable its owner to face life as it comes, rather than to look back as it recedes" (Klein, 2013, p. 223). Indeed, despite it cannot be denied that some parts of the memory system must depend on certain past events, this dependence does not mean logically that memory, either in the retrieval phase or from an adaptive point of view, is linked in a specific way to the past. It would be an erroneous conceptualization of memory function and purpose to claim that since memory depends on the past, it is necessarily about the past.

The view that Klein (2013) proposed is that memory *depends* on the past but is not *about* the past. The contrast between *depending* on and being *about* is more than a grammatical or semantic problem. From this point of view, the fallacy of memory-as-a-passive-store becomes evident when one considers retention as the adaptive function of memory. Obviously, there is some kind of storage function that guides memory processes and that refers explicitly to the past (e.g., to know the way between home and work), but this will always be used at a future point in time (e.g., its purpose is to be able to go to work).

In order to further restructure this concept, as Klein (2013) did, it can be useful to consider memory as an object and distinguish its capacities – what we can do with it – from its function – the activity that it evolved for (e.g., Anderson, 1991; Cosmides & Tooby, 1992; Klein, 2007; Klein, Cosmides, Tooby, & Chance, 2002; Williams, 1966). The fact that an object or a system can execute a certain process does not mean that it has evolved for it.

The argument that Klein (2013) presents is quite strong: the fact that human beings are capable of remembering does not mean that this function is the principal purpose of

memory; it could simply be a collateral result of another process. This restructuring of the theme is then incorporated in an adaptive perspective: it is clear that memory constitutes an extremely evolved system, refined by natural selection, and that it exists in its current form because it enables the organism in some way to adapt to the environment more effectively (e.g., Glenberg, 1997; Howe, 2011; Klein et al., 2002; Nairne, 2005; Sherry & Schacter, 1987). The evolutionary process that it underwent in past millennia resulted in a functional organization that was particularly adaptive and therefore contributed to the survival and reproduction of the individuals that possessed it (e.g., Barkow, Cosmides, & Tooby, 1992; Howe, 2011; Klein, 2014; Mayr, 2001; Nairne, 2005; Sherry & Schacter, 1987; Williams, 1966). By this reasoning, possessing a precise knowledge of a certain fact but without any link with the environment (e.g., the day I was attacked by a bear) does not make sense biologically, but to possess some knowledge about how a certain state of things tends to express itself (e.g., how a bear moves when it wants to attack) would enable an individual to change its own behavior and to anticipate what is going to happen. The effects of classical conditioning and associative learning provide the simplest examples of a form of memory that works this way (e.g., Ginsburg & Jablonka, 2007), whose principal functionality would be linked to anticipation – and therefore to prediction – rather than recalling the past.

As Klein (2013) argued, one of the factors that may have influenced the current conception of memory is the subjective experience that we have of remembering, especially of the episodic component. However, if one considers the temporal orientation of procedural memory, there is considerable ambiguity; the fact that it cannot be immediately verbalized makes it in a certain sense disconnected from the past and difficult to express other than through a present behavior that is oriented toward the future. In this category belong all those actions that we do automatically, and which have very little to do anymore with the moment in which they were learned, but which serve only to guarantee a motor response that is adapted to the environment.

In parallel, semantic memory too, once it is used to guide behavior – and hence to predict – can be oriented toward the future. As indicated before, maintaining a piece of information without any application would be useless, but to know the way of functioning of a certain entity or a state of affairs – extracted from repeated exposure – enables one to

predict how it is going to behave. In this sense, the pioneering studies of Bartlett (1932) on schemas and extraction of statistical regularities, as well as more recent findings (e.g., Dudai, Karni, & Born, 2015; Ghosh & Gilboa, 2014; Nadel, Hupbach, Gomez, & Newman-Smith, 2012), call into question the link between memory and temporality. Furthermore, it has been repeatedly shown how the retention of episodic memories involves, in the intervening days, a radical transformation of the material and leads to a semantic storage (for a review, see Dudai et al., 2015).

Episodic memory falls outside of this conception only apparently: while on the one hand one can assert that, in light of the distortions that it undergoes (for a review, see Loftus, 2005, 2013), its work consists in the collateral manifestation of a system whose purpose is to construct semantic knowledge, on the other hand one can bring the argument from the adaptive point of view. It has been shown that episodic memory and prospection share common neural substrates (e.g., Addis et al., 2007, 2009; Thakral et al., 2017), which could suggest that during evolution *Homo sapiens* managed to use for adaptive purposes structures that initially had no clear predictive functioning.

Within this context we proposed a memory model which describes both accurate retrieval and memory transformation (Vecchi & Gatti, 2020). There would exist two systems of memory, linked to the way the material is encoded. First, there is the normal memory system, whose primary purpose is to construct models of predictive interaction with the environment; it performs its functions through transformation of memories by automating certain aspects and integrating them into preexisting schemas or scripts. The second is a sensory-like system, characterized by a greater involvement of sensory, motor, and emotion areas during recall; contrary to the first memory system, the second memory system tends to retain less important contextual details of the event with high fidelity, is relatively vivid, and has little adaptive value.

Note also that accurate retention of information constitutes a maladaptive property only when it concerns long-term memories. Having highly accurate short-term memory that can retain information with high fidelity is highly adaptive because this enables forming new knowledge where necessary. This system is shown to be maladaptive only when one is

incapable of subsequently integrating the new memory into the existing semantic schemes, and instead it remains a separate unit. In this sense, childhood memory would be similar to systems of this kind: while semantic knowledge is being constructed, it is necessary to maintain with accuracy a certain number of elements, so as to proceed to extract statistical regularities; this childhood memory system would normally gradually disappear in the course of cognitive development as semantic models become operational.

It may appear radical, but, following this line of reasoning, we can postulate that the use we make of memory in particular contexts (e.g., legal testimony, instruction, social relations, etc.) builds on a collateral function of a system that functions for a completely opposite purpose (Klein, 2013). This conclusion, however, does not take anything away from the importance of autobiographical memory for the construction of identity and autonoesis, even if this relationship is nowadays heavily debated (Medved & Brockmeier, 2015).

In conclusion, experimental evidence accumulated in recent years constitutes an important challenge for accounts of memory as an information retention system; instead, we proposed a new perspective in which the principal property of the system, valuable in an adaptive sense, would be that we remember because we must predict.

2.2

Is there a future for false memory?

In the previous section we outlined a new perspective on remembering and argued that memory is not a real memory system, but rather a cognitive system mainly involved in predictive functioning. Within this framework, false memory plays a relevant role, it is the real cornerstone of this work. False memory itself is the key to the future of memory.

Therefore, it could seem odd asking if there is a future for false memory, but it is not. An impressive number of studies investigated false memory across a large number of conditions and provided groundbreaking evidence regarding its pervasiveness.

For example, almost a century ago, in a series of pioneering studies on memory, Bartlett (1932) demonstrated that humans use their semantic knowledge to encode, store and remember information, generally adapting it to their own conceptual schemas and expectations, with systematic errors that may occur when the same individuals have to recall the presented material. Hence, participants would forget the precise features of the material that has been previously presented in favor of an extraction of the gist of the information: during these phases, memory distortions (i.e., false memories) have been repeatedly shown to occur (e.g., Bartlett, 1932; Brewer & Treyens, 1981; Sulin & Dooling, 1974). Similarly, in the 70s, Elizabeth Loftus (e.g., Loftus, 1975, 1977; Loftus, Burns, & Miller, 1978; Loftus & Palmer, 1974) created and probed an experimental paradigm, known as the *postevent misinformation effect paradigm*, which is a method to reproduce a witness setting in a laboratory and induce false memories in the participants. Other studies have shown that it is possible to experimentally induce never-happened childhood events through the suggestion

that parents remembered them (e.g., being lost in the supermarket when they were a child); this was enough to produce recall of that memory through particular methods and suggestions (Hyman, Husband, & Billings, 1995; Loftus & Pickrell, 1995; Porter, Yuille, & Lehman, 1999). More recently, it has been shown that it is possible to induce false memories of logically or objectively impossible events (e.g., Braun, Ellis, & Loftus, 2002; Mazzoni & Memon, 2003; Thomas & Loftus, 2002; Wade, Garry, Read, & Lindsay, 2002) or very unlikely ones, such as proposing marriage to a Pepsi machine (Seamon, Philbin, & Harrison, 2006).

However, the perhaps most widely used tool to investigate false memories is the Deese–Roediger–McDermott task (DRM; Deese, 1959; Roediger & McDermott, 1995), which is the main focus of the next section and of the experimental studies presented in this dissertation. In the DRM task participants are asked to memorize several lists of words and then to recognize them after a brief distracting task. The semantic / associative / orthographical / phonological relationships among the to-be-studied words are generally manipulated to investigate how such processes contribute to false recognitions (and recalls) of non-showed words.

From a cognitive standpoint, false memories obtained through the DRM task have been generally traced back to gist extraction from studied information, and thus to semantic and associative processes (Brainerd & Reyna, 2002; Gallo & Roediger, 2002; Reyna & Brainerd, 1995; Roediger, Watson, McDermott, & Gallo, 2001).

At this point, we could ask another a relevant question: *does the production of false memories have an adaptive basis, or does it reflect only a maladaptive aspect of memory?*

One key to answering this question is to analyze the consequences of memory distortions for behavior. Some studies have shown that participants who were more prone to developing false memories tended also to update their behavior as a result (Berkowitz, Laney, Morris, Garry, & Loftus, 2008; Bernstein, Laney, Morris, & Loftus, 2005a, 2005b).

In two studies conducted by Bernstein and colleagues (2005a, 2005b), in particular, participants who developed a false memory about eating problems during childhood, in conjunction with a particular type of food, were significantly more reluctant to appreciate this food than a control group. Geraerts and colleagues (2008) obtained the same results and showed that the avoidance behavior was maintained even 4 months later.

The existence of phenomena such as false memories brings forth critical questions about memory and how it works. As indicated previously, if the ultimate purpose of memory is to retain information, these phenomena imply that the system performs this purpose in a very limited fashion. However, another point of view could be that these characteristics guarantee a better adaptation to the environment (e.g., Schacter et al., 2011), and if so, this implies a new framework for memory research.

In particular, what emerges from several decades of investigation of false memories is that these can be induced in a large percentage of the population and that, furthermore, they influence the future behavior of these individuals; in this way memory would be conceptualized as a predictive system. For example, it is possible to hypothesize that the inclusion of memories during experiments using the DRM paradigm reflects an attempt to anticipate a particular element by recalling the corresponding semantic schema, and in the same way the ease with which false memories are created reflects the plasticity of a system that enables the best possible updating of the existing memories.

So, in the end, yes, there is a future for false memory, maybe a bright future detached from the past. There is a future for research in false memory, and there is a future for false memory itself: a place cognitively located forward in time. A place that links distortion and anticipation, a place in which memory transformation is not an error – something that cause frustration – but the evidence of an healthy memory functioning.

2.3

The DRM task: procedure and theories

In the previous sections we argued that memory should be conceived as a system dealing with the future, rather than with the past and outlined how and why false memory should be considered as the cornerstone of this perspective.

In the present section, then, we will describe an experimental paradigm, the Deese–Roediger–McDermott task (DRM; Deese, 1959; Roediger & McDermott, 1995), which is one of the most popular tasks used to investigate false memory.

The DRM task is typically divided in two phases: in the encoding phase, participants are presented once with several lists of words that have to be memorized (within each list, the words are generally semantically / associatively related to a non-shown target word, named *critical lure*; e.g., mad, fear, hate, rage, temper, fury, etc. – critical lure: anger); in the second phase, participants have to indicate whether a given word was part of the memorized lists or not (i.e., recognition task). During the recognition task, participants tend to report as "old" the critical lures (i.e., as if they were part of the memorized lists), although these stimuli were never presented in the previous experimental phase (for a review, see Gallo, 2010).

The false recognition of the critical lure is not the only memory distortion that can occur in the DRM task. That is, other studies also employed as lures words that are semantically weakly related to the presented words (generally presenting shorter lists of words and using the excluded words as *weakly related lures*; e.g., Roediger & McDermott,

1995) and, despite in a lower proportion compared to critical lures, participants misrecognized them as "old". Similarly, participants may sometimes even report as "old" stimuli considered as (completely) unrelated to the studied words, this category being associated with the lowest occurrence of false memories (e.g., McKelvie, 2003, 2004). The distinction of lure into somewhat separate categories, i.e. critical lures, weakly related lures and unrelated words, has been often adopted in previous studies (e.g., Kloft et al., 2020; Roediger & McDermott, 1995), as it reflects how the DRM task is created. These categories are generally obtained from norms derived from human ratings. At times the creation of the DRM test starts by selecting some target words (i.e., the critical lures) and then asking participants to indicate the words mostly associated with each target word. The associated words will be next ordered based on their frequency of response occurrence (e.g., Iacullo & Marucci, 2016). During the encoding phase of the DRM task this order will determine the sequential presentation of the words to be remembered. Generally, the weakly related lures will be selected from last positions of the list (i.e., usually between the 13th and the 15th positions, for 15-items lists), whereas the unrelated words are taken from other lists not presented in the encoding phase. Alternatively, it is possible to build lists using norms that provide a language-based measure of the strength of the association between one word (i.e., the prime) and another word (i.e., the target). In this way, lists are built around a target word, placing the words in a list in decreasing order of associative strength (e.g., Nelson, McEvoy, & Schreiber, 1998).

Two main theories have been proposed to explain participants' performance in the DRM taks, the *fuzzy-trace theory* (FTT – Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995) and the *activation-monitoring framework* (AMF – Gallo & Roediger, 2002; Roediger et al., 2001). These two theories outline different cognitive processes that may create memory distortions, but both imply a role of semantic memory in false memories formation.

According to the FTT, during the encoding phase, two memory traces are encoded: a verbatim trace, linked to perceptive features of the stimuli, and a gist trace, linked to the semantic theme of the list (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). Both traces would be involved in the correct recognition of previously showed words, but the verbatim trace would be specifically involved in correct rejection of new stimuli, while only the gist

trace would be responsible for the production of false memories. On the other hand, according to the AMF, the critical lure would be activated during the encoding phase and this would be responsible for the production of false alarms, while successful source-monitoring processes would account for both correct rejections and recognition of studied items (Gallo & Roediger, 2002; Roediger et al., 2001). Thus, activation is thought to enhance false memories, while monitoring is thought to reduce them. That is, during the study phase, the presentation of each word would cause an over-activation, via spread of activation between preexisting conceptual representations in a mental lexicon, of the critical lure (Roediger, Balota, & Watson, 2001). Activation is not related to a single linguistic process but includes any process that mentally activates the related lure from a semantic, phonological and syntactic point of view (Gallo 2010). Indeed, it has been shown that the backward associative strength (i.e., the association strength from the word that composes each list to the critical lure; BAS) is a strong predictor of participants' performance in the DRM task (Roediger et al., 2001). On the contrary, monitoring is described as any memory process that helps to detect the origins of the activated information (Gallo, 2010). Monitoring can be divided into diagnostic and disqualifying: the former describes memory processes relying on expectations, while the latter describes memory processes relying on collateral information (Gallo, 2004).

Finally, while FTT assumes that list words and critical lures share semantic features that underlie false recognition (i.e., false recognition is semantic in nature), AMF assumes that false recognition occurs due to the associative link between the critical lure and the list words (i.e., false recognition is associative in nature). The difference might seem subtle, since very often two words that are associatively related are also semantically related (Grossman & Eagle, 1970; Thompson-Schill, Kurtz, & Gabrieli, 1998; but cfr also: Brainerd et al., 2008). Yet, such differences appear critical when trying to describe the cognitive processes involved in the DRM task (see Gallo, 2013). For instance, the FTT better accounts for false memory in a wide range of ecological and non-ecological tasks (Brainerd & Reyna, 2002), despite the definition of semantic processes intervening is not precisely operationalized. On the contrary, the AMF, by quantitatively describing the association via the BAS, directly account for false recognition in the DRM task (Roediger et al., 2001), as well in several manipulations of this

task, such as with the use of hybrid lists of semantic and phonological associates (Watson, Balota, & Roediger, 2003).

Given its structure and its reproducibility, the DRM is probably one of the best paradigms to investigate how false memory emerges, how it is stored and retrieved and how it interacts with other memory and cognitive processes in general.

2.4

Overview of the studies

In Study 1 we predict participants' performance in the DRM task by taking advantage from recent distributional semantic models. In particular, we develop a new method to compute semantic similarity among the to-be-recognized words in the DRM task and their reference list.

In Study 2 we apply the same method developed in Study 1 and we investigate how such semantic predictor predict participants' performance in the DRM as measured by mouse-tracking dependent variables.

In Study 3 we adopt an individual differences approach to investigate how episodic and semantic memory are involved in the DRM task. This approach is critical within a task such as the DRM in which both memory systems are thought to be involved, but only semantic components can be predicted at the item level.

In Study 4 we merge an individual differences approach with the inferences made possible through the use of distributional semantic models. This approach demonstrates how the method developed in the first study can be applied on the behavioral level.

Finally, in Study 5 we investigate the role of the cerebellum in semantic memory processes administering online transcranial magnetic stimulation while participants perform the DRM task. We predict participants' performance using the above-mentioned method extracted from distributional semantic models and we show how it can be applied also on more complex paradigms, such as brain stimulation ones.

3. Behavioral evidence

3.1

Decomposing the semantic processes underpinning veridical and false memories

The present study is published in its extended and definitive version on *Journal of Experimental Psychology: General*. To cite it:

Gatti, D., Rinaldi, L., Marelli, M., Mazzoni, G., & Vecchi, T. (2021). Decomposing the semantic processes underpinning veridical and false memories. *Journal of Experimental Psychology: General*.

3.1.1 Introduction

Human memory is far from being an accurate recorder, as retrieval of long-term memories involves active rather than passive reproduction (Schacter, 2021; Vecchi & Gatti, in press 2020). Almost a century ago, in a series of pioneering studies on memory, Bartlett (1932) demonstrated that humans use their semantic knowledge to encode, store and remember information, generally adapting it to their own conceptual schemas and expectations, with systematic errors that may occur when the same individuals have to recall the presented material. Hence, participants would forget the precise features of the material that has been previously presented in favor of an extraction of the gist of the information: during these phases, memory distortions (i.e., false memories) have been repeatedly shown to occur (e.g., Bartlett, 1932; Brewer & Treyens, 1981; Sulin & Dooling, 1974).

From a cognitive standpoint, false memories obtained with some specific procedures have been generally traced back to gist extraction from studied information, and thus to semantic and associative processes (Brainerd & Reyna, 2002; Gallo & Roediger, 2002; Reyna & Brainerd, 1995; Roediger, Watson, McDermott, & Gallo, 2001). The perhaps most widely used tool to investigate false memories is the Deese–Roediger–McDermott task (DRM; Deese, 1959; Roediger & McDermott, 1995). The DRM task is typically divided in two phases: in the encoding phase, participants are presented once with several lists of words that have to be memorized (within each list, the words are semantically/associatively related to a non-shown target word, named *critical lure*; e.g., mad, fear, hate, rage, temper, fury, etc. – critical lure: anger); in the second phase, participants have to indicate whether a given word was part of the memorized lists or not (i.e., recognition task). During the recognition task, participants tend to report as "old" the critical lures (i.e., as if they were part of the memorized lists), although these stimuli were never presented in the previous experimental phase (for a review, see Gallo, 2010).

The false recognition of the critical lure is not the only memory distortion that can occur in the DRM task. That is, other studies also employed as lures words that are semantically weakly related to the presented words (generally presenting shorter lists of words and using the excluded words as *weakly related lures*; e.g., Roediger & McDermott, 1995) and, despite in a lower proportion compared to critical lures, participants misrecognized them as "old". Similarly, participants may sometimes even report as "old" stimuli considered as (completely) unrelated to the studied words, this category being associated with the lowest occurrence of false memories (e.g., McKelvie, 2003, 2004). The distinction of lure into somewhat separate categories, i.e. critical lures, weakly related lures and unrelated words, has been often adopted in previous studies (e.g., Kloft et al., 2020; Roediger & McDermott, 1995), as it reflects how the DRM task is created. These categories are generally obtained from norms derived from human ratings. At times the creation of the DRM test starts by selecting some target words (i.e., the critical lures) and then asking participants to indicate the words mostly associated with each target word. The associated words will be next ordered based on their frequency of response occurrence (e.g., Iacullo & Marucci, 2016). During the encoding phase of the DRM task this order will determine the

sequential presentation of the words to be remembered. Generally, the weakly related lures will be selected from last positions of the list (i.e., usually between the 13th and the 15th positions, for 15-items lists), whereas the unrelated words are taken from other lists not presented in the encoding phase. Alternatively, it is possible to build lists using norms that provide a language-based measure of the strength of the association between one word (i.e., the prime) and another word (i.e., the target). In this way, lists are built around a target word, placing the words in a list in decreasing order of associative strength (e.g., Nelson, McEvoy, & Schreiber, 1998). Yet, such a categorical distinction in critical lures, weakly related lures and unrelated words may be arbitrary and a more complex model, including also other predictors, might better capture the semantic relationship between the words composing the DRM task, especially considering that the norms used to build the different categories of lures are generally derived from specific linguistic contexts (i.e., individuals from specific geographical areas and with unknown sociodemographic characteristics) and may thus lack generalizability even within the same language.

One element that has not been fully examined is that the distinction between critical lures, weakly related lures and unrelated words may not fully reflect the complex structure of the mental lexicon, especially considering that recent models of semantic memory conceive word meanings as distributed representations populating a continuous mental space (Jones, Willits, & Dennis, 2015). Some hints supporting a continuous gradient of semantic involvement in false memories formation came from computational models representing the associative connections between words in memory by means of association norms collected from humans. By applying scaling methods on free-association norms, these former computational models were indeed able to replicate some memory-related phenomena, including the occurrence of false memories in the DRM paradigm (Kimball, Smith, & Kahana, 2007; Steyvers, Shiffrin, & Nelson, 2005). This represented a crucial step, as these models provided initial insights on the processes subserving memory distortions (for a similar approach see also Montefinese, Zannino & Ambrosini, 2015).

Yet, a more comprehensive theoretical framework on semantic memory most likely provides a better account for the cognitive structure responsible for false memories; that is, an ideal framework should also unveil that these cognitive phenomena adhere to the specific

structure of human semantic memory. Initial evidence for this comes from other computational studies employing distributional semantic models (DSMs) that represent word meanings as high-dimensional numerical vectors, extracted from large amounts of natural language data. DSMs, indeed, fulfil the criteria to be considered as psychological models of the nature of semantic representations and the structure of semantic memory (Günther, Rinaldi & Marelli, 2019).

Notably, evidence for these models as theories of human semantic representations is not lacking (Günther et al., 2019). That is, DSMs fulfill the criteria to be considered as psychological models of the nature of semantic representations and the structure of semantic memory (see Jones et al., 2015). Such a theoretical assumption is also supported from an empirical standpoint, as DSMs have been shown to be impressively high-performing across a wide range of semantic tasks (for a review on the recent prediction-based class of models, see Mandera et al., 2017). For instance, they can achieve perfect scores on multiple-choice tests (Bullinaria & Levy, 2012), above 95% in word categorization (Baroni & Lenci, 2010), and correlations around .8 with human word similarity ratings – a score comparable with that of human inter-rater agreements (Bruni, Tran, & Baroni, 2014). Moreover, they have been shown to perform extremely well in synonym tests (Bullinaria & Levy, 2012; Landauer & Dumais, 1997). Indeed, in DSMs two words will be similar in meaning not because of their mutual co-occurrence score but rather because they have similar global distributional patterns. Hence, words that never occur together (i.e., in the same linguistic contexts) can nevertheless end up with very similar meaning representations. Essentially, the distributional similarity between two words as extracted from language will amount to their degree of mutual substitutability: in language usage, instructor can be more easily substituted with teacher as compared to scientist, with this propriety being efficiently captured by DSMs.

Perhaps more critically, recent DSMs, such as word-embeddings, are also mathematically related to psychologically plausible learning models. Indeed, recent DSMs based on a prediction principle (neural networks that learn to predict a target word on the basis of its lexical contexts) are consistent with relatively simple, psychologically grounded associative learning mechanisms (Günther et al., 2019; Louwerse, 2018; Rinaldi & Marelli, 2020).

Interestingly, building upon these models, some studies (Johns & Jones, 2009; Johns, Jones, & Mewhort, 2012) reproduced the typical false memory patterns described in seminal studies. Perhaps more interestingly, a very recent study (Osth, Shabahang, Mewhort, & Heathcote, 2020) applied to different recognition memory databases a DSM based on psychologically plausible learning mechanisms (BEAGLE; Jones & Mewhort, 2007). Specifically, by re-analyzing data from previous works, Osth and colleagues (2020) found that the higher the semantic similarity between the lure and the words composing the lists, the greater the performance impairment (i.e., occurrence of false memories).

Here, we build on these pioneering studies to systematically investigate the specific semantic processes involved in a typical false memory task (i.e., DRM), by taking advantage from a recent family of DSMs, namely, word-embeddings. Word-embeddings (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, Chen, Corrado & Dean, 2013) are based on a predictive component (i.e., for this very reason they are also referred to as "predict" models; Baroni, Dinu, & Kruszewski, 2014), inducing word vectors using a neural network architecture with one hidden layer, which is optimized to match a target word. Briefly, these models are trained on large collections of texts that document natural language use. Nodes in the input and output layers represent words, and a neural network learns to predict a target word on the basis of the lexical contexts in which it appears (i.e., the words it co-occurs with in the text), incrementally updating a set of weights by minimizing the difference between model predictions and observed data at each learning event (i.e., every occurrence of the target word). The estimated sets of weights will eventually capture word meanings. These distributed representations, or vectors, can be quantitatively compared by measuring their distance in a multidimensional space, which in turn is thought to capture semantic similarity between words (Günther et al., 2019): similar words will occur in similar contexts, ending up being associated with vectors that are geometrically closer. Importantly, word-embeddings have been shown to be high-performing across a wide range of semantic tasks (for a review on the recent prediction-based class of models, see e.g., Baroni et al., 2014). Moreover, they are equivalent to psychologically grounded associative learning models (Günther et al., 2019; Mandera, Keuleers, & Brysbaert, 2017). For these reasons, they are the ideal tool to capture the semantic bases of false memories (for a graphical representation

of the semantic similarity structure for a DRM list of words as extracted from DSMs, see Figure 1).

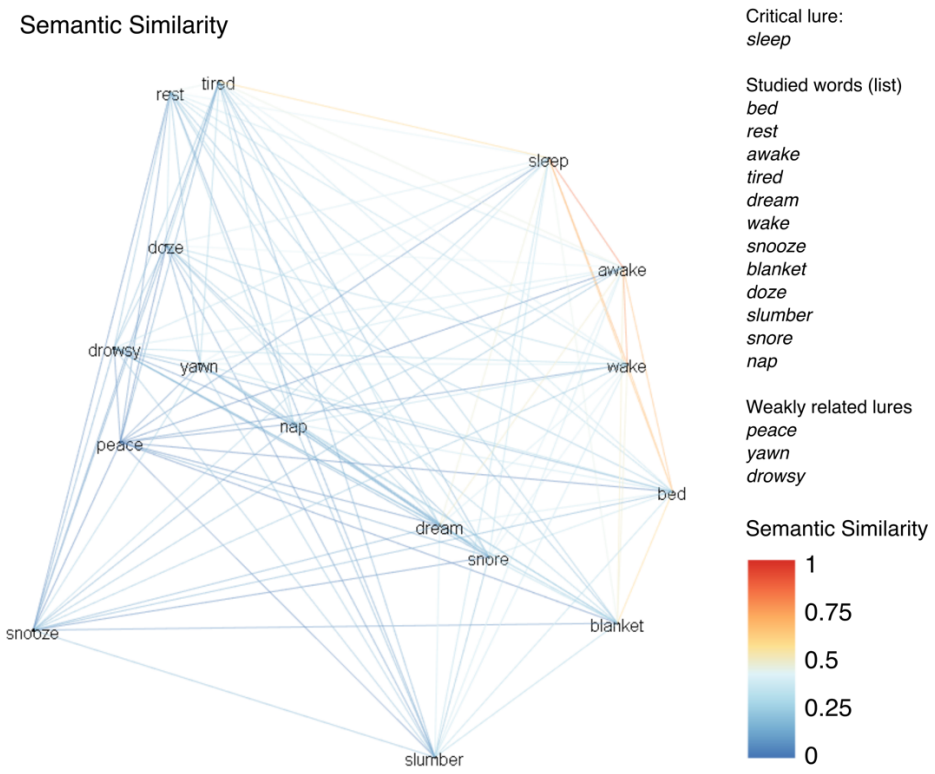


Figure 1. A two-dimensional projection of the semantic similarity structure among the words composing the list *sleep* (i.e., critical lure; words list taken from Roediger et al., 2001) in an actual semantic space obtained using the package *LSAfun* (Günther, Dudschig, & Kaup, 2015) with *R-Studio* (RStudio Team, 2015). Warmer red colors represent higher semantic similarity. Note that in this case we plotted all the words in the list (i.e., including the possible weakly related lures).

On these grounds, in the present study we used semantic similarity estimates as extracted from word embeddings to uncover the involvement of semantic processes in a typical DRM task. Participants performed an Italian version of the DRM task, in which they were asked to memorize several lists of words and then to perform a recognition task. We thus predicted false memories from a word-embeddings model based on Italian language (Marelli, 2017).

First, using a continuous predictor (i.e., semantic similarity), we directly tested the above-mentioned theories about semantic-based false memories in the DRM task. We expect that, if semantic memory is involved in memory distortions, false memories should increase

for words with higher semantic similarity as compared to those with lower semantic similarity. Second, by means of this approach, we aimed at disentangling the exact semantic strategies adopted by participants and, in particular, whether they employed a "local" or "global" approach on the task at hand. This point in our opinion is crucial because in the DRM task participants are asked to memorize several words that can be semantically grouped in clusters (i.e., the lists that compose the task) but are presented one list after the other without interruptions, thus without clues that could help the participants to cluster the words (and, hence, as if they were presented in a whole single list) (e.g., Díez, Gómez-Ariza, Díez-Álamo, Alonso, Fernandez, 2017). It is thus possible that performance for new words is driven either by global (i.e., semantic similarity between each new word in the recognition task and all the words studied during the encoding phase) or local semantic processes (i.e., semantic similarity between each new word in the recognition task and the cluster of the words studied during the encoding phase in each list). Indirect evidence supporting the role of global semantic processes comes from a very recent study showing that false memories can be predicted as a function of the semantic similarity between the lure and the studied words in a large range of list lengths (i.e., spanning from 28 to 150; Osth et al., 2020). We expect that, if participants adopt a global semantic approach, false memories should be better predicted by the semantic similarity between the target word in the recognition task and all the studied words in the encoding phase. On the other hand, we expect that, if participants adopt a local approach, false memories should be better predicted by the semantic similarity with the studied words that composed each semantic cluster.

Second, we also tested the extent to which semantic processes are similarly engaged in false and veridical memory recognitions. Indeed, according to the FTT, the semantic trace would be involved in both false and veridical recognitions, while a verbatim trace linked to perceptive features of the stimuli would be specifically involved in veridical recognition (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). Following this theory, we may expect that participants' veridical memory recognition would be predicted to a lower extent by semantic similarity, compared to false recognitions, as the former would benefit also from the verbatim trace.

3.1.2 Methods

Participants

Forty-four Italian students (8 males, *M* age = 22 years, *SD* = 1.6) participated in the experiment. All participants were native Italian speakers, had normal or corrected-to-normal vision and were naïve to the purpose of the study. Informed consent was obtained from all participants before the experiment. The study protocol was approved by the local ethical committee of the University of Pavia and all procedures were in accordance with the Declaration of Helsinki.

Stimuli

We used the DRM task (Deese, 1959; Roediger & McDermott, 1995), a typical false memories paradigm. Participants were first instructed to remember several lists of words presented continuously in a single session (i.e., without any clear marker in-between lists; encoding phase) and then to perform a recognition task. The words that composed each list were associatively related to a non-shown word (called *critical lure*).

For the encoding phase, we selected 16 lists of words out of 24 from the normative data for the Italian DRM test (Iacullo & Marucci, 2016; see Appendix A), which were arbitrarily divided in two sets of 8 lists. Each list was originally composed by 15 words: we selected the first 12 words (96 words for each set), while 2 of the 3 remaining words were used as weakly related lures (see below). Each participant was randomly asked to memorize one of the two sets of lists.

The recognition phase was composed of 72 words, 32 of which had been presented in the previous phase (i.e., studied words) and 40 of which had not been previously presented (i.e., new words). The 32 studied words presented in this experimental phase were those in serial positions 1, 4, 8 and 11[1] in the studied lists[2]. Of the 40 new words, 8 were the critical lures

[1] Given the structure of the task (i.e., in which the words occurring in the first positions of the list are more related to the critical lure), we opted for selecting these positions to have a wider range of relatedness between the studied words and the critical lure.

[2] In the list *giustuzia* (justice), the studied word *salto* (jump; in 2nd position) was removed and replaced with the next word because in our opinion the associative link between the two words was absent.

from the studied lists (i.e., the non-shown words mostly associated with the words composing each list), 16 were weakly related lures and 16 were unrelated words. The weakly related lures were 2 of the 3 words of the studied lists that were not presented in the list, generally those in position 13 and 14[3]. The unrelated words were those in serial positions 3, 6, 9 and 12 in the 8 non-used lists; this criterion was established arbitrarily.

Procedure

Participants were seated comfortably at a distance of 60 cm from a 17″ computer monitor. Stimuli were displayed on a computer monitor using Matlab (Mathworks, Inc.) and the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997).

Participants were told that they would take part to a memory experiment: they performed an encoding phase in which they were required to study several lists of words, immediately after they performed a distracting task in which they were required to solve math operations, and finally a recognition phase.

During the encoding phase, participants were showed the 12 words that composed each of the 8 lists in descending forward associative strength (i.e., the association strength from the critical lure to the word that compose the list). The order of the lists was random, but the words were blocked by list (see Roediger & McDermott, 1995). Each trial started with a central fixation cross (presented for 1000 ms) followed by a word (presented for 1500 ms); then the next trial was presented.

At the end of the encoding phase, participants were requested to solve as many arithmetical operations as they could for 2 minutes to interfere with recency memory processes (Roediger & McDermott, 1995). The arithmetical tasks were presented as a paper-and-pencil version.

In the recognition phase, participants were shown one word at a time and were instructed to respond if the word showed was old or new, that is, if they had been presented with the

[3] This criterion was violated in only two occasions: the list *giustizia* (justice) had only two extra words (due to the abovementioned exclusion of the word *salto* (jump)); in the list *freddo* (cold), we excluded the non-shown word *corrente* (stream) because this word was also a studied word in the list *fiume* (river).

word in the encoding phase ("old") or not ("new"). Participants were asked to respond as fast and as accurately as possible by pressing two buttons of a standard keyboard (e.g., J and K) using the index and middle fingers of their right hand; the response keys assignment was counterbalanced among participants (i.e., half of the participants classified "old" words with the button J and "new" words with the button K, with the opposite assignment for the other half of the participants). Each trial started with a central fixation cross (presented for 3000 ms) followed by a word (presented until response); after participant's response, the next trial began.

Word-embeddings

Vector representations for the words used in this study were extracted from a semantic space obtained by inducing word embeddings using the Continuous Bag of Words (CBOW) method, an approach originally proposed by Mikolov and colleagues (Mikolov, Chen et al., 2013). The model, released by Marelli (2017), was trained on itWaC, a free Italian text corpus based on web-collected data and consisting of about 1.9 billion tokens. The model used is set on the following parameters: *9-word co-occurrence window*, *400-dimension vectors*, negative sampling with $k = 10$, subsampling with $t = 1e^{-5}$. This set of parameters defines the learning procedure used to induce word vectors (Mikolov, Chen et al., 2013). CBOW indicates the applied learning procedure: when using CBOW, the obtained vector dimensions capture the extent to which a target word is reliably predicted by the contexts in which it appears. Co-occurrence window size indicates how large the considered lexical contexts are; in our case, a *9-word window* indicates that we estimated predictions concerning 4 words on the left and 4 words on the right of the target word. The number of vector dimensions indicates how many nodes are included in the hidden layer, representing the result of the dimensionality reduction process implicitly applied by the network. Negative sampling estimates the probability of a target word by learning to distinguish it from draws from a noise distribution; the parameter k specifies the amount of these draws. The subsampling parameter t specifies a threshold-based procedure that limits the impact of very frequent, uninformative words.

From this semantic space, we extracted vector representations for the words used in this study[4]. Specifically, for each word pair it is possible to obtain a semantic-similarity index (hence SSim) based on the cosine of the angle formed by vectors representing the meanings of these words. In particular, the higher the SSim value, the more semantically similar the words should be as estimated by the model. A two-dimensional projection of the semantic similarity structure among the words composing a DRM list is represented in Figure 1.

Computation of semantic similarity values

First, we extracted from the DSM the SSim between critical lures and each of the 12 words composing their corresponding lists (SSim *lure*).

Next, for each new word (16 critical lures, 32 weakly related lures and 32 unrelated words) we computed two SSim values: a *global* and a *local* index. The *global* index (Figure 2A) was computed as the frequency-weighted average SSim (for a similar approach see: Marelli & Amenta, 2018) between each word in the recognition phase and each of the 96 studied words presented in the encoding phase (i.e., the whole set of lists), using the following formula:

$$global\ SSim = \frac{\sum_{i=1}^{k} SSim_i \times F_i}{\sum_{i=1}^{k} F_i}$$

where $SSim_i$ refers to the semantic similarity between a new word and each of the i studied word, while F_i is the frequency of each studied word as extracted from the Italian SUBTLEX (http://crr.ugent.be/subtlex-it/).

The *local* index (Figure 2B) was computed as the frequency-weighted average SSim between each word in the recognition phase and each of the 12 words that composed its

[4] The relative semantic similarity indexes used in this Experiment can be reproduced through the freely available user-friendly interface at: http://meshugga.ugent.be/snaut-italian/
Note that these indexes were subtracted from 1 to transform the values on a proximity scale.

relative list. For unrelated words we computed the *local* index randomly matching each word with a list.

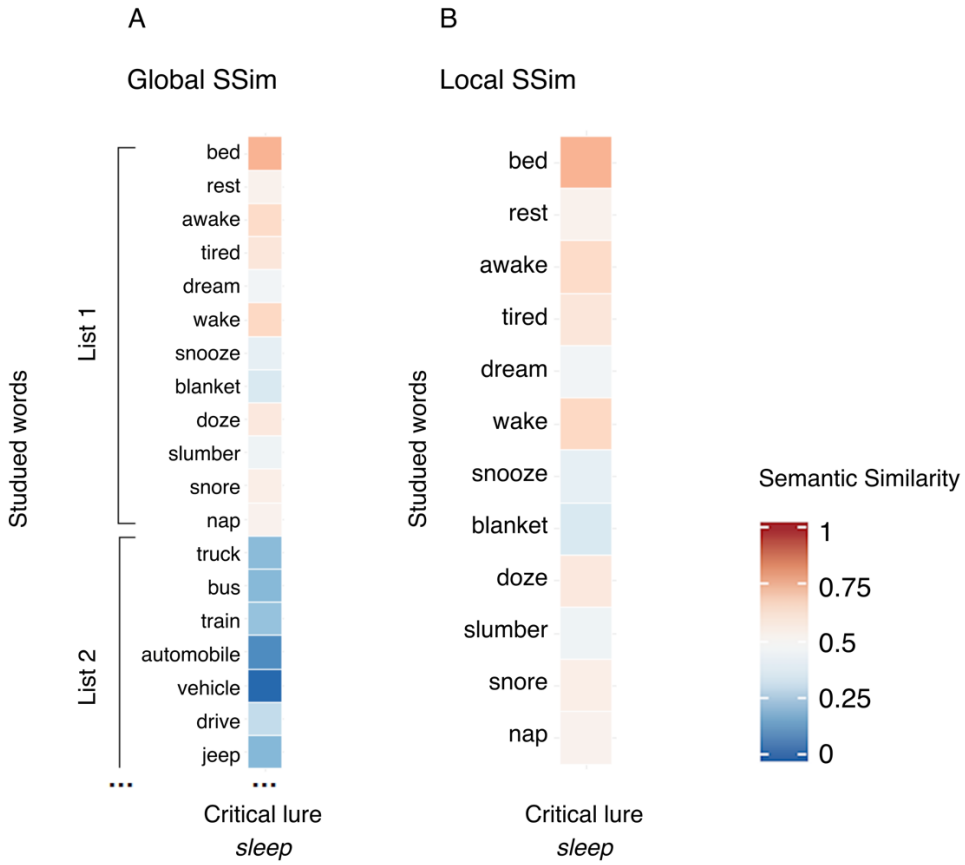


Figure 2. A graphical representation of the semantic indexes used in this study. The *global* SSim was computed as the mean semantic similarity between each new word presented in the recognition phase (i.e., whether critical lures, weakly related lures or unrelated words) and all the words studied from all lists presented to the participant (A). The *local* SSim was computed as the mean semantic similarity between each new word (i.e., critical or weakly related lures) and all the words composing its relative list (B). Note that in these graphical examples we opted for plotting semantic similarity values for a words list taken from an English study by Roediger and colleagues (2001). For coherence with the methods of Experiment 1, however, we only plotted values relative to 12 words in the list. Warmer red colors represent higher semantic similarity. The single images were obtained using the package *ggplot2* (Wickham, 2016) with *R-Studio* (RStudio Team, 2015). Note also that here we reported the data extracted from the DSM, but the *global* and *local* SSim values used as predictors in the present study were weighted using the frequency of the studied words.

Following the same rationale, for the studied words we computed two indexes, one *global* and one *local*. In this case, for the *global* index, we computed the frequency-weighted average SSim between each studied word presented in the recognition phase and each of the

other 95 studied words presented in the encoding phase. For the *local* index, we computed the frequency-weighted average SSim between each studied word presented in the recognition phase and each of the other 11 words that composed its relative list.

Data analysis

All the analyses were performed using *R-Studio* (RStudio Team, 2015). ANOVAs and t-tests were run using the *stats* package (R Core Team, 2019). Linear mixed models (LMMs) and generalized linear mixed models (GLMMs) were run using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015). The graphs reported were obtained using the *effects* package (Fox, 2003; Fox & Weisberg, 2019).

As a first sanity check, we tested in a series of analyses if DSMs can thoroughly account for the structure of the DRM task. In the DRM test, words with higher positions in the lists (i.e., those presented as first) should be more semantically related with the critical lure. We, therefore, explored whether the position of the words in the lists could be predicted by the SSim between each word in the list and the relative critical lure. In particular, we ran a LMM with the word position as dependent variable, SSim as predictor and list as random intercept. We also probed whether SSim describes arbitrary definition of new words in categories (critical lures, weakly related lures and unrelated words). In this case, we ran a one-way ANOVA with category (critical lures, weakly related lures, unrelated words) as between-subject factor. As dependent variable, we used the *global* SSim between each new word and the words studied in the whole session. Similarly, we ran a one-way ANOVA comparing *local* SSim between critical lures, weakly related lures and unrelated words.

Then, we explored the semantic processes subserving false memories formation. We analyzed separately participants' performance with new words (i.e., critical lures, weakly related lures and unrelated words) at the *global* vs. *local* levels. We thus ran two separate GLMMs, both having false memories (i.e., "new" responses were scored as 0, whereas "old" responses as 1) as dependent variable and subjects and items as random intercepts. In particular, in a first GLMM we included the global index as continuous predictor, while in a second GLMM we included the *local* index as predictor. The two models were then compared looking at the

Akaike information criterion (AIC), which returns an estimation of the quality of the model (Akaike, 1973). AIC allows to select the model that gives the most accurate description of the data. Models with smaller AIC values are to be preferred (Wagenmakers & Farrel, 2004). In this case, the model with the smaller AIC would better capture the semantic processes subserving false memories. After having identified the best model, we also explored any possible additional effect of SSim with respect to the typical categorical model (including the *a priori* distinction in critical lures vs. weakly related lures vs. unrelated words). We thus ran a new GLMM including as predictors the factor from the smaller AIC model and the categorical predictor, with subjects and items included as random intercepts.

We also explored whether SSim values could predict veridical recognition of studied words. Following a similar rationale as for the previous set of analyses, we explored whether the performance with studied words was better accounted by a *global* vs. *local* strategy. We thus ran two separate GLMMs, both having veridical recognition (i.e., "new" responses were scored as 0, "old" as 1) as dependent variable and subjects and items as random intercepts. Yet, in a first GLMM we included the *global* index as continuous predictor, while in a second GLMM we included the *local* index as predictor.

Finally, we tested the possible interaction between SSim and the type of stimuli (new words including critical lures, weakly related lures and unrelated words vs. studied words). We estimated a new GLMM with recognition (i.e., "new" responses were scored as 0, "old" as 1) as dependent variable and *local* SSim (i.e., as this resulted to be the best predictor in the previous analyses) and type of stimuli as predictors; subjects and items were included as random intercept.

3.1.3 Results

1. Do DSMs capture the structure of the DRM task?

The first LMM showed a significant effect of the position of the studied words in predicting SSim *lure*, $F(1,175) = 14.005$, $p = .0002$, *Pseudo-*R^2 (total) = .22, $b = -.01$. In particular, the higher the position of the studied word in the lists, the higher the SSim *lure* value (Figure 3A).

An ANOVA revealed a non-significant main effect of category on the *global* SSim, $F(2,77) = 2.34$, $p = .10$ (Figure 3B). The second ANOVA revealed a significant main effect of category on the on the *local* SSim, $F(2,77) = 26.82$, $p < .001$ (Figure 3C). Post-hoc comparisons using Bonferroni correction showed that the *local* SSim was higher for critical lures ($M = .19$, $SD = .08$) compared to both weakly related lures ($M = .10$, $SD = .08$), $t(77) = 4.18$, $p < .001$, and to unrelated words ($M = .04$, $SD = .04$), $t(77) = 7.27$, $p < .001$; similarly, *local* SSim was higher for weakly related lures compared to unrelated words, $t(77) = 3.77$, $p < .001$. Hence, the *local* SSim index seem to better replicate the distinctions in categories of the DRM task and represent valid candidates in possibly explaining participants' performance, as compared to the *global* SSim index, which in turn does not account for the categorical distinction among words.

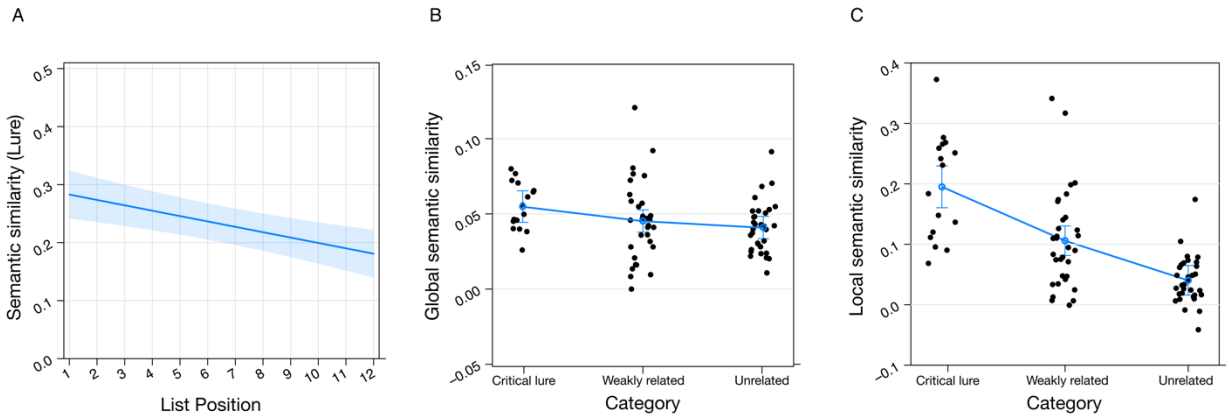


Figure 3. The structure of DRM task as captured by DSMs. The significant relationship between semantic similarity and words position in the list; in particular, the higher the semantic similarity between the critical lure and the words in the list, the higher their actual position (A). The distinction in categories (critical lures, weakly related lures and unrelated words) was not reflected by the *global* (B), but only by the *local* SSim indexes (C).

2. New words (false recognitions)

2.1 *Does SSim predict false recognition on a global level?*

The effect of *global* SSim on false recognition was found to be significant, $z = 2.31$, $p = .01$, *Pseudo-*R^2 (total) = .54. In particular, the higher the *global* SSim (i.e., the higher the

semantic similarity between the new word and all the studied words), the higher the chances of making false memories (Figure 4A).

On the contrary, the effect of *global* SSim on false memories was not found to be significant when considering only critical lures and weakly related lures, $z = 1.50$, $p = .13$, *Pseudo-R^2* (total) = .53.

2.2 *Does SSim predict false recognition on a local level?*

The effect of *local* SSim on false recognition was found to be significant, $z = 6.75$, $p < .001$, *Pseudo-R^2* (total) = .53. In particular, the higher the *local* SSim (i.e., the higher the semantic similarity between the new word and the studied words of each relative list), the higher the chances of making false memories (Figure 4B).

The effect of *local* SSim on false memories was found to be significant also when considering only critical lures and weakly related lures, $z = 4.04$, $p < .001$, *Pseudo-R^2* (total) = .51. In particular, the higher the *local* SSim, the higher the chances of making false memories.

2.3 *Model comparison*

Since both *global* and *local* SSim significantly predicted false recognition, we next compared the two AIC values to identify the best fitting model. The resulting AICs were AIC_{global} = 1406.53 and AIC_{local} = 1375.80. Hence, the model with *local* SSim outperformed the model with *global* SSim with a Δ *AIC* = 30.73. No comparisons were performed for the models including only critical lures and weakly related lures, since only the *local* model was found to be significant.

Having selected the best model in predicting participants' performance, we further ensured whether SSim – and not simply the distinction between arbitrary categories that is indirectly captured by SSim (see the previous results section, *Do DSMs capture the structure of the DRM task?*) – would account for false memories. We thus ran a GLMM with false recognition as dependent variable, *local* SSim and category (critical lures vs. weakly related lures vs. unrelated words) as additive predictors and subjects and words as random intercepts. We

found significant effects for both *local* SSim and category (Table 1). The former replicates the effect of *local* SSim previously described, while the latter indicates higher chances of false memories for critical lures as compared to weakly related lures.

Table 1. Results of the GLMM with false recognition as dependent variable, SSim and category (critical lures vs. weakly related lures vs. unrelated words) as predictors.

FIXED EFFECTS	***z*-value**	***p*-value**
(Intercept)	-.59	.55
local SSim	2.36	**.02**
Type – Unrelated word	-7.33	**< .001**
Type – Weakly related lure	-6.59	**< .001**

3. Studied words (veridical recognitions)

3.1 *Does SSim predict veridical recognition on a global level?*

The effect of *global* SSim on veridical recognition was not significant, $z = 0.76$, $p = .45$, *Pseudo-*R^2 (total) = .23 (Figure 4C).

3.2 *Does SSim predict veridical recognition on a local level?*

The effect of *local* SSim on veridical recognition was significant, $z = 2.96$, $p = .003$, *Pseudo-*R^2 (total) = .23. Thus, the higher the *local* SSim (i.e., the higher the semantic similarity between each word and remaining words studied in its relative list), the better the participants' veridical recognition (Figure 4D).

For studies words, therefore, no model comparison between *global* and *local* SSim was performed, as only the latter was found to predict participants' performance.

4. Interaction between local SSim and the type of stimuli

We finally investigated the extent to which veridical and false memories differently rely on a semantic basis. We therefore tested whether semantic similarity differently predicted veridical and false memory recognitions. In a GLMM including recognition (i.e.,

"new" responses were scored as 0, "old" as 1) as dependent variable, *local* SSim and type of stimuli as predictors, we found significant effects of both variables on recognition (*local* SSim: $z = 10.83$, $p < .001$; type of stimuli: $z = 7.52$, $p < .001$; *Pseudo-R^2* (total) = .50). Critically, the interaction *local* SSim by type of stimuli was significant, z = -2.33, p = .02. From an inspection of Figure 5, this significant interaction indicates that the effect of *local* SSim was smaller for veridical recognitions compared with false recognitions.

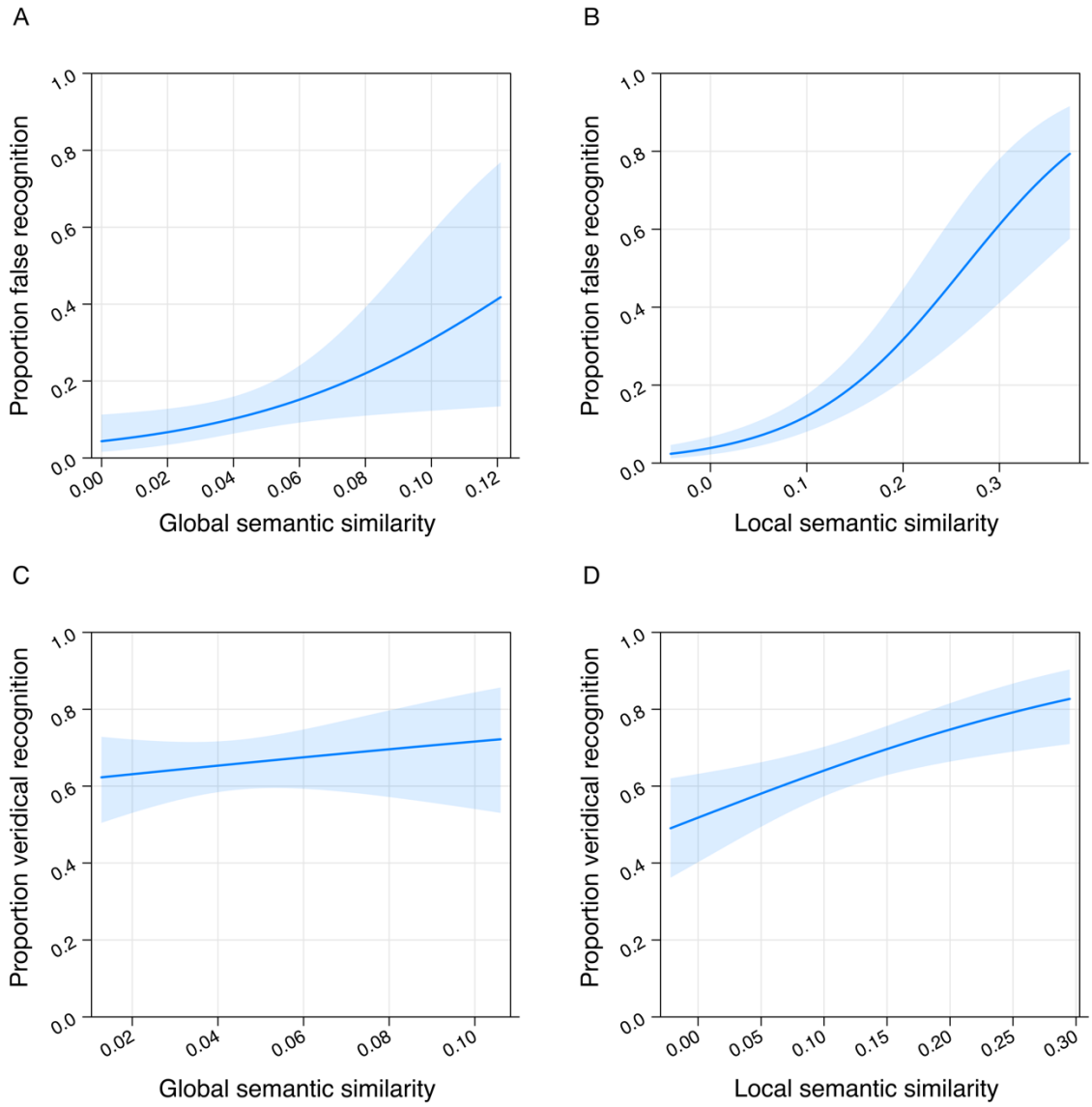


Figure 4. Results from the four GLMMs tested in Experiment 1 using *global* (A, C) and *local* SSim (B, D), illustrating the positive relationship between SSim and false (A, B) and veridical recognition (C, D). In both recognition types the model employing *local* SSim better described participants' performance as demonstrated by the lower AIC

value. Please note that the local SSim indexes are differently computed for false and veridical recognitions (see the section on the computation of semantic similarity values).

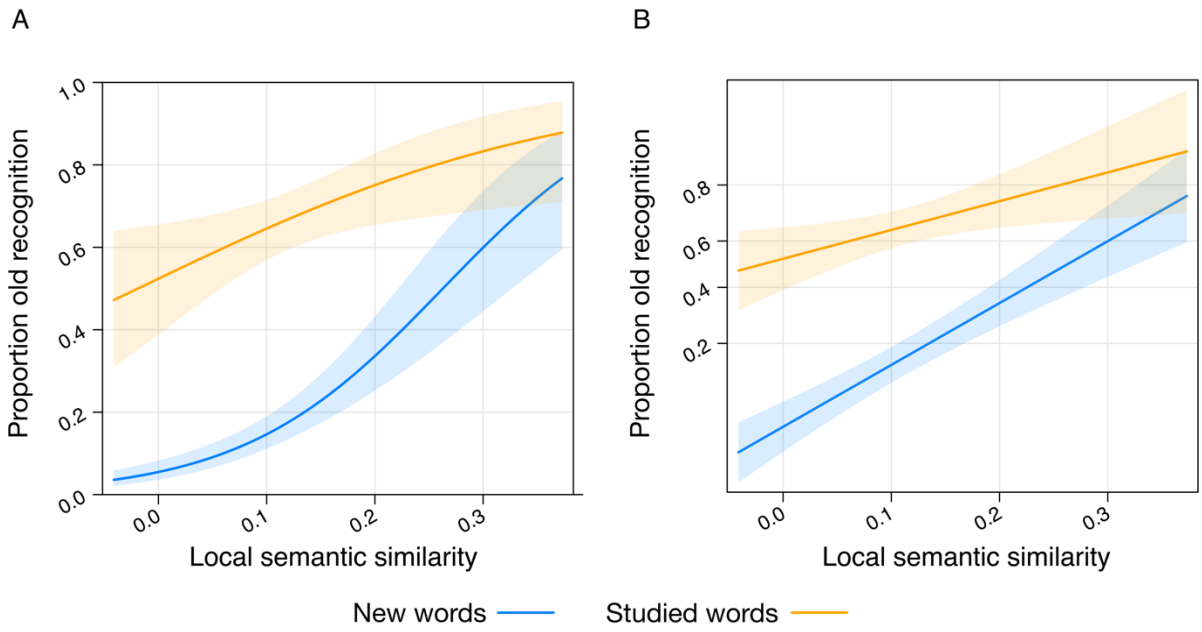


Figure 5. Results from the GLMM including the interaction between *local* SSim and type of stimuli on participants' proportion of old response using two different vertical axis specifications (type = "response", A; type = "rescale", in which is easier to interpret the interaction, B).

3.1.4 Discussion

In the present study, we investigated the semantic processes underpinning false and veridical memories taking advantage of recent machine-learning techniques from computational linguistics. In particular, we used a distributional semantic model (namely, word-embeddings) to directly probe the alleged involvement of semantic memory processes in false memory formation in the DRM task. Our findings indicate that higher memory distortions occur for more similarly related words. The origins of such distortions hark back to a *local* approach to the task at hand: the single lists composing the whole DRM test would activate specific semantic clusters responsible for the occurrence of false memories. In particular, we found that false memories occurred significantly more for higher sematic

similarity between the lure (i.e., the false memory item) and the words in the relative list. The adoption of a *local* strategy was substantiated by the fact that the effect was significant also when including the classical categorical predictor in an additive model. Notably, because the participants are presented with a stream of words that are not segmented by any breaks in the overall experimental session (i.e., no clear marker signaling the composition of each list), this clustering process would operate in a purely automatic fashion and would rely on semantic memory processes.

Results showed that, for newly presented words, participants' false memories were predicted by semantic relatedness both when computed as the mean of the semantic similarity between any new word and all the studied words (*global* SSim) or the studied words of its reference list (*local* SSim). Yet, a comparison between these the models including these indexes further indicated that false memories were better accounted by the adoption of a *local* strategy. This indicates that the higher the semantic similarity between the new words and the words studied in the relative lists, the higher the probability that participants produced false memories. Notably, this pattern of results was observed for both critical lures and weakly related words, thus further strengthening the view that semantic relatedness is a key responsible for false memories formation. In fact, even when only considering critical lures, false memories are more likely to occur for those critical lures that are more semantically related to the words of their relative list. The same argument applies for weakly related lures. Thus, whereas previous studies documented the occurrence of false memories on a categorical basis (i.e., more false memories for critical lures compared with weakly related lures; McKelvie, 2003, 2004; Roediger & McDermott, 1995, for a review: Gallo, 2010), here we demonstrate that, in addition to that effect, a different gradient of false memories can be detected within the same categories as a function of semantic similarity. Consistent with previous evidence on memory recognition (Osth et al., 2020), the same effect of *local* SSim was documented also for veridical recognition.

During the encoding task, participants were shown the lists of words without any clues that clustered the words. Critically, our findings – with the model including *local* SSim outperforming the one including *global* SSim – show that participants implicitly and automatically activated the semantic trace for each list and then used it during the

recognition task for new words. As a possible interpretative account for our findings, we propose that the sequential presentation of each word in the list would incrementally activate a cluster composed by the same words (for a graphical depiction see Figure 1). The semantic similarity between these words and the new word (i.e., whether critical lures or new words) would then affect (false) memory formation: the more the cluster of words composing each list would be semantically close to the new word, the higher the probability these new words would be misidentified as studied words.

Participants' performances in the DRM task have been mainly explained using the *activation-monitoring framework* (Gallo & Roediger, 2002; Roediger et al., 2001) or the *fuzzy-trace theory* (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). Despite being grounded on different cognitive processes, both these theories maintain a common semantic memory involvement in false memories formation and false recognition. In particular, according to the activation-monitoring framework, the semantic over-activation of the critical lure during the presentation of the words list in the encoding phase would be responsible for the production of false alarms, and successful source-monitoring processes would account for veridical recognition of studied words and correct rejections of new words (Gallo & Roediger, 2002; Roediger et al., 2001). According to the fuzzy-trace theory, false recognition would be tracked back to the gist trace (i.e., linked to semantic features of the studied words) as extracted during the presentation of the word lists and veridical recognition would be determined by both gist and verbatim trace (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). Compatibly with both these theories, our study shows that semantic memory is involved in both false recognition of new words and veridical recognition of studied words in the DRM task. Perhaps more crucially, our findings provide empirical support to the idea that semantic memory is differently involved in veridical and false memories, as predicted by the fuzzy-trace theory. Indeed, our results showed that veridical memories were predicted to a lower extent than false memories by the *local* semantic similarity index. In particular, for words with lower *local* SSim (that is, for words with a weaker gist trace) participants' proportion of 'old' responses was higher for studied words compared to new words, while this difference decreased as *local* SSim increased. We interpret this effect as a possible benefit given by the verbatim trace for studied words, which would be maximized when the gist

trace (here indexed by *local* SSim) is rather weak. Two other possible alternative explanations can account for this effect. First, since for studied words the verbatim trace is available, participants would rely on a lesser extent on the gist trace. Second, studied words that are more weakly associated might be more distinctive in their being qualitatively different compared with the other words comprising the list (e.g., when studying highly related words, such as *pipe*, *cigar*, *cigarette*, *tobacco*, the word *pollution* could stand out more), thus yielding a mnemonic advantage (by virtue of a sort of distinctiveness effect).

The present data have relevant implications also for the lively theoretical debate on the nature of semantic representations, and the relative influence of semantic vs. associative context on word processing, which has been mostly empirically grounded on findings from priming tasks (Hutchinson, 2003). In this task, participants are presented with a prime word immediately followed by either a related or an unrelated target word; typically, faster responses are observed when the target word is preceded by a related as compared to an unrelated word. Interestingly, both associative and semantic priming have been shown to occur in this task: *associative priming* occurs between words that are associated (i.e., hence reflecting word use, such as "spider-web"), while *semantic priming* occurs between words that share many semantic features (i.e., hence reflecting word meaning, such as "horse-pony"). Critically, these types of priming effects have been demonstrated to be independent from one another (Ferrand & New, 2003; Hutchinson, 2003).

Notably, semantic similarity metrics as extracted from DSMs have been successful in predicting priming effects: that is, the higher the semantic similarity index, the stronger the priming effect (i.e., faster reaction times from target words preceded by related primes) (Günther, Dudschig & Kaup, 2016). Moreover, some DSMs predict both associative and semantic priming effects (Günther et al., 2016; Jones, Kintsch & Mewhort, 2006). In line with these observations, word-embeddings can excellently predict human association data (i.e., word association norms) as well as human judgments of semantic similarity and relatedness (Mandera et al., 2017), which is also confirmed by the positive relationship found in the current study between BAS and the semantic similarity index. The fact that data from distributional semantic models can account for performance across a large number of associative and semantic tasks, including the DRM task used here (which taps on both

components), may indicate that the difference between associative and semantic processing could be fuzzier than previously assumed. Indeed, in the present study, the semantic similarity index was computed based on word embeddings from existing corpora, thus possibly including both natural associative and semantic information between words. Yet, when building word representations from language usage, the model used here adopts the very same computational operation (i.e., predicting a target word from the linguistic context in which it typically appears or vice versa) for all the words, not being geared towards capturing specific kinds of relations (differently from some count-models; Sahlgren, 2006). As a consequence, these two types of information are ultimately synthesized into a single vector and, hence, into a single word representation. The fact that the same process can thoroughly account for both associative and semantic information can be taken as an indirect evidence that, in a natural setting, these two types of information are likely to be interdependent when learning the meaning of words and may converge into partially overlapping structural representations of human memory. Associative and semantic processes can be indeed dissociated on the experimental level, but such a dissociation can be ascribed to the contingent task demands: for instance, guessing at associates or defining a concept would necessarily force participants to focus on different aspects of words (for a discussion see, Maki & Buchanan, 2008). Yet, in a natural context, it is almost impossible to find words that are purely semantically or associatively related (Jones et al., 2006).

It is worth noting that other previous studies have attempted to account for false memories formation starting from data obtained from distributional semantic models. First, while the topic model has been successfully used to predict participants' performance in the DRM paradigm and, more specifically, as a possible account of gist-based memory (Griffiths, Steyvers, & Tenenbaum, 2007), we argue that such evidence has been limited to free recall and based on aggregated data. Here, we do not only extend such predictive power to the recognition task, but also show that DSMs can be capable of predicting false memories at the individual item level (including a much larger set of studied and non-studied words and hence having, in turn, a much finer gradient of semantic similarity). Moreover, our model (word-embeddings) has the clear advantage of incorporating psychologically plausible learning mechanisms to create word representations (Murdock, 1982; Rescorla & Wagner,

1972). Indeed, word-embeddings are based on general-domain associative learning mechanisms compatible with human learning (Gunther et al., 2019; Rinaldi & Marelli, 2020). Second, our approach is clearly different from previous studies adopting a much more complex computational architecture, combining structured semantic representations with neural synchronization and information accumulation mechanisms specifically tuned for memory recognition (Jones et al., 2012). Our approach is different, as it is, on purpose, not developed to capture the process of memory retrieval. Rather, our findings show that false memories formation can be accounted by a model that solely replicates the structure of human semantic memory. The fact that a more parsimonious model can account for false memories formation indicates that the structure of the semantic network is in itself responsible for their occurrence in the DRM task. Finally, while previous models were entirely based on a single language (e.g., English), our findings extend their predictive power cross-linguistically by using freely-available data.

In this study, we demonstrate that memory distortions can be predicted by the semantic relatedness between words. In particular, our findings indicate that the presentation of a cluster of words (as the one presented when participants have to memorize a list in a DRM task) automatically activates a set of surrounding words in the human semantic memory system. This activation, captured by the semantic similarity index extracted from DSMs, is stronger for those words that are closer in the semantic space to the cluster of words presented and can lead to specific patterns of false memories. Our study, therefore, offers an empirically-driven explanation to account for false memories, tracing back their origins in the structure itself of human semantic memory. This, in turn, corroborates the view that the semantic memory system can be conceived as a continuous semantic space populated by elements (i.e., words), whose activation can be predicted by the way this system is structured.

Our study has relevant implications from a methodological standpoint. DRM lists are assembled starting from normative data, in which a group of representative individuals is asked to indicate the words most associated to several target words, and then by computing the associative strength from the lure to studied words or from studied words to the lure (respectively, forward associative strength or backward associative strength; Arndt,

2012; Brainerd, & Wright, 2005). This preliminary step has so far been generally required in the development of the task. By showing that computational models trained on natural language data can thoroughly replicate the structure of the DRM task, our findings indicate that such step may not be necessary, thus facilitating the construction of the test. Thus, future studies may use semantic similarity as extracted from these models to construct the lists. That is, the selection of the studied words and of the relative critical lures can be based uniquely on SSim. This applies as well to the selection of weakly related lures and unrelated words. Such data-driven implementation of the test could be particularly useful for languages that have no normative DRM data available, while for languages with associative data available a mixed approach could be more suitable (e.g., Steyvers et al., 2005). This, in turn, would also allow to systematically set analyses based continuous predictors.

Finally, just as there are well documented effects of semantics on false memories, robust levels of false recall and false recognition have been obtained with lists of phonological associates. Accordingly, in DRM literature, several authors provided evidence for false memories formation following orthographic/phonological lists (e.g., Dewhurst & Robinson, 2004; Finley, Sungkhasettee, Roediger, & Balota, 2017; Griffin, & Schnyer, 2020; Watson, Balota, & Roediger, 2003). To account for these findings, different approaches other than DSMs may be more suitable. That is, future studies investigating orthographically and phonologically driven false memories could predict participants' performance using metrics of string edit distance such as the Levenshtein Distance, which quantifies the orthographic similarity between words (Yarkoni, Balota, & Yap, 2008), or with neural network model of visual word recognition (e.g., Davis, 2001). Perhaps more interestingly, these effects can be predicted as well with hybrid approaches positing that word processing is influenced by the relative distribution of form and meaning in the lexicon (Marelli, Amenta, & Crepaldi, 2015).

In conclusion, using DSMs, we outlined the specific semantic processes that underlie both veridical and false memory formation. Our findings well complement previous theories accounting for performance in the DRM task and also provide insights on the cognitive operations that subserve memory distortions.

3.2

Hands on false memory: a mouse-tracking study on the DRM task

The present study is currently under review. To cite it:

Gatti, D., Marelli, M., Mazzoni, G. Vecchi, T., & Rinaldi, L. (under review). Hands on false memory: a mouse-tracking study on the DRM task.

3.2.1 Introduction

Human memory is not simply a precise tape recorder, but a system that encodes and ultimately represents knowledge through a process of active reconstruction rather than passive reproduction (Schacter, 2021; Vecchi & Gatti, 2020). This was first documented in Bartlett's pioneering studies on memory (1932) and supported by a large body of subsequent research, in which it was demonstrated that participants would forget the precise features of the stimuli they have memorized, in favor of an extraction of the gist of the information (for a review: Chang & Brainerd, 2021; Brainerd, Reyna & Ceci, 2008). That is, humans would use their semantic memory to encode, store and retrieve information, adapting new information to what has been previously memorized, with systematic errors that may occur during these phases (Brewer & Treyens, 1981; Sulin & Dooling, 1974).

The origins of inaccurate or false memories have therefore gradually become a matter of great scientific interest. Different experimental paradigms have been developed to account for false memories' formation, with the Deese–Roediger–McDermott (DRM; Deese, 1959;

Roediger & McDermott, 1995) task being one of the most widely used method in the verbal domain. In this task, participants are typically first presented once with several lists of words that have to be memorized (within each list, the words are related to a non-shown target word, named critical lure; e.g., word list: *door*, *glass*, *pan*, *shade*, *ledge*, etc. – critical lure: *window*) and then, after a brief distracting task, they are asked to perform a recognition task in which they have to indicate whether a given word was part of the memorized lists or not. Interestingly, during this latter phase, participants tend to erroneously report as "old" the critical lures, (i.e., they recognize them as if they were part of the memorized lists, although these words were never presented during the encoding phase; for a review, see: Gallo, 2010).

To explain participants' performance in the DRM task, two main theories have been proposed: activation-monitoring framework (Gallo & Roediger, 2002; Roediger, Watson, McDermott, & Gallo, 2001) and fuzzy-trace theory (Brainerd & Reyna, 2002; Reyna & Brainerd, 1995). According to activation-monitoring framework, the critical lure would be hyperactivated by the presentation of the studied words related to it, thus leading to high levels of false recognition (Roediger et al., 2001). Conversely, according to fuzzy-trace theory, while studying the words participants would encode a memory trace – called gist trace – linked to the semantic content of each list, which would be responsible for the production of the false recognitions (Brainerd & Reyna, 2002; Reyna & Brainerd, 1995).

Consistent with these perspectives, previous studies have successfully predicted false memory occurrence on associative (i.e., associative relationships reflect word use, such as "spider-web") and semantic (i.e., semantic relationships reflect overlap of conceptual features between words, such as "horse-pony") bases (Brainerd, Yang, Reyna, Howe, & Mills, 2008; Roediger et al., 2001). In particular, seminal studies have shown that the association strength between the words that compose each list and the critical lure (i.e., the backward associative strength, BAS) is a central factor in determining false memories (Roediger et al., 2001) and that multiple semantic sub-components underlie false memories (Cann, McRae, & Katz, 2011). In addition to this, recent evidence indicates that participants' memory performance follows a continuous semantic gradient, with higher false recognitions occurring for higher semantic similarity between the new words presented in the recognition phase and the words previously studied (Gatti, Rinaldi, Marelli, Mazzoni, & Vecchi, *in press*).

However, currently little is known about how the decision process unfolds when participants accept or reject words in the DRM task. That is, accuracy and reaction times – the classical dependent variables used in memory research – although informative of some explicit and implicit cognitive components, are associated with the final state of the decision process, and hence cannot provide a direct measure of how this process unfolds or cannot directly quantify potential conflicts in the response (Freeman, 2018; Stillman, Shen, & Ferguson, 2018). Alternative methods, such as drift diffusion models (e.g., Krajbich & Rangel, 2011; Ratcliff, 1978; for evidence on the DRM task see: Huff & Aschenbrenner, 2018), can be used to isolate certain decision components, but "the complexity of these approaches makes interpretation less straightforward" (Stillman et al., 2018, p. 537). These cognitive components are generally measured through other methodological approaches such as mouse-tracking, a paradigm which is particularly reliable in isolating the dynamics of response conflict and indecision, as well as the evolution of the choice (Freeman, 2018). Accordingly, it has also shown that mouse-tracking measures outperform reaction times in predicting participants performance in decisions involving risk (Stillman et al., 2020). In recent years, mouse-tracking has been indeed successfully used to investigate how participants' decision unfolds across several cognitive domains such as language (Lins, & Schöner, 2019; Spivey, Grosjean, & Knoblich, 2005), social cognition (Freeman, Dale, & Farmer, 2011; Freeman, Pauker, & Sanchez, 2016), recognition memory (Papesh, & Goldinger, 2012; Papesh, Hicks, & Guevara Pinto, 2019), semantic memory (Gatti, Marelli & Rinaldi, 2021), and also to detect faking-good behavior when responding to personality questionnaires (Mazza et al., 2020).

In mouse-tracking paradigms, participants are required to make decisions by moving their mouse from a starting position (typically placed in the middle-bottom part of the screen) to one of the two options presented (typically placed in the two upper corners of the screen). It is assumed that motor outputs (i.e., hand movements) are executed in parallel with the decision that participants are required to make (Freeman, & Ambady, 2010), thus allowing for the quantification of the conflict of the choice and its evolution, which cannot be directly assessed using only reaction times (Stillman et al., 2018). Through mouse-tracking packages (e.g., Kieslich, Henninger, Wulff, Haslbeck, & Schulte-Mecklenbeck, 2019) it is

indeed possible to extract several dependent variables that are informative about decision-making processes. Generally, decision conflict is quantified by computing the maximum deviation from the direct path (MD; i.e., the furthest point on the actual trajectory from the idealized straight trajectory between the starting point and the selected stimulus), while the decision evolution is quantified by computing the sample entropy, which measures the irregularity and unpredictability degree of the trajectory (for a complete discussion on other possible indexes see, Freeman & Ambady, 2010; Stillman et al., 2018). For both measures the higher the value, the higher the conflict and the level of indecision. Additionally, mouse-tracking paradigms allow for more refined decision time indexes, such as the computation of the time at which the trajectory reaches the maximum distance (the time at which the decision takes place).

Here, building upon this evidence, we take advantage of mouse-tracking paradigms to explore the decisional stages subserving recognition memory in the DRM task. Specifically, we applied an already established method to compute the semantic similarity between the new words and the studied words (for a complete discussion see: Gatti et al., *in press*), by employing indexes extracted from distributional semantic models (DSMs). DSMs induce words meanings from large databases of natural language data, representing them as high-dimensional numerical vectors: these models are indeed thought to well capture the structure of semantic memory (Günther, Rinaldi & Marelli, 2019; Jones, Willits, & Dennis, 2015). In particular, here we used word-embeddings that are based on a predictive component: these DSMs induce word vectors using a neural network architecture with one hidden layer, which is optimized to match a target word (Baroni, Dinu, & Kruszewski, 2014; Mikolov, Chen, et al., 2013; Mikolov, Sutskever, Chen, Corrado & Dean, 2013). Briefly, these models are trained on large collections of texts that document natural language use. Nodes in the input and output layers represent words, and a neural network learns to predict a target word on the basis of the lexical contexts in which it appears (i.e., the words it co-occurs with in the text), incrementally updating a set of weights by minimizing the difference between model predictions and observed data at each learning event (i.e., every occurrence of the target word). The estimated sets of weights will eventually capture word meanings. These distributed representations, or vectors, can be quantitatively compared by measuring their

distance in a multidimensional space, which in turn is thought to capture semantic similarity between words (Günther et al., 2019): similar words will occur in similar contexts, ending up being associated with vectors that are geometrically closer. Importantly, word-embeddings have been shown to be high-performing across a wide range of semantic tasks (for a review on the recent prediction-based class of models, see e.g., Baroni et al., 2014). Moreover, they are equivalent to psychologically grounded associative learning models (Günther et al., 2019; Mandera, Keuleers, & Brysbaert, 2017).

While previous studies predicted participants' performance in the DRM task mainly adopting human based measures (e.g., backward associative strength – BAS; Roediger et al., 2001), here we thus employed a measure not computed on human ratings, but rather automatically extracted from natural language. It should be noted that BAS and DSMs-based metrics in the DRM task have been shown to be correlated (i.e., $r = .50$, see: Gatti et al., *in press* for an in depth discussion regarding such relationship). However, the adoption of an independent-source measure such as data from DSMs may be preferable: that is, predicting human performance using data from association norms in a task that is necessarily tapping on the cognitive processes generating such norms (i.e., as most DRMs are explicitly constructed from free-association norms) may lead to explanatory circularity (Westbury, 2016). In line with this view, here we aimed to predict participants' behavior in the DRM task starting from independent models that replicate the structure of semantic memory by applying a psychologically-plausible learning model to environmental regularities (i.e., word co-occurrences) (Günther et al., 2019).

Participants were asked to study several lists of words from a classical DRM task and then, in the recognition phase, they were asked to indicate using their mouse if the words showed were "old" (i.e., presented in the encoding phase) or "new" (i.e., not previously presented). The spatial and temporal measures extracted from mouse movements were then predicted using a semantic index extracted from a DSM. This method allowed us to investigate whether the decision process differs depending on the position of the new (and studied) words in the semantic space (i.e., whether words in the recognition phase are more semantically similar or not to the studies words).

3.2.2 Methods

Participants

Seventy-one students participated in the study (28 males, *M* age = 24.4 years, *SD* = 3.63, age range = 19 – 35). All participants were native Italian speakers, had normal or corrected to normal vision and were naïve to the purpose of the study. Informed consent was obtained from all participants before the experiment. The protocol was approved by the psychological ethical committee of the University of Pavia and participants were treated in accordance with the Declaration of Helsinki.

Stimuli

We used the DRM task (Deese, 1959; Roediger & McDermott, 1995), a typical false memories paradigm. Participants were first instructed to memorize several lists of words and then to perform a recognition task. The words that composed each list were associatively related to a non-shown word (called critical lure).

For the encoding phase, we selected 12 lists of words out of 24 from the normative data for the Italian DRM test (Iacullo & Marucci, 2016). Each list was originally composed by 15 words: we selected the first 12 words (144 words in total), while 2 of the 3 remaining words were used as weakly related lures.

The recognition phase was composed of 96 words, 48 of which had been presented in the previous phase (i.e., studied words) and 48 of which had not been previously presented (i.e., new words). The 48 studied words presented in this experimental phase were those in serial positions 1, 4, 7 and 10 in the studied lists. Of the 48 new words, 12 were the critical lures from the studied lists (i.e., the non-shown words mostly associated with the words composing each list), 24 were weakly related lures and 12 were unrelated words. The weakly related lures were 2 of the 3 words of the studied lists that were not presented in the list, specifically those in position 13 and 14. The unrelated words were chosen randomly among

the words of the excluded lists; this criterion was established arbitrarily (for a similar method, see Gatti, Rinaldi, Marelli, Mazzoni & Vecchi, in press).

Procedure

Participants were tested using Psychopy (Pierce, 2007, 2009; Pierce & MacAskill, 2018; Pierce et al., 2019) through the online platform Pavlovia (https://pavlovia.org/) and answered using an external mouse.

During the first part of the task, participants had to memorize a series of words (i.e., they were required to study 12 lists of words without interruptions). Participants were shown the 12 words that composed each of the 12 lists in descending forward associative strength (FAS; i.e., the association strength from the critical lure to the word that compose the list). The order by which the lists were presented was random, while the order of the words within each list was fixed according to the FAS (see Iacullo & Marucci, 2016). Each trial started with a central fixation cross (presented for 500 ms) followed by a word (presented for 1500 ms) and a blank screen (presented for 300 ms), then the script moved automatically to the next fixation cross.

At the end of the encoding phase, participants were required perform an attention task (i.e., a modified version of the go-no go) as a distracting task for 2 minutes.

Then participants were asked to perform the recognition phase. Participants were instructed to make old/new judgments using their mouse: participants had to first click on a square presented below the "START" label by pressing the mouse left button, and next to move to the selected option (old or new), again pressing the left button to make their decision. Each trial began with a fixation cross presented at the center of the screen (500 ms), then a red "START" button appeared at the bottom-center of the screen (Arial font, button located at: x = 0, y = -.35); after the participants clicked on it the word to be recognized appeared (showed for 1500 ms; Arial font, x = 0, y = 0), while "old" and "new" buttons appeared in the upper right and upper left of the screen (locations counterbalanced across participants, Arial font for the old/new; buttons located at: x = ± .50, y = .25; see

Figure 6). To ensure valid trajectories, participants were asked to initiate their physical movement as fast as possible.

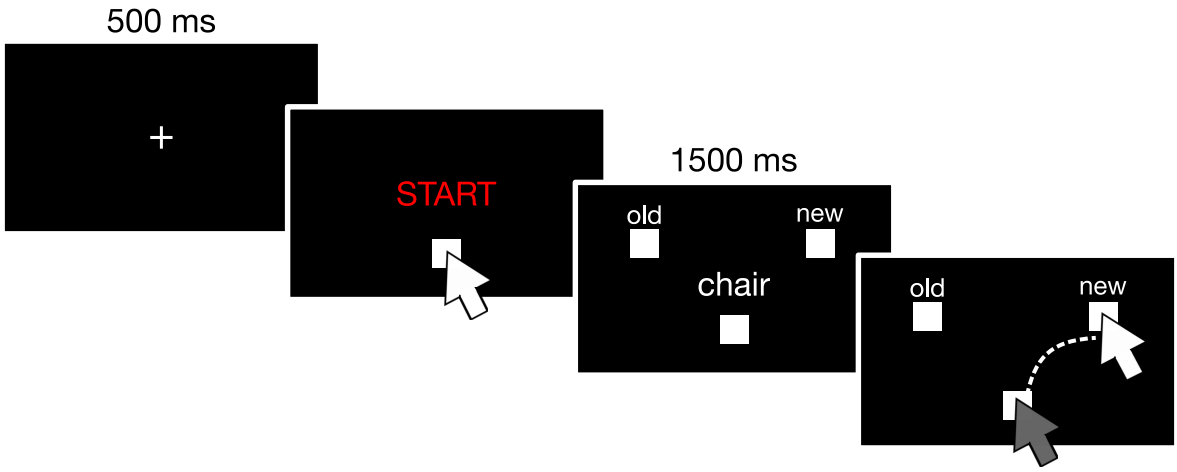


Figure 6. Recognition task. Participants were asked to make old/new judgments using their mouse. After clicking on the START square, they were shown a word and were required to make the memory judgement.

Word-embeddings

Vector representations for the words used in this study were extracted from a semantic space obtained by inducing word embeddings using the Continuous Bag of Words (CBOW) method, an approach originally proposed by Mikolov and colleagues (Mikolov, Chen et al., 2013). The model, released by Marelli (2017), was trained on itWaC, a free Italian text corpus based on web-collected data and consisting of about 1.9 billion tokens. The model used is set on the following parameters: *9-word co-occurrence window*, *400-dimension vectors*, negative sampling with $k = 10$, subsampling with $t = 1e^{-5}$. This set of parameters defines the learning procedure used to induce word vectors (Mikolov, Chen et al., 2013). CBOW indicates the applied learning procedure: when using CBOW, the obtained vector dimensions capture the extent to which a target word is reliably predicted by the contexts in which it appears. Co-occurrence window size indicates how large the considered lexical contexts are; in our case, a *9-word window* indicates that we estimated predictions concerning 4 words on the left and 4 words on the right of the target word. The number of

vector dimensions indicates how many nodes are included in the hidden layer, representing the result of the dimensionality reduction process implicitly applied by the network. Negative sampling estimates the probability of a target word by learning to distinguish it from draws from a noise distribution; the parameter k specifies the amount of these draws. The subsampling parameter t specifies a threshold-based procedure that limits the impact of very frequent, uninformative words.

From this semantic space, we extracted vector representations for the words used in this study[5]. Specifically, for each word pair it is possible to obtain a semantic-similarity index (hence SSim) based on the cosine of the angle formed by vectors representing the meanings of these words. In particular, the higher the cosine the more semantically similar the words should be. A heatmap matrix of the semantic similarity structure among the words composing a DRM list is represented in Figure 7.

Computation of semantic similarity values

For each new word (12 critical lures, 24 weakly related lures and 12 unrelated words) we computed a semantic similarity index (SSim). That is, SSim was computed as the frequency-weighted average SSim (for a similar approach see: Gatti, Rinaldi, Marelli, Mazzoni & Vecchi, *in press*; Marelli & Amenta, 2018) between each word in the recognition phase and each of the 12 words that composed its relative list. For unrelated words, we computed the index randomly matching each word with a list. The formula used was:

$$SSim = \frac{\sum_{i=1}^{k} SSim_i \times F_i}{\sum_{i=1}^{k} F_i}$$

where $SSim_i$ refers to the semantic similarity between a new word and each of the i studied word composing its list, while F_i is the frequency of each studied word as extracted from the

[5] The relative semantic similarity indexes used in this Experiment can be reproduced through the freely available user-friendly interface at: http://meshugga.ugent.be/snaut-italian/ .
Note that these indexes were subtracted from 1 to transform the values on a proximity scale.

Italian SUBTLEX (http://crr.ugent.be/subtlex-it/). Following the same rationale, we computed a SSim index for studied words.

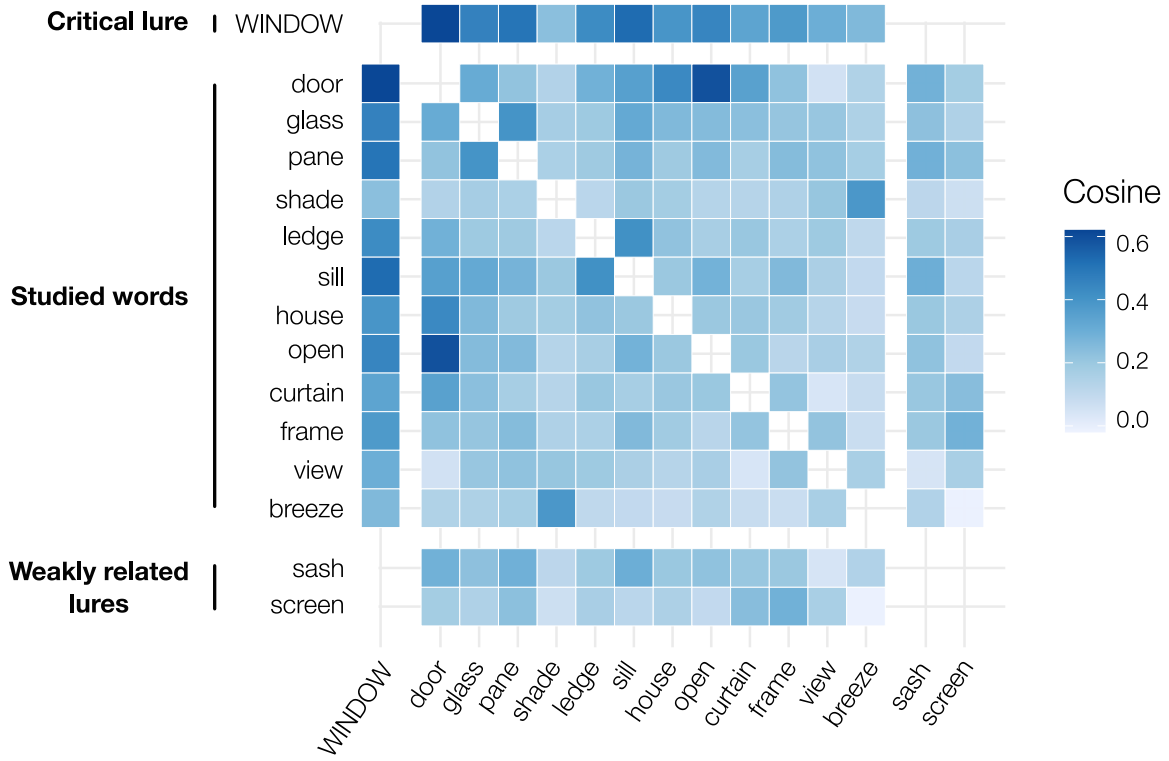


Figure 7. A heatmap matrix of the cosine values among the words composing the list *window* (i.e., critical lure; words list taken from Roediger & McDermott, 1995). Darker blue colors represent higher cosine values (and, hence, account for more semantically related words as predicted by DSMs). Note that in this case we plotted all the words in the list (i.e., including the possible weakly related lures) and that we computed the cosine values through an English DSM, available here: http://meshugga.ugent.be/snaut-english/ (see also: Mandera, Keuleers, & Brysbaert, 2017).

Data analysis

All the analyses were performed using R-Studio (RStudio Team, 2015). Data were analyzed through a mixed-effects approach, which incorporates both fixed-effects and random-effects (i.e., associated to statistical units as participants and task stimuli) and allow for the specification of predictors at both participants and/or item level. All the generalized linear mixed models (GLMMs) and linear mixed models (LMMs) were run using the *lme4*

package (Bates, Maechler, Bolker, & Walker, 2015). The graphs reported were obtained using the *effects* package (Fox, 2003; Fox & Weisberg, 2019).

Besides proportion of "old" responses, our main dependent variables were: maximum deviation from direct path (MD), sample entropy, initiation reaction times (RTs) and maximum deviation from direct path time (MDtime).

MD is defined as the furthest point on the actual trajectory from the idealized straight trajectory and is thought to quantify the conflict in the choice. Sample entropy is defined as the degree of irregularity and unpredictability in movement across the x-axis and is thought to measure the evolution of the choice (i.e., with higher values indicating higher irregularity). Initiation RT is computed as the time elapsed between the click on the START button and the first hand-movement, while MDtime indexes the time at which the trajectory reaches the MD. These two dependent variables are considered to measure two different time windows of decision-making, with the former (initiation RTs) indicating the time at which the decisional process starts and the latter (MDtime) indicating the time at which a decision is finally achieved (see for a complete discussion regarding mouse-tracking variables: Stillman et al., 2018).

All the mouse-tracking related dependent variables were computed using the *mousetrap* R package (Kieslich, et al., 2019). All trajectories were normalized into 101 time-steps and remapped symmetrically in order to allow for direct comparison of trajectories which differed in duration and number of data points. Initiation RTs, MDtime and sample entropy were all log-transformed.

First, we aimed to replicate the significant interaction SSim by type of stimuli (new vs. old) on participants' proportion of "old" responses as previously reported (for a complete discussion of this analysis see: Gatti et al., in press). Proportion of "old" responses were thus analyzed estimating a GLMM with SSim, type of stimuli (new vs. old) and their interaction as predictors; subjects and items were included as random intercept.

Then, following a similar approach, we included the SSim predictor in the analyses on mouse-tracking dependent variables estimating four LMMs in which we included SSim, type of stimuli (new vs. old), participant's response (new vs. old; i.e., whether the participant

judged each word as showed or not) and their interactions as predictors; subjects and items were included as random intercept. In case of singularity issues the random model was simplified removing the intercept of the item and then refitted. In particular, in *lme4* syntax, the models tested were:

$$DV \sim SSim * Type * Response + (1|Participant) + (1|Item)$$

Then, to exclude the impact of overly influential outliers, after having fitted the model, data points were removed on the basis of a threshold of 2.5 *SD* standardized residual errors (model criticism; Baayen, 2008). Results based on the refitted models are reported.

3.2.3 Results

Trials in which overall reaction times were faster than 300 ms or slower than 5000 ms were excluded from the analysis (.7% of the trials excluded). Aberrant movements were detected in 6% of additional trials and were discarded. Trials in which MD and sample entropy were ± 3SD from the mean of the participants were removed from the analysis (2% of additional trials excluded).

1. The differential effect of SSim on false and veridical recognitions

In the GLMM including recognition responses (i.e., "new" responses were scored as 0, "old" as 1) as dependent variable, SSim and type of stimuli (studied vs. new words) as predictors, we found significant effects of both variables on recognition (SSim: $z = 9.84$, $p < .001$; type of stimuli: $z = 3.36$, $p < .001$; *Pseudo-*R^2 (total) = .47). Critically, the interaction SSim by type of stimuli was significant, $z = -3.24$, $p = .001$.

The significant interaction indicates that the effect of SSim was higher for false recognitions, $z = 6.43$, $p < .001$, compared with veridical recognitions, $z = .90$, $p = .37$.

2. Analyses on mouse-tracking dependent variables

2.1 Descriptive statistics

Descriptive statistics for the four variables are reported in Table 2. The correlation matrix between the four dependent variables is reported in Table 3. Overall, the correlations ranged from very low, as the one between MD and MDtime (r = -.06), to high, as the one between MD and sample entropy (r = .59). Note that the correlations were computed on raw data, hence on a total of 6203 trials.

Table 2. Descriptive statistics for the four dependent variables analyzed. The descriptive statistics of the two variables expressing temporal processes (initation RTs and MDtime) are reported in milliseconds.

	Initiation RTs	MD	Sample entropy	MDtime
Mean	330	.22	.10	917
SD	197	.12	.03	204
Min – Max	51 – 907	-.01 – .48	.05 – .90	554 – 1402

Table 3. Correlation matrix between the four dependent variables included in the current study. The correlation values ranged from very low to moderate (6201 degrees of freedom, all ps < .01).

	Initiation RTs	MD	Sample entropy	MDtime
Initiation RTs	1			
MD	-.20	1		
Sample entropy	-.24	.59	1	
MDtime	.42	-.06	.08	1

2.2 Initiation reaction times

In the LMM including Initiation RTs as dependent variable, SSim, type of stimuli and participant's response as predictors, we found no significant effects, all ps > .20.

2.3 Maximum deviation from direct path

In the LMM including MD as dependent variable, SSim, type of stimuli and participant's response as predictors, we found significant effects of SSim, $F(1,6040) = 4.91$, $p = .02$, participant's response, $F(1,6047) = 10.11$, $p = .001$, the interaction SSim by participant's response, $F(1,6044) = 13.97$, $p < .001$, the interaction type of stimuli by participant's response, $F(1,6045) = 7.22$, $p = .007$, and, critically, of the interaction SSim by type of stimuli by participant's response, $F(1,6043) = 4.71$, $p = .02$, *Pseudo-*R^2 (total) = .27 (Figure 8).

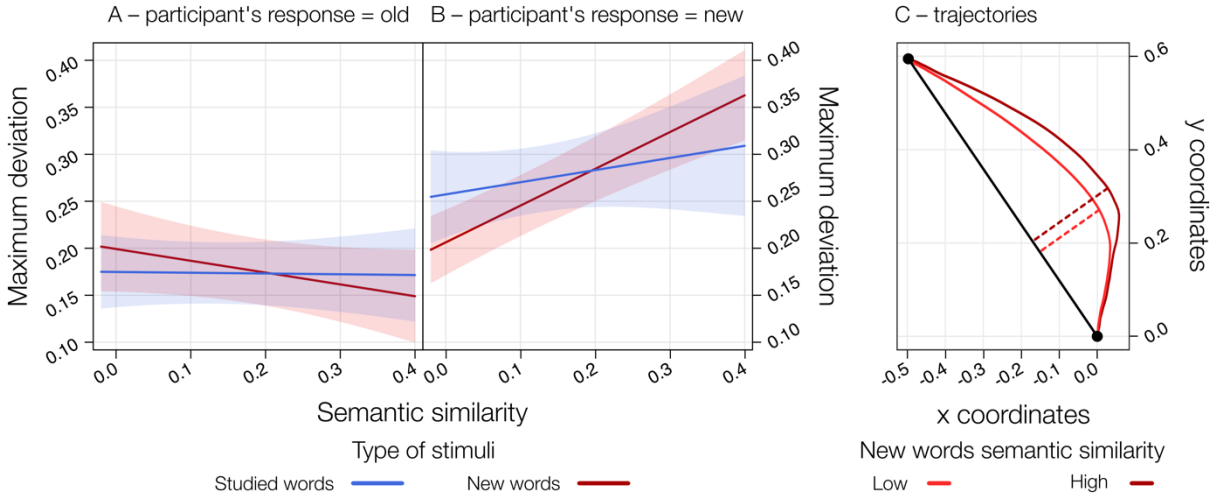


Figure 8. Results from the LMM on MD including the interaction between SSim, type of stimuli divided across participant's response = old (A) and participant's response = new (B), showing the positive relationship between SSim and MD when participants correctly rejected new words. A pictorial representation of how the maximum deviation (in dashed lines) from the direct path (black straight line) varied across two hypothetical levels of SSim (i.e., low and high SSim; the two categories were created by selecting the half of words with lower SSim and the other half with the higher SSim) of new words when participants correctly rejected them (C).

The significant triple interaction indicates that, when participants rejected new words, SSim significantly predicted MD, $t(6041) = 6.36$, $p < .001$, $b = .39$. In particular, the higher the SSim between the new word and those studied, the higher participants' degree of indecision when correctly rejecting it (Figure Y). No effects of SSim were found when participants reported as old both new, $t(6041) = -1.54$, $p = .12$, $b = -.12$, and studied words,

$t(6040)$ = -.11, p = .90, b = -.008, or when rejecting studied words, $t(6043)$ = 1.05, p = .29, b = .12.

2.4 Sample entropy

In the LMM including sample entropy as dependent variable, SSim, type of stimuli and participant's response as predictors, we found significant effects of participant's response, $F(1,6083)$ = 4.98, p = .02, of the interaction type of stimuli by participant's response, $F(1,6082)$ = 6.80, p = .009 and, critically, of the interaction SSim by type of stimuli by participant's response, $F(1,6080)$ = 4.68, p = .03, *Pseudo-*R^2 (total) = .27 (Figure 9).

The significant triple interaction indicates that, when participants rejected new words, SSim significantly predicted sample entropy, $t(6079)$ = 6.79, p < .001, b = .50. In particular, the higher the SSim between the new word and those studied, the higher participants' degree of indecision when correctly rejecting it. No effects of SSim were found when participants reported as old both new, $t(6079)$ = -1.26, p = .21, b = -.22, and studied words, $t(6078)$ = .91, p = .36, b = .14, or when rejecting studied words, $t(6080)$ = .18, p = .86, b = .05.

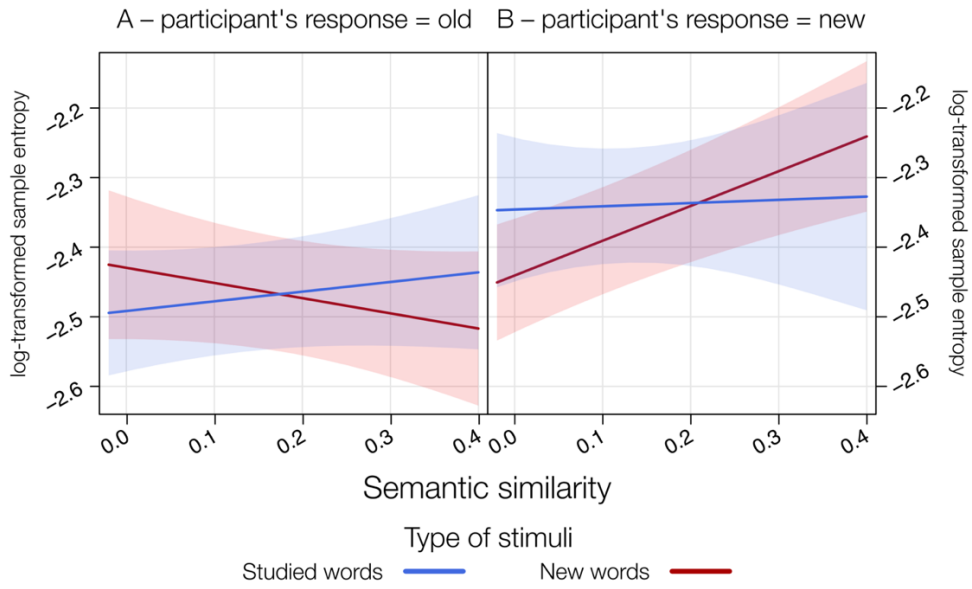


Figure 9. Results from the LMM on log-transformed entropy including the interaction between SSim, type of stimuli divided across participant's response = old (A) and participant's response = new (B), showing the positive relationship between SSim and log-transformed entropy when participants correctly rejected new words.

2.5 MDtime

In the LMM including MDtime as dependent variable, SSim, type of stimuli and participant's response as predictors, we found significant effects of participant's response, $F(1,4783) = 6.28$, $p = .01$, of the interaction type of stimuli by participant's response, $F(1,4783) = 8.05$, $p = .004$ and of the interaction SSim by participant's response, $F(1,4899) = 10.94$, $p < .001$, *Pseudo-*R^2 (total) = .39 (Figure 10). On the contrary, the interaction SSim by type of stimuli by participant's response was not significant, $F(1,4912) = 1.29$, $p = .25$.

The significant interaction SSim by participant's response indicates that, regardless of the type of stimuli, the higher the SSim, the faster participants were in deciding that a word was old, $t(280) = -2.10$, $p = .04$, $b = -.26$; conversely, the higher the SSim, the slower participants were in rejecting a word (i.e., identifying it as new), $t(153) = 2.22$, $p = .03$, $b = .22$.

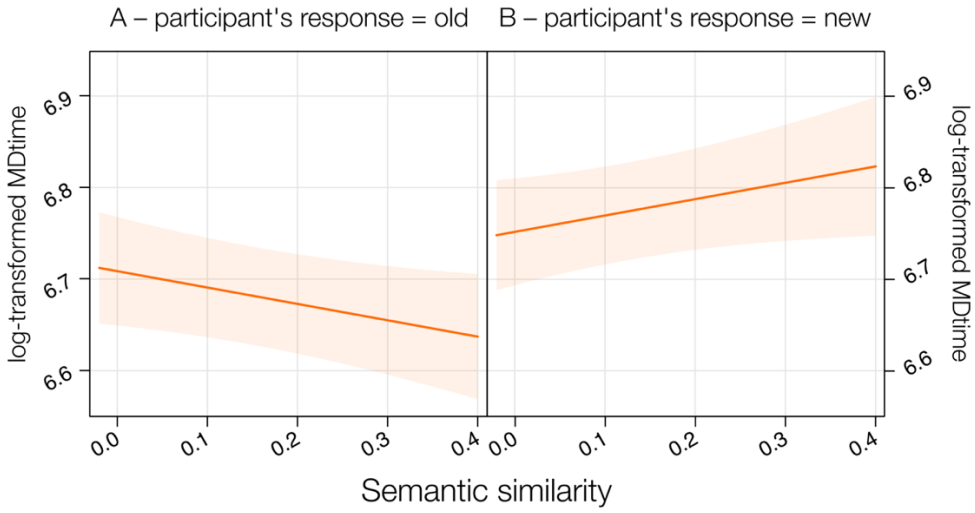


Figure 10. Results from the LMM on log-transformed MDtime including the interaction between SSim, type of stimuli divided across participant's response = old (A) and participant's response = new (B), showing the negative relationship between SSim and MDtime when participants recognized as old both types of stimuli (left) and the positive relationship between SSim and MDtime when participants rejected both types of stimuli (right).

3.2.4 Discussion

In the present study, we explored the decisional stages subserving recognition memory in the DRM task taking advantage of mouse-tracking and of distributional semantic models. Participants were asked to memorize several lists of words in a classical DRM task and then to recognize them among new words using their mouse. The decision-making processes were indexed through different variables computed from mouse trajectories and predicted through an item-level semantic metrics extracted from distributional semantic models (for a complete discussion see: Gatti et al., *in press*). Overall, our findings indicate that semantic memory can affect human behavior at the level of motor control, testifying its pervasive influence on online decision-making. Specifically, our findings indicate that mouse trajectories are affected by the semantic similarity between each word in the recognition phase and the previously studied words. That is, the higher the semantic similarity, the higher the conflict driving the choice and the irregularity in the trajectory (respectively measured with the maximum deviation from direct path and with sample entropy) when correctly rejecting new words. Conversely, on the temporal evolution of the decision, our findings indicate that semantic similarity predicts complex temporal measures indexing the online decision processes subserving task performance. More specifically, we found that regardless of the type of stimuli (old or new), when responding that a word was "old", the higher the semantic similarity, the earlier the stage at which the decision was achieved; on the contrary, when rejecting a word (i.e., when responding that a word was "new"), the higher the semantic similarity, the later the stage at which the decision was achieved.

These findings well complement the key assumptions of the two main theories accounting for false memory in the DRM task, namely activation-monitoring framework and fuzzy-trace theory. Indeed, both theories trace back the origin of false recognitions to associative/semantic mechanisms, with adequate episodic and source memory processes that would counter them and enhance the occurrence of veridical recognition (Brainerd, & Reyna, 2002; Gallo & Roediger, 2002; Reyna & Brainerd, 1995; Roediger et al., 2001; and for individual differences evidence see: Gatti, Rinaldi, Mazzoni, & Vecchi, 2021). In interpreting our findings in light of these theories, we first note that the dependent variable used to measure the conflict in the decision (maximum deviation from direct path) quantifies it as a

measure of the attractiveness of the unselected option (Freeman, & Ambady, 2010). Thus, the higher the attractiveness of the unselected option, the higher the maximum deviation value, because mouse trajectories would be associated with a larger curvature. Accordingly, our results show that even when participants correctly rejected new words (i.e., by selecting the "new" button), the "old" button exerted high attractivity. Notably, the level of attractiveness varied as a function of the semantic similarity, with greater level of conflict for more semantically similar new words. Hence, while previous studies have shown that semantic memory is involved in the production of false memories (Gatti et al., *in press*; see also: Montefinese, Zannino, & Ambrosini, 2015), here we demonstrate that in the DRM task semantic processes participate also when correctly rejecting new words (i.e., the false memory items). Such an interpretation holds as well for the degree of irregularity and unpredictability of mouse movements, as the evolution of the choice was similarly affected by semantic similarity. That is, when correctly rejecting new words, movement irregularity was higher for more semantically similar new words.

The increased conflict in the rejection of new words as a function of their semantic similarity can be interpreted in terms of the alleged conflict emerging when participants are requested to judge if a new word was actually studied or not. In particular, activation-monitoring framework assumes that the critical lure is associatively hyperactivated in the encoding phase and then, in the recognition phase, such hyperactivation would be responsible for the false recognition (Gallo & Roediger, 2002; Roediger et al., 2001). On the other hand, fuzzy-trace theory assumes that while studying the word lists the participants would encode two memory traces: a semantic one, linked to the semantic content of each list; and an episodic one, linked to the contextual and perceptual features. In the recognition of a new word, these two traces would therefore counter each other (i.e., the semantic trace would increase the likelihood of a false recognition, while the episodic trace would operate in the opposite direction), while for studied words they would both boost veridical recognition (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). The enhanced conflict and uncertainty in participants' rejection of new words with increasing semantic similarity observed here documents for the first time the online decision processes underpinning false memories, as maintained by both activation-monitoring framework and fuzzy-trace theory. That is, for

the new words in which the episodic trace is lacking, the conflict and the uncertainty would increase with increased semantic similarity of the new word.

Previous studies investigating associative and semantic involvement in the DRM task have shown that backward associative strength is a major predictor of false memories (Roediger et al., 2001; but see also: Brainerd, Chang, & Bialer, 2020), that multiple semantic sub-components underlie false memories (Cann et al., 2011) and that memory performance follows a continuous semantic gradient (Gatti et al., *in press*). This last finding was replicated in this study, by observing that for new words, the higher the semantic similarity value, the higher the occurrence of false memories. Critically, here we extended this evidence by showing that the semantic similarity between the words presented in the recognition phase and those previously studied affects not only the explicit memory judgements (i.e., "yes" and "no" recognition responses), but also more implicit measures extracted from participants' motor outputs. These findings support previous evidence, in that they suggest that, while memorizing the words, participants would implicitly and automatically activate the semantic trace for each list and would then use this trace during the recognition task when judging new words (Gatti et al., *in press*). This effect has been explained arguing that, since during the encoding task participants were shown the lists of words without any clues that clustered the words, the sequential presentation of each word within each list would have incrementally activated a meaning cluster composed by the same words in semantic memory (Gatti et al., *in press*). Here, extending upon these findings, we further show that semantic similarity between studied and new words affects decision-making in memory retrieval: the more the cluster of words composing each list is semantically close to the new word, the higher the participants' conflict and uncertainty when correctly rejecting new words. This indicates, therefore, that the structure of semantic memory and the activation of specific clusters of words affect memory retrieval, with the degree of overlap between the vectors representing new and studied words accounting for the level of conflict in the decision-making process.

These findings can be framed also within the debate regarding the nature of false memory in the DRM task (for recent evidence see: Brainerd et al., 2020), that is whether false memory can be considered as the result of *associative* (i.e., hence reflecting word use,

such as "spider-web") or *semantic* (i.e., hence reflecting word meaning, such as "horse-pony") processing (which are independent, Ferrand & New, 2003; Hutchinson, 2003). DSMs-based metrics have been indeed successful in predicting both associative and semantic priming effects (Günther, Dudschig, & Kaup, 2016; Jones, Kintsch & Mewhort, 2006). Such evidence, indicating that the same process (i.e., predicting a target word from the linguistic context in which it typically appears) can explain both associative and semantic processing, suggests that these two processes are likely interdependent in natural settings and may converge into (partially) overlapping structural representations of human memory. That is, the dissociation between associative and semantic processes would be possible on an experimental level by forcing participants to rely on a specific component given certain task demands (e.g., defining a concept or guessing at associates necessarily forces participants to rely on different components of human memory; for a discussion see, Maki & Buchanan, 2008); yet, in natural contexts it is almost impossible to isolate such components and thus to find words that are purely semantically or associatively related (Jones et al., 2006).

For the main timing measure (i.e., the time at which the mouse trajectory finally deviates), we found that the semantic similarity affected participants' performance in both veridical and false memories. In particular, the higher the semantic similarity, the faster participants were in deciding that a word was old, and the opposite for "new" judgments. Hence, the time at which the decision occurred was influenced in opposite directions by semantic similarity, indicating that this variable overall impacted on the temporal dynamics subserving task performance. This dissociation may suggest that semantic memory involvement in the DRM task could affect differently temporal and spatial measures of decision making, thus dissociating the time needed to decide if a word was old or new from deeper decision-making components, such as the conflict and indecision underlying memory judgements. As maintained by activation-monitoring framework and fuzzy-trace theory, different cognitive processes (i.e., associative/semantic and episodic) come simultaneously into play in the DRM task, generating in turn different outcomes (Brainerd, & Reyna, 2002; Gallo & Roediger, 2002; Reyna & Brainerd, 1995; Roediger et al., 2001). Specifically, the conflict observable in the spatial measures that can be traced back to the interplay between associative/semantic and episodic traces was not observed in the main timing measure,

suggesting that two main decision components (i.e., spatial and temporal) are active in parallel during memory retrieval in the DRM task. This dissociation can be explained through dynamical decision-making frameworks arguing that several explicit and implicit processes simultaneously compete when making a decision (Freeman & Ambady, 2011; Melnikoff & Bargh, 2018).

Our findings have relevant implications from both theoretical and methodological points of view. On a theoretical level, our results clarify semantic memory involvement in a complex memory task such as the DRM when participants' performance is measured through fine-grained hand movements. In particular, here we provide evidence for a possible differential involvement of semantic memory across time, conflict and uncertainty of participants' decisions. Additionally, by successfully predicting mouse-tracking measures using a semantic predictor extracted from a distributional semantic model, we provide further support to the idea that these models are extremely efficient in capturing the structure of human semantic memory (Günther et al., 2019). Indeed, while previous studies have predicted participants' performance using distributional semantic models across a wide range of semantic tasks, such as multiple-choice tests (Bullinaria & Levy, 2012), word categorization (Baroni & Lenci, 2010), word relatedness ratings (Bruni, Tran, & Baroni, 2014), word naming and lexical decision (Marelli & Amenta, 2018), semantic priming (Günther et al., 2016), recognition memory (Gatti et al., *in press*; Gatti, Vecchi, & Mazzoni, 2021), as well as using mouse-tracking (Gatti et al., 2021), our study is the first to report its effect also on mouse-tracking variables in a complex memory task such as the DRM. Furthermore, on a methodological level, we show that by pairing distributional semantic models with mouse-tracking it is possible to investigate deep decision-making features of human behavior, thus opening new avenues for probing the detailed processes subserving human memory.

In conclusion, using distributional semantic models combined with mouse-tracking, we document the decision-making semantic processes underpinning false memories. Our findings are consistent with previous theories on participants' behavior in the DRM task and provide novel insights on the impact of semantic memory on different decision-making components.

3.3

Semantic and episodic processes differently predict false memories

A preprint of the present study is available on *PsyArXiv*. To cite it:

Gatti, D., Rinaldi, L., Mazzoni, G., & Vecchi, T. (2021). Semantic and episodic processes differently predict false memories in the DRM task.

3.3.1 Introduction

Human memory cannot be simply conceived as a recorder that passively encodes, stores and faithfully retrieves information, but rather as a system that actively reconstructs information from memory traces (Schacter, 2021; Vecchi & Gatti, 2020). Indeed, beyond accurate remembering, it has been shown that humans use their semantic memory to encode, store and remember information adapting it to their own previous knowledge, favoring consequently a gist extraction: in such a process, memory distortions and false memories may thus occur (Bartlett, 1932; Sulin & Dooling, 1974).

One of the most widely used task to create false memories is the Deese–Roediger–McDermott task (DRM; Deese, 1959; Roediger & McDermott, 1995). During the DRM task, participants are first asked to encode several words divided in lists (i.e., within each list, the words are semantically/associatively related to a non-shown target word, named critical lure; e.g., *table*, *sit*, *legs*, *seat*, *couch*, *desk*, etc. – critical lure: *chair*) and then, after a brief distracting task, participants are asked to perform a recognition task (i.e., they have to

indicate whether a given word was part of the memorized lists or not). Typically, during the recognition task, participants report as "old" (i.e., as previously memorized) a fairly large number of critical lures, although these words have not been presented before (for a review see, Gallo 2010).

The two main theories proposed to explain participants' performance in the DRM task have traced back false memories occurrence to semantic processing. In particular, according to the activation-monitoring framework (AMF – Gallo & Roediger, 2002; Roediger, Watson, McDermott, & Gallo, 2001), the critical lure would be associatively hyperactivated by the presentation of the studied words (i.e., spreading activation), leading to high levels of false recognitions. Alternatively, according to the fuzzy-trace theory (FTT – Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995), participants would encode two different memory traces: a trace linked to episodic and perceptive features of the studied items, called verbatim trace, and a trace linked to the semantic content of each list, called gist trace, which would be responsible for the production of the false memories. Besides predicting a semantic basis for false memory production, both theories also predict that adequate source or episodic memory processes can reduce the occurrence false recognitions. That is, according to the AMF, if participants can successfully distinguish between words presented and words hyperactivated but not presented, the production of false memories would decrease (Gallo, 2010). Alternatively, according to the FTT, the verbatim trace would be involved in the correct rejection of non-presented words, since no episodic memory trace is available for those stimuli (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995).

In line with these theoretical perspectives, it has been shown that false memory occurrence can be predicted on a semantic and associative basis (Brainerd, Yang, Reyna, Howe, & Mills, 2008; Roediger et al., 2001). In particular, the backward associative strength (BAS, i.e., the association strength from the words that compose each list to the critical lure) is thought to be the central factor in determining false recall and false recognition (Roediger et al., 2001). Furthermore, recent studies also showed that the semantic component that underlies false memories can be decomposed into various sub-components (Cann, McRae, & Katz, 2011) and that higher false recognition rates occur for new words with higher semantic similarity with the ones studied (Gatti, Rinaldi, Marelli, Mazzoni, &

Vecchi, *in press*). At the same time, it has been shown that inducing a higher monitoring process warning participant about the false memory effect in the DRM (Gallo, Roberts, & Seamon, 1997) or instructing them to focus on the distinctive aspects of each presented word (Westerberg & Marsolek, 2006) would reduce false recognitions, thus pointing to the possible opponent role of episodic memory processes in countering memory distortions. Yet, the experimental manipulations introduced by prior research do not fully clarify the extent to which semantic and episodic memory processes differently contribute to false memories.

Previous studies have been indeed able to predict participants' memory performance in the DRM task (e.g., Roediger et al., 2001), but this evidence is mainly limited to semantic processing and based only on manipulations at the item level (i.e., the type of stimuli used). Here, to directly probe the role of semantic and episodic processes in memory distortions, we adopted an individual differences approach. An advantage of an individual differences framework is that it allows to conceive individuals along continuous dimensions reflecting their semantic and episodic memory abilities, and to relate such natural variabilities with false memories production. Notably, such an approach has proven to be successful in demonstrating that that higher working memory abilities (Holden, Conway, & Goodwin, 2020; Unsworth & Brewer, 2010; Watson, Bunting, Poole, & Conway, 2005) or high memory self-efficacy (Iacullo, Marucci, & Mazzoni, 2016) are associated with a lower occurrence of false memories. Similarly, age differences have been linked to false memories, with false memories increasing in older individuals (e.g., Balota et al., 1999; Norman & Schacter, 1997; Tun, Wingfield, Rosen, & Blanchard, 1998; Watson, Balota, & Sergent-Marshall, 2001; but cfr. also: Pansuwan et al., 2020). Moreover, other studies investigating the link between creativity abilities such as convergent and divergent thinking have shown that the former, but not the latter, is associated with increased false memories (Dewhurst, Thorley, Hammond, & Ormerod, 2011). Finally, individual differences in need for cognition, an index that characterizes individuals' preferences for engaging in effortful information processing (Cacioppo & Petty, 1982), can predict memory performance in the DRM task: in particular, participants with a high need for cognition, thus more likely to engage in effortful information processing, typically show a greater occurrence of false recalls (Leding, 2011) and false recognitions (Graham, 2007).

Here, following the main theoretical accounts linking false memories with semantic and episodic processes, we aim to systematically dissociate the possible contribution of each memory system to memory distortions. We thus expect participants' false recognitions of the critical lures to be positively associated with their semantic abilities (i.e., higher number of false memories for individuals with better semantic memory abilities) and negatively related with their episodic abilities (i.e., lower number of false memories for individuals with better episodic memory abilities). In addition to this, since the FTT posits that the verbatim trace is linked to both episodic and perceptive features of the studied items (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995), we also investigated participants' source memory abilities. Participants were asked to perform a semantic task and an episodic-source memory task, from which we extracted the relative memory scores. We then predicted false and veridical memories occurrence in the DRM task using the scores extracted.

3.3.2 Methods

Participants

Fifty Italian students (41 females, *M* age = 23.10 years, *SD* = 3.56) participated in the study. All participants were native Italian speakers, had normal or corrected to normal vision and were naïve to the purpose of the study. Informed consent was obtained from all participants before the experiment. The protocol was approved by the psychological ethical committee of the University of Pavia and participants were treated in accordance with the Declaration of Helsinki.

Stimuli and procedure

Participants were tested online using Psychopy (Pierce, 2007, 2009; Pierce & MacAskill, 2018; Pierce et al., 2019) through the online platform Pavlovia (https://pavlovia.org/). Participants performed the three tasks in three different days (i.e., the three sessions had to be completed within 1 week) at the same time of the day. The order of the tasks was counterbalanced across participants.

Semantic memory task

The task used was an associative/semantic priming task, in which participants were shown two stimuli (i.e., a prime stimulus and a target one, sequentially presented on the screen) and then were asked to judge if the second one (i.e., the target) was a word or a pseudoword. Generally, participants reaction times tend to be predicted by the degree of semantic relationship between prime and target words, with faster reaction times occurring for more related pairs. This speeding up is thought to reflect the enhanced involvement of semantic processing (Hutchison et al., 2013).

Primes were 120 words selected from the Italian database provided by Montefinese and colleagues (Montefinese, Ambrosini, Fairfield, & Mammarella, 2013). For each prime, a target word was chosen using the distributional semantic model SNAUT (Mandera, Keuleers, & Brysbaert, 2017; Marelli, 2017; http://meshugga.ugent.be/snaut-italian/), by selecting among the most semantically similar words and avoiding repetitions of the same target word across the task (for a review on distributional semantic models, see: Günther, Rinaldi, & Marelli, 2019). Distributional semantic models have been shown to be high-performing across a wide range of semantic tasks (e.g., Baroni et al., 2014), and they are equivalent to psychologically grounded associative learning models (Günther et al., 2019; Mandera et al., 2017). Critically, the semantic similarity index extracted from these databases is thought to involve both associative and semantic processes (for a similar approach with the DRM task, see Gatti et al., *in press*). Then, the 120 word-pairs obtained were divided into two sets of 60 word-pairs (i.e., related and unrelated). In the unrelated set, primes and targets were pseudo-randomly mixed in order to remove the semantic link between the words (i.e., characterizing in turn the related set). From the same Italian database 120 additional words were extracted and transformed into pseudowords (i.e., reversing two letters: *paper – pepar*). Pseudowords were readable, but meaningless.

Each prime appeared twice, one time followed by a word, and another time by a pseudoword. Related and unrelated pairs were comparable in terms of length and logarithm

of the frequency of the first and second word, as well as in total length of the two words paired together and their paired logarithm of the frequency (all $ps > .44$, all BFs $< .24$).

Participants were shown two letter strings stimuli presented sequentially one after the other, and were required to judge if the second stimulus was a word or not. Participants were instructed to silently read the first letter string and to respond only to the second one as fast and as accurately as possible by pressing the left/right key (A and L) using the left and right index fingers, respectively; the response keys were counterbalanced across participants. The trials were shown in random order.

Each trial started with a central fixation cross (presented for 500 ms) and was followed by a first word (presented for 200 ms) and then by a second word (presented for 500 ms). Participants' response ended the trial and moved to the fixation cross of the next trial. See Figure 11 for a schematic representation of the semantic memory task.

Episodic-source memory task

We used an episodic-source memory task similar to the one recently employed by Chen and colleagues (Chen, Lo, Liu, & Cheng, 2016). In a first phase, participants were asked to study several words printed in different colors and then, after a distracting task, to discriminate old/new words and to retrieve the color in which they have been originally shown. This task is thought to reflect general episodic abilities (i.e., the old/new judgements) and deeper source-memory processing (i.e., the color recognition).

We selected 100 concrete words from the Italian database provided by Della Rosa and colleagues (Della Rosa, Catricalà, Vigliocco, & Cappa, 2010). The 100 words were divided into two sets of 50 words. The first 50 were divided into two subsets, presented respectively in red and blue fonts. These words were presented during the memorization phase and were therefore "old" items in the subsequent recognition task. The remaining 50 words were not shown in the study phase and were used as "new" items in the subsequent recognition phase.

The two sets (i.e., as well as the two subsets) were comparable in terms of concreteness, imageability, familiarity, age of acquisition, context availability, abstractness, mode of acquisition, number of letters and logarithm of the frequency (all *p*s $> .32$, all BFs $< .33$).

During the first part of the task (i.e., encoding phase), participants were instructed to memorize 50 words: 25 words were colored in red and 25 words were colored in blue. Each trial started with a central fixation cross (presented for 500 ms) followed by a word (presented for 1500 ms) presented in either red or blue, followed by a blank screen (presented for 300 ms), then the script moved automatically to the next fixation cross. The words were shown in random order. At the end of the study phase, participants were required to perform an attention task (i.e., a modified version of the go-no go task) as a brief distracting task for 2 minutes.

Then participants had to complete the recognition phase. Participants were instructed to make old/new judgments and to respond as fast and as accurately as possible by pressing the left/right keys (A and L); the response keys were counterbalanced across participants. In case of "old" judgements (i.e., when participants judged the word as already presented in the memorization phase), the script moved to the next screen showing two question marks (one in the left side and one in the right side, one in red and the other in blue, with the position counterbalanced across participants); the participants were asked to make a source judgment by identifying the correct color in which the word was presented in the memorization phase. On the contrary, in case of a "new" judgement, participants were instructed to press the space bar, then the script moved automatically to the next trial. The trials were shown in random order. In this phase, each trial started with a central fixation cross (presented for 1000 ms) followed by a word (presented for 2500 ms) and, after participants' response, by the source judgement. See Figure 11 for a schematic representation of the episodic-source memory task.

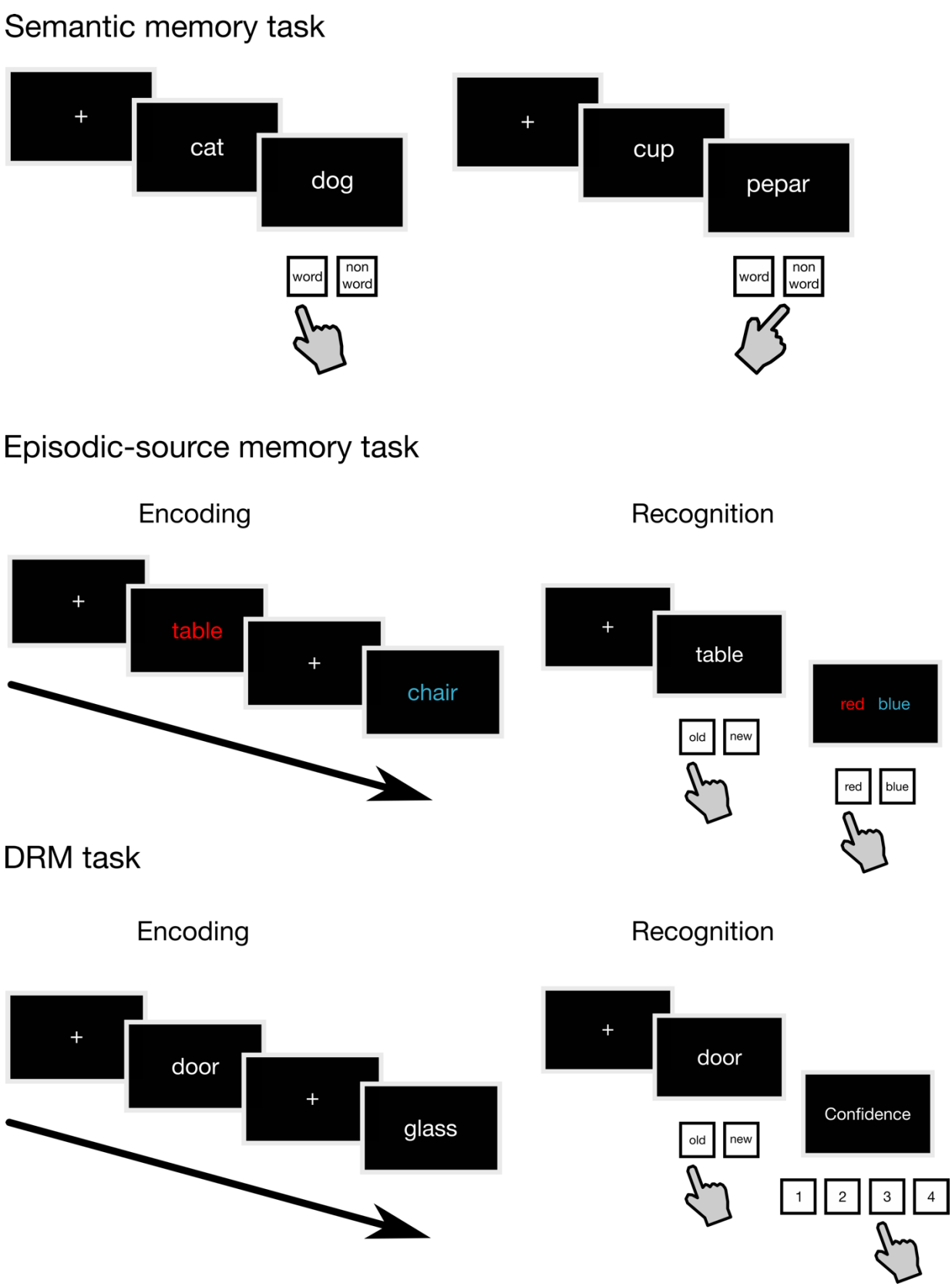


Figure 11. Schematic representation of the semantic priming task used to measure participants' semantic abilities (top), of the task used to measure participants' episodic and source memory abilities (middle), and of the two phases of the DRM task (bottom).

DRM task

We used a typical version of the DRM task (Roediger & McDermott, 1995). In a first phase, participants were asked to memorize several words grouped by lists (within each list, the words were semantically/associatively related to a non-shown target word, named critical lure; see below) and then, after a distracting task, to discriminate old/new words. Typically, participants report as "old" (i.e., as previously memorized) a large number of critical lures, although they were not in the studied items (for a review see, Gallo 2010).

We selected 10 lists of words from the normative data for Italian DRM reported by Iacullo and Marucci (2016). Each list was composed by 15 words associatively related to a non-shown word (called *critical lure*). In this study, the words in position 13 and 14 were excluded and used as weakly related lures during the recognition phase, while the 15th word was not used; thus, participants studied 12 words lists. From the remaining lists reported by Iacullo and Marucci (2016) – the ones not used in the present study – we extracted the words used as control stimuli.

The recognition phase was composed of 80 words, half of which had been studied and half of which had not. The 40 studied words presented in the recognition phase were those in positions 1, 4, 7 and 10 in the studied lists. Of the 40 non-studied words, 10 were the critical lures of the studied lists, 20 were weakly related lures, and 10 were unrelated words. The unrelated words were chosen randomly from the above-mentioned excluded lists. For example, for the list with *slow* as critical lure, the words included in the list were: *fast, snail, trend, dance, train, adagio, elderly, calm, delay, waltz, clear and tortoise*; the words used in the recognition phase as studied words were: *fast, dance, elderly and waltz*; the words used as weakly related lures were *peaceful* and *ant*; the unrelated word was *lemon*.

During the first part of the task, participants had to memorize a series of words (i.e., they were required to study 10 lists of words without interruptions). Participants were shown the 12 words that composed each of the 10 lists in descending forward associative strength (FAS; i.e., the association strength from the critical lure to the word that compose the list). The order by which the lists were presented was random, while the order of the words within each list was fixed (see Roediger & McDermott, 1995). Each trial started with a central

fixation cross (presented for 500 ms) followed by a word (presented for 1500 ms) and a blank screen (presented for 300 ms), then the script moved automatically to the next fixation cross.

At the end of the encoding phase, participants were required to perform an attention task (i.e., a modified version of the go-no go, different from the one employed during the episodic task) as a distracting task for 2 minutes.

Then participants were asked to perform the recognition phase. Participants were instructed to make old/new judgments and to respond as fast and accurately as possible by pressing the left/right key (A and L) using both hands; the response keys were counterbalanced across participants. After the old/new judgment, the script moved to the next screen and participants were asked to make a confidence judgement about their response using the keys 1, 2, 3, and 4 (i.e., with the key 1 representing the lowest levels of confidence and the key 4 representing the highest levels of confidence). The trials were shown in random order.

Each trial started with a central fixation cross (presented for 1000 ms) followed by a word (presented for 2500 ms) and then by the confidence judgement. The confidence judgement ended the trial and the fixation cross of the next trial was presented. See Figure 11 for a schematic representation of the DRM task.

Computation of the memory components scores

Signal detection theory measures were calculated using R-Studio (RStudio Team, 2015), by means of the *psycho* package (Makowski, 2018).

The semantic memory score was computed as the priming effect z-transformed score (i.e., in order to get homogeneous values from different memory scores used as predictors in the final analyses) induced by the prime on the processing of the target word (i.e., the speeding up in participants' correct reaction times; RTs). That is, for each participant, the semantic score was computed by subtracting the mean RTs for related prime-target words pairs to the mean RTs for unrelated prime-target words pairs (i.e., positive values indicate a facilitation in the processing of the target word as induced by the related prime as

compared to the unrelated prime and, consequently, a stronger activation of the semantic memory network). In analogy with previous studies using associative and semantic priming tasks (Argyropoulos & Muggleton, 2013), we analyzed responses with RTs < 1200 ms (5% of the trials excluded), since slower responses are thought to reflect distraction or low familiarity with the words showed, rather than lexical access (Perea & Gotor, 1997). Similarly, responses judged too fast (RTs < 300 ms) were excluded (2% of the trials excluded). Critically, and as expected, participants' RTs for related words were overall significantly faster compared to RTs for unrelated words, $t(49) = -2.44$, $p = .01$, thus reflecting a facilitation due to semantic relatedness.

The episodic memory score was computed as participants' *d'* (Stanislaw & Todorov, 1999), that is, the z-value of the hit-rate ("yes" response when the correct response is "yes") minus the z-value of the false-alarm rate ("yes" response when the correct response is "no") in old/new judgements of the episodic task. Participants' *d'* was computed only on responses with RTs > 300 ms and RTs < 3000 ms (6% of the trials excluded).

The source memory score was computed as participants' *d'* in the source judgment of the episodic task. In this case, we considered only trials in which participants responded "old" to old words (i.e., only trials in which there was actually a color source to be accessed to), since including also the other items would have caused an overestimation in participants' error rate. To compute the source memory score, we considered as hits those trials in which participants correctly identified the correct color, while we considered as false alarms those in which participants indicated the wrong color. We included only trials with RTs > 150 ms and RTs < 3000 ms (2% of the trials excluded).

The semantic memory score was not correlated with the episodic memory score, $r = .03$, $p = .78$, nor with the source memory score, $r = .05$, $p = .68$. However, the episodic memory score and the source memory score were moderately correlated, $r = .34$, $p = .01$. The significant correlation between episodic and source memory scores is consistent with the theoretical view according to which they would represent two sub-components of the same memory system (Baddeley, Eysenck, & Anderson, 2009).

Finally, in the DRM task, trials with RTs < 300 ms and RTs > 3000 ms were excluded from the analyses (4% of the trials excluded for old/new judgements; 11% of the trials excluded for confidence judgements). The analyses on confidence judgements were performed including only the trials in which participants responded "old" (i.e., hits for studied words and false alarms for critical lures, weakly related lures and unrelated words).

Data analysis

All the analyses were performed using R-Studio (RStudio Team, 2015). Data were analyzed through a mixed-effects approach, which incorporate both fixed-effects and random-effects (i.e., associated to statistical units as participants and task stimuli) and provide more detailed information about relationships among predictors and outcome variables compared with Pearson correlation (which simply measures the strength of the linear relationship between each selected pair of variables independent of the others; Koerner & Zhang, 2017). Generalized linear mixed models (GLMMs) were run using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015), while cumulative link mixed models (CLMMs) were run using the *ordinal* package (Christensen, 2019). The graphs reported were obtained using the *effects* package (Fox, 2003; Fox & Weisberg, 2019).

First, we explored the memory processes subserving false and veridical memory. We ran a GLMM having old/new judgements in the DRM task (i.e., "new" responses were scored as 0, whereas "old" responses as 1) as the dependent variable and subjects and items as random intercepts. In particular, in our statistical model, we additively included as predictors the semantic, episodic and source memory scores along with their interaction with the Type of stimuli (critical lures vs. weakly related lures vs. unrelated words vs. studied words). That is, in the *lme4* syntax, we tested the following model:

$$Response \sim (Semantic + Episodic + Source) * Type + (1|Participant) + (1|Item)$$

Second, we explored the memory processes subserving confidence judgements when making veridical and false memories. We ran a CLMM with confidence judgements (i.e., from low to high confidence, 1 vs. 2 vs. 3 vs. 4, as a factor) as the dependent variable and subjects and items as random intercepts. The semantic, episodic and source memory scores were additively included along with their interaction with the type of stimuli (critical lures vs. weakly related lures vs. unrelated words vs. studied words; i.e., the model is analogous to the one reported above for old/new judgements).

In the Results section we report only the analyses on old/new judgements, since no significant effect was found for confidence judgements (i.e., the results on confidence judgements are reported as Supplementary Materials of the main manuscript).

3.3.3 Results

The *Pseudo-R^2* (total) of the model estimated was = .48. In particular, the effects of episodic memory score, $\chi^2(1) = 8.20$, $p = .004$, semantic memory score, $\chi^2(1) = 4.11$, $p = .04$, and type of stimuli, $\chi^2(1) = 36.62$, $p < .001$, were significant. Conversely, the effect of source memory score was not significant, $\chi^2(1) = 3.34$, $p = .07$. Critically, the interactions semantic memory score by type of stimuli and episodic memory score by type of stimuli were found to be significant, respectively, $\chi^2(4) = 9.59$, $p = .02$; $\chi^2(4) = 28.12$, $p < .001$.

The significant interaction semantic memory scores by type of stimuli (Figure 12A) indicates that for critical lures the higher the semantic memory score, the higher the chances of making false memories, $z = 2.03$, $p = .04$. No effect was found for studied words, $z = -.22$, $p = .82$, weakly related lures, $z = -.88$, $p = .38$, nor for unrelated words, $z = -1.06$, $p = .29$.

The significant interaction episodic memory scores by type of stimuli (Figure 12B) indicates that for critical lures, $z = -2.87$, $p = .004$, weakly related lures, $z = -2.47$, $p = .01$, and unrelated words, $z = -2.32$, $p = .01$, the higher the episodic memory score, the lower the chances of making false memories. No effect was found for studied words, $z = .91$, $p = .36$.

Conversely, the interaction source memory score by type of stimuli was not significant, $\chi^2(4) = 2.42$, $p = .48$ (Figure 12C).

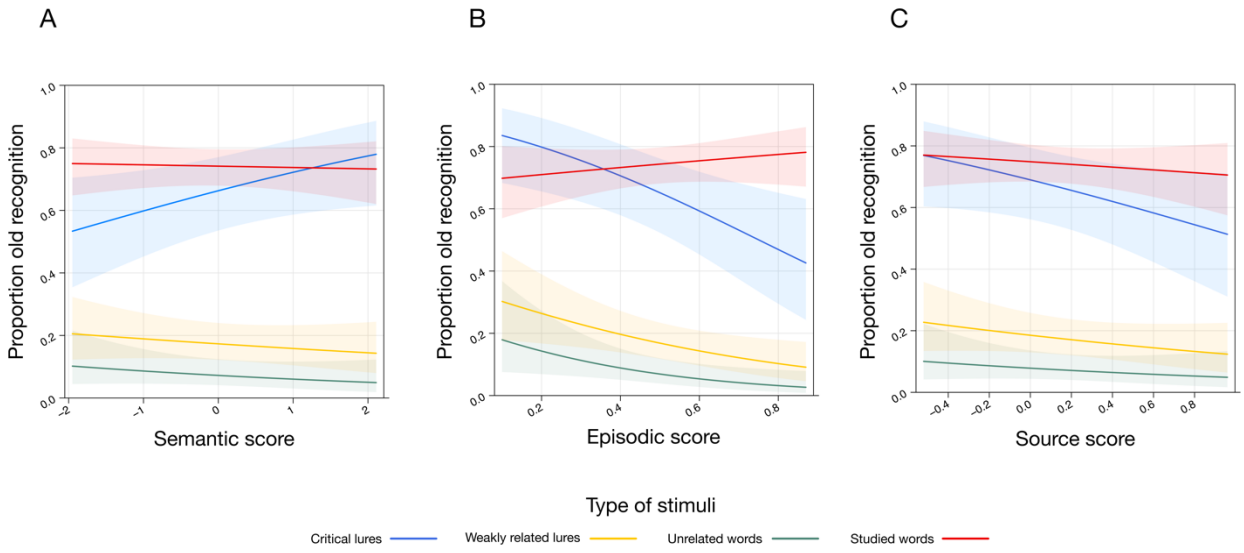


Figure 12. Results from the GLMM including the interaction type of stimuli by semantic memory score (A), episodic memory score (B) and memory source score (C) as predictors and the proportion of "old" responses (divided in critical lures, weakly related lures, unrelated words and studied words) as dependent variable. The results show that the occurrence of false recognitions of the critical lures is predicted by both semantic and episodic memory scores, albeit with different directions (i.e., more false memories for individuals with better semantic memory abilities; less false memories for individuals with worse episodic abilities). On the contrary, the source memory score does not have any effect on false memories.

3.3.4 Discussion

In the present study, we investigated individual differences in memory processes subserving false recognitions in a classic DRM task. In particular, we assessed participants' semantic, episodic and source memory abilities and explored whether these three memory components differently predicted performance in the DRM task. Our findings indicate that the occurrence of false memories is predicted by both semantic and episodic memory, yet to a different extent. That is, higher semantic memory abilities predicted a *larger* number of false memories, while higher episodic abilities predicted a *smaller* number of false memories. Interestingly, the effect of semantic memory was specific for critical lures (i.e., better semantic memory abilities were associated with an increased likelihood of false recognitions), while the effect of episodic memory was found for all the three types of new stimuli, namely critical lures, weakly related lures and unrelated words (i.e., better episodic memory abilities were associated with a reduced likelihood of memory errors).

These findings have potential implications for our understanding of false memories formation. First, previous studies investigated semantic memory involvement in the DRM task only at the item level (e.g., Roediger et al., 2001). That is, these studies manipulated the level of semantic memory involvement during the DRM task, by including word lists that could be differently associated with a critical lure (i.e., hence having lists more semantically related with critical lures, Brainerd et al., 2008). Typically, the more the word lists are semantically related with the critical lure, the higher the occurrence of false memories (Brainerd et al., 2008; Cann et al., 2011; Gatti et al., *in press*; Roediger et al., 2001). Our findings, by showing that also differences in participants' semantic abilities play a role in the formation of false memories, indicate that these two components (i.e., the semantic content of the words processed and the extent to which participants rely on semantic memory) are possibly interacting during the DRM task. Second, by showing that also participants' episodic abilities contribute to both veridical and false memories, our data suggest that episodic memory is actively involved while judging if a word was studied or not. Such an effect is likely reflecting the recall-to-reject process (Tulving, 1983) in its diagnostic component (Gallo, 2004). That is, in the recognition phase, participants would recall the words studied and compare them with the ones actually presented: if the episodic trace for the word actually presented is weaker compared to the ones studied, the participant is less likely to accept the word as "old" (Gallo, 2004, Israel & Schacter, 1997). Notably, the contribution of episodic processes would be hard to be operationalized at the item level, again testifying the advantages of the methodological approach employed here.

Conversely, source memory – the participants' ability to retrieve the exact perceptive features (i.e., the color of the studied words) – did not have any impact on the DRM task. These findings, indicate that source memory abilities (at least in the way we operationalized them, but for an alternative perspective on source memory see: Gallo, Korthauer, McDonough, Teshale, & Johnson, 2011) might not play a major role in true and false memories formation. To account for this null finding we note that, in the DRM task, the studied words are typically presented from the same source (e.g., in the same ink color or read aloud by the same experimenter). In the studies in which the source of the information was manipulated (e.g., which experimenter read the study list; Payne, Elie, Blackwell, &

Neuschatz, 1996; but see also: Hicks, & Marsh, 1999) participants certainly noticed the difference, but this did not substantially affect the rate of false memories reported. Our operationalization of source memory ability thus may tap on aspects that are not related to the specific cognitive requirements of the DRM task. Additionally, the lack of a source memory involvement in correctly rejecting new stimuli could also be specifically related to the fact that here both the encoding and recognition phases were presented in the visual modality (while a mixed auditory-visual modality has been employed in some previous research; e.g., Roediger & McDermott, 1995). That is, the sensory facilitation occurring in the visual modality could have induced participants to rely on a lesser extent on source memory.

More generally, these data support the utility of individual differences for investigating the factors involved in false memory production and formation. Previous studies have indeed investigated participants' performance in the DRM task using an individual differences approach showing that several variables underly some of the individual variation in susceptibility to memory illusions. For example, participants' performance in the DRM task has been predicted from working memory abilities (Holden et al., 2020; Unsworth & Brewer, 2010; Watson et al., 2005), age (Norman & Schacter, 1997; Watson et al., 2001), memory self-efficacy (Iacullo et al., 2016), creativity (Dewhurst et al., 2011), or need for cognition (Graham, 2007; Leding, 2011). Here, for the first time, we employed individuals' semantic and episodic memory to predict responses in the DRM task. Our results support the specific contribution of episodic and semantic processes in the production of false memories, thus corroborating previous theoretical evidence accounting for the DRM task.

In particular, these findings directly support the two mainstream theories accounting for false memory, the AMF and FTT theories, which predict the occurrence of false recognitions on an associative/semantic basis, while adequate episodic memory processes should counter them (Brainerd, & Reyna, 2002; Gallo & Roediger, 2002; Reyna & Brainerd, 1995; Roediger et al., 2001). The main difference between AMF and FTT is that the former assumes that critical lures are falsely recognized due to the associative link between that and the list words, while according to the latter theory critical lures and list words would share semantic features that underlie false recognition. Yet, disentangling the associative from the

semantic framework may be sometimes difficult, as often two words that are associatively related are also semantically related (Grossman & Eagle, 1970; Thompson-Schill, Kurtz, & Gabrieli, 1998; but see Brainerd et al., 2008). This is also reflected in the task used here to assess semantic memory, which necessarily involves both semantic and associative processes and was selected on purpose to be representative for both AMF and FTT hypotheses.

On the one hand, according to the AMF, during the encoding phase, the presentation of the list words would associatively hyperactivate the critical lure (Gallo & Roediger, 2002; Roediger et al., 2001). Consistent with this possibility, Roediger and colleagues (2001) have shown that the more the list words are associatively related to the critical lure, the more the participants are prone to false memories and recognitions. Crucially, the AMF hypothesis also posits that, during the recognition phase, successful monitoring processes (i.e., the ability to distinguish whether retrieved information refers to past events or not; Johnson & Raye, 1981) can counteract false recognitions and enhance veridical memories (Gallo & Roediger, 2002; Roediger et al., 2001). In line with this possibility, here we showed that the higher the participants' episodic memory ability, the lower the chances of making false memories. Critically, this effect was not limited to critical lures, but spread as well to other types of new words (i.e., weakly related lures and unrelated words), indicating that participants could adopt the same strategies in rejecting the various types of new stimuli.

On the other hand, according to the FTT participants' false memories would rely on a semantic trace, called gist trace and linked to the semantic content of each list (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). Such a theoretical framework is as well in line with our findings: participants with higher semantic memory abilities are thought to have stronger abilities in forming the gist trace, resulting in turn in a higher false recognition occurrence. Furthermore, according to the FTT, correct rejections would depend on the verbatim trace, linked to both episodic and perceptive features of the studied words (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). Here, however, we could only find evidence for a relationship between episodic memories and false recognitions, but not with source memory (i.e., defined in our study as the ability to correctly retrieve the perceptual features of studied words).

Beyond the two main theoretical hypotheses accounting for false memory, previous evidence suggests that false recognitions follow a continuous semantic gradient in terms of backward associative strength (Roediger et al., 2001) and semantic similarity (Gatti et al., *in press*) between the studied words and the critical lures. Similarly, it has been shown that the semantic component that should elicit false recognitions can be decomposed into various sub-components (situation features, synonyms, antonyms, and taxonomic relations; Cann et al., 2011). The results of the present study extend these findings by showing that also individual differences in semantic and episodic memory play a crucial role in determining false memory production. These results also point out the need for investigating semantic and episodic processes while analyzing participants' performance in the DRM task in order to link false recognition proportion to episodic vs. semantic modulation. For example, employing this approach in neuromodulation studies would allow for a better comprehension of the underlying memory processes also on a neural level (i.e., if enhanced false recognition depends on impaired episodic memory on enhanced semantic memory; for alternative approaches see also: Diez, Gomez-Ariza, Diez-Alamo, Alonso, & Fernandez, 2017; Gatti, Vecchi, & Mazzoni, 2021).

Finally, one possible point of concern could be related to the lack of effects of semantic and episodic memory on veridical recognition. To account for this pattern of results, we first note that, besides unrelated words, we also used as control stimuli weakly related words; these weakly related words could have induced a shift in participants' response criterion favoring in certain cases a conservative response bias (i.e., recognizing that also weakly related words were shown in the recognition task could have induced some participants to respond "no" more often to less semantically related studied words). Second, as the semantic features of the studied words underlying false recognitions comprise several sub-components (Cann et al., 2011), participants' semantic memory could be decomposed into different sub-systems involved in veridical and false recognitions. Consequently, extracting the semantic score from another task could better account for veridical recognition, with this approach being particularly promising for future research in order to distinguish between the AMF and FTT theories.

In conclusion, by taking an individual differences approach, in this study we show that various memory systems are differently involved in veridical and false memories production. Our findings thus point to the importance of investigating individual differences linked to false memory production, possibly through more ecological paradigms, since this has ultimately links with the quality of justice trials and witness evaluation (Schacter & Loftus, 2013).

3.4

The relationship between theory of mind and false memories: an individual differences approach

The present study is currently in preparation. To cite it:

Gatti, D., Stagnitto, S., Basile, C., Mazzoni, G., Vecchi, T., Rinaldi, L., & Lecce, S. (in preparation). The relationship between theory of mind and false memories: an individual differences approach.

3.4.1 Introduction

Human memory and the ability to understand/infer other's mental states and intentions (generally labeled as *Theory of Mind*, ToM, Premack & Woodruff, 1978; Wimmer & Perner, 1983; and for a recent overview see: Devine & Lecce, 2021) are closely related across the whole lifespan. As Perner (1991; Perner et al., 2007) proposed, ToM could be a key factor in memory development, for example allowing to understand the difference between one's memory and reality (Hoerl, 2018) or the difference between one's subjective experience of the event, at the moment in which the event has occurred, and the memory about that event (Martin, 2001). However, although several neuroimaging studies have shown that the areas active during ToM functions and autobiographical memory are largely overlapped (Buckner & Carroll, 2007; Rabin et al., 2010), this relationship cannot be defined clearly, since other studies have reported intact ToM abilities in amnesic patients

(Rosenbaum et al., 2007). Recent studies have also investigated the relationship between ToM abilities and language components, such as pragmatics (e.g., Del Sette et al., 2020; 2021; Lecce et al., 2019), but no studies have directly investigated the role of ToM in semantic memory processes.

An indirect link between ToM and semantic memory processes comes from studies that investigated the performance of individuals with schizophrenia or autism, which generally show impaired ToM abilities (Brüne, 2005; Frith, 2004; Happé et al., 2017), in memory tasks such as the DRM task (Deese, 1959; Roediger & McDermott, 1995). In the DRM task participants are first presented once with several lists of words that have to be memorized (within each list, the words are semantically/associatively related to a non-shown target word, named *critical lure*; e.g., door, glass, pan, shade, ledge, etc. – critical lure: window) and then, after a brief distracting task, they are asked to perform a recognition task in which they have to indicate whether a given word was part of the memorized lists or not. Interestingly, during this last phase, participants tend to report as "old" the critical lures (i.e., the recognize them as if they were part of the memorized lists), although these words were never presented in the encoding phase (for a review, see Gallo, 2010).

The DRM is a semantic task as shown by previous studies successfully predicting false memory occurrence using several predictors related to semantic and associative processes (Brainerd et al., 2008; Cann et al., 2011; Gatti et al., *in press*; Roediger et al., 2001).

Specifically, using the DRM task, it has been shown that the processing of semantic relationships as well as the extraction of the meaning is impaired in individuals with schizophrenia (Lee et al., 2007; Paz-Alonso et al., 2013). Similarly, individuals with autism are less susceptible to the false memory effect of the DRM both in the verbal (Beversdorf et al., 2000; Wojcik et al., 2018) and in the visual (Hillier et al., 2007) versions of the task. These results can be interpreted using two complementary frameworks explaining information processing in autism, such as the Fuzzy Trace Theory (FTT; which happens to be widely used also to explain typical individuals' performance in the DRM task, Reyna &

Brainerd, 1995) and the Weak Central Coherence hypothesis (WCC; Frith, 1989; Frith & Happé, 1994; Happé & Frith, 2006).

FTT assumes that, during childhood, individuals' memory is essentially based on detailed encoding of information as it was presented and, as they grow up, they develop memory processes based on the meanings associated to the original input (Brainerd & Reyna, 1992, Reyna & Brainerd, 1991a, Reyna & Brainerd, 1991b), with both of these components playing a relevant role in memory performance across the whole lifespan. Similarly, the WCC hypothesis assumes that individuals with autism are biased toward a detailed-focused processing (i.e., hyper encoding of the verbatim trace) and an impaired global processing (i.e., hypo encoding of the gist trace) (Happé & Frith, 2006).

Using the FTT it is indeed possible to explain typical individuals' performance in the DRM. During the encoding phase, participants encode two memory traces: a verbatim trace, linked to perceptive features of the stimuli, and a gist trace, linked to the semantic theme of the list (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). Both traces would be involved in the correct recognition of previously showed words, but the verbatim trace would be specifically involved in correct rejection of new stimuli, while the gist trace would be responsible for the production of false memories. According to FTT and WCC, the pattern observed in individuals with autism can be traced back to their lower reliance on gist information in favor of a verbatim-biased information processing style, thus protecting against the classical false memory pattern observed in typical individuals (for an in-depth discussion see: Miller et al., 2014). Similarly, the impaired extraction of the meaning in individuals with schizophrenia can be traced back to impaired gist-based processes (e.g., Lee et al., 2007).

Here, building on these perspectives linking ToM abilities and semantic processing, we aimed to directly investigate such relationship in typical adults using the DRM task. Specifically, we applied an already established method to compute the semantic similarity between the words in the recognition phase and the studied ones (for a complete discussion see: Gatti et al., *in press*; but see also: Gatti et al., 2021a, 2021b), by employing indexes extracted from distributional semantic models (DSMs). DSMs induce words meanings from

large databases of natural language data, representing them as high-dimensional numerical vectors: these models indeed are thought to well capture the structure of semantic memory (Günther et al., 2019; Jones et al., 2015). Participants were asked to perform the DRM task and three other tasks: the Triangles task (Abell et al., 2000; Castelli et al., 2000, 2002) and the Reading the Mind in the Eyes (RMET; Baron-Cohen et al., 2001), which are respectively used to measure cognitive and affective ToM components; and the Vocabulary subtest of the Wechsler Adult Intelligence Scale (Wechsler, 2008), which is generally used as a covariate when employing the Triangles task (Wilson, 2021). Following the FTT and the WCC, we predicted participants' reliance on semantic memory in recognizing the false memory items of the DRM task as extracted from DSMs to be predicted by participants' score in one of the two ToM measures.

3.4.2 Methods

Participants

One hundred and three Italian students (30 males; one participant reported being not binary, *M* age = 22.96 years, *SD* = 3.39, participants' age ranged between 19 and 35) participated in the study. All participants were native Italian speakers, had normal or corrected to normal vision and were naïve to the purpose of the study. Informed consent was obtained from all participants before the experiment. The protocol was approved by the psychological ethical committee of University of Pavia and participants were treated in accordance with the Declaration of Helsinki.

DRM task: stimuli and procedure

Participants performed the DRM task (Deese, 1959; Roediger & McDermott, 1995), a typical false memories paradigm, in which participants are instructed to remember several lists of words and then, after a brief distracting task, to perform a recognition task. The words that compose each list are associatively related to a non-shown word (called critical

lure) and this association is thought to be responsible of participants' false memory (Roediger et al., 2001; but see also: Gatti et al., *in press*).

The task is composed by two consecutive phases: an encoding phase and a recognition phase. For the encoding phase, we selected 12 lists of words out of 24 from the normative data for the Italian DRM test (Iacullo & Marucci, 2016). Each list was originally composed by 15 words: we selected the first 12 words (144 words in total), while 2 of the 3 remaining words were used as weakly related lures (see below).

The recognition phase was composed of 96 words, 48 of which had been presented in the previous phase (i.e., studied words) and 48 of which had not been previously presented (i.e., new words). The 48 studied words presented in this experimental phase were those in serial positions 1, 4, 7 and 10 in the studied lists. Of the 48 new words, 12 were the critical lures from the studied lists (i.e., the non-shown words mostly associated with the words composing each list), 24 were weakly related lures and 12 were unrelated words. The weakly related lures were 2 of the 3 words of the studied lists that were not presented in the list, those in position 13 and 14. The unrelated words were chosen randomly among the words of the excluded lists; this criterion was established arbitrarily.

During the encoding phase, participants were required to study 12 lists of words. Participants were shown the 12 words that composed each of the 12 lists in descending forward associative strength (FAS; i.e., the association strength from the critical lure to the word that compose the list). The order by which the lists were presented was random, while the order of the words within each list was fixed (see: Roediger & McDermott, 1995). Each trial started with a central fixation cross (presented for 500 ms) followed by a word (presented for 1500 ms) and a blank screen (presented for 300 ms), then the script moved automatically to the next fixation cross. At the end of the twelfth list, participants were requested to perform a distracting task (i.e., to solve as many arithmetical operations as they could) for 2 minutes.

Then participants were asked to perform the recognition task. Participants were instructed to make old/new judgments and to respond as fast and accurately as possible by pressing the left/right key (A and L) using both hands; the response keys were

counterbalanced across participants. After the old/new judgment, the script moved to the next trial. The trials were shown in random order.

Each recognition trial started with a central fixation cross (presented for 1000 ms) followed by a word (presented for 2500 ms), followed by the confidence judgement. The confidence judgement ended the trial and the fixation cross of the next trial was presented.

The DRM task was administered online using Psychopy (Pierce, 2007, 2009; Pierce & MacAskill, 2018; Pierce et al., 2019) through the online platform Pavlovia (https://pavlovia.org/).

Psychological and cognitive measures: stimuli and procedure

After giving informed consent and filling in a first questionnaire assessing sociodemographic information (i.e., age, gender, years of education), participants performed the DRM task and then, in counterbalanced order, completed the other tasks investigating vocabulary competencies and ToM (e.g., reading the mind eyes and triangles tasks). Sociodemographic information and these last three tasks were collected online using Google Forms.

Vocabulary

In order to examine the vocabulary knowledge, we used the Vocabulary subtest of the Wechsler Adult Intelligence Scale – Fourth Edition (Wechsler, 2008), in the Italian standardized version proposed by Orsini & Pezzuti (2013). Participants were presented with 27 words (4-30 item, Wechsler, 2008) and were asked to provide the meaning of each word. We adapted the original oral modality to an online form, in which participants were presented with the written words one by one and were asked to read each of them and write down what each word means (i.e., providing a definition).

Each item was scored by using a value from 0 to 2. The score of 2 indicates a good comprehension of the word, in which the participant provided an exhaustive definition, with one or more main or definitive features. The score of 1 indicates a correct answer but poor

in content (e.g., providing a vague or irrelevant synonym or just an example of the term). Finally, the score of 0 indicates clearly wrong or absent answers that do not show any correct comprehension of the proposed word. The total score of each participant is given by the sum of the scores obtained in all the 27 items.

Reading the mind in the eyes

Participants were presented with the reading the mind in the eyes task (RMET) stimuli proposed by Baron-Cohen and colleagues (Baron-Cohen, Richler, Bisarya, Gurunathan, & Wheelwright, 2003; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), in the Italian version adapted by Serafin & Surian (2004). The RMET is composed by 37 black and white photos, with the first as practice trial and 36 task stimuli. Each image depicts the eye-region of different people, of different age and of both sexes, and they are taken from English newspapers and magazines (Baron-Cohen, 2003).

Each image is surrounded by 4 different adjectives or terms regarding emotional states/expressions, placed one at each corner of the picture, as in the original English version (Baron-Cohen, 2003). Participants were required to look at the image, read each of the four terms proposed for that picture and to select the term that best described what the person in the photo was thinking or feeling (i.e., a forced choice task between four response options). Instead of the classical pen-and-paper form, we adapted the test to an online modality, where participants were presented with the image surrounded by the four terms at the corners (as the original task) and, below it, a multiple-choice format, where participants had to select among the four options presented in column.

In order to minimize vocabulary differences among participants, they were told to ask to the examiner for the vocabulary definition whenever they had uncertainties regarding the meaning of the proposed terms: the examiner reads the definitions proposed in the glossary, proposed by the Italian version of the test (Serafin & Surian, 2004).

Each response is scored as 1 or 0. The value 1 indicates correct response, while 0 indicates wrong response. The total score of each participant is given by the sum of the

scores obtained in all the 36 items, excluding the first test example trial for which the examiner provided the feedback.

Triangles task

Participants were shown 9 silent animations, lasting 35-45 seconds each, depicting two triangles: a big blue one and a small orange one are moving about in a framed white landscape (Abell, Happe, & Frith, 2000; Castelli, Happé, Frith, & Frith, 2000, 2002). Participants were instructed to watch each video just once and, after it, to answer a question, which is the same for all the videos: *"In your opinion, what happened in this video?"*.

The animations were divided in two typologies: 1) Action *clips* (3 videos), in which triangles moved in a goal-directed fashion (e.g., chasing, fighting); 2) ToM clips (6 videos), in which triangles moved interactively with implied intentions (e.g., coaxing, tricking).

Participants' answers were scored on the basis of the index of intentionality (range 0-5), which reflects the degree to which the subject describes complex, intentional mental state. It is calculated analyzing the content of each response. Two scores of intentionality, one for the Action clips and one for the ToM clips, were computed by summing scores in each answer. Interrater agreement (based on double-coding of 25% of the responses) was good (*Cohen's* $K = .83$, $p < .001$).

Word-embeddings

In the present study we applied a method developed by us capitalizing on seminal studies on distributional semantics (for a review see: Günther, Rinaldi, & Marelli, 2019) to compute semantic similarities values between the words in the recognition phase and those in the encoding phase of the DRM task (for an extended discussion see Gatti et al., in press; see also below *Computation of semantic similarity values*). This method allows to quantify at the item level the degree of semantic similarity between the to be recognized words and the encoded ones building word representations from language usage (i.e., predicting a target word from the linguistic context in which it typically appears).

Vector representations for the words used in the DRM task were extracted from a semantic space obtained by inducing word embeddings using the Continuous Bag of Words (CBOW) method, an approach originally proposed by Mikolov and colleagues (Mikolov, Chen et al., 2013). The model, released by Marelli (2017), was trained on itWaC, a free Italian text corpus based on web-collected data and consisting of about 1.9 billion tokens. The model used is set on the following parameters: *9-word co-occurrence window*, *400-dimension vectors*, negative sampling with $k = 10$, subsampling with $t = 1e^{-5}$. This set of parameters defines the learning procedure used to induce word vectors (Mikolov, Chen et al., 2013). CBOW indicates the applied learning procedure: when using CBOW, the obtained vector dimensions capture the extent to which a target word is reliably predicted by the contexts in which it appears. Co-occurrence window size indicates how large the considered lexical contexts are; in our case, a *9-word window* indicates that we estimated predictions concerning 4 words on the left and 4 words on the right of the target word. The number of vector dimensions indicates how many nodes are included in the hidden layer, representing the result of the dimensionality reduction process implicitly applied by the network. Negative sampling estimates the probability of a target word by learning to distinguish it from draws from a noise distribution; the parameter k specifies the amount of these draws. The subsampling parameter t specifies a threshold-based procedure that limits the impact of very frequent, uninformative words.

From this semantic space, we extracted vector representations for the words used in this study. Specifically, for each word pair it is possible to obtain a semantic-similarity index (hence SSim) based on the cosine of the angle formed by vectors representing the meanings of these words. In particular, the higher the SSim value, the more semantically similar the words should be as estimated by the model. A heatmap matrix of the semantic similarity structure among the words composing a DRM list is represented in Figure 13.

Computation of semantic similarity values

For each new word (12 critical lures, 24 weakly related lures and 12 unrelated words) we computed a semantic similarity index (SSim). That is, SSim was computed as the

frequency-weighted average SSim (for a similar approach see: Gatti, Rinaldi, Marelli, Mazzoni & Vecchi, in press; Marelli & Amenta, 2018) between each new word in the recognition phase and each of the 12 words that composed its relative list. For unrelated words we computed the index randomly matching each word with a list. The formula used was:

$$\mathrm{SSim}(nw) = \frac{\sum_{i=1}^{k} cos(\overrightarrow{nw}, \overrightarrow{sw_i}) \times Fsw_i}{\sum_{i=1}^{k} Fsw_i}$$

where *nw* is a new word showed during the recognition task, cos refers to the cosine of the angle formed by the vectors representing a new word ($\overrightarrow{nw}$) and each of the k studied words ($\overrightarrow{sw}$) composing its list (e.g., for critical lures, we computed all the cosines between the critical lure and each word of its list), while Fsw_i is the frequency of each studied word as extracted from the Italian SUBTLEX (http://crr.ugent.be/subtlex-it/).

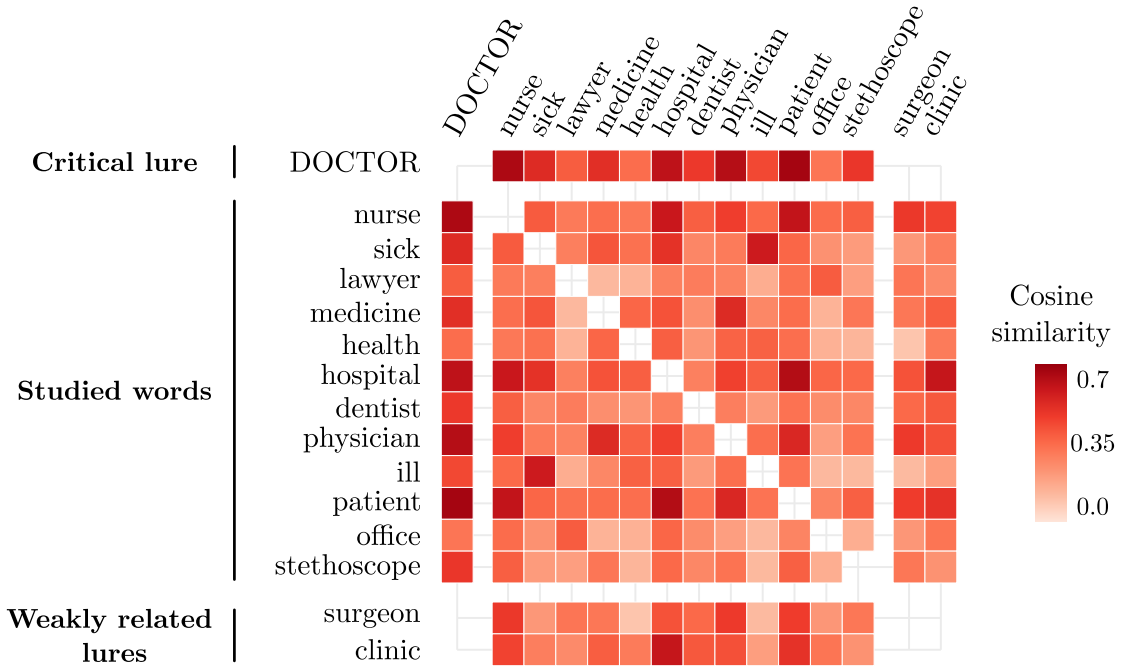


Figure 13. A heatmap matrix of the cosine values among the words composing the list *window* (i.e., critical lure; words list taken from Roediger & McDermott, 1995). Darker blue colors represent higher cosine values (and, hence, account for more semantically related words as predicted by DSMs). Note that in this case we plotted all the words in the list (i.e., including the possible weakly related lures) and that we computed the cosine values through an English DSM, available here: http://meshugga.ugent.be/snaut-english/ (see also: Mandera, Keuleers, & Brysbaert, 2017).

Data analysis

All the analyses were performed using *R-Studio* (RStudio Team, 2015). Generalized linear mixed models (GLMMs) were run using the *lme4* R package (Bates, Maechler, Bolker, & Walker, 2015), while linear models were run using the *stats* R package (R Core Team, 2019). The graphs reported were obtained using the *effects* R package (Fox, 2003; Fox & Weisberg, 2019).

The analyses can be divided into two main parts. First, before testing the main study hypothesis, we aimed to replicate the positive effect of SSim on participants' old responses for new words (for a complete discussion of this analysis see also: Gatti et al., *in press*). We, thus, estimated a GLMM having participants' responses ("old" responses scored as 1 and "new" responses scored as 0) for new words (i.e., critical lures, weakly related lures and unrelated words) as dependent variable and SSim as continuous predictor; subjects and items were included as random intercept and the effect of SSim was included as random slope across participants. Then, we aimed to quantify for each participant an estimate of the effect of the SSim predictor on new words while performing the DRM task. We thus extracted the conditional modes of the random effects estimated for each subject in the GLMM, namely the individual-level random slopes. Extracting the individual-level slopes for subjects allows us to observe the inter-participant variability in the sensitivity to the congruency effect. This index (labeled hereafter as SSim sensitivity) can be conceptualized as the SSim effect on participants' performance as extract by the *ranef* R command, which provides the conditional modes of the random effects from a fitted model object, that is the set of differences between the population-level average predicted response for a given set of fixed-effect values (SSim in our case) and the response predicted for each participant (i.e., thus each value describes how much each participant's slope differ from the slope of the total sample; see: Bates et al., 2015).

Secondly, in order to examine how individual differences in the effect of SSim while performing the DRM task are associated with Vocabulary and ToM variables, we estimated a linear model including the previously extracted SSim sensitivity index as dependent

variable and the psychological and cognitive measures (Triangles task ToM clips, Triangles task Action clips, RMET and Vocabulary) in the other tasks as continuous predictors.

3.4.3 Results

1. The effect of SSim on recognition in the DRM task

Trials in which overall RTs were faster than 300 ms or slower than 5000 ms (2% of the trials) were excluded from the analysis.

As expected, the effect of SSim on new words was significant, $z = 9.94$, $p < .001$, *Pseudo-*R^2 (total) = .44 (Figure 2), indicating that the higher the SSim (i.e., the higher the semantic similarity between the to be recognized new word and the studied words of each relative list), the higher the chances of recognizing it as "old" (i.e., the higher the chance of making a false recognition). From this model, we extracted the participants' random slopes, namely SSim sensitivity.

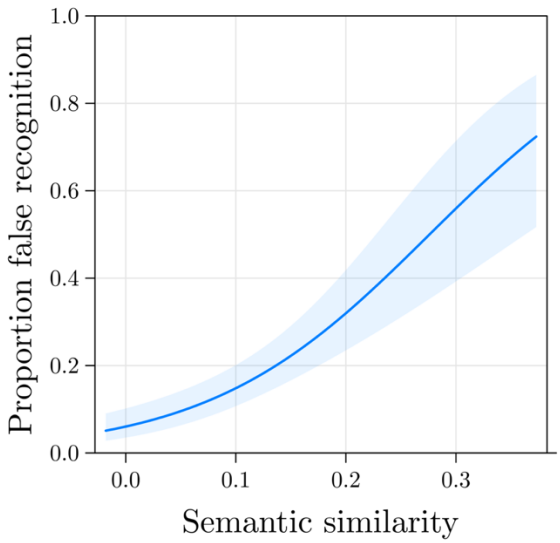


Figure 14. Results from the GLMM estimated using SSim as continuous predictor illustrating the positive relationship between SSim and recognition.

2. Descriptive statistics for the ToM and vocabulary tasks

Descriptive statistics for the four predictors are reported in Table 4.

Table 4. Descriptive statistics for the Triangle task (ToM and Action clips), Reading the Mind in the Eyes (RMET) task and the Vocabulary ability.

	Triangles task ToM clips	Triangles task Action clips	RMET	Vocabulary
Mean	21.97	7.60	26.90	40.17
SD	3.72	1.86	3.27	5.76
Possible range	0 – 30	0 – 15	0 – 36	0 – 69
Actual range	12 – 28	4 – 12	18 – 33	26 – 51

The correlation matrix between the four predictors and the dependent variable (SSim) is reported in Table 2. Overall, the correlations between the four predictors ranged from very low, as the one between Triangles ToM clips and RMET (r = -.02), to low, as the one between RMET and Vocabulary (r = .26), which was the only significant one, p = .007. Additionally, here we found a significant correlation (r = .28) between SSim sensitivity and Triangles ToM clips. However, these correlations do not allow to understand the overall relationships among SSim sensitivity and the various measures as corrected using the others (e.g., once corrected for the cognitive ToM measure, also the affective one could be relevant).

Table 5. Correlation between four predictors included in the current study.

	SSim	Triangles task ToM clips	Triangles task Action clips	RMET	Vocabulary
SSim	1				
Triangles task ToM clips	.25*	1			
Triangles task Action clips	-.07	.10	1		
RMET	-.02	.04	-.03	1	
Vocabulary	.05	.11	-.04	.26*	1

Note: * $p < .05$;

3. Does ToM variables predict participants' semantic performance in the DRM task?

The effects of the linear model having SSim sensitivity as dependent variable and the three ToM (i.e., Triangles task ToM clips, Triangles task Action clips and RMET) and vocabulary variables as continuous predictors are reported in Table 3 and Figure 3. Globally the model explained the 9% of the variance, R^2 = .09. Results also showed that only the effect of Triangles task ToM clips was significant, indicating that the higher participants'

score in the Triangles task ToM clips, the higher participants' reliance on semantic memory while performing the DRM task. No other significant effects were found.

Table 6. Results of the linear model on participants' semantic performance in the DRM task including TOM and Vocabulary predictors.

	b	*t-value*	*p-value*
(Intercept)	-.42	-.80	.38
Triangles task ToM clips	**.03**	**2.65**	**.01**
Triangles task Action clips	-.02	-1.05	.30
RMET	-.005	-.49	.63
Vocabulary	.002	.38	.70

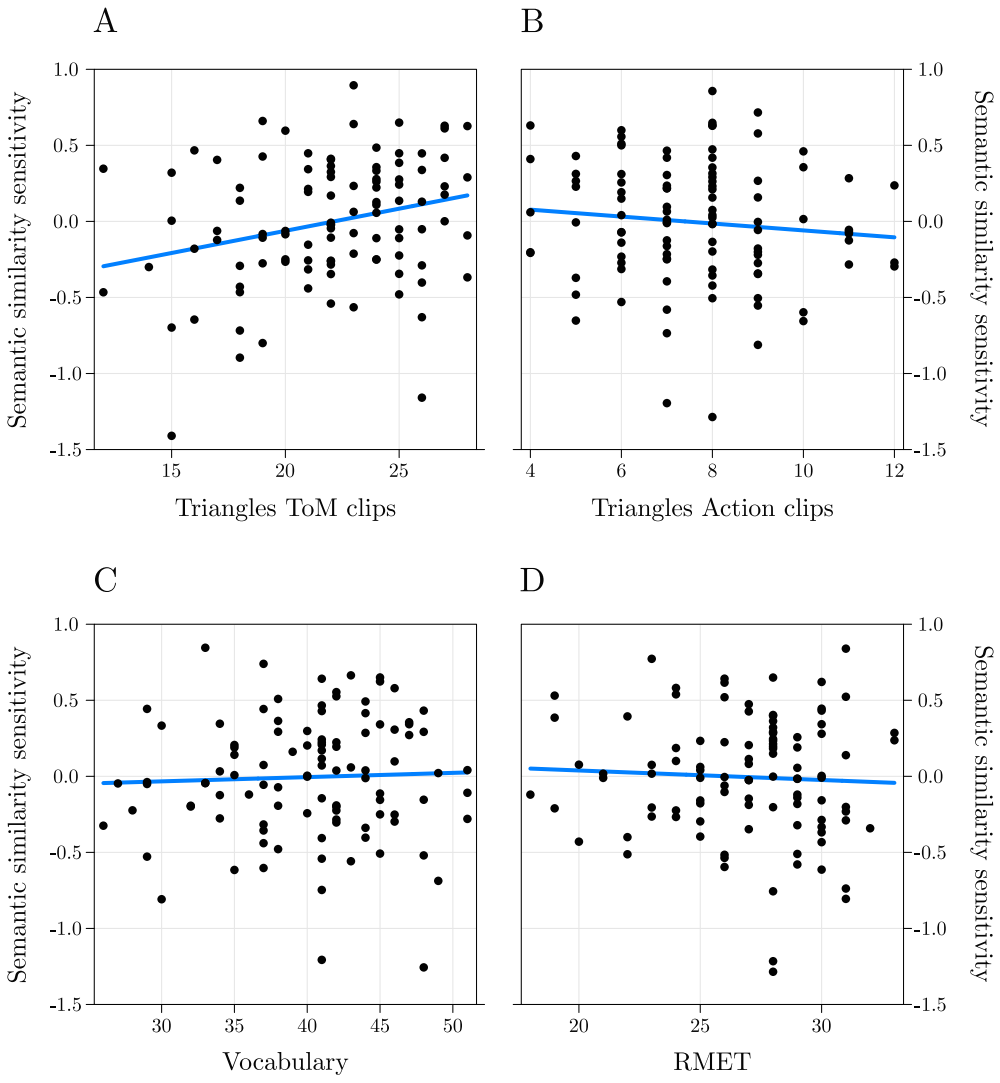


Figure 15. Results from the linear model including the effect of Triangles task ToM clips (A), Triangles task Action clips (B), RMET (C) and Vocabulary (D) on participants' reliance on semantic memory while performing the DRM task. The effect of Triangles TOM was significant, no other effects were found.

3.4.4 Discussion

In the present study, we investigated the link between Theory of Mind (ToM) abilities and semantic processes underlying false memory in the DRM task by taking advantage of recent machine-learning techniques from computational linguistics. In particular, from each participant's performance we extracted an index measuring its reliance on semantic memory processes while falsely recognizing new words in the DRM task, and then, the obtained index was predicted using ToM and vocabulary measures. Our findings indicate that participants' semantic performance is predicted only by the cognitive ToM component over and above the other predictors. That is, the higher participants' cognitive ToM, the higher participants' reliance on semantic memory while falsely recognizing new words the DRM task.

Specifically, it should be noted that the semantic index computed using the distributional semantic model was extracted from participants' responses for new words. This choice was made because, according to seminal theories, semantic processing is involved in both false and veridical recognition, but with different involvement of episodic processes (Brainerd & Reyna, 1998; 2002; Reyna & Brainerd, 1995; Roediger et al., 2001), and previous studies have shown that this semantic measure is reliable in predicting participants' performance for both studied and new words, but in the latter case the effect is higher (Gatti et al., *in press*). Thus, this index provides a general measure of how much each participant's false recognition was predicted by the semantic similarity between the new words showed in the recognition phase and the studied ones.

These findings have potential implications for our understanding the relationship between ToM and semantic memory. These two functions are indeed closely related during the entire lifespan. Within typical children, for example, it has been shown that at increasing age both ToM abilities (Devine & Lecce, 2021) and false semantic memories in the DRM task (e.g., Brainerd, 2013; Brainerd et al., 2008) increase. Here, we provide for the first time evidence that these two abilities are directly related during adulthood.

The dissociation found between cognitive (i.e., Triangles task) and affective (i.e., RMET) ToM indexes is supported by previous developmental, clinical, lesional and neuroimaging studies (e.g., Bottiroli et al., 2016; Corradi-Dell'Acqua et al., 2020; Schlaffke et al., 2015; Stietz et al., 2019) showing that these two measures are not correlated. The positive relationship between cognitive ToM and semantic memory can be interpreted by considering the two main frameworks accounting for the performance of individuals with autism in the DRM task: the Fuzzy Trace Theory (FTT; Reyna & Brainerd, 1995) and the Weak Central Coherence hypothesis (WCC; Frith, 1989; Frith & Happé, 1994; Happé & Frith, 2006). Individuals with autism indeed show impaired ToM and memory performance significantly different from typical adults in the DRM task (e.g., Griego et al., 2019).

According to the FTT, individuals' memory performance is based on two different memory traces: a verbatim one, linked to specific perceptual details of the encoded information, and a gist one, linked to the semantic themes of the information (Brainerd & Reyna, 1992). Specifically, during the lifespan, individuals' memory shifts from being essentially verbatim based to rely on both traces (Brainerd & Reyna, 1992, Reyna & Brainerd, 1991a, Reyna and Brainerd, 1991b), with this notion explaining the relation between false memory in the DRM task and age (e.g., Brainerd, 2013; Brainerd et al., 2008). Following this perspective, the diminished false memory effect in the DRM task observed in individuals with autism can be traced back to an unbalanced reliance on the verbatim trace, at the expenses of the gist trace (for an in-depth discussion see: Miller et al., 2014). Similarly, according to the WCC, individuals with autism are more likely to focus their attention on the local level of the incoming information, without integrate them in a general framework or global level (Frith, 1989; Frith & Happé, 1994). Consistent with this it has been proposed that a deficit in central coherence (i.e., WCC) may induce individuals to focus on details, rather than on the global level, also in other domains, such as memory performance (Smith et al., 2007).

Our results are consistent with both perspectives and integrate previous evidence regarding diminished false memory effect in the DRM task in individuals with autism (Beversdorf et al., 2000; Hillier et al., 2007; Wojcik et al., 2018). Specifically, our findings suggest that both cognitive ToM and semantic memory rely on the same underlying

integrative process. That is, on the one hand, ToM can be seen as the ability to adopt a general and global picture of the surrounding (social and mental) environment, "[a] cohesive interpretative device par excellence: it forces together complex information from totally disparate sources" (Frith, 1989, p.174). On the other hand, the use of semantic memory in the DRM task is related to participants' ability to build upon the global meaning of the words studied and to use it in the subsequent recognition phase (Brainerd & Reyna, 1998; 2002; Reyna & Brainerd, 1995). Based on these inferences, the result of the present study about the relation between ToM and semantic memory may be seen as evidence that both abilities rely on the same cognitive process, that is, one's ability to create a global view, integrating a variety of information.

Several limitations should be acknowledged. Among these, first, in studying individual differences linked to false memory production and semantic memory in general, more ecological paradigms could be adopted, in order to provide more direct evidence from the practical point of view of justice trials and witness evaluation (Schacter & Loftus, 2013). Second, as measures of ToM, participants were required to complete only one task for each ToM component: thus the result should be replicated using different ToM tasks (e.g., Osterhaus et al., 2016) and different DRM task variants (e.g., asking also to freely recall after each list, see: Roediger & McDermott, 1995), in order to verify that it is consistent independently of the ToM task adopted.

Future studies are required in order to replicate the present findings and to provide support to the theoretical explanations proposed. In particular, a question that remains to be addressed in future research is whether the observed relationship among typical individuals, described in the present study, and among individuals with autism, described in the literature, are qualitatively similar. Notwithstanding these limitations, the present research offers some important advancing in the field, demonstrating for the first time in the general population the existence of a positive association between ToM abilities and the tendency to rely on semantic memory.

In conclusion, the present study investigated the role of several ToM components in predicting the semantic performance in the DRM task. Results showed that cognitive ToM

predicted the semantic performance over and above the other predictors. These results are consistent with theories that explained ToM deficits and memory performance across clinical populations, such as individuals with autism, and indicate that a common cognitive process could underlie these two abilities.

4. Neurostimulation evidence

4.1

Cerebellum and semantic memory: a TMS study using the DRM paradigm

The present study has been published in its extended and definitive version on *Cortex*. To cite it:

Gatti, D., Vecchi, T., & Mazzoni, G. (2021). Cerebellum and semantic memory: A TMS study using the DRM paradigm. *Cortex*, 135, 78-91.

4.1.1 Introduction

Traditionally, the cerebellum has been linked to motor functions, but recent evidence suggest that it is also involved in a wide range of cognitive processes (e.g., language, timing, working memory; for reviews: Adamaszek et al., 2016; Baumann et al., 2015; D'Angelo & Casali, 2013; Koziol et al., 2014; Manto et al., 2012; Mariën et al., 2014; Van Overwalle et al., 2020; and for a general discussion see: D'Angelo, 2019). Given the uniformity of cerebellar cortex microstructure (Ito, 1993; Ramnani, 2006; Voogd, & Glickstein, 1998), it has been proposed that the same computational process (e.g., *dysmetria of thought* - Schmahmann, 1991) might underlie cerebellar involvement in both motor and cognitive functions (Schmahmann, 2019). Besides cortical uniformity, cerebellar participation to both motor,

cognitive and affective processes would be granted by extensive cerebro-cerebellar connections (e.g., Buckner, Krienen, Castellanos, Diaz, & Yeo, 2011; Halko, Farzan, Eldaief, Schmahmann, & Pascual-Leone, 2014; Granziera et al., 2009; Guell, et al., 2020; for a review see, Guell & Schmahmann, 2020), which are segregated (Kelly & Strick, 2003; Krienen & Buckner, 2009; Sokolov, Erb, Grodd, & Pavlova, 2014) and thus would allow specific cerebellar areas to participate in specific cognitive functions (Schmahmann, 2019). Consistent with this perspective, lesion studies showed that cognitive and affective symptoms that arise after cerebellar dysfunction follow a similar pattern of abnormality compared to motor symptoms (Guell, Hoche, & Schmahmann, 2015; Hoche, Guell, Sherman, Vangel, & Schmahmann, 2016; Schmahmann, Weilburg, & Sherman, 2007; for a review: Guell, Gabrieli, & Schmahmann, 2018). Similarly, intermittent theta-burst stimulation over two different cerebellar nodes resulted in a similar impact on the temporal complexity of cortical changes (Farzan, Pascual-Leone, Schmahmann, & Halko, 2016).

Following this evidence, recently it was also hypothesized that cerebellar participation in cognition could involve both procedural and declarative memory processes (Vecchi & Gatti, 2020). Consistent with this perspective, on a motor level, cerebellar impairments affect procedural memory (Elyoseph, Mintz, Vakil, Zaltzman, & Gordon, 2020; Molinari et al., 1997; Pascual-Leone et al., 1997; Torriero, Oliveri, Koch, Caltagirone, & Petrosini, 2004) and associative learning (Bracha et al., 2000; Hoffland et al., 2012; McCormick & Thompson, 1984; Monaco, Casellato, Koch, & D'Angelo, 2014). Interestingly, patients with cerebellar lesions also showed impaired semantic associative learning (Timmann et al., 2010), that is, compared with control participants, they were slower in learning associations between colors and numbers and their recognition memory performance was impaired (Drepper, Timmann, Kolb, & Diener, 1999; Timmann et al., 2002, 2004).

Furthermore, on a declarative level, neuroimaging studies reported cerebellar activations during episodic (Andreasen et al., 1995; 1999; Fliessbach, Trautner, Quesada, Elger, & Weber, 2007; Kim, Daselaar, & Cabeza, 2010), semantic (Andreasen et al., 1995; Desmond, Gabrieli, & Glover, 1998) and autobiographic memory tasks (Addis, Moloney, Tippett, Roberts, & Hach, 2016), and neurostimulation studies reported improved episodic memory encoding following cerebellar theta stimulation (Dave, VanHaerents, & Voss, 2020).

However, despite neurostimulation evidence about cerebellar involvement in various semantic and language domains, such as semantic prediction (D'Mello, Turkeltaub, & Stoodley, 2017; Lesage, Morgan, Olson, Meyer, & Miall, 2012; Miall et al., 2016), semantic integration (Gatti, VanVugt, & Vecchi, 2020) or associative priming (Argyropoulos, 2011; Argyropoulos & Muggleton, 2013; Gilligan, & Rafal, 2019), no direct evidence of cerebellar involvement in semantic memory is available.

To test whether the cerebellum is causally involved in semantic memory, we carried out an experiment using TMS. Participants performed the Deese-Roediger-McDermott task (DRM; Deese, 1959; Roediger & McDermott, 1995) while TMS was administered. Specifically, participants were instructed to remember several lists of words in each of the blocks that composed the experimental session and then to perform a recognition task during the administration of TMS. The words that composed each list were associatively related to more non-shown words (named *critical lures*, *weakly related lures*, and *unrelated words*). Previous studies showed that, during the recognition task of the DRM paradigm, participants falsely recognize the critical lures as if they were actually studied (for a review: Gallo, 2010).

The DRM paradigm was chosen because participants' memory performance in this task follows an associative semantic gradient, that is, the higher the semantic relation between a word showed during the recognition task and those studied, the higher the probability that participants recognize it as studied (Gatti, Rinaldi, Marelli, Mazzoni, & Vecchi, in press; Johns, Jones, & Mewhort, 2012; Montefinese, Zannino, & Ambrosini, 2015). TMS was administered over the right posterior cerebellar hemisphere, the coordinate chosen corresponded to the cerebellar loci of activation when correctly recognizing a studied word and falsely recognizing a critical lure (McDermott, Gilmore, Nelson, Watson, & Ojemann, 2017). Our choice to target the right cerebellum was driven by two established evidence in literature. Firstly, verbal and semantic processing tends to be right-lateralized in the cerebellum (Stoodley & Schmahmann, 2009), whereas in the cerebral cortex, it is left-lateralized (Habib, Nyberg, & Tulving, 2003; Tulving, Kapur, Craik, Moscovitch, & Houle, 1994), reflecting crossed cerebro-cerebellar connections (Middleton & Strick, 1994; Palesi et al., 2017). Secondly, associative models (e.g., Anderson, 1983; Collins & Loftus, 1975; Steyvers & Tenenbaum, 2005) proposed that semantic memory is organized by similarity

and co-occurrence in language and evidence has been reported for a role of the right cerebellum in processing words that co-occur in language (e.g., *soap-cleaning*, Argyropoulos, 2011; Argyropoulos & Muggleton, 2013) or generating related verbs for given nouns (e.g., *eat-cake*, Fiez, Petersen, Cheney, & Raichle, 1992; Frings et al., 2006; Gebhart, Petersen, & Thach, 2002; Petersen et al., 1989). Therefore, given the role of associative semantic connections among the words presented in the DRM lists, and with the non-presented lures, we expected right cerebellar TMS to affect participants' discriminability for words more related to those studied without affecting non-related control stimuli (i.e., unrelated words).

4.1.2 Methods

Participants

Thirty-two Italian students (7 males, *M* age = 21.6 years, *SD* = 1.3) participated in the experiment. All participants were fluent Italian speakers, had normal or corrected to normal vision and were naïve to the purpose of the study. Prior to the experiment, each participant filled in a questionnaire (translated and adapted from: Rossi, Hallett, Rossini, & Pascual-Leone, 2011) to evaluate their compatibility with TMS before undergoing the experiment. None of the participants reported neurological problems or history of seizures, none was taking medications that could interfere with neuronal excitability. Written informed consent was obtained from all participants before the experiment. The protocol was approved by the psychological ethical committee of the University of Pavia and participants were treated in accordance with the Declaration of Helsinki.

Stimuli

We adapted the DRM paradigm (Deese, 1959; Roediger & McDermott, 1995), a classical false memories paradigm, to a within participants TMS design. Specifically, participants were instructed to remember several lists of words in each of the blocks that composed the experimental session and then to perform a recognition task during the administration of TMS. The words that composed each list were associatively related to a

non-shown word (called *critical lures*). It has been shown that during both recall and recognition of the studied words, participants tend to falsely recognize the critical lures, although they were never studied (for a review: Gallo, 2010).

We selected 16 lists of words from the normative data for Italian DRM reported by Iacullo and Marucci (2016). Each list was composed by 12 words. Eight lists reported by Iacullo and Marucci (2016) were not used in the present study, some of the words composing those lists were used as control stimuli (see below). In the list *giustuzia* (justice), the studied word *salto* (jump) was removed and replaced with the next word because in our opinion the associative link between the two words was absent.

The two recognition blocks were composed of 72 words each, 32 which had been studied and 40 of which had not. The 32 studied words were those in serial positions 1, 4, 8 and 11 in the studied lists of that block. Of the 40 non-studied words, 8 were the critical lures from the studied lists, 16 were weakly related lures, and 16 were unrelated words. The weakly related lures were the words non-shown in position 13 and 14 of the studied lists. This criterion was violated in only two occasions: the list *giustizia* (justice) had only two extra words (due to the abovementioned exclusion of the word *salto* (jump)); and in the list *freddo* (cold) we excluded the non-shown word *corrente* (stream) because this word was also a studied word in the list *fiume* (river). The unrelated words were those in serial positions 3, 6, 9 and 12 in the non-used lists. This criterion was established arbitrarily. No differences were found between the length of the words that composed each condition or session (all *p*s $> .09$).

Procedure

Participants were seated comfortably at a distance of 60 cm from a 17″ computer monitor. Stimuli were displayed on a computer monitor using Matlab (Mathworks, Inc.) and the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997).

Participants were told that they would take part to a memory experiment. Each participant performed the task once (i.e., one encoding task, one distracting task and one

recognition task), and within the encoding and the recognition task were displayed the stimuli composing the two sets (labeled A and B) (Figure 16). In this way, one participant with randomization A-B and order of sites of stimulation Right Cerebellum-Vertex, in the encoding task studied the words of the set A, then the ones of the set B; in the recognition task, during cerebellar TMS, recognized the words of the set A and then, during vertex stimulation, the words of the set B. Each part of both tasks was performed without interruptions. Sets, lists within each set, TMS sites, order of words during the recognition task and response keys were randomized across participants.

For the encoding task, each trial started with a central fixation cross (presented for 1000 ms) followed by a word of the list (presented for 1500 ms), then the script moved automatically to the next fixation cross. The words that composed the lists were showed in descending forward associative strength (FAS).

For the recognition task, each trial started with a central fixation cross (presented for 3000 ms) followed by a word (until response), participants' response ended the trial and moved to the fixation cross of the next trial. TMS was delivered at the onset of each stimulus. During the recognition task, at the end of the first set, the experimenter moved the coil to the other site of stimulation and the participant continued the task.

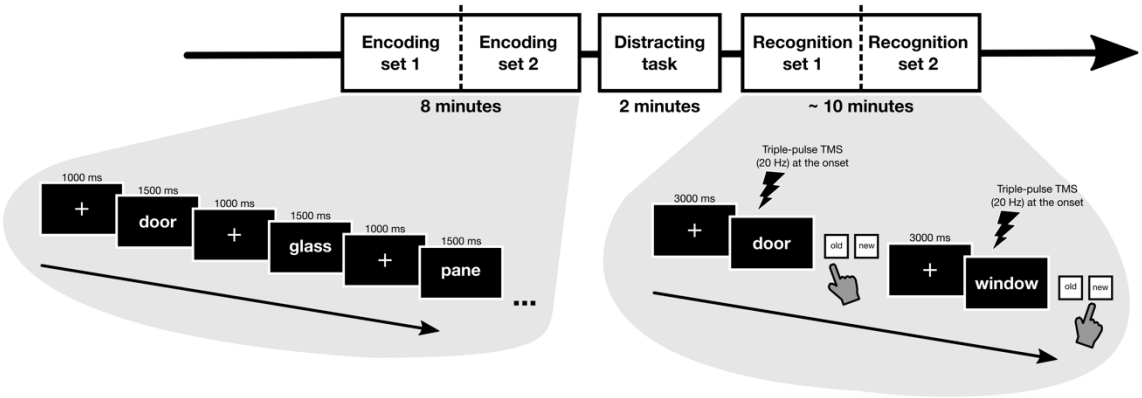


Figure 16. Participants performed the DRM, 20 Hz triple-pulse TMS was delivered at the onset of each word in the recognition task. Participants were stimulated over one of the two sites during the first half of the recognition task and then the coil was moved over the other site (counterbalanced order).

Transcranial Magnetic Stimulation (TMS)

Online neuronavigated TMS was performed using a Magstim Rapid^2 stimulator (Magstim Co., Ltd, Whitland, UK) connected to a 70-mm butterfly coil. At the beginning of each session, single pulse TMS was applied at increasing intensities to determine individual motor threshold (MT). MT is lowest TMS intensity capable of evoking a muscle twitch in the contralateral hand in 5/10 consecutive trials (cf. Hanajima et al., 2007, for methodological details). During the experiment, participants were stimulated at 100% of their MT (*M* TMS intensity delivered: 48.8% of the maximum stimulator output, *SD* = 2.3%). Triple-pulse 20 Hz TMS was delivered at the onset of each stimulus.

TMS was delivered over the right cerebellum and the vertex (control condition). The right cerebellum was localized by means of stereotaxic navigation on individual estimated magnetic resonance images (MRI) obtained fitting a high-resolution MRI template with the participant's scalp model and craniometric points (Softaxic, EMS, Bologna, Italy). Anatomical MNI coordinates used for neuronavigation were x = 32, y = -66, z = -34 for the right cerebellum, corresponding to Crus I/Crus II, cerebellar loci of activation reported in a previous neuroimaging study investigating true and false recognition (McDermott et al., 2017). The vertex was localized as the point falling half the distance between the nasion and the inion. The coil was placed tangentially to the scalp with the handle pointing superiorly during the stimulation over the right cerebellum and with the handle pointing backward during the stimulation over the vertex. Since repetitive TMS over posterior cerebellar areas can induce muscular twitches, prior to the experiment a few TMS pulses over the right cerebellar hemisphere were administered in order to familiarize participants with the skin sensations.

Data analysis

Following previous neurostimulation studies investigating false memories (Díez, Gómez-Ariza, Díez-Álamo, Alonso, & Fernandez, 2017a), we performed non-parametric signal-detection analyses (SDT), computing A' and B"$_D$ values (Donaldson, 1992; Stanislaw & Todorov, 1999). A' provides a measure of discriminability (i.e., the ability to discriminate

the signal - the words actually studied - from the noise) comprised between 0 and 1, with values around 0.5 indicating chance performance and values around 1 good performance. B''_D provides a response bias estimate (comprised between –1 and 1), with values greater than 0 indicating conservative bias (i.e., tendency to answer *no* when uncertain) and lesser than 0 indicating liberal bias (i.e., tendency to answer *yes* when uncertain). For each participant, A' and $B"_D$ values were calculated using the proportion of hits (*old* response when the item was old), correct-rejections (*new* response when the item was new), misses (*new* response when the item was old) and false alarms (*old* response when the item was new).

We first compared general measures of A' and $B"_D$, computed using data for studied words as signal and data for new items (critical lures + weakly related lures + unrelated words) as noise. Then, to assess if stimulation differentially affected participants' performance in discrimination of studied words vs. different types of new items (critical lures vs. weakly related lures vs. unrelated words) we compared A' and $B"_D$ for the types of new stimuli, computed using data for studied words as signal and data for critical lures, weakly related lures or unrelated words as noise. Following these criteria, for each participant, we obtained two general measures of A' and $B"_D$ (right cerebellum vs. vertex) and four measures of A' and $B"_D$ related to the new items (2x3, TMS site * type of new items).

To assess if stimulation affected discriminability or response bias for studied words compared to new items, we estimated two linear mixed models with TMS site (right cerebellum vs. vertex) as fixed effect and intercepts as random coefficients across participants.

Then, to assess if stimulation affected discriminability or response bias for studied words compared with the two types of new items, we estimated two linear model with the area stimulated, the type of new items (critical lures vs. weakly related lures vs. unrelated words) and their interaction as fixed effects and intercepts as random coefficients across participants.

To further investigate the effect of cerebellar TMS on memory performance, we estimated a linear mixed model using the difference between vertex stimulation and

cerebellar TMS (with higher values indicating higher impairment caused by cerebellar TMS) in the three types of new items computed as a continuous variable. The computation the three types of new items computed as a continuous variable was performed based on their semantic relatedness with the studied words (unrelated words = 1; weakly related lures = 2; critical lures = 3). The linear mixed model was estimated with the type of new items (critical lures vs. weakly related lures vs. unrelated words) as continuous predictor and intercepts as random coefficients across participants.

Following the same rationale, we analyzed the raw data from participants' performance on new words (i.e., critical lures, weakly related lures and unrelated words; for a similar approach see: Gatti et al., in press). We ran a mixed logistic model having participants' memory performance (i.e., "new" responses were scored as 0, whereas "old" responses as 1) as dependent variable and subjects and items as random intercepts; TMS site and a semantic index (see below) and their interaction were added as fixed effects.

To compute the semantic index, we extracted semantic similarity values (SSim) for each new word (16 critical lures, 32 weakly related lures and 32 unrelated words) from an Italian distributional semantic model (DSM, for a review on DSMs, see: Günther, Rinaldi & Marelli, 2019) released by Marelli (2017). Then, the semantic index was computed as the frequency-weighted average SSim (for a similar approach see: Marelli & Amenta, 2018) between each word in the recognition phase and each of the 12 words that composed its relative list. The relative semantic similarity indexes used in this Experiment can be reproduced at: http://meshugga.ugent.be/snaut-italian/. Note also that these indexes were subtracted from 1 to transform the values on a proximity scale. For unrelated words we computed the semantic index randomly matching each word with a list. To compute the semantic index, we used the following formula:

$$\text{semantic index} = \frac{\sum_{i=1}^{k} SSim_i \times F_i}{\sum_{i=1}^{k} F_i}$$

where $SSim_i$ refers to the semantic similarity between a new word and each of the i studied word composing its list, while F_i is the frequency of each studied word as extracted from the Italian SUBTLEX (http://crr.ugent.be/subtlex-it/). Following the same rationale, we computed a SSim index for studied words. This semantic index is thought to capture semantic similarity between words (Günther et al., 2019), with higher values indicating higher semantic similarity.

SDT measures were calculated using R-Studio (RStudio Team, 2015) and *psycho* package (Makowski, 2018). All the models were estimated using *lme4* package (Bates, Maechler, Bolker, & Walker, 2015); the graphs reported were obtained using the *effects* package (Fox, 2003; Fox & Weisberg, 2019).

4.1.3 Results

The analyses on general SDT measures revealed that the effect of TMS site was significant for A', $F(1,31) = 6.84$, $p = .01$, indicating that participants' discriminability was lower during cerebellar TMS ($M = .75$, $SE = .02$) compared to vertex stimulation ($M = .79$, $SE = .02$); the effect of TMS site for B''_D was not significant (M right cerebellum $= .16$, $SE = .03$; M vertex $= .16$, $SE = .03$), $F(1,31) = .004$, $p = .94$.

Then, in order to assess if TMS differently affected participants' memory performance or response bias within the different type of new items, we performed two linear mixed models on the A' and on the B''_D of new items. The analysis on A' (Figure 17A) revealed a significant main effect of TMS site, $F(1,155) = 10.31$, $p = .001$. The main effect of type of new items was significant, $F(2,155) = 180.48$, $p < .0001$, showing that participants' discriminability was lower for critical lures compared to both weakly related lures and unrelated words, and for weakly related lures compared to unrelated words (all ps $< .01$, Bonferroni corrected); the interaction TMS site by type of new item was not significant, $F(2,155) = 1.32$, $p = .26$.

The analysis on B''_D revealed a significant main effect of type of new items, $F(2,155) = 157.19$, $p < .0001$, indicating that participants' responses were more conservative for unrelated words compared to both weakly related lures and critical lures, and for weakly

related lures compared to critical lures (all *p*s < .0001, Bonferroni corrected), but not of TMS site, $F(1,155) = .005$, $p = .94$; the interaction TMS site by type of new item was not significant, $F(2,155) = .53$, $p = .58$.

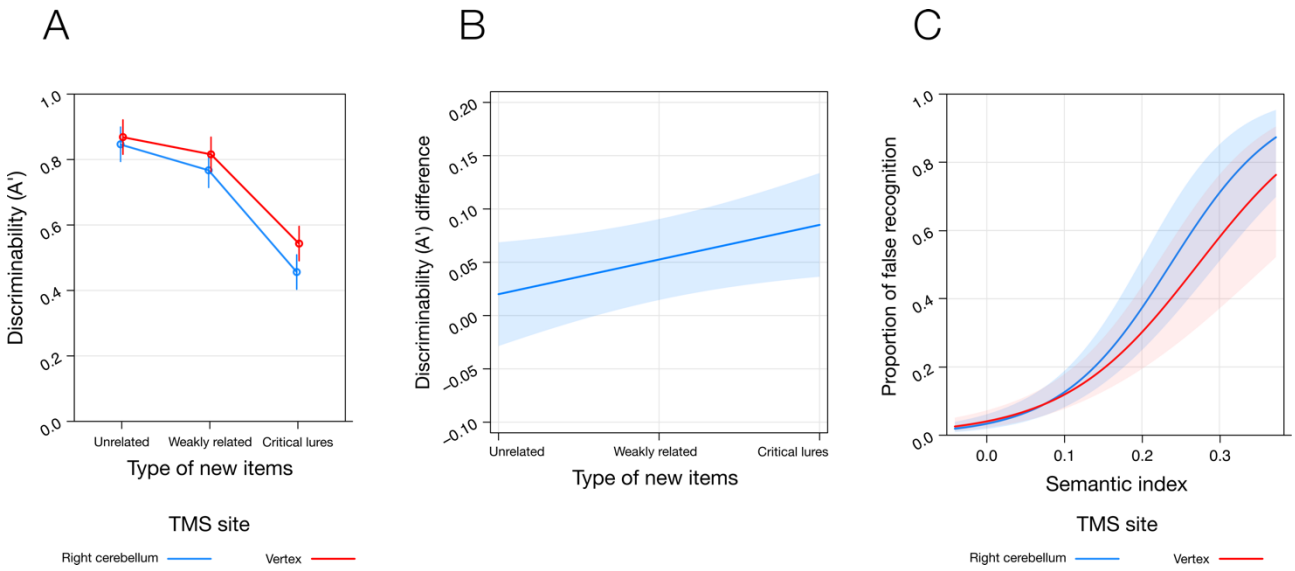


Figure 17: Cerebellar TMS effects followed a semantic gradient with higher impairments occurring for critical lures and lower impairments occurring for unrelated words (A). Results from the linear mixed model illustrating the positive relationship between the type of new items and the effect of cerebellar TMS (B). Results from the mixed logistic model including the interaction between TMS site and the semantic index computed (C). Error bars represent 95% confidence interval.

The analysis on the difference of A' revealed a significant main effect of the type of new items, $F(1,63) = 4.50$, $p = .03$, indicating that, higher was the semantic relatedness between the type of items and the studied words, higher was the difference in discriminability (Figure 17B); thus resulting in greater impairment caused by cerebellar TMS (*M* A' for unrelated words = .02; *M* A' for weakly related lures = .05; *M* A' for critical lures = .08).

Finally, the mixed logistic model on false recognition revealed a significant effect of the semantic index on false recognition, $z = 7.44$, $p < .001$, indicating that the higher was the semantic relation between the new word and the studied words composing its list, the higher the chances of making false memories. The effect of TMS site was not significant, $z = .93$, $p = .35$. Critically, the interaction TMS site by semantic index was significant, $z = -1.97$, $p = .05$. From an inspection of Figure 17C, this significant interaction indicates the

higher the semantic relation between the new word and the studied words composing its list, the higher the false alarm rate during right cerebellar TMS compared to vertex stimulation.

4.1.4 Discussion

In the present study, we investigated the role of the cerebellum in semantic memory retrieval. Participants were asked to encode several words from Italian normative DRM lists (Iacullo & Marucci, 2016) and then to perform a recognition task while TMS was administered over the right cerebellum or over a control site.

We tested the hypothesis that cerebellar involvement in semantic memory follows a semantic gradient of association by including weakly related lures in addition to unrelated words and critical lures. Consistent with our hypotheses, we found that the higher was the semantic association between new and studied words, the higher was the memory impairment caused by the TMS. Cerebellar TMS impaired participants' discriminability (i.e., the ability to correctly discriminate studied words from new words) resulting in higher false alarms rate, that here we interpret as an enhancement of the activation of associative links in semantic memory. This perspective is in line with previous brain stimulation evidence concerning the increase of associative (but not of categorical) priming size after right cerebellar theta-burst stimulation (Argyropoulos, 2011; Argyropoulos & Muggleton, 2013; Gilligan, & Rafal, 2019). It should be noted that the DRM lists that we employed were built by means of associative (and not categorical) relations among the critical lures and the studied words (Iacullo & Marucci, 2016), thus our results corroborate previous evidence about cerebellar involvement in associative semantic processing (Argyropoulos, 2011; Argyropoulos & Muggleton, 2013; Gilligan, & Rafal, 2019).

While it is an established evidence that right cerebellum is active during various linguistic tasks (e.g., Cullum et al., 2019; D'Mello et al., 2017; Moberget, Gullesen, Andersson, Ivry, & Endestad, 2014; Xiang et al., 2003; for a review: Stoodley & Schmahmann, 2009), our study is the first TMS study to report a direct cerebellar involvement in semantic memory retrieval, consistent with previous neuroimaging evidence (Desmond et al., 1998). Previous brain stimulation studies that investigated semantic and

language processing in the cerebellum employed tasks such as semantic prediction task (D'Mello et al., 2017), in which participants were asked to judge the appropriateness of the last word of a sentence based on the context (e.g., two – plus – two – is – *apple* vs. two – plus – two – is – *four*), associative priming tasks (Argyropoulos, 2011; Argyropoulos & Muggleton, 2013; Gilligan, & Rafal, 2019), speech production tasks (Runnqvist et al., 2016) or integrative processes (Gatti et al., 2020). Crucially, none of the previous studies that investigated cognitive processing in the cerebellum employed a task such as the DRM in which participants are asked to encode stimuli in a first phase and then to perform a recognition task. Here, administering TMS over the right cerebellum during the recognition phase of the DRM, we found that cerebellar involvement in semantic memory follows a semantic gradient of association, with higher impairment caused by the TMS occurring for non-presented words associatively more related to the studied ones.

Two main theories have been proposed to explain participants' performance in the DRM paradigm, both describing semantic memory involvement in veridical and false recognition. According to the activation-monitoring framework (AMF) participants' performance would be affected by the semantic relation (i.e., spreading semantic activation) between the words studied during the encoding phase and those seen in the recognition task, and source-monitoring would counter false recognition (Gallo & Roediger, 2002; Roediger et al., 2001). That is, according to the AMF, the critical lure would be hyperactivated by the presentation of the studied words, leading to high levels of false recognitions. However, if participants can successfully distinguish (via source monitoring) between words actually presented and words hyperactivated but not presented, the production of false recognitions decreases (Gallo, 2010). Consistent with this view, it has been shown that inducing a higher monitoring process (i.e., warning participants about the false memory effect in the DRM) increases participants' memory accuracy (Gallo, Roberts, & Seamon, 1997; Westerberg & Marsolek, 2006). On the other hand, according to the fuzzy-trace theory (FTT), during the encoding phase, participants would encode two different memory traces: a verbatim one, linked to perceptive features of the studied words, and a gist one, linked to the semantic content of each list (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). During veridical recognition both traces would be involved, while the gist trace would be specifically

responsible for false recognition (Brainerd, & Reyna, 2002; Reyna & Brainerd, 1995). Within the AMF, the memory impairment occurring during cerebellar TMS would be linked to interference in source-monitoring processes. Consistent with this view, previous clinical and neuroimaging studies reported right cerebellar involvement in error processing and response inhibition (e.g., performance monitoring in verbal working memory, for a review: Peterburs & Desmond, 2016) and source memory (Tamagni et al., 2010). Conversely, within FTT, the memory impairment occurring during cerebellar TMS would be linked to the modulation of the gist trace. Consistent with this second perspective, in a previous neuroimaging study it has been shown that the right cerebellum is involved in the search for responses in semantic memory (Desmond et al., 1998). While some results of the current study, in particular the results of Experiment 1, might be interpreted through either theory, we believe that the FTT theory can account better for the findings that we report here. Firstly, the impairment caused by cerebellar TMS can be traced back to the modulation of the gist trace. This modulation could be direct, implying that cerebellar TMS affects directly the semantic network. However, the modulation can also be indirect, as TMS might also affect the episodic network, degrading the veridical memory trace and forcing participants to rely to a greater extent on the gist trace. Both perspectives are consistent with the observed gradient of semantic modulation during cerebellar TMS and direct evidence has been reported for cerebellar involvement in semantic (Argyropoulos & Muggleton, 2013; Gilligan, & Rafal, 2019) and episodic processes (Dave et al., 2020, but cfr. also: Rami et al., 2003, that found no evidence of cerebellar involvement). However, more studies are needed to disentangle cerebellar participation in memory processes; future studies should address this issue by directly investigating cerebellar involvement in semantic and episodic memory employing different tasks and stimulation techniques or adopting new paradigms (e.g., Cardona, Rodriguez-Fornells, Nye, Rifà-Ros, & Ferreri, 2020). Secondly, the significant relation between degree of association of non-presented items with studied words and degree of impairment in discrimination produced by cerebellar TMS is likely to reflect processing of the gist trace rather than source-monitoring impairment. In case of source-monitoring impairment, we would expect similar effects across the types of new words.

The relation between the association of new words with studied words and the modulation caused by the TMS could also be accounted for by cerebellar involvement in executive processes (Bellebaum, & Daum, 2007; Koziol, Budding, & Chidekel, 2012), since cerebellar activation increases with increasing task demands (Küper et al., 2016; Xiang et al., 2003). In this case, the effect of cerebellar TMS would be modulated by the degree of activation of the cerebellum: higher cerebellar involvement would occur for more difficult conditions (i.e., discriminating studied words from critical lures), resulting in bigger effects. Within this perspective, cerebellar participation to semantic processes could be related to the application of a uniform cerebellar computation across both motor and cognitive domains. As Schmahmann proposed, "in the same way as the cerebellum regulates the rate, force, rhythm, and accuracy of movements, so may it regulate the speed, capacity, consistency, and appropriateness of mental or cognitive processes" (Schmahmann, 1991, p. 1183). Thus, the memory impairment observed during cerebellar stimulation would be caused by TMS perturbation of cognitive coordination related to cerebellar activity. Similarly, following cerebellar lesions, patients show an impairment of the ability to judge facial emotional expressions (Hoche et al., 2016) described as an impairment in the implicit, automatic modulation of emotions (Schmahmann et al., 2007). Here, we propose that the detrimental effect of cerebellar TMS on participants' performances might be linked to an impairment in the implicit, automatic modulation of semantic memory.

However, the mechanisms underlying cerebellar participation to semantic memory in the DRM remain to be fully clarified. Using TMS it has been shown that the cerebellum has an inhibitory effect on motor areas (Ugawa, Uesaka, Terao, Hanajima, & Kanazawa, 1995), thus it is also possible that the cerebellum indirectly modulates the activity of other brain areas linked to semantic processing, such as temporal areas (Chadwick et al., 2016; for a review on cerebral cortical activity following non-invasive cerebellar stimulation: Fernandez et al., 2020). Alternatively, it is possible that the cerebellum is directly integrated within the network involved in semantic memory. Previous neurostimulation studies employing the DRM paradigm that administered transcranial direct current stimulation (tDCS) or TMS during or before the encoding phase found reduced false recognition following temporal stimulation (Boggio et al., 2009; Díez et al., 2017a; Gallate, Chi, Ellwood, & Snyder, 2009)

and reduced false recall after medial prefrontal stimulation (Berkers et al., 2017). On the other hand, the studies that administered tDCS during the recognition phase found increased false recognition following parietal stimulation (Pergolizzi, & Chua, 2015) and no effects following temporal stimulation (Díez, Gómez-Ariza, Díez-Álamo, Alonso, & Fernandez, 2017b). The cerebellum is functionally connected to frontal, parietal and temporal areas (e.g., Allen et al., 2005; Habas et al., 2009; Pascual et al., 2015) and may interact with this set of areas during semantic associative processing contributing at different points in time or performing different computations at the same time.

As stated above, cerebellar interaction with frontal and temporo-parietal areas during memory processes could be similar to that observed for motor domains (i.e., universal cerebellar transform, for a review see, Guell et al., 2019). A relevant model that accounts for the interaction among frontal lobes, temporo-parietal areas and the cerebellum during motor and cognitive functions, and that complements this perspective, has been proposed by Ito (2008). Frontal cortices are thought to act as a controller (i.e., executive functions) and then to perform functions such as conscious control of actions and thought; controlled objects, named mental models (e.g., long-term memories, see: Johnson-Laird, 1983), would be stored in temporo-parietal areas, while the cerebellum would represent both inverse and forward internal models that would allow the manipulation of the objects (Ito, 2008; and for a review on inverse and forward models in the cerebellum: Ishikawa, Tomatsu, Izawa, & Kakei, 2016). Previous studies showed that frontal perturbation affects participants' performance during recall but not recognition (Berkers et al., 2017), while temporo-parietal stimulation affects participants' performance during (false) recognition (Boggio et al., 2009; Diez et al., 2017a; Gallate et al., 2009; Pergolizzi & Chua, 2015). Here, we integrate previous evidence reporting a recognition memory modulation during cerebellar stimulation. Consistent with Ito's model, we can interpret the results reported by Berkers and colleagues (2017) as reflecting higher frontal executive involvement in free recall compared to recognition; similarly, the effect reported after temporo-parietal stimulation would reflect the modulation of semantic long-term memories stored. Finally, consistent with Ito's model, we interpret the effect of cerebellar TMS reported in the present study as a modulation occurring during mental models manipulation in the cerebellum and efforts during retrieval. This manipulation could

consist in the coordination of the mental simulation (i.e., the memory process) required by the task when judging if a word was previously presented. This is in line with what proposed by Koziol, Budding and Chidekel (2012, p. 519): "as the cerebellum informs motor regions about the most efficient way of executing behaviors, it instructs the prefrontal cortex how to manipulate ideas for problem solving". However, other studies are needed to shed light on the interaction among these areas. Future studies might systematically manipulate timing and areas targeted by the stimulation in order to better understand the interplay during the various phases (i.e., encoding, consolidation, recognition) and the role played by the different areas comprised in the network during retrieval itself.

Another possible interpretation of our results is related to the link between memory and prediction (for a review: Vecchi & Gatti, 2020). Adaptive perspectives on human memory proposed that the purpose of storage and retrieval is not to allow to remember the past, but rather to predict what is going to happen (Klein, 2013; 2014). Thus, memory could also be studied as a predictive function by investigating areas linked to predictive functioning (Bubic, Von Cramon, & Schubotz, 2010), such as the cerebellum (Sokolov, Miall, & Ivry, 2017). Indeed, several studies showed that memory and prospection share common neural substrates (Addis, Wong, & Schacter, 2007; for a review: Schacter et al., 2012) and that the cerebellum is comprised within the network active during both retrieval and prediction (Addis, Pan, Vu, Laiser, & Schacter 2009; Thakral, Benoit, & Schacter, 2017). Similarly, previous studies showed that the right cerebellum is involved in semantic prediction (D'Mello et al., 2017; Moberget et al., 2014) and our results suggest that the right cerebellum is involved in the processing of associations during semantic memory retrieval. Cerebellar involvement in this predictive framework of memory could follow the above-mentioned dynamics about internal model manipulation (Ito, 2008) or be more radical due to its major role in predictive cognition (Sokolov et al., 2017). However, despite the evidence about semantic processing and right cerebellum, little evidence is available about episodic processes (Dave et al., 2020; Rami et al., 2003).

Finally, a few limitations need to be acknowledged. Firstly, despite the MNI coordinate we chose was specific for right Crus I/Crus II and the magnetic field generated by a figure-of-eight coil is more focal than other TMS devices, it is likely that the stimulation

spread over a larger part of the right cerebellar posterior lobule. Cerebellar structures are deep and the more TMS signal goes in depth, the less focal it is (Deng, Lisanby, & Peterchev, 2013), making it impossible to directly link our results to the involvement of one specific cerebellar lobule. Secondly, the parameters adopted during the stimulation could not be tailored on the individual differences in participants' scalp and skull morphology (Fox, Liu, & Pascual-Leone, 2013; Stokes et al., 2007). Similarly, although neuronavigated TMS with estimated MRIs is widely employed in cognitive neuroscience (Carducci & Brusco, 2012) individual differences in brain morphology could not be fully accounted. Thirdly, the semantic gradient of new words employed in the present study is fairly limited; in other words, the distinction between unrelated words, weakly related lures and critical lures may not reflect the complex and continuous structure of semantic memory (Jones, Willits, & Dennis, 2015). Future studies might address this issue by modulating a priori semantic relatedness between new and studied words through distributional semantic models (Günther et al., 2019), thus allowing more structured analyses with semantic relatedness as continuous predictor.

In conclusion, our findings indicate that the right cerebellum is causally involved in the retrieval of semantic associations. These results are consistent with theories that proposed cerebellar participation to cognitive functions via internal models manipulation, as well with recent perspectives about cerebellar involvement in memory and predictive cognition.

5. Conclusions

We started from a main question, asking *why do we remember?* and arguing that human memory is not actually a memory system, but rather a predictive system adaptively shaped (for a complete discussion: Vecchi & Gatti, 2020). This perspective is supported by a large number of studies showing both on a cerebral and a cognitive level, that memory traces undergo transformations that render the memory inaccurate and therefore make a testimony that is based on it unreliable (Schacter & Loftus, 2013). On the other hand, there are conditions in which memory is not only durable but also extraordinarily accurate (Hirst & Phelps, 2016; Julian et al., 2009; Luminet, 2009; Wright, 2009). Interestingly, these latter conditions happen as a consequence of abnormal events and in more extreme cases constitute maladaptive phenomena in the daily life of individuals. Thus, if the answer is not in the past, it should be in the future.

Indeed, predictive memory processes – and not purely retrieval processes – are consistent with a large number of characteristics of human memory, from the ease of forgetting, passing for the update of memory, to false memory in general.

For example, reconsolidation processes (for a review of the experimental evidence, see Lee, Nader, & Schiller, 2017), which cause the alteration of a reactivated memory as a result of a pharmacological or behavioral intervention, are "ideally placed to enable memories to be updated with new information" (Lee et al., 2017, p. 532; but see also: Agren, 2014; Exton-McGuinness, Lee, & Reichelt, 2015; Nader & Hardt, 2009). Indeed, following the reactivation of a stored memory, the original trace can be modified, even radically, leading to incorporation of new material, updating the original memory (Lee et al., 2017).

The purpose of this kind of transformations that memory continuously undergoes would therefore be to enable the updating of memory traces at the cost of their accuracy, thereby ensuring that they maintain a certain relevance in an environment that constantly changes (Dudai, 2004, 2006; Lee, 2009; Sara, 2000; Schiller & Phelps, 2011; Tronson & Taylor, 2007). The relevance of the memory trace is obviously established by how adaptive it is, that is, the relevance is based on its future usefulness in a predictive phase (e.g., Klein,

2013), and therefore the memory would transform itself so that it can yield a more precise and up-to-date prediction.

This framework is particularly desirable within the topic of memory distortion and false memory. Indeed, our next question was *does the production of false memories have an adaptive basis, or does it reflect only a maladaptive aspect of memory?*

The findings reported here can help to further answer this question. In Study 1 we showed that participants' memory performance for both studied and new items follows a continuous semantic gradient, with a higher number of "remember" judgements occurring for items with a higher semantic similarity with the studied ones. In Study 2 we showed that semantic memory plays a role in participants' performance even when correctly rejecting a new item, with increased conflict in the choice for new items with higher semantic similarity with the studied ones. Then, in Study 3 and Study 4 we adopted an individual differences approach and showed that participants' episodic and semantic memory scores differently predict false memory, as well as that participants' reliance on semantic memory when falsely recognizing new words is predicted by theory of mind indexes. Finally, in Study 5 we showed that the cerebellum is causally involved in semantic processes and that cerebellar perturbation through TMS can over-activate semantic trace and thus increase the number of false recognitions.

False memory, indeed, cannot be simply considered as a misbehavior of a system entirely devoted to retrieval processes. False memory should be considered as a product of a healthy predictive system. That is, human memory is essentially a semantically-based system (Vecchi & Gatti, 2020), in which a large number of events are devoted to the construction of general-level memories that can be used in the present. Semantic memory exerts an extensive influence on our behavior, as showed in Study 1 and Study 2. Additionally, Study 3 and Study 4 showed that participants with better semantic memory and better theory of mind tend to commit a higher number of false recognitions. Both these functions – semantic memory and theory of mind – are highly adaptive, since both allow for a better adaptation to the environment. This is also indirectly supported by evidence showing that individuals

with autism or poor mnestic abilities tend to commit a lower number of false recognitions as compared with typical individuals (Wojcik et al., 2018).

However, it should be noted that increased false recognition has also been observed in atypical individuals, such those with confabulation (Ciaramelli et al., 2006), and thus does not automatically constitute an adaptive advantage. That is, as stated above, the relevance of the memory trace is based on its future usefulness in a predictive phase (e.g., Klein, 2013). Indeed, also false memory should be somehow "controlled" in order to be useful and, consistent with this, in Study 5 we showed that cerebellar perturbation interferes with this control system and cause an increased number of false memories.

Overall, the studies presented in this Thesis point to the need to build a more global view of memory and of memory ultimate function itself. Indeed, the classic view that conceives memory as a mere storage system does not seem to take into account actual memory characteristics. Human memory works through updating and transforming information so that it can form the basis for a better prediction system. Accuracy is not the ultimate goal of the system and, to some extent, a precise memory may produce maladaptive functioning. We remember worse so that we can predict better.

Mi sveglierò, starò su da letto, i monti negli occhi, il vento nell'orecchio, i cani nell'aia [...], accendere il fuoco, faccende e affanni. Bella la mia gente: la pena che ci tormenta è come avere una brace nel cuore, un tizzone stretto in mano e attorno, tutt'in tondo, ciarlatani e baccano. E attorno, tutt'in tondo, ciarlatani e baccano.

GIOVANNI LINDO FERRETTI

6. References

Abell, F., Happe, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1-16.

Adamaszek, M., D'Agata, F., Ferrucci, R., Habas, C., Keulen, S., Kirkby, K.C., ... & Orsi, L. (2017). Consensus paper: cerebellum and emotion. *The Cerebellum*, 16, 552-576.

Addis, D.R., Moloney, E.E.J., Tippett, L.J., Roberts, R.P., & Hach, S. (2016). Characterizing cerebellar activity during autobiographical memory retrieval: ALE and functional connectivity investigations. *Neuropsychologia*, 90, 80-93.

Addis, D.R., Pan, L., Vu, M.A., Laiser, N., & Schacter, D.L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia*, 47, 2222-2238.

Addis, D.R., Wong, A.T. & Schacter, D.L. (2007). Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, 45, 1363-1377.

Agren, T. (2014). Human reconsolidation: a reactivation and update. Brain research bulletin, 105, 70-82.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the Second International Symposium on InformationTheory* (pp. 267-281). Budapest: Akademiai Kiado.

Allen, G., McColl, R., Barnard, H., Ringe, W.K., Fleckenstein, J., & Cullum, C.M. (2005). Magnetic resonance imaging of cerebellar-prefrontal and cerebellar–parietal functional connectivity. *NeuroImage*, 28(1), 39-48.

Anderson, J.R. (1983). Retrieval of information from long-term memory. *Science*, 220(4592), 25-30.

Anderson, J.R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14(3), 471–517.

Andreasen, N.C., O'Leary, D.S., Cizadlo, T., Arndt, S., Rezai, K., Watkins, G.L., ... Hichwa, R.D. (1995). Remembering the past: two facets of episodic memory explored with positron emission tomography. *American Journal of Psychiatry*, 152, 1576-1585.

Andreasen, N.C., O'Leary, D.S., Paradiso, S., Cizadlo, T., Arndt, S., Watkins, G.L., ... Hichwa, R.D. (1999). The cerebellum plays a role in conscious episodic memory retrieval. *Human Brain Mapping*, 8, 226-234.

Argyropoulos, G.P. (2011). Cerebellar theta-burst stimulation selectively enhances lexical associative priming. *The Cerebellum*, 10, 540-550.

Argyropoulos, G.P., & Muggleton, N.G. (2013). Effects of cerebellar stimulation on processing semantic associations. *The Cerebellum*, 12(1), 83-96.

Arndt, J. (2012). The influence of forward and backward associative strength on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(3), 747.

Baayen, R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baddeley, A., Eysenck, M.W., & Anderson, M.C. (2009). *Memory*. Psychology Press.

Balota, D.A., Cortese, M.J., Duchek, J.M., Adams, D., Roediger, H.L., McDermott, K.B., & Yerys, B.E. (1999). Veridical and false memories in healthy older adults and in dementia of the Alzheimer's type. *Cognitive Neuropsychology*, 16(3-5), 361-384.

Barkow, J.H., Cosmides, L., & Tooby, J. (Eds.). (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., & Wheelwright, S. (2003). The systemizing quotient: an investigation of adults with Asperger syndrome or high–functioning autism, and normal sex differences. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1430), 361-374.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241-251.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36, 673–721.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 238–247). Stroudsburg, PA: Association for Computational Linguistics.

Bartlett, F.C. (1932). *Remembering*. Cambridge University Press.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Baumann, O., Borra, R.J., Bower, J.M., Cullen, K.E., Habas, C., Ivry, R.B., ... & Paulin, M.G. (2015). Consensus paper: the role of the cerebellum in perceptual processes. *The Cerebellum*, 14, 197-220.

Baym, C.L., & Gonsalves, B.D. (2010). Comparison of neural activity that leads to true memories, false memories, and forgetting: An fMRI study of the misinformation effect. *Cognitive, Affective, & Behavioral Neuroscience*, 10(3), 339–348.

Bellebaum, C., & Daum, I. (2007). Cerebellar involvement in executive control. *The Cerebellum*, 6(3), 184-192.

Berkers, R.M., van der Linden, M., de Almeida, R.F., Müller, N.C., Bovy, L., Dresler, M., ... & Fernandez, G. (2017). Transient medial prefrontal perturbation reduces false memory formation. *Cortex*, 88, 42-52.

Berkowitz, S.R., Laney, C., Morris, E.K., Garry, M., & Loftus, E.F. (2008). Pluto behaving badly: False beliefs and their consequences. *The American Journal of Psychology*, 121(4), 643–660.

Bernstein, D.M., Laney, C., Morris, E., & Loftus, E.F. (2005b). False memories about food can lead to food avoidance. *Social Cognition*, 23(1), 11–34.

Bernstein, D.M., Laney, C., Morris, E.K., & Loftus, E.F. (2005a). False beliefs about fattening foods can have healthy consequences. *Proceedings of the National Academy of Sciences*, 102(39), 13724–13731.

Beversdorf, D. Q., Smith, B.W., Crucian, G.P., Anderson, J.M., Keillor, J.M., Barrett, A.M., Hughes, J.D., Felopulos, G.J., Bauman, M.L., Nadeau, S.E., & Heilman, K.M. (2000). Increased discrimination of "false memories" in autism spectrum disorder. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(15), 8734-8737.

Beversdorf, D.Q., Smith, B.W., Crucian, G.P., Anderson, J.M., Keillor, J.M., Barrett, A. M., ... & Heilman, K.M. (2000). Increased discrimination of "false memories" in autism spectrum disorder. *Proceedings of the National Academy of Sciences*, 97(15), 8734-8737.

Boggio, P.S., Fregni, F., Valasek, C., Ellwood, S., Chi, R., Gallate, J., ... & Snyder, A. (2009). Temporal lobe cortical electrical stimulation during the encoding and retrieval phase reduces false memories. *PloS One*, 4(3).

Bottiroli, S., Cavallini, E., Ceccato, I., Vecchi, T., & Lecce, S. (2016). Theory of Mind in aging: Comparing cognitive and affective components in the faux pas test. *Archives of Gerontology and Geriatrics*, *62*, 152-162.

Bracha, V., Zhao, L., Irwin, K.B., & Bloedel, J.R. (2000). The human cerebellum and associative learning: Dissociation between the acquisition, retention and extinction of conditioned eyeblinks. *Brain Research*, 860(1–2), 87–94.

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

Brainerd, C.J. (2013). Developmental reversals in false memory: A new look at the reliability of children's evidence. *Current Directions in Psychological Science*, *22*(5), 335-341.

Brainerd, C.J., & Reyna, V. F. (1992). Explaining "memory free" reasoning. *Psychological Science*, *3*(6), 332-339.

Brainerd, C.J., & Reyna, V.F. (1998). Fuzzy-Trace Theory and Children's False Memories. *Journal of Experimental Child Psychology*, *71*(2), 82-129.

Brainerd, C.J., & Reyna, V.F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*(5), 164-169.

Brainerd, C.J., Chang, M., & Bialer, D.M. (2020). From association to gist. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Brainerd, C.J., & Wright, R. (2005). Forward association, backward association, and the false-memory illusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3), 554.

Brainerd, C.J., Reyna, V.F., & Ceci, S.J. (2008). Developmental reversals in false memory: A review of data and theory. *Psychological Bulletin*, 134(3), 343.

Brainerd, C.J., Yang, Y., Reyna, V.F., Howe, M.L., & Mills, B.A. (2008). Semantic processing in "associative" false memory. *Psychonomic Bulletin & Review*, 15(6), 1035-1053.

Braun, K.A., Ellis, R., & Loftus, E.F. (2002). Make my memory: How advertising can change our memories of the past. *Psychology & Marketing*, 19(1), 1–23.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. Cognitive psychology, 13(2), 207-230.

Brüne, M. (2005). "Theory of mind" in schizophrenia: a review of the literature. *Schizophrenia Bulletin*, *31*(1), 21-42.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal dis- tributional semantics. *Journal of Artificial Intelligence Research*, 49, 1–47.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977-990.

Bubic, A., Von Cramon, D.Y., & Schubotz, R I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4, 25.

Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., & Yeo, B.T. (2011) The organization of the human cerebellum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106 (2011), 2322-2345.

Bullinaria, J.A., & Levy, J.P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39, 510–526.

Cacioppo, J.T., & Petty, R.E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42, 116–131.

Calvillo, D.P., & Parong, J.A. (2016). The misinformation effect is unrelated to the DRM effect with and without a DRM warning. *Memory*, 24(3), 324-333.

Cann, D.R., McRae, K., & Katz, A.N. (2011). False recall in the Deese–Roediger–McDermott paradigm: The roles of gist and associative strength. *Quarterly Journal of Experimental Psychology*, 64(8), 1515-1542.

Cardona, G., Rodriguez-Fornells, A., Nye, H., Rifà-Ros, X., & Ferreri, L. (2020). The impact of musical pleasure and musical hedonia on verbal episodic memory. *Scientific Reports*, 10(1), 1-13.

Carducci, F., & Brusco, R. (2012). Accuracy of an individualized MR-based head model for navigated brain stimulation. *Psychiatry Research - Neuroimaging*, 203(1), 105-108.

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain*, *125*(8), 1839-1849.

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, *12*(3), 314-325.

Chadwick, M.J., Anjum, R.S., Kumaran, D., Schacter, DL., Spiers, H.J., & Hassabis, D. (2016). Semantic representations in the temporal pole predict false memories. *Proceedings of the National Academy of Sciences*, 113(36), 10180-10185.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., & Blei, D.M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288-296).

Chang, M., & Brainerd, C.J. (2021). Semantic and phonological false memory: A review of theory and data. *Journal of Memory and Language*, 119, 104210.

Chen, N.F., Lo, C.M., Liu, T.L., & Cheng, S.K. (2016). Source memory performance is modulated by transcranial direct current stimulation over the left posterior parietal cortex. *NeuroImage*, 139, 462-469.

Ciaramelli, E., Ghetti, S., Frattarelli, M., & Làdavas, E. (2006). When true memory availability promotes false memory: evidence from confabulating patients. *Neuropsychologia*, 44(10), 1866-1877.

Coane, J.H., McBride, D.M., Termonen, M.L., & Cutting, J.C. (2016). Categorical and associative relations increase false memory relative to purely associative relations. *Memory & Cognition*, 44(1), 37-49.

Collins, A.M., & Loftus, E.F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407-428.

Collins, A.M., & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 240–247.

Corradi-Dell'Acqua, C., Ronchi, R., Thomasson, M., Bernati, T., Saj, A., & Vuilleumier, P. (2020). Deficits in cognitive and affective theory of mind relate to dissociated lesion patterns in prefrontal and insular cortex. *Cortex*, *128*, 218-233.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

Cullum, A., Hodgetts, W.E., Milburn, T.F., & Cummine, J. (2019). Cerebellar Activation During Reading Tasks: Exploring the Dichotomy Between Motor vs. Language Functions in Adults of Varying Reading Proficiency. *Cerebellum*, 1-17.

D'Angelo, E. (2019). The cerebellum gets social. *Science*, 363(6424), 229-229.

D'Angelo, E., & Casali, S. (2013). Seeking a unified framework for cerebellar function and dysfunction: from circuit operations to cognition. *Frontiers in Neural Circuits*, 6, 116.

D'Mello, A.M., Turkeltaub, P.E., & Stoodley, C.J. (2017). Cerebellar tDCS Modulates Neural Circuits during Semantic Prediction: A Combined tDCS-fMRI Study. *Journal of Neuroscience*, 37(6), 1604-1613.

Dave, S., VanHaerents, S., & Voss, J.L. (2020). Cerebellar theta and beta noninvasive stimulation rhythms differentially influence episodic memory versus semantic prediction. *Journal of Neuroscience*, 40(38), 7300-7310.

Davis, C.J. (2001). *The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition* (Doctoral dissertation, ProQuest Information & Learning).

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17.

Del Sette, P., Bambini, V., Bischetti, L., & Lecce, S. (2020). Longitudinal associations between theory of mind and metaphor understanding during middle childhood. *Cognitive Development*, *56*, 100958.

Del Sette, P., Ronchi, L., Bambini, V., & Lecce, S. (2021). Longitudinal associations between metaphor understanding and peer relationships in middle childhood. *Infant and Child Development*, e2232.

Della Rosa, P.A., Catricalà, E., Vigliocco, G., & Cappa, S.F. (2010). Beyond the abstract—concrete dichotomy: mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, 42, 1042-1048.

Deng, Z.D., Lisanby, S.H., & Peterchev, A.V. (2013). Electric field depth–focality tradeoff in transcranial magnetic stimulation: simulation comparison of 50 coil designs. *Brain Stimulation*, 6(1), 1-13.

Desmond, J.E., Gabrieli, J.D., & Glover, G.H. (1998). Dissociation of frontal and cerebellar activity in a cognitive task: evidence for a distinction between selection and search. *NeuroImage*, 7, 368-376.

Devine, R. T., & Lecce, S. (Eds.). (2021). *Theory of Mind in Middle Childhood and Adolescence: Integrating Multiple Perspectives*. Routledge.

Dewhurst, S.A., & Robinson, C.A. (2004). False memories in children: Evidence for a shift from phonological to semantic associations. *Psychological Science*, 15(11), 782-786.

Dewhurst, S.A., Thorley, C., Hammond, E.R., & Ormerod, T.C. (2011). Convergent, but not divergent, thinking predicts susceptibility to associative memory illusions. *Personality and Individual Differences*, 51(1), 73-76.

Díez, E., Gómez-Ariza, C.J., Díez-Álamo, A.M., Alonso, M.A., & Fernandez, A. (2017a). The processing of semantic relatedness in the brain: Evidence from associative and categorical false recognition effects following transcranial direct current stimulation of the left anterior temporal lobe. *Cortex*, 93, 133-145.

Díez, E., Gómez-Ariza, C.J., Díez-Álamo, A.M., Alonso, M.A., & Fernandez, A. (2017b). Encoding/retrieval dissociation of false recognition with transcranial direct current stimulation (tDCS) of the left temporal lobe. *Brain Stimulation*, 10(2), 372.

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General*, 121, 275-277.

Drepper, J., Timmann, D., Kolb, F.P., & Diener, H.C. (1999). Non-motor associative learning in patients with isolated degenerative cerebellar disease. *Brain*, 122(1), 87–97.

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, 55, 51–86.

Dudai, Y. (2006). Reconsolidation: The advantage of being refocused. *Current Opinion in Neurobiology*, 16, 174–178.

Dudai, Y., Karni, A., & Born, J. (2015). The consolidation and transformation of memory. *Neuron*, 88(1), 20–32.

Elyoseph, Z., Mintz, M., Vakil, E., Zaltzman, R., & Gordon, C.R. (2020). Selective Procedural Memory Impairment but Preserved Declarative Memory in Spinocerebellar Ataxia Type 3. *Cerebellum*, 1-9.

Exton-McGuinness, M.T., Lee, J.L., & Reichelt, A. C. (2015). Updating memories—the role of prediction errors in memory reconsolidation. *Behavioural Brain Research*, 278, 375-384.

Farzan, F., Pascual-Leone, A., Schmahmann, J.D., & Halko, M. (2016). Enhancing the temporal complexity of distributed brain networks with patterned cerebellar stimulation. *Scientific Reports*, 6, 23599.

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Fernandez, L., Rogasch, N.C., Do, M., Clark, G., Major, B.P., Teo, W.P., ... & Enticott, P. G. (2020). Cerebral Cortical Activity Following Non-invasive Cerebellar Stimulation – a Systematic Review of Combined TMS and EEG Studies. *Cerebellum*, 1-27.

Ferrand, L., & New, B. (2003). Associative and semantic priming in the mental lexicon. In P. Bonin (Ed.), *The mental lexicon: Some words to talk about words* (pp. 26– 43). New York, NY: Nova Science.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th web as corpus workshop* (WAC-4) *can we beat Google* (pp. 47–54).

Ferrari, C., Cattaneo, Z., Oldrati, V., Casiraghi, L., Castelli, F., D'Angelo, E., & Vecchi, T. (2018). TMS Over the Cerebellum Interferes with Short-term Memory of Visual Sequences. *Scientific reports*, 8.

Fiez, J.A., Petersen, S.E., Cheney, M.K., & Raichle, M.E. (1992). Impaired non-motor learning and error detection associated with cerebellar damage: A single case study. *Brain*, 115, 155-178.

Finley, J.R., Sungkhasettee, V.W., Roediger, H.L., & Balota, D.A. (2017). Relative contributions of semantic and phonological associates to over-additive false recall in hybrid DRM lists. *Journal of Memory and Language*, 93, 154-168.

Fliessbach, K., Trautner, P., Quesada, C.M., Elger, C.E., & Weber, B. (2007). Cerebellar contributions to episodic memory encoding as revealed by fMRI. *NeuroImage*, 35, 1330-1337.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, *8*(15), 1-27.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd Edition. Thousand Oaks, CA.

Fox, M.D., Liu, H., & Pascual-Leone, A. (2013). Identification of reproducible individualized targets for treatment of depression with TMS based on intrinsic connectivity. *NeuroImage*, 66, 151-160.

Freeman, J.B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, 27(5), 315-323.

Freeman, J.B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, 42(1), 226-241.

Freeman, J.B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, 118(2), 247.

Freeman, J.B., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, 2, 59.

Freeman, J.B., Pauker, K., & Sanchez, D.T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science*, 27(4), 502-517.

Frings, M., Dimitrova, A., Schorn, C.F., Elles, H.G., Hein-Kropp, C., Gizewski, E.R., ... & Timmann, D. (2006). Cerebellar involvement in verb generation: an fMRI study. *Neuroscience Letters*, 409, 19-23.

Frith, C.D. (2004). Schizophrenia and theory of mind. *Psychological medicine*, *34*(3), 385-389.

Frith, U. (1989). Autism: Explaining the enigma. *British journal of developmental psychology*, *3*, 465-468.

Frith, U., & Happé, F. (1994). Autism: beyond "theory of mind". *Cognition*, *50*(1-3), 115-132.

Gallate, J., Chi, R., Ellwood, S., & Snyder, A. (2009). Reducing false memories by magnetic pulse stimulation. *Neuroscience Letters*, 449(3), 151-154.

Gallo, D.A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory and Cognition*, *38*(7), 833–848.

Gallo, D.A. (2004). Using recall to reduce false recognition: diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1), 120.

Gallo, D.A. (2004). Using recall to reduce false recognition: Diagnos tic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 30, 120-128.

Gallo, D.A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, 38(7), 833-848.

Gallo, D.A. (2013). *Associative illusions of memory: False memory research in DRM and related tasks*. Psychology Press.

Gallo, D.A., & Roediger, H. L. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language*, 47(3), 469-497.

Gallo, D.A., Korthauer, L.E., McDonough, I.M., Teshale, S., & Johnson, E.L. (2011). Age-related positivity effects and autobiographical memory detail: Evidence from a past/future source memory task. *Memory*, 19(6), 641-652.

Gallo, D.A., Roberts, M.J., & Seamon, J.G. (1997). Remembering words not presented in lists: Can we avoid creating false memories?. *Psychonomic Bulletin & Review*, 4(2), 271-276.

Garoff, R.J., Slotnick, S.D., & Schacter, D.L. (2005). The neural origins of specific and general memory: The role of the fusiform cortex. *Neuropsychologia*, 43(6), 847–859.

Gatti, D., Marelli, M., & Rinaldi, L. (2021, April 7). Predicting hand movements with distributional semantics: evidence from mouse-tracking. https://doi.org/10.31234/osf.io/aw9vb

Gatti, D., Marelli, M., Mazzoni, G., Vecchi, T., & Rinaldi, L. (2021a; PsyArXiv preprint). Hands-on false memories: A combined study with distributional semantics and mouse-tracking.

Gatti, D., Rinaldi, L., Marelli, L., Mazzoni, G., Vecchi, T. (*in press*). Decomposing the semantic processes underpinning veridical and false memories. *Journal of Experimental Psychology: General*.

Gatti, D., Rinaldi, L., Mazzoni, G., & Vecchi, T. (2021, March 19). Semantic and episodic processes differently predict false memories in the DRM task. https://doi.org/10.31234/osf.io/59asx

Gatti, D., VanVugt, F., & Vecchi, T. (2020). A causal role for the cerebellum in semantic integration: a transcranial magnetic stimulation study. *Scientific Reports*, 10, 18139, 1-12.

Gatti, D., Vecchi, T., & Mazzoni, G. (2021). Cerebellum and semantic memory: a TMS study using the DRM paradigm. *Cortex*.

Gebhart, A.L., Petersen, S.E., & Thach, W.T. (2002). Role of the posterolateral cerebellum in language. *Annals of the New York Academy of Sciences*, 978, 318-333.

Geraerts, E., Bernstein, D.M., Merckelbach, H., Linders, C., Raymaekers, L., & Loftus, E.F. (2008). Lasting false beliefs and their behavioral consequences. *Psychological Science*, 19(8), 749–753.

Ghosh, V.E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, 53, 104–114.

Gilligan, T.M., & Rafal, R.D. (2019). An Opponent Process Cerebellar Asymmetry for Regulating Word Association Priming. *The Cerebellum*, 18(1), 47-55.

Ginsburg, S., & Jablonka, E. (2007). The transition to experiencing: I. Limited learning and limited experiencing. *Biological Theory*, 2(3), 218–230.

Glenberg, A.M. (1997). What memory is for. *The Behavioral and Brain Sciences*, 20(1), 1–55.

Graham, L.M. (2007). Need for cognition and false memory in the Deese-Roediger-McDermott paradigm. *Personality and Individual Differences*, 42, 409–418.

Granziera, C., Schmahmann, J.D., Hadjikhani, N., Meyer, H., Meuli, R., Wedeen, V., & Krueger, G. (2009). Diffusion spectrum imaging shows the structural basis of functional cerebellar circuits in the human cerebellum in vivo. *PloS One*, 4(4), e5101.

Griego, A. W., Datzman, J. N., Estrada, S. M., & Middlebrook, S. S. (2019). Suggestibility and false memories in relation to intellectual disability and autism spectrum disorder: a meta-analytic review. *Journal of Intellectual Disability Research*, 63(12), 1464-1474.

Griffin, N.R., & Schnyer, D.M. (2020). Memory distortion for orthographically associated words in individuals with depressive symptoms. Cognition, 203, 104330.

Griffiths, T.L., Steyvers, M., & Tenenbaum, J.B. (2007). Topics in semantic representation. *Psychological Review*, 114(2), 211-244.

Grossman, L., & Eagle, M. (1970). Synonymity, antonymity, and association in false recognition responses. *Journal of Experimental Psychology*, 83(2p1), 244.

Guell, X., & Schmahmann, J.D. (2020). Cerebellar functional anatomy: a didactic summary based on human fMRI evidence. *The Cerebellum*, 19(1), 1-5.

Guell, X., D'Mello, A.M., Hubbard, N.A., Romeo, R.R., Gabrieli, J.D., Whitfield-Gabrieli, S., ... & Anteraper, S.A. (2020). Functional territories of human dentate nucleus. *Cerebral Cortex*, 30(4), 2401-2417.

Guell, X., Gabrieli, J.D., & Schmahmann, J.D. (2018). Embodied cognition and the cerebellum: Perspectives from the Dysmetria of Thought and the Universal Cerebellar Transform theories. *Cortex*, 100, 140-148.

Guell, X., Hoche, F., & Schmahmann, J.D. (2015). Metalinguistic deficits in patients with cerebellar dysfunction: empirical support for the dysmetria of thought theory. *The Cerebellum*, 14(1), 50-58.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun: An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, 47(4), 930–944.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, 69, 626-653.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006-1033.

Habas, C., Kamdar, N., Nguyen, D., Prater, K., Beckmann, C.F., Menon, V., & Greicius, M.D. (2009). Distinct cerebellar contributions to intrinsic connectivity networks. *Journal of Neuroscience*, 29(26), 8586-8594.

Habib, R., Nyberg, L., & Tulving, E. (2003). Hemispheric asymmetries of memory: the HERA model revisited. *Trends in Cognitive Sciences*, 7(6), 241-245.

Halko, M.A., Farzan, F., Eldaief, M.C., Schmahmann, J.D., & Pascual-Leone, A. (2014). Intermittent theta-burst stimulation of the lateral cerebellum increases functional connectivity of the default network. *Journal of Neuroscience*, 34(36), 12049-12056.

Hanajima, R., Wang, R., Nakatani-Enomoto, S., Hamada, M., Terao, Y., Furubayashi, T., ... & Ugawa, Y. (2007). Comparison of different methods for estimating motor threshold with transcranial magnetic stimulation. *Clinical Neurophysiology*, 118(9), 2120.

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *36*(1), 5-25.

Happé, F., Cook, J. L., & Bird, G. (2017). The Structure of Social Cognition: In (ter) dependence of sociocognitive processes. *Annual Review of Psychology*, *68*, 243-267.

Harris, Z. (1954). Distributional structure. *Word*, 10, 146–162

Hicks, J.L., & Marsh, R.L. (1999). Attempts to reduce the incidence of false recall with source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5), 1195.

Hilbe, J.M. (2011). *Negative binomial regression*. Cambridge University Press.

Hillier, A., Campbell, H., Keillor, J., Phillips, N., & Beversdorf, D. Q. (2007). Decreased false memory for visually presented shapes and symbols among adults on the autism spectrum. *Journal of Clinical and Experimental Neuropsychology*, *29*(6), 610-616.

Hirst, W., & Phelps, E. A. (2016). Flashbulb memories. *Current Directions in Psychological Science*, 25(1), 36-41.

Hoche, F., Guell, X., Sherman, J.C., Vangel, M.G., & Schmahmann, J.D. (2016). Cerebellar contribution to social cognition. *Cerebellum*, 15(6), 732-743.

Hoerl, C. (2018). Episodic memory and theory of mind: A connection reconsidered. *Mind and Language*, *33*(2), 148-160.

Hoffland, B.S., Bologna, M., Kassavetis, P., Teo, J.T.H., Rothwell, J.C., Yeo, C.H., ... & Edwards, M.J. (2012). Cerebellar theta burst stimulation impairs eyeblink classical conditioning. *The Journal of Physiology*, 590(4), 887–897.

Holden, L.R., Conway, A.R.A., & Goodwin, K.A. (2020). How Individual Differences in Working Memory and Source Monitoring matter in Susceptibility to False Memory. https://doi.org/10.31234/osf.io/h48bv

Howe, M.L. (2011). *The nature of early memory*. New York: Oxford University Press.

Huff, M.J., & Aschenbrenner, A.J. (2018). Item-specific processing reduces false recognition in older and younger adults: Separating encoding and retrieval using signal detection and the diffusion model. *Memory & Cognition*, 46(8), 1287-1301.

Hutchison, K.A. (2003). Is semantic priming due to association strength or feature overlap? A micro-analytic review. *Psychonomic Bulletin & Review*, 10, 785–813.

Hutchison, K.A., Balota, D.A., Neely, J.H., Cortese, M.J., Cohen-Shikora, E. R., Tse, C.S., ... & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, 45(4), 1099-1114.

Hyman, I.E., Husband, T.H., & Billings, F.J. (1995). False memories of childhood experiences. *Applied Cognitive Psychology*, 9(3), 181–197.

Iacullo, V.M., & Marucci, F.S. (2016). Normative data for Italian Deese/Roediger–McDermott lists. *Behavior Research Methods*, *48*(1), 381-389.

Iacullo, V.M., Marucci, F.S., & Mazzoni, G. (2016). Inducing false memories by manipulating memory self-efficacy. *Learning and Individual Differences*, 49, 237-244.

Irish, M., & Vatansever, D. (2020). Rethinking the episodic-semantic distinction from a gradient perspective. *Current Opinion in Behavioral Sciences*, 32, 43-49.

Ishikawa, T., Tomatsu, S., Izawa, J., & Kakei, S. (2016). The cerebro-cerebellum: Could it be loci of forward models?. *Neuroscience Research*, 104, 72-79.

Israel, L., & Schacter, D.L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review*, 4(4), 577-581.

Ito, M. (1993). Movement and thought: identical control mechanisms by the cerebellum. *Trends in Neurosciences*, 16(11), 448-450.

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience*, 9(4), 304-313.

Johns, B.T., & Jones, M.N. (2009). False recognition through semantic amplification. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 31(31), 2795-2800.

Johns, B.T., Jones, M.N., & Mewhort, D.J. (2012). A synchronization account of false recognition. *Cognitive Psychology*, 65(4), 486-518.

Johnson-Laird, P.N. (1983). *Mental models*. Cambridge University Press.

Johnson, M.K., & Raye, C.L. (1981). Reality monitoring. *Psychological Review*, 88(1), 67.

Jones, M.N. (2019). When does abstraction occur in semantic memory: insights from distributional models. *Language, Cognition and Neuroscience*, 34(10), 1338-1346.

Jones, M.N., & Mewhort, D.J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1), 1.

Jones, M.N., Hills, T.T., & Todd, P.M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, 122(3), 570-574.

Jones, M.N., Kintsch, W., & Mewhort, D.J.K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55, 534–552.

Jones, M.N., Willits, J., & Dennis, S. (2015). Models of semantic memory. *Oxford Handbook of Mathematical and Computational Psychology*, 232-254.

Kelly, R.M., & Strick, P.L. (2003). Cerebellar loops with motor cortex and prefrontal cortex of a nonhuman primate. *Journal of Neuroscience*, 23(23), 8432-8444.

Kieslich, P.J., Henninger, F., Wulff, D.U., Haslbeck, J.M.B., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: A practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods* (pp. 111-130). New York, NY: Routledge.

Kim, H., & Cabeza, R. (2006). Differential contributions of prefrontal, medial temporal, and sensory-perceptual regions to true and false memory formation. *Cerebral Cortex*, 17(9), 2143–2150.

Kim, H., Daselaar, S.M., & Cabeza, R. (2010). Overlapping brain activity between episodic memory encoding and retrieval: roles of the task-positive and task-negative networks. *NeuroImage*, 49, 1045-1054.

Kimball, D.R., Smith, T.A., & Kahana, M.J. (2007). The fSAM model of false recall. *Psychological Review*, 114(4), 954.

Klein, S.B. (2013). The temporal orientation of memory: It's time for a change of direction. *Journal of Applied Research in Memory and Cognition*, 2(4), 222-234.

Klein, S.B. (2007). Phylogeny and evolution: Implications for understanding the nature of a memory system. In H.L. Roediger, Y. Dudai, & S. Fitzgerald (Eds.), *Science of memory: Concepts* (pp. 377–381). New York: Oxford University Press.

Klein, S.B. (2014). Evolution, memory, and the role of self-referent recall in planning for the future. In B.L. Schwartz, M.L. Howe, M.P. Toglia, & H. Otgaar (Eds.), *What is adaptive about adaptive memory?* (pp.11-34). New York: Oxford University Press.

Klein, S.B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review*, 109(2), 306–329.

Kleiner, M., Brainard, D., Pelli, D. (2007). What's new in Psychtoolbox-3?. *Perception*, 36 ECVP Abstract Supplement.

Kloft, L., Otgaar, H., Blokland, A., Monds, L.A., Toennes, S.W., Loftus, E.F., & Ramaekers, J.G. (2020). Cannabis increases susceptibility to false memory. *Proceedings of the National Academy of Sciences*, 117(9), 4585-4589.

Koerner, T.K., & Zhang, Y. (2017). Application of linear mixed-effects models in human neuroscience research: a comparison with Pearson correlation in two auditory electrophysiology studies. *Brain Sciences*, 7(3), 26.

Koutstaal, W., Verfaellie, M., & Schacter, D.L. (2001). Recognizing identical versus similar categorically related common objects: Further evidence for degraded gist representations in amnesia. *Neuropsychology*, 15(2), 268–289.

Koziol, L.F., Budding, D.E., & Chidekel, D. (2012). From movement to thought: executive function, embodied cognition, and the cerebellum. *Cerebellum*, 11(2), 505-525.

Koziol, L.F., Budding, D., Andreasen, N., D'Arrigo, S., Bulgheroni, S., Imamizu, H., ... & Pezzulo, G. (2014). Consensus paper: the cerebellum's role in movement and cognition. *Cerebellum*, 13, 151-177.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33), 13852-13857.

Krienen, F.M., & Buckner, R.L. (2009). Segregated fronto-cerebellar circuits revealed by intrinsic functional connectivity. *Cerebral Cortex*, 19(10), 2485-2497.

Kubota, Y., Toichi, M., Shimizu, M., Mason, R.A., Findling, R.L., Yamamoto, K., & Calabrese, J.R. (2006). Prefrontal hemodynamic activity predicts false memory—A near-infrared spectroscopy study. *NeuroImage*, 31(4), 1783–1789.

Küper, M., Kaschani, P., Thürling, M., Stefanescu, M.R., Burciu, R.G., Göricke, S., ... & Timmann, D. (2016). Cerebellar fMRI activation increases with increasing working memory demands. *Cerebellum*, 15(3), 322–335.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.

Laney, C., & Loftus, E.F. (2010). False memory. *The Cambridge Handbook of Forensic Psychology*, 187.

Lecce, S., Ronchi, L., & Devine, R.T. (2021). Mind what teacher says: Teachers' propensity for mental-state language and children's theory of mind in middle childhood. *Social Development*.

Lecce, S., Ronchi, L., Del Sette, P., Bischetti, L., & Bambini, V. (2019). Interpreting physical and mental metaphors: Is Theory of Mind associated with pragmatics in middle childhood?. *Journal of Child Language*, *46*(2), 393-407.

Lecce, S., Ronchi, L., Del Sette, P., Bischetti, L., & Bambini, V. (2019). Interpreting physical and mental metaphors: Is Theory of Mind associated with pragmatics in middle childhood?. *Journal of Child Language*, 46(2), 393-407.

Leding, J.K. (2011). Need for cognition and false recall. *Personality and Individual Differences*, 51(1), 68-72.

Lee, J.L., Nader, K., & Schiller, D. (2017). An update on memory reconsolidation updating. *Trends in Cognitive Sciences*, 21(7), 531-545.

Lee, Y.S., Iao, L.S., & Lin, C.W. (2007). False memory and schizophrenia: evidence for gist memory impairment. *Psychological Medicine*, *37*(4), 559-567.

Lesage, E., Morgan, B.E., Olson, A. C., Meyer, A.S., & Miall, R.C. (2012). Cerebellar rTMS disrupts predictive language processing. *Current Biology*, 22(18), R794-R795.

Lins, J., & Schöner, G. (2019). Computer mouse tracking reveals motor signatures in a cognitive task of spatial language grounding. *Attention, Perception, & Psychophysics*, 81(7), 2424-2460.

Loftus, E.F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, 7(4), 560–572.

Loftus, E.F. (1977). Shifting human color memory. *Memory & Cognition*, 5(6), 696–699.

Loftus, E.F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4), 361–366.

Loftus, E.F. (2013). 25 years of eyewitness science ... finally pays off. *Perspectives on Psychological Sciences*, 8(5), 556–557.

Loftus, E.F., & Palmer, J.C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior,* 13(5), 585–589.

Loftus, E.F., & Pickrell, J.E. (1995). The formation of false memories. *Psychiatric Annals*, 25(12), 720–725.

Loftus, E.F., Burns, H.J., & Miller, D.G. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, 4(1), 19–31.

Louwerse, M.M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, 10, 573-589.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203-208.

Maki, W.S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, 15(3), 598-603.

Makowski, D. (2018). The psycho Package: an Efficient and Publishing-Oriented Workflow for Psychological Science. *Journal of Open Source Software*, 3(22), 470.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of meaning distance based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57-78.

Manto, M., Bower, J.M., Conforto, A.B., Delgado-García, J.M., Da Guarda, S.N.F., Gerwig, M., ... & Molinari, M. (2012). Consensus paper: roles of the cerebellum in motor control—the diversity of ideas on cerebellar involvement in movement. *Cerebellum*, 11, 457-487.

Marelli, M. (2017). Word-embeddings Italian Semantic spaces: a semantic model for psycholinguistic research. *Psihologija*, *50*(4), 503-520.

Marelli, M., & Amenta, S. (2018). A database of orthography-semantics consistency (OSC) estimates for 15,017 English words. *Behavior Research Methods*, *50*(4), 1482-1495.

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The effect of Orthography-Semantics Consistency on word recognition. *Quarterly Journal of Experimental Psychology*, 68(8), 1571-1583.

Mariën, P., Ackermann, H., Adamaszek, M., Barwood, C.H., Beaton, A., Desmond, J., ... & Leggio, M. (2014). Consensus paper: language and the cerebellum: an ongoing enigma. *Cerebellum*, 13, 386-410.

Martin, M.G.F. (2001). Out of the past: Episodic recall as retained acquaintance. *Time and Memory: Issues in Philosophy and Psychology*, 257-284.

Mayr, E. (2001). *What evolution is*. New York: Basic Books.

Mazza, C., Monaro, M., Burla, F., Colasanti, M., Orrù, G., Ferracuti, S., & Roma, P. (2020). Use of mouse-tracking software to detect faking-good behavior on personality questionnaires: an explorative study. *Scientific Reports*, 10(1), 1-13.

Mazzoni, G., & Memon, A. (2003). Imagination can create false autobiographical memories. *Psychological Science*, 14(2), 186–188.

McCormick, D.A., & Thompson, R.F. (1984). Cerebellum: Essential Involvement in the Classically Conditioned Eyelid Response. *Science*, 223(4633), 296–298.

McDermott, K.B., Gilmore, A.W., Nelson, S.M., Watson, J.M., & Ojemann, J. G. (2017). The parietal memory network activates similarly for true and associative false recognition elicited via the DRM procedure. *Cortex*, 87, 96-107.

McKelvie, S.J. (2004). False recognition with the Deese-Roediger-McDermott-Reid-Solso procedure: A quantitative summary. *Perceptual and Motor Skills*, 98(3), 1387-1408.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

Medved, M., & Brockmeier, J. (2015). When memory goes awry. In A. Tota & T. Hagen (Eds.), *Routledge international handbook of memory studies* (pp. 445–457). London: Routledge.

Melnikoff, D.E., & Bargh, J A. (2018). The mythical number two. *Trends in Cognitive Sciences*, 22(4), 280-293.

Miall, R.C., Antony, J., Goldsmith-Sumner, A., Harding, S.R., McGovern, C., & Winter, J.L. (2016). Modulation of linguistic prediction by TDCS of the right lateral cerebellum. *Neuropsychologia*, 86, 103-109.

Middleton, F.A., & Strick, P.L. (1994). Anatomical evidence for cerebellar and basal ganglia involvement in higher cognitive function. *Science*, 266, 458-461.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 26, pp. 3111–3119). Red Hook, NY: Curran Associates.

Miller, H. L., Odegard, T.N., & Allen, G. (2014). Evaluating information processing in autism spectrum disorder: The case for fuzzy trace theory. *Developmental Review*, *34*(1), 44-76.

Moberget, T., Gullesen, E.H., Andersson, S., Ivry, R., & Endestad, T. (2014). Generalized role for the cerebellum in encoding internal models: evidence from semantic processing. *Journal of Neuroscience*, 34, 2871-2878.

Molinari, M., Leggio, M.G., Solida, A., Ciorra, R., Misciagna, S., Silveri, M.C., & Petrosini, L. (1997). Cerebellum and procedural learning: evidence from focal cerebellar lesions. *Brain*, 120(10), 1753-1762.

Monaco, J., Casellato, C., Koch, G., & D'Angelo, E. (2014). Cerebellar theta burst stimulation dissociates memory components in eyeblink classical conditioning. *European Journal of Neuroscience*, 40(9), 3363–3370.

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods*, 45(2), 440-461.

Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological Research*, 79(5), 785-794.

Murdock, B.B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89, 609–626

Nadel, L., Hupbach, A., Gomez, R., & Newman-Smith, K. (2012). Memory formation, consolidation and transformation. *Neuroscience and Biobehavioral Reviews*, 36(7), 1640–1645.

Nader, K., & Hardt, O. (2009). A single standard for memory: the case for reconsolidation. Nature Reviews Neuroscience, 10(3), 224-234.

Nairne, J.S. (2005). The functionalist agenda in memory research. In A.F. Healy (Ed.), *Experimental cognitive psychology and its applications: A Festschrift in honor of Lyle Bourne, Walter Kintsch and Thomas Landaur* (pp. 115–126). Washington, DC: American Psychological Association.

Nelson, D.L., McEvoy, C.L., & Schreiber, T.A. (1998). The University of South Florida word association, rhyme, and word fragment norms.

Newman, E.J., & Lindsay, D.S. (2009). False memories: What the hell are they for?. *Applied Cognitive Psychology*, 23(8), 1105-1121.

Norman, K.A., & Schacter, D.L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, 25, 838–848.

Okado, Y., & Stark, C.E. (2005). Neural activity during encoding predicts false memories created by misinformation. *Learning & Memory*, 12(1), 3–11.

Orsini, A., & Pezzuti, L. (2013). *Wechsler Adult Intelligence Scale*—Forth Edition [WAIS-IV] *Edizione Italiana*. Giunti, OS: Firenze, Italy.

Ost, J., Blank, H., Davies, J., Jones, G., Lambert, K., & Salmon, K. (2013). False memory≠ false memory: DRM errors are unrelated to the misinformation effect. *PloS one*, 8(4), e57939.

Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of Advanced Theory-of-Mind Tasks. *Child Development. 87*(6), 1971-1991.

Osth, A.F., Shabahang, K.D., Mewhort, D.J., & Heathcote, A. (2020). Global semantic similarity effects in recognition memory: Insights from BEAGLE representations and the diffusion decision model. *Journal of Memory and Language*, 111, 104071.

Palesi, F., De Rinaldis, A., Castellazzi, G., Calamante, F., Muhlert, N., Chard, D., ... & Gandini Wheeler-Kingshott, C.A.M. (2017). Contralateral cortico-ponto-cerebellar pathways reconstruction in humans in vivo: implications for reciprocal cerebro-cerebellar structural connectivity in motor and non-motor areas. *Scientific Reports*, 7, 1-13.

Pansuwan, T., Breuer, F., Gazder, T., Lau, Z., Cueva, S., Swanson, L., ... & Morcom, A. M. (2020). Evidence for adult age-invariance in associative false recognition. *Memory*, 28(2), 172-186.

Papesh, M.H., & Goldinger, S.D. (2012). Memory in motion: Movement dynamics reveal memory strength. *Psychonomic Bulletin & Review*, 19(5), 906-913.

Papesh, M.H., Hicks, J.L., & Guevara Pinto, J.D. (2019). Retrieval dynamics of recognition and rejection. *Quarterly Journal of Experimental Psychology*, 72(9), 2328-2341.

Pascual-Leone, A., Grafman, J., Clark, K., Stewart, M., Massaquoi, S., Lou, J.S., & Hallett, M. (1993). Procedural learning in Parkinson's disease and cerebellar degeneration. *Annals of Neurology*, 34(4), 594-602.

Pascual, B., Masdeu, J. C., Hollenbeck, M., Makris, N., Insausti, R., Ding, S.L., & Dickerson, B.C. (2015). Large-scale brain networks of the human left temporal pole: a functional connectivity MRI study. *Cerebral Cortex*, 25(3), 680-702.

Payne, D.G., Elie, C.J., Blackwell, J.M., & Neuschatz, J.S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language*, 35(2), 261-285.

Paz-Alonso, P.M., Ghetti, S., Ramsay, I., Solomon, M., Yoon, J., Carter, C.S., & Ragland, J.D. (2013). Semantic processes leading to true and false memory formation in schizophrenia. *Schizophrenia research*, *147*(2-3), 320-325.

Peirce, J.W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, 162 (1-2), 8-13.

Peirce, J.W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2 (10), 1-8.

Peirce, J.W., & MacAskill, M.R. (2018). *Building Experiments in PsychoPy*. London: Sage.

Peirce, J.W., Gray, J.R., Simpson, S., MacAskill, M.R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*.

Peirce, J.W., Gray, J.R., Simpson, S., MacAskill, M.R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195-203.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437-442.

Perea, M., & Gotor, A. (1997). Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming. *Cognition*, 62(2), 223-240.

Pergolizzi, D., & Chua, E.F. (2015). Transcranial direct current stimulation (tDCS) of the parietal cortex leads to increased false recognition. *Neuropsychologia*, 66, 88-98.

Perner, J. (1991). *Understanding the representational mind*. The MIT Press.

Perner, J., Kloo, D., & Gornik, E. (2007). Episodic memory development: Theory of mind is part of re-experiencing experienced events. *Infant and Child Development*, *16*(5), 471-490.

Peterburs, J., & Desmond, J. E. (2016). The role of the human cerebellum in performance monitoring. *Current Opinion in Neurobiology*, 40, 38-44.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., & Raichle, M.E. (1989). Positron emission tomographic studies of the processing of singe words. *Journal of Cognitive Neuroscience*, 1, 153-170.

Porter, S., Yuille, J.C., & Lehman, D.R. (1999). The nature of real, implanted, and fabricated memories for emotional childhood events: Implications for the recovered memory debate. *Law and Human Behavior*, 23(5), 517–537.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515-526.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabin, J. S., Gilboa, A., Stuss, D. T., Mar, R. A., & Rosenbaum, R. S. (2010). Common and unique neural correlates of autobiographical memory and theory of mind. *Journal of Cognitive Neuroscience*, *22*(6), 1095-1111.

Rami, L., Gironell, A., Kulisevsky, J., Garcıa-Sánchez, C., Berthier, M., & Estevez-Gonzalez, A. (2003). Effects of repetitive transcranial magnetic stimulation on memory subtypes: a controlled study. *Neuropsychologia*, 41(14), 1877-1883.

Ramnani, N. (2006). The primate cortico-cerebellar system: anatomy and function. *Nature Reviews Neuroscience*, 7, 511-522.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59.

Renoult, L., & Rugg, M. D. (2020). An historical perspective on Endel Tulving's episodic-semantic distinction. *Neuropsychologia*, 139, 107366.

Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From knowing to remembering: the semantic–episodic distinction. *Trends in Cognitive Sciences*, 23(12), 1041-1057.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F.

Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Reyna, V.F., & Brainerd, C.J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making*, *4*(4), 249-262.

Reyna, V.F., & Brainerd, C.J. (1991). Fuzzy-trace theory and children's acquisition of mathematical and scientific concepts. *Learning and Individual Differences*, *3*(1), 27-59.

Reyna, V.F., & Brainerd, C.J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*(1), 1-75.

Rinaldi, L., & Marelli, M. (2020). Maps and space are entangled with language experience. *Trends in Cognitive Sciences*, 24, 853-855.

Roediger, H.L., & McDermott, K.B. (1995). Creating False Memories: Remembering Words Not Presented in Lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803.

Roediger, H.L., Balota, D.A., & Watson, J.M. (2001). Spreading activation and the arousal of false memories. In H.L. Roediger, J.S. Nairne, I. Neath, & A.M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95-115). Washington, DC: American Psychological Association.

Roediger, H.L., Watson, J.M., McDermott, K.B., & Gallo, D.A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, 8, 385-407.

Rosenbaum, R. S., Stuss, D. T., Levine, B., & Tulving, E. (2007). Theory of mind is independent of episodic memory. *Science*, *318*(5854), 1257-1257.

Rossi, S., Hallett, M., Rossini, P.M., & Pascual-Leone, A. (2011). Screening questionnaire before TMS: an update. *Clinical Neurophysiology*, 122, 1686.

RStudio Team (2015). *RStudio: Integrated Development for R.* RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8, 627–633.

Runnqvist, E., Bonnard, M., Gauvin, H.S., Attarian, S., Trébuchon, A., Hartsuiker, R.J., & Alario, F. X. (2016). Internal modeling of upcoming speech: A causal role of the right posterior cerebellum in non-motor aspects of language production. *Cortex*, 81, 203-214.

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di Linguista*, 20, 33–53.

Sara, S.J. (2000). Retrieval and reconsolidation: Toward a neurobiology of remembering. *Learning & Memory*, 7(2), 73–84.

Schacter, D.L. (2021). The seven sins of memory: an update. Memory, 1-6.

Schacter, D.L., & Loftus, E.F. (2013). Memory and law: What can cognitive neuroscience contribute? *Nature Neuroscience*, *16*(2), 119-123.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: remembering, imagining, and the brain. *Neuron*, 76(4), 677-694.

Schacter, D.L. (2001). *The seven sins of memory: How the mind forgets and remembers*. Mifflin and Company.

Schacter, D.L. (2012). Constructive memory: past and future. *Dialogues in Clinical Neuroscience*, 14, 7-18.

Schacter, D.L., Guerin, S.A., & St. Jacques, P.L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences*, 15(10), 467–474.

Schiller, D., & Phelps, E.A. (2011). Does reconsolidation occur in humans? *Frontiers in Behavioral Neuroscience*, 5, 24.

Schlaffke, L., Lissek, S., Lenz, M., Juckel, G., Schultz, T., Tegenthoff, M., ... & Brüne, M. (2015). Shared and nonshared neural networks of cognitive and affective theory-of-mind: A neuroimaging study using cartoon picture stories. *Human Brain Mapping*, *36*(1), 29-39.

Schmahmann, J.D. (1991). An Emeging Concept: The Cerebellar Contribution to Higher Function. *Neurological Review*, 48, 1178-1187.

Schmahmann, J.D. (2019). The cerebellum and cognition. *Neuroscience Letters*, 688, 62-75.

Schmahmann, J.D., Weilburg, J.B., & Sherman, J.C. (2007). The neuropsychiatry of the cerebellum—insights from the clinic. *Cerebellum*, 6(3), 254-267.

Seamon, J.G., Luo, C.R., & Gallo, D.A. (1998). Creating false memories of words with or without recognition of list items: Evidence for nonconscious processes. *Psychological Science*, 9(1), 20-26.

Seamon, J.G., Philbin, M.M., & Harrison, L.G. (2006). Do you remember proposing marriage to the Pepsi machine? False recollections from a campus walk. *Psychonomic Bulletin & Review*, 13(5), 752–756.

Serafin, M., & Surian, L. (2004). Il Test degli Occhi: uno strumento per valutare la" teoria della mente". *Giornale italiano di psicologia*, *31*(4), 839-862.

Sherry, D.F., & Schacter, D.L. (1987). The evolution of multiple memory systems. *Psychological Review*, 94(4), 439–454.

Smith, B. J., Gardiner, J. M., & Bowler, D. M. (2007). Deficits in free recall persist in Asperger's Syndrome despite training in the use of list-appropriate learning strategies. *Journal of Autism and Developmental Disorders*, *37*(3), 445-454.

Sokolov, A.A., Erb, M., Grodd, W., & Pavlova, M.A. (2014). Structural loop between the cerebellum and the superior temporal sulcus: evidence from diffusion tensor imaging. *Cerebral Cortex*, 24(3), 626-632.

Sokolov, A.A., Miall, R.C., & Ivry, R. (2017). The Cerebellum: Adaptive Prediction for Movement and Cognition. *Trends in Cognitive Sciences*, 21, 313-332.

Spivey, M.J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, 102(29), 10393-10398.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31, 137-149.

Steyvers, M. (2000). *Modeling semantic and orthographic similarity effects on memory for individual words. Unpublished doctoral dissertation*, Indiana University.

Steyvers, M., & Tenenbaum, J.B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1), 41-78.

Steyvers, M., Shiffrin, R.M., & Nelson, D.L. (2005). Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In A.F. Healy (Ed.), *Decade of behavior. Experimental cognitive psychology and its applications* (p. 237–249). American Psychological Association.

Stietz, J., Jauk, E., Krach, S., & Kanske, P. (2019). Dissociating empathy from perspective-taking: Evidence from intra-and inter-individual differences research. *Frontiers in Psychiatry*, *10*, 126.

Stillman, P.E., Krajbich, I., & Ferguson, M.J. (2020). Using dynamic monitoring of choices to predict and understand risk preferences. *Proceedings of the National Academy of Sciences*, 117(50), 31738-31747.

Stillman, P.E., Shen, X., & Ferguson, M.J. (2018). How mouse-tracking can advance social cognitive theory. *Trends in Cognitive Sciences*, 22(6), 531-543.

Stokes, M.G., Chambers, C.D., Gould, I.C., English, T., McNaught, E., McDonald, O., & Mattingley, J.B. (2007). Distance-adjusted motor threshold for transcranial magnetic stimulation. *Clinical Neurophysiology*, 118(7), 1617-1625.

Stoodley, C.J., & Schmahmann, J.D. (2009). Functional topography in the human cerebellum: a meta-analysis of neuroimaging studies. *NeuroImage*, 44, 489-501.

Sulin, R.A., & Dooling, D.J. (1974). Intrusion of a Thematic Idea in Retention of Prose. *Journal of Experimental Psychology*, 103(2), 255–262.

Tamagni, C., Mondadori, C. R., Valko, P. O., Brugger, P., Schuknecht, B., & Linnebank, M. (2010). Cerebellum and source memory. *European Neurology*, 63(4), 234-236.

Thakral, P.P., Benoit, R.G., & Schacter, D.L. (2017). Imagining the future: The core episodic simulation network dissociates as a function of timecourse and the amount of simulated information. *Cortex*, 90, 12-30.

Thomas, A.K., & Loftus, E.F. (2002). Creating bizarre false memories through imagination. *Memory & Cognition*, 30(3), 423–431.

Thompson-Schill, S.L., Kurtz, K.J., & Gabrieli, J.D. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory and Language*, 38(4), 440-458.

Timmann, D., Drepper, J., Calabrese, S., Bürgerhoff, K., Maschke, M., Kolb, F.P., ... & Diener, H.C. (2004). Use of sequence information in associative learning in control subjects and cerebellar patients. *Cerebellum*, 3(2), 75–82.

Timmann, D., Drepper, J., Frings, M., Maschke, M., Richter, S., Gerwig, M., & Kolb, F.P. (2010). The human cerebellum contributes to motor, emotional and cognitive associative learning. A review. *Cortex*, 46(7), 845–857.

Timmann, D., Drepper, J., Maschke, M., Kolb, F.P., Böring, D., Thilmann, A.F., & Diener, H.C. (2002). Motor deficits cannot explain impaired cognitive associative learning in cerebellar patients. *Neuropsychologia*, 40(7), 788–800.

Torriero, S., Oliveri, M., Koch, G., Caltagirone, C., & Petrosini, L. (2004). Interference of left and right cerebellar rTMS with procedural learning. *Journal of Cognitive Neuroscience*, 16, 1605-1611.

Tronson, N.C., & Taylor, J.R. (2007). Molecular mechanisms of memory reconsolidation. *Nature Reviews. Neuroscience*, 8(4), 262–275.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic Press.

Tulving, E. (1983). *Elements of episodic memory*. London: Oxford University Press.

Tun, P.A., Wingfield, A., Rosen, M.J., & Blanchard, L. (1998). Response latencies for false memories: gist-based processes in normal aging. *Psychology and Aging*, 13(2), 230.

Ugawa, Y., Uesaka, Y., Terao, Y., Hanajima, R., & Kanazawa, I. (1995). Magnetic stimulation over the cerebellum in humans. *Annals of Neurology*, 37, 703-713.

Unsworth, N., & Brewer, G.A. (2010). Individual differences in false recall: A latent variable analysis. *Journal of Memory and Language*, 62, 19-34

Van Overwalle, F., Manto, M., Cattaneo, Z., Clausi, S., Ferrari, C., Gabrieli, J.D., ... & Michelutti, M. (2020). Consensus Paper: Cerebellum and Social Cognition. *Cerebellum*, 1-36.

Vecchi, T. & Gatti, D. (2020). *Memory as prediction: From looking back to looking forward*. MIT Press.

Verfaellie, M., Schacter, D.L., & Cook, S.P. (2002). The effect of retrieval instructions on false recognition: Exploring the nature of the gist memory impairment in amnesia. *Neuropsychologia*, 40(13), 2360–2368.

Voogd, J., & Glickstein, M. (1998). The anatomy of the cerebellum. *Trends in Cognitive Sciences*, 2(9), 307-313.

Wade, K.A., Garry, M., Read, J.D., & Lindsay, D.S. (2002). A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic Bulletin & Review*, 9(3), 597–603.

Wagenmakers, E.J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*. 11(1), 192-196.

Watson, J.M., Balota, D.A., & Roediger, H.L. (2003). Creating false memories with hybrid lists of semantic and phonological associates: Over-additive false memories produced by converging associative networks. *Journal of Memory and Language*, 49(1), 95-118.

Watson, J.M., Balota, D.A., & Sergent-Marshall, S. D. (2001). Semantic, phonological, and hybrid veridical and false memories in healthy older adults and in individuals with dementia of the Alzheimer Type. *Neuropsychology*, 15, 254–267.

Watson, J.M., Bunting, M.F., Poole, B.J., & Conway, A.R. (2005). Individual differences in susceptibility to false memory in the Deese-Roediger-McDermott paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 76.

Wechsler, D. (2008). WAIS-IV: Wechsler Adult Intelligence Scale. San Antonio, TX: Pearson.

Wellman, H.M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, *15*(6), 728-755.

Westbury, C. (2016). Pay no attention to that man behind the curtain: Explaining semantics without semantics. *The Mental Lexicon*, 11(3), 350-374.

Westerberg, C.E., & Marsolek, C.J. (2006). Do instructional warnings reduce false recognition?. *Applied Cognitive Psychology*, 20(1), 97-114.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Williams, G.C. (1966). *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton: Princeton University Press.

Wilson, A.C. (2021). Do animated triangles reveal a marked difficulty among autistic people with reading minds?. *Autism*, *25*(5), 1175-1186

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128.

Winocur, G., Moscovitch, M., & Sekeres, M. (2007). Memory consolidation or transformation: context manipulation and hippocampal representations of memory. *Nature Neuroscience*, 10(5), 555-557.

Wojcik, D.Z., Díez, E., Alonso, M.A., Martín-Cilleros, M.V., Guisuraga-Fernández, Z., Fernández, M., ...& Fernandez, A. (2018). Diminished false memory in adults with autism spectrum disorder: Evidence of identify-to-reject mechanism impairment. *Research in Autism Spectrum Disorders*, *45*, 51-57.

Xiang, H,. Lin, C., Ma, X., Zhang, Z., Bower, J.M., Weng, X., & Gao, J.H. (2003). Involvement of the cerebellum in semantic discrimination: an fMRI study. *Human Brain Mapping*, 18(3), 208-214.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5), 971-979.

Zhu, B., Chen, C., Loftus, E. F., Lin, C., & Dong, Q. (2013). The relationship between DRM and misinformation false memories. *Memory & Cognition*, 41(6), 832-838.