UNIVERSIDADE DE LISBOA

FACULDADE DE PSICOLOGIA

LEARNING WITH THE TEST: ADAPTING ENCODING STRATEGIES TO THE RETRIEVAL REQUISITES

Pedro André Ribeiro Marques

Orientador: Prof. Doutor Leonel Garcia Marques

Tese especialmente elaborada para obtenção do grau de Doutor em Psicologia

(Psicologia Cognitiva)

2017

UNIVERSIDADE DE LISBOA
FACULDADE DE PSICOLOGIA

LEARNING WITH THE TEST: ADAPTING ENCODING STRATEGIES TO THE RETRIEVAL REQUISITES

Pedro André Ribeiro Marques

Orientador: Prof. Doutor Leonel Garcia Marques

Tese especialmente elaborada para obtenção do grau de Doutor em Psicologia (Psicologia Cognitiva)

Júri:

Presidente: Doutora Isabel Maria de Santa Bárbara Teixeira Nunes Narciso Davide

Vogais:

- Doutor Emanuel Pedro Viana Barbas Albuquerque, Professor Associado da Escola de Psicologia da Universidade do Minho;
- Doutora Teresa Maria Freitas Teixeira de Morais Garcia Marques, Professora Catedrática da Unidade de Investigação em Psicologia Cognitiva, do Desenvolvimento e da Educação do ISPA - Instituto Universitário de Ciências Psicológicas, Sociais e da Vida;
- Doutor Leonel Garcia Marques, Professor Catedrático da Faculdade de Psicologia da Universidade de Lisboa, orientador;
- Doutora Ana Luísa Nunes Raposo, Professora Auxiliar da Faculdade de Psicologia da Universidade de Lisboa.

Investigação financiada pela Fundação para a Ciência e a Tecnologia através da bolsa doutoral SFRH/BD/78697/2011

2017

Acknowledgements

This dissertation documents what I have learned and presents the output of the research I developed along with an inspiring group of bright people during the last four years. As will be noticeable along the pages of this thesis, it would be literally impossible for me to learn, understand (?) and investigate the ideas here presented if I was alone in this endeavor. Thus, some important acknowledgements are due.

First, I have to thank Leonel for all his guidance, friendship, encouragement, and also for teaching me the importance of learning from the requisites of the tasks. He always challenged me to pursuit new and original ideas, to look at problems from different angles, and to look for inspiration and solutions in less obvious sources, promoting a wider understanding of how things work, while making it fun and compelling. I grew a lot as a researcher and a person as a result of these learnings, and can't wait to see what future collaborations will bring. Thank you.

Secondly, I would also like to thank to all the colleagues that contributed to this work and thus to my development. To those who directly contributed for the empirical work here presented – Ludmila, Paula, Yana, Diana, Tomás, Ana Luísa – a big thank you. I hope our paths will cross often; there is still a lot of science to be done. Also, I would like to thank those colleagues that were hugely influential along these years: João, Sofia, Jerônimo, Joana, Fábio, Ana, Mário, Ana Sofia, Sara, Tânia, and everyone from the CO2 group. I couldn't ask for better company during this time.

This work is also dedicated to the memory of Prof. J. Frederico Marques, who had a profound impact in my development as a scientist.

Importantly, a special thank you to Profs. Bob Bjork and Elizabeth Bjork for receiving me in their lab, and going through great lengths to ensure that my visit was fruitful. My stay at the Bjork Learning & Forgetting Lab at the UCLA left lasting

marks on my knowledge and understanding about the science of memory, and left me eager to know more. Extended thanks go to the rest of the lab members, and to the participants of CogFog meetings, where good ideas never stopped flowing.

I would also like to thank Nick and Melanie Soderstrom for their friendship and hospitality during my stay in the big city of Los Angeles.

This list would be incomplete without mentioning my parents and my brother, always an inspiration and a source of love and support. Thank you.

To all my friends – the oldest ones from Tondela, 'maltinha' from Faculdade de Psicologia, people of the 'Nhabez': you make life better, and I continuously learn from you. Big hug.

Finally, I would like to dedicate this dissertation to Júlia, who accompanied me on most of this journey, and made it meaningful. Thank you.

Lisboa, 30th of October, 2016

Abstract

Retrieving information from memory is usually, and obviously, the final objective of encoding it. In that line, we propose that we encode and store information as a function of the particular ways we have used similar information in the past. More specifically, we contend that the experience of retrieval can serve as a powerful cue to the most effective ways to encode similar information in comparable future learning episodes. This ubiquitous characteristic of human memory and its effects will be explored along six studies, modifications of the repeated retrieval paradigm, where we keep encoding conditions constant, but manipulate the requisites of memory tests and observe the gains and costs in learning across several study-test cycles, along with manifestations of learning based on the structures and requisites of the tests (and not of the studied materials) in final surprise-tasks.

In this dissertation we will introduce the evolution of the study of the impact of retrieval in memory functioning, and establish possible parallels with conceptual learning. In three empirical chapters we will explore the adaptive aspects, benefits and costs of repeated retrieval. In Chapter II we will show how the nature of the distracters in recognition tests will, by itself, impact subsequent encoding strategies of new similar information: relational when when distracters are not related to the lists, and distinctiveness-based when they are related to the study lists and conceptual knowledge is not useful for the task. In Chapter III we will present converging evidence to this phenomenon, showing that if memory tests demand responding to spatial cues, instead of semantic cues, the activation of the item's semantic networks will be hindered, suggesting that along several study-test cycles we learn to disregard stimuli dimensions that are irrelevant to the tasks. In Chapter IV we show that even if

the structure of memory tests encourage and allow for the use of conceptual knowledge (and not episodic) to respond, conceptual processing of information only present at test is hindered.

We will discuss interactions, similarities and differences between memory and conceptual learning, and also how acts of retrieval in a given context can impact further attention to specific stimuli attributes. We will advance some theoretical considerations on the present research, along with some practical applications.

Keywords: Memory, Adaptation, Retrieval, Learning

Resumo

Recuperar informação da memória é, obviamente, o objectivo final da sua codificação. Pensando neste propósito final, propomos que codificamos e armazenamos informação em função das formas específicas como fizemos uso de informação semelhante no passado. Especificamente, argumentamos que a experiência da recuperação pode servir como uma forte pista para as formas mais eficientes de codificar informação semelhante noutras situações equiparáveis de aprendizagem. Esta característica ubíqua da memória humana e os seus efeitos serão explorados a partir dos dados de seis estudos - modificações do paradigma de recuperação repetida - em que mantemos as condições de codificação constantes, mas manipulamos os requisitos dos testes de memória. Vamos analisar os ganhos e custos na aprendizagem ao longo de vários ciclos estudo-teste, bem como as manifestações de aprendizagens baseadas nas estruturas e requisitos dos testes (e não nos materiais estudados) em tarefas-surpresa finais.

Nesta dissertação introduziremos a evolução do estudo do impacto da recuperação no funcionamento mnésico, e possíveis paralelismos com a aprendizagem conceptual. Em três capítulos empíricos iremos explorar os aspectos adaptativos, benefícios e custos da recuperação repetida de informação. No Capítulo II mostramos como o tipo de distractores presente em testes de reconhecimento vai, por si, impactar estratégias de codificação subsequentes para informação nova, tornando-se estas mais relacionais quando os distractores não estão relacionados com as listas, ou mais baseadas na distintividade quando são relacionadas com as listas de estudo, e o conhecimento conceptual não é útil na tarefa. No Capítulo III apresentamos evidências convergentes para este fenómeno, mostrando que se testes

de memória exigirem resposta a pistas espaciais, em vez de semânticas, a ativação das redes semânticas das palavras será atenuada, sugerindo que ao longo de ciclos de estudo-teste aprendemos a negligenciar dimensões irrelevantes dos estímulos. No Capítulo IV mostramos como apesar de a estrutura de testes de memória poder encorajar e permitir o uso de conhecimento conceptual (por oposição a episódico) para responder, o processamento conceptual de informação presente no teste é deficitário.

Serão discutidas interações, semelhanças e diferenças entre memória e aprendizagem conceptual, bem como a influência dos atos de recuperação em contexto na atenção a atributos específicos dos estímulos. Serão ainda avançadas algumas considerações teóricas e aplicações práticas da investigação aqui apresentada.

Palavras-chave: Memória, Adaptação, Recuperação, Aprendizagem

Resumo alargado

Nesta dissertação propomos que codificamos e armazenamos informação em função da utilização de informação semelhante no passado. Especificamente, sustentamos que experienciar a recuperação pode servir como uma poderosa pista para as formas mais adequadas de codificar informação semelhante em episódios de aprendizagem futuros.

Apesar de terem sido durante muito tempo considerados eventos neutros que apenas serviriam objectivos de avaliação do conhecimento, sem qualquer influência na representação, codificação e recuperação futura de informação, os atos de recuperação têm vindo a ocupar um papel central na investigação em memória. A recuperação é agora vista como impactante na própria representação da informação (e.g., Bjork, 1975), e evidências têm-se acumulado mostrando que recuperar repetidamente informação pode levar a melhor retenção a longo-prazo (um efeito "directo" da recuperação repetida; e.g., Karpicke & Roediger, 2006a) e até potenciar episódios subsequentes de estudo dessa mesma informação (um efeito "indirecto";e.g., Arnold & McDermott, 2013; Izawa, 1970).

A investigação aqui apresentada propõe-se a explorar um efeito "indirecto" da recuperação repetida, mostrando como alguns paralelismos entre memória e aprendizagem conceptual (e.g., Bruner, Goodnow & Austin, 1956) podem ser observados em condições onde são experienciados vários ciclos de estudo-teste: a utilização de conhecimento e estratégias de codificação conceptual, bem como a adaptação destas a relações entre os estímulos e os requisitos específicos dos testes memória aprendidas pela recuperação repetida. Numa observação pertinente sobre os aspectos adaptativos da memória humana, Higham e Brooks (1997) mostraram como

os participantes em experiências de memória e de categorização são sensíveis a invariantes do mundo, ainda que tacitamente, ao exibirem efeitos típicos de categorização em tarefas de memória de reconhecimento (e.g., produzir mais falsos alarmes com distractores que partilham as regras de selecção de estímulos das listas de estudo, do que com distractores que violam essas regras), e também efeitos típicos de tarefas de memória de reconhecimento em tarefas de categorização (e.g., itens estudados são mais prontamente categorizados como seguindo uma regra do que itens não-estudados que também seguem essa mesma regra).

No Capítulo II desta dissertação, mostramos como condições em que há sobreposição entre o estatuto episódico dos itens e o seu estatuto conceptual, os participantes são encorajados a utilizar estratégias de codificação (e recuperação) conceptual, enquanto condições em que o conhecimento conceptual não é diagnóstico do estatuto episódico dos itens vão desencorajar este tipo de processamento, levando os participantes a adoptar estratégias focadas na distintividade dos itens. Para isso, construímos duas experiencias onde os participantes passam por quatro ciclos de estudo-teste em que as listas de estudo são sempre exemplares de categorias (i.e., de uma categoria distinta por cada ciclo), sendo apenas manipulada a natureza dos distractores nos testes de reconhecimento. Os teste ora incluíam apenas distractores da mesma categoria da lista de estudo ou distractores não-relacionados provenientes de várias categorias. Experienciar testagem repetida levou os participantes a evitar estratégias de processamento conceptual (i.e., atendendo aos aspectos comuns e relacionais dos itens das listas) quando este tipo de conhecimento não era diagnóstico para a resposta ao teste, ou seja, quando os testes incluíam distractores da mesma categoria das listas. Assim, apesar de os participantes cujos testes incluíam distractores não-relacionados obterem melhor performance que os seus homólogos

(cujos testes incluíam distractores relacionados), esta vantagem inverteu-se quando os requisitos mudaram abruptamente num teste final, igual para todos os participantes, que só incluía distractores relacionados (Experiência 1). Na Experiência 2 deste capítulo mostramos ainda esta inversão dos níveis de performance quando o teste final passa a ser de recordação-livre: os participantes testados com distractores relacionados, apesar de pior performance nos primeiros ciclos, desenvolveram uma estratégia baseada nos itens (por oposição a baseada na categoria) que beneficiou a sua performance em recordação-livre (e.g., Hunt & Einstein, 1981; Nairne, 2006).

No sentido de investigar a ubiquidade deste fenómeno de adaptação de estratégias de processamento subsequente aos requisitos do teste, estendemos também a nossa hipótese a uma outra dimensão (Underwood, 1969) dos estímulos: a localização. A localização dos estímulos é um aspecto central da informação que pode ser codificado e usado como pista de recuperação (e.g., Rajaram, 1998; Underwood, 1969), sendo que o seu processamento é tido como sendo largamente automático (e.g. Hasher & Zacks, 1988) mas ainda assim susceptível a optimização através da prática (e.g., Caldwell & Masson, 2001). No Capítulo III mostramos os resultados de duas experiencias em que, tal como no Capítulo II, mantivemos as condições de codificação constantes e manipulámos os requisitos dos testes. Numa adaptação do Jogo da Memória (também conhecido como Concentração), os participantes tinham que estudar pares de exemplares de uma categoria que eram apresentados sob a forma de cartas que continham cada palavra, colocados em duas de dezasseis posições no ecrã, sendo depois manipulado entre-participantes o tipo de teste de memória a que tinham que responder: ou recebiam pistas semânticas (uma das cartas estudadas era revelada, bem como a localização do outro membro do par, e os participantes deveriam escrever a palavra correspondente), ou recebiam pistas de localização (uma

das cartas estudadas era revelada, bem como a outra palavra-membro, e os participantes deveriam indicar a localização dessa palavra) ao longo de quatro ciclos de estudo-teste. Enquanto que na condição de pistas semânticas uma estratégia mais profunda e baseada no significado favoreceria a performance, na condição de pistas de localização uma estratégia perceptiva e focada nas posições relativas seria a mais adequada. Na Experiência 1, quando invertemos os requisitos abruptamente no último teste, as estratégias que antes favoreciam a performance agora prejudicá-la-iam, mostrando os resultados numa queda da performance, principalmente para participantes testados com pistas de localização. Na Experiência 2, após os quatro ciclos de estudo-teste, os participantes tinham que gerar associados-livres a palavras experienciadas no último ciclo e palavras novas (ver Hourihan e Macleod, 2007). Os participantes testados com pistas semânticas mostraram uma facilitação nas respostas a itens antigos em relação a itens novos (com tempos de resposta significativamente mais rápidos), enquanto que este padrão foi inexistente para os participantes testados com pistas de localização, sugerindo que estes aprenderam a negligenciar a dimensão semântica dos itens, e assim a minorar a ativação das redes semânticas correspondentes, levando tanto tempo a gerar um associado para um item antigo como para um novo. Estes resultados sugerem que os participantes aprenderam a negligenciar ou inibir o processamento de dimensões tidas como irrelevantes para os testes experienciados, num padrão que se assemelha ao encontrado em tarefas de aprendizagem conceptual (e.g., Bruner, Goodnow, & Austin, 1956; Deng & Sloutsky, 2015).

Os resultados destas experiências sugerem que memória e aprendizagem conceptual parecem coexistir e interagir ao nível das tarefas, tornando, por vezes, tarefas de reconhecimento e de categorização intercambiáveis (Capítulo II). Contudo,

propomos uma diferença fundamental entre elas: o ato de recuperação envolve uma redirecção da atenção para o passado, em detrimento de possíveis novas aprendizagens num momento de teste. De forma a caracterizar condições-limite para o fenómeno aqui apresentado, e a conciliar estas ideias com investigação recente que mostra um défice em aprender informação nova apenas presente no momento do teste (e.g., Davis & Chan, 2015; Finn & Roediger, 2013), construímos duas experiencias – adaptações do paradigma utilizado no Capítulo II – em que manipulamos, entre e intra participantes, o *locus* da informação conceptual: nas listas de estudo, ou no teste como distractores. Na Experiência 1 as listas de estudo são exemplares de categorias e os distractores são palavras não-relacionadas – excepto dois distractores que são exemplares da categoria estudada – ou vice-versa – excepto dois distractores que são palavras não-relacionadas. Na Experiência 2 as listas de estudo são exemplares de duas categorias, e no teste os alvos são exemplares de uma só delas e os distractores são exemplares de uma categoria nova. Ao longo de quatro ciclos de estudo-teste os participantes mostraram altos níveis de falsos alarmes coerentes com o tipo de lista de estudo (Experiência 1) bem como altos níveis de performance, sugerindo aprendizagem conceptual. Num teste final de reconhecimento acerca de todas listas de estudo e de teste anteriores, a condição em que os distractores eram membros de uma categoria (Experiência 1) e as categorias que só apareceram no teste (Experiência 2) exibiram níveis baixos de falsos alarmes, sugerindo um défice de processamento conceptual no teste. De notar que este défice é relativo, pois neste caso (especialmente na Experiência 2) leva a melhor desempenho.

Estes ideias serão enquadradas em teorias recentes de memória (e.g., Bjork, 2011) e inibição (e.g., Hasher & Zacks, 2007) e serão apontadas implicações gerais para o estudo do papel da recuperação mnésica na cognição humana, bem como

algumas clarificações acerca do impacto da investigação sobre os efeitos dos testes em contextos educativos.

Overview

Since Ebbinghaus (1885) human memory as a delineated subject in experimental research has gathered an impressive amount of data and theorizing, revealing a fascinatingly complex system that more often than not resists fitting in general and stable laws, despite the myriad robust and replicable effects (see Roediger, 2008). Some of the core characteristics of memory research include: clearly separated study (encoding; where most learning would occur) and test (retrieval; where few learning would happen, serving assessment purposes only) phases; insulation against world knowledge when assessing episodic memory (e.g., using distractors that match targets in all dimensions except their episodic status, and one-time study and test opportunities), and the notion that measures of memory performance differ in sensitivity only. In this work, we will build upon a host of studies that suggest that going beyond these classic characteristics of the experimental study of memory can uncover new aspects of our cognitive functioning, or at least allow for a better investigation of its functional or procedural facets (e.g. Kolers, 1973, 1979; Neisser, 1995).

In this dissertation we will explore how people can capitalize world knowledge to respond to memory prompts, if past retrieval experiences encourage them to do so as the moment of retrieval can allow one to learn about the specific requisites of memory tasks, thus promoting adaptation in encoding and responding. We will also show how during retrieval, besides the fostering in learning of task requisites and guiding of subsequent encoding, the learning of new information that is deemed retrieval-irrelevant can be hindered. These ideas derived especially from three lines of research: 1) the *testing effects* literature (e.g., Roediger & Karpicke, 2006a; Roediger & Butler, 2011), showing that repeated retrieval of information will result in

better long-term retention than restudy; 2) literature that shows that people are sensitive to the contingencies between stimuli characteristics and task structure (e.g. Higham & Brooks, 1997), namely when target items and distractors differ in a given conceptual feature, giving rise to responses that adhere to this perceived difference (such can be one interpretation of the classic DRM effect; see Gallo, 2010) and 3) the functional value of forgetting or inhibiting information in memory as to allow for new learning and memory updating (e.g., Bjork & Bjork, 1992; Bjork, 2011).

Our general hypothesis is that repeated retrieval allows for the learning of the task structure, thus impacting the way we will encode subsequent information in new but similar study-test episodes, in ways similar to concept learning (e.g., Bruner, Goodnow & Austin, 1956). Moreover, we will show how this idea is compatible with recent data evidencing retrieval-based learning impairments of new information. In the introduction we will review relevant literature on testing effects (both on its direct and indirect facets), its tentative overlap with conceptual learning effects, its impact on several stimuli dimensions, and its associated learning costs.

In the introduction we will review the course of research on retrieval processes and their impact on learning, retention and representation, going from its characterization as a neutral aspect of memory, to a more central role in memory theorizing and its applications. We will also discuss the tentative overlap between retrieval processes and conceptual learning as part of the adaptive character of memory. Moreover, we will integrate these ideas with related effects in false memory research, and explain how our hypothesis fit the successful memory framework of the New Theory of Disuse (Bjork & Bjork, 1992), as a functional way to be adapted to specific current contexts in order to keep our knowledge updated. The following three sections will be three empirical chapters, in the form of research articles (one

published, and two submitted to publication), where we present data supporting our claims and further discussion on the effects of retrieval in learning. In Chapter II we will try to establish the phenomenon of adaptation of encoding strategies to the retrieval requirements in two experiments where the relationship between the distracters in recognition tests and study-lists is manipulated; the results indicate that when distracters are related to the study-lists further processing becomes more item-based, while when they are unrelated further processing becomes relational. In Chapter III we show further evidence, and extend this effect to another stimuli dimension (location) on two experiments; our results lead us to argue that when the test requisites are location-based, semantic activation of words can be hindered. In Chapter IV we show how while promoting the learning of the test requisites and guiding encoding, retrieval can have undesirable side effects in terms of a conceptual processing deficit at test. In the General Discussion these results will be summarized and integrated, and we will draw theoretical and practical consequences of this research.

Table of Contents

CHAPTER I. Introduction

1. On the role of retrieval in memory

1.1 Early research

The act of retrieving information from one's memory (or testing our memory) has been in and out of the spotlights in cognitive psychology research over the years. Perhaps coupled with common practices in daily life, and especially in educational settings, memory retrieval was for some time regarded as a mere byproduct of knowledge assessment, that would leave the contents of our memory (if they were there in the first place) untouched, and that wouldn't bear any influence on future learning or testing episodes. While some notable thinkers (like Aristotle, Francis Bacon or William James; see Roediger & Karpicke, 2006a) had already speculated that retrieving information from memory (or rehearsing the to-be-remembered materials) would result in stronger and longer-lasting memories when compared to passive restudy, empirical studies addressing this issue were scarce. Thus, this notion was not part of societal beliefs, nor the interests of the then-infant field of experimental psychology.

The first documented results showing evidence for long-term storage benefits of repeated retrieval relative to restudy were the works developed by Edwina Abbott (1908, 1909) at the University of Illinois, but these studies suffered from small sample sizes and lax experimental control, and were not influential at the time. Nonetheless, empirical interest on the phenomenon was sparked and on the succeeding years large-scale studies both in lab and in the classroom accumulated early evidence showing a clear overall advantage of repeated retrieval over repeated studying in later retention. Three early works can be considered as the foundations of the rigorous study of this

phenomenon: Gates (1917), Jones (1923/1924) and Spitzer (1939). Gates (1917) was the first to show the advantage of repeated retrieval over restudying with both nonsense syllables and educationally relevant materials, using a larger sample of students from a wide range of grades; he manipulated the time participants had to recite the to-be-learned materials, and found a clear benefit of the time spent reciting on a final recall test. Jones (1923/1924) provided further evidence for the phenomenon using college students (now the "industry standard" of samples) both in the lab and in the classroom. Spitzer (1939) conducted a very-large scale study using over three thousand 6th graders as participants, where students were to study the materials and were then subject to differing retrieval practice (testing) and restudy schedules without feedback, before a final test administered two months later. His findings further established the superiority of retrieving information from memory over its restudy in long-term retention. Despite some methodological issues with these studies (discussed ahead), the results proved to be robust and were replicated by other researchers (e.g. Forlano, 1936; Sones & Stroud, 1940). Nonetheless, and perhaps due to shift on the *zeitgeist* to forgetting phenomena around that time (see Crowder, 1976), further research directly addressing the mnemonic benefits of retrieval was scarce, with some influential authors even advising against the use of repeated retrieval paradigms as they contaminated forgetting research results by halting the forgetting functions (e.g., Deese, 1958).

As often happens in science, the nuisances and contaminating factors at the time (for forgetting research) later became the focal points of interest, and methodological tools to study further memory phenomena. By then, besides the general acceptance that active retrieval would lead to better long-term retention, little else was known about the underlying processes and boundaries of this phenomenon.

Even lab studies were conducted under an educational framework (i.e., interest was on general conditions that would lead to better performance, as measured by some kind of final test), and the manipulation of interest was mainly the time spent rehearsing (retrieving) the to-be-remembered information, or the number of administered test (with more time spent rehearsing, or more tests, leading to better retention). The typical experiment would involve students reading a prose passage (or a list of nonsense syllables; Ebbinghaus, 1885/1964) in silence (i.e., study it) on a first phase, followed by differing schedules of rehearsing (e.g., Gates, 1917) or multiple-choice testing (e.g. Spitzer, 1939). Despite the robust evidence for the retrieval superiority in long-term retention (the few authors that investigated it replicated the effect), the employed paradigm had some problems, such as lax (or unspecified) experimental control and lack of proper control (e.g., restudy) conditions (thus not disentangling mere re-exposure and retrieval *per se* as the factors guiding the effect).

1.2 Retrieval promotes learning, and is a memory modifier

Some years after, and in the midst of a heated discussion on whether learning was an incremental or an all-or-none process (Rock, 1957; for an historical review of this discussion see Roediger & Arnold, 2012), emerged an article by Endel Tulving (1967) that while not directly addressing the role of retrieval in memory performance, provided a paradigm and data that stimulated interest in the retrieval superiority phenomenon (at least for some time). This article called into question an assumption from experimenters at the center of the incremental/all-or-none debate: the mentioned learning was to happen only during exposure to the materials, i.e., the study phases, while the moments of retrieval, i.e., testing phases, were neutral events that served merely an assessment purpose. Tulving's position was that retrieving information

from memory, at least as part of a chain of encoding and retrieval episodes, would impact the participants performance as much as studying that information. To test this hypothesis, he devised a multi-trial design where participants were to study a list of words and have a free-recall test on it on three distinct schedules: standard [three 4-trial cycles, where studying (S) and testing (T) were intercalated – STST x 3]; repeated-study (three 4-trial cycles where participants studied the materials three times and were tested once – SSST x 3); and repeated-testing (three 4-trial cycles where participants studied the materials and were then tested four times – STTT x 3), with the time spent studying and recalling equated (two minutes). If learning was to occur only at study, the STTT condition participants should exhibit a lower performance than their counterparts, and the SSST participants should outperform the one from the other conditions. Surprisingly, the learning curves for the three conditions were very similar, with no clear advantage of any of the schedules, indicating that retrieval could be a potent learning event at least to the same extent as encoding. This effect was subject to both direct and conceptual replications on the following years (e.g. Birnbaum & Eichner, 1971; Donaldson, 1971; Hogan & Kintsch, 1971; Rosner, 1970), establishing the somewhat counterintuitive finding that under some conditions testing in the absence of studying can result in equally good performance in a subsequent test. Despite the robustness and the novelty of this effect, only few researchers (the aforementioned replications) dedicated their efforts in further exploring it. A contemporary exception was Chizuku Izawa's work (e.g., Izawa, 1966, 1970; for a review see Izawa, Maxwell, Hayden, Matrana, & Izawa-Hayden, 2005) on what she called test potentiation effects: she and her collaborators started off testing the hypothesis that while no learning nor forgetting were to occur during retrieval attempts (tests), these attempts would have a positive effect on a

subsequent encoding (study) episode (e.g. Izawa & Estes, 1965); these notions will be addressed further in this introduction, as they share some similarity with the hypothesis that we explore in this work, but follow a somewhat separate research tradition (more focused on developing a mathematical model of repeated study-test cycles on paired-associates learning).

An important theoretical point was made a few years later by Robert A. Bjork (1975), one that despite not having an immediate impact, served as a strong basis for subsequent theorizing on the role of retrieval processes in memory (e.g. Bjork & Bjork, 1992). In a period where Craik and Lockhart's (1972) *Levels of Processing* (LoP) approach was dominating the literature [see Craik (2010, 2002), Roediger & Gallo (2002) for reviews], Bjork (1975) argued that it was not only the ways how we encode information that impact its retention and representation but also the ways we retrieve it from memory, even comparing this phenomenon to Heisenberg principle-derived effects. Based on results on the *negative recency* literature (the last words of a study-list being better recalled on immediate tests, but worse on delayed tests; e.g. Craik, 1970; Bjork, 1968), and on results showing that long-term recall benefits more from an initial retrieval attempt than long-term recognition (e.g. Hogan & Kintsch, 1971; Whitten, 1974) he argued that retrieval, especially in more effortful ways, would alter the way the information was represented, and subsequently its retention as measured by delayed tests. As already mentioned, despite the accumulation of data and theory, interest on effects of retrieval remained scarce.

1.3. Establishing the effect

It was only after more than a decade later, and more data showing the powerful effects of active retrieval on long-term retention (e.g. Cuddy & Jacoby,

1982; Jacoby, 1978; Runquist, 1986), that an article focused on directly investigating the causes, processes and boundaries of this phenomenon had broader impact on field, and started to grab researchers' attention. In an aptly titled work ("The 'testing' phenomenon: Not gone but nearly forgotten"), Glover (1989) sought to test two explanations of the effect: the *amount of processing* hypothesis (an interim test would provide participants a mere extra amount of time for processing of the information before the final test; e.g. Kolers, 1973) and the *number of complete retrieval events* hypothesis (stating that it was the specific processes involved in retrieval that would affect the item's representation in memory, augmenting the probability of later recall). If the *retrieval* hypothesis was more adequate to explain the effect, then 1) taking more interim tests should result in better performance, and 2) tests with differing requisites (e.g. free recall, cued recall, recognition) should also result in differential final performance. To test this, in his experiments, participants were to read a prose passage, and two days later they took either a free recall, cued recall, or recognition test; another two days later they returned to the lab and took a final test (again, either free-recall, cued-recall or recognition). He also manipulated the number of interim tests, and the study-test schedules (maintaining the time of exposure to the items constant). His results showed that taking an initial free-recall test lead to better performance on the final test, regardless of its format, followed by taking an initial cued recall test, which lead to better performance on both cued recall and recognition final tests. Moreover, the results also showed that both taking more tests and following different schedules (capitalizing on the spacing effects; e.g., Landauer & Bjork, 1978; Cepeda, Vul, Rohrer, Wixted & Pashler, 2006) also resulted in better performance. This pattern seemed to fit the *retrieval* hypothesis better than the *amount of processing* hypothesis, in that different test formats, schedules and number

of tests impacted long-term retention, while the amount of time of exposure to the items was kept constant. The publication of this article indeed served to renew the field's interest in testing phenomena, both in theoretical and methodological grounds, while calling for an increased implementation of present knowledge about the effect in educational settings. At the theoretical level, it was in line with general views of memory that incorporated effects of retrieval on memory representations and the role of its difficulty and completeness (e.g. Anderson, 1985; Bjork, 1975); on the methodological level it started to examine the role of the task requisites on the effect (an aspect that is central to our current work). Despite this, the use of prose passages and the specific characteristics of its cued recall tests lead to results that were inconsistent with what were known and robust regularities of human memory: participants performed better on free-recall than on cued-recall, in stark contrast with most memory studies (e.g. Tulving & Pearlstone, 1966), leading some authors to advise caution in interpreting these data, due to peculiarities of the materials and tests (e.g., Roediger & Karpicke, 2006a).

Shortly after, Carrier and Pashler (1992) carried out the most thorough examination of the testing effect by then, investigating the guiding research question of whether it was recall *per se* that caused the testing effects (*cf.* Izawa, 1970), that is, whether retrieval of an item would strengthen its memory trace. The authors pointed to three main methodological problems in the literature that were obscuring critical tests to this hypothesis in studies that directly compared retrieval episodes with re-presentations of the items: 1) the number of items in retrieval conditions (e.g., cued-recall tests) was inferior to the number of items in re-presentation conditions, thus alleviating attentional load and interference (e.g., Allen, Mahler & Estes, 1969); 2) when using free-recall tests as the retrieval condition against which re-presentation

will be compared, participants in the retrieval episode can simply rehearse the recalled items and strengthen inter-item associations, which would be a post-retrieval explanation for the testing effect (e.g., Wenger, Thompson & Bartling, 1980); 3) the way re-presentation conditions were devised (e.g., passive listening of the items from a tape recorder) could encourage participants to slack off, compared to the retrieval conditions, where participants would have to be actively engaged on the task (e.g., Wenger et al., 1980); the authors also suggested that the results showing retrieval leading to poorer performance than re-presentation (e.g. McDaniel & Masson, 1985) could in part be explained by the low retrieval success in the interim tests, which would render them incomparable to the full re-presentation episodes. The authors used paired-associates learning as their basic paradigm (nonsense word-numbers pairs – Exp. 1 – or pairs of English-Yupik translation words – Exp. 2), and compared two schedules of study-test episodes on their benefits in performance in a final test: *pure study* (ST), where both members of the pair were presented together for 10 seconds, and *test trial/study trial* (TTST), where one member appeared on its own for 5 seconds before the target member was presented along for another 5 seconds. In their design, participants would learn all pairs in a first phase (same as a ST trial), after which all pairs would appear three times in ST or TTST conditions (half pairs in each), followed by a filler task, and two final cued-recall tests (5min later, and 24h later). During the interim learning phase, participants were instructed to say out loud the corresponding target stimulus (nonsense/English word member) as fast as they could, which for ST trials involved merely reading the words, while for TTST trials it involved retrieving (or generating; Jacoby, 1978) them. The results showed increased performance in the final tests for TTST trials items, thus demonstrating the *testing effect* and discarding the explanation that retrieval merely reflects an effect of re-

presentation of the materials (ST trials involved longer exposure to the stimuli). Interestingly, the authors also tried to discard another type of testing effect as contributing to the phenomenon: participants could attain knowledge on the item's recallability in the TTST trials, thus adapting their encoding strategies to by allocating more time studying or rehearsing items deemed less recallable, or by rehearsing only TTST items and neglecting the ST ones (see Slamecka & Katsaiti, 1987; deWinstanley & Bjork, 2004). By placing TTST trials at the end of the experiment (Exp. 3; thus discarding the "recallability knowledge" explanation) or by using pure lists in final sublists' study-test cycles (Exp. 4; thus discarding "selective rehearsal" explanation), the authors separated what are now dubbed "direct" and "indirect" testing effects (see Arnold & McDermott, 2013; Roediger & Karpicke, 2006a; for a similar distinction, see Pastötter & Bäuml, 2014).

This article was of upmost importance for the study of the testing phenomena, setting the use of paired-associates with foreign words and their English translation as the preferred paradigm to study the effects of retrieval due to its simplicity and (to a certain point) similarity with some educational materials (e.g., learning the vocabulary of a new language), and also by highlighting the role of the direct effects of retrieval (e.g. Bjork, 1975). Nonetheless, one aspect of Carrier and Pashler's (1992) design could still be confounding the results (at least if the goal was to investigate the "direct" effects of testing): on the testing (TTST) trials, tested materials were always re-presented (i.e., the ST part of TTST trials, where both members of the pair were presented), thus allowing for their restudy after the retrieval attempt (actually, an enhanced re-study opportunity; Izawa et al., 2005). Kuo and Hirshman (1996; Exp. 1) provided a solution for this confound by using a modified Brown-Peterson paradigm (Brown, 1958; Peterson & Peterson, 1959), where three-word lists were presented,

after an initial study episode, in a 16-condition within-participants design, as a restudy (S) or a test (T) trial, with a filler task following each trial, and a final critical test (that was never mentioned in the instructions). After going through these lengths to separate the direct and indirect effects of testing, they replicated Carrier and Pashler's (1992) results, with overall better performance for words in T trials, including when comparing SSSSS and STTTT conditions (evidence that enhanced re-study was not enough to explain the effect). Moreover, more T trials also resulted in better performance, and also, when comparing conditions where the T trials were spaced (e.g., STSST) with conditions where these trials were blocked (e.g., STTSS, SSTTS, SSSTT), a clear benefit of spacing also emerged, suggesting that retrieval in fact involves different processing than encoding, leading to increased retention.

2. Direct and indirect effects of retrieval

2.1. The direct effects of retrieval

Both Carrier and Pashler (1992) and Kuo and Hirshman (1996) provided robust evidence of the superiority of testing over re-study in long-term retention, due to the processing peculiarities of retrieval. For present purposes, it is also interesting to note how they sought to eliminate influences of the test trials on subsequent study trials, the so-called indirect testing effects (although Kuo and Hirshman stated that this was an interesting phenomenon on its own).

The next big wave of interest on the power of testing occurred a decade later, with the publication of an extensive review and an impactful empirical paper by Henry Roediger and Jeffrey Karpicke (2006a,b), focused mainly in the direct (or 'unmediated') effects of repeated retrieval.

In Roediger and Karpicke's (2006b) Experiment 2, participants were to study (relatively) complex prose passages on educationally relevant topics and were to subsequently either re-study them (S) or perform a free-recall test (T) about them, according to three experimental conditions: SSSS, STTT, SSST [this notation differs from Carrier and Pashler's (1992) and Kuo and Hirshman's (1996), as each letter now represents a single study/test trial]. Thus, participants either re-studied the passage three times, were tested on it three times, or re-studied the passage two times and were tested on it once. Afterwards, they were given a final critical test (again free-recall) either 5 minutes or 1 week later. The results showed an interesting and striking pattern: when tested immediately after (the 5 minutes condition), participants in the SSSS condition outperformed their counterparts, followed by those in SSST condition; but when tested a week later, this pattern inverted, with STTT participants largely outperforming SSSS participants, who also performed worse than participants who were tested once (SSST condition). This clearly indicated that retrieval opportunities could afford as much or more long-term learning then encoding opportunities, with the nuance that even one test (this is referring to the SSST condition) could boost performance when compared to purely re-studying the materials. Further evidence was presented in Karpicke and Roediger (2008), where, now using pairs of English-Swahili words [akin to Carrier and Pashler (1992)'s use of English-Yupik pairs], participants were assigned to four 'dropout' conditions: after initial study (same for all conditions), participants would complete further study-test cycles, where the correctly retrieved items would be present or absent of study lists and tests. In the standard condition (ST) participants would study a list of pairs, then would respond to a cued-recall test, then study the full list again, then respond to the same test, for four study-test cycles; in the other conditions, the correctly recalled

items could either be dropped from the study list but appear on the tests (S_NT condition), appear on the study list but be dropped from the test (ST_N condition), or be dropped both from study lists and tests (S_NT_N condition). A final test was administered one week later. Strikingly, results showed overwhelming advantage for the conditions where full testing was employed (ST and S_NT; $M \sim .80$) over the ones where the recalled items were dropped from testing (ST_N and S_NT_N; $M \sim .35$) even though participants in all conditions exhibited similar cumulative learning curves, and achieved maximum performance on the last interim trial. One of the most surprising aspects of these results was that additional re-study didn't seem to produce any benefits in long-term retention (e.g. ST and S_NT yielded similar final results, despite the fact that in ST participants had approximately 80 more study opportunities), thus describing retrieval as an extremely powerful way to boost learning.

For present purposes, it is interesting to notice, though, some oddities about Karpicke & Roediger (2008)'s results. One of them is the near ceiling mean performance rates on the "testing" (ST and S_NT) conditions compared to the "study" (ST_N and S_NT_N) conditions; another one is the inexistence of differences between the ST_N and S_NT_N conditions, as it goes against evidence showing test potentiation effects (i.e., testing should enhance subsequent study episodes, and this would be particulary noticeable in the ST_N condition where participants had opportunities to study the whole set of items; e.g. Izawa, 1966; Arnold & McDermott, 2013; Soderstrom & Bjork, 2014). In fact, Soderstrom, Kerr & Bjork (2016) reasoned that the between-participants design employed by Karpicke and Roedgier (2008) could be obscuring the role of one potent mediated/indirect testing effect, spacing (e.g. Cepeda et al., 2006), due to the different schedules of study and test episodes and their frequency, resulting from the dropout of items. By employing a within-participants version of

Karpicke and Roediger (2008)'s design – items were treated as ST, S_NT, ST_N and S_NT_N within each trial – Soderstrom et al., (2016) still obtained results showing the potent effects of testing on long-term retention (ST and S_NT items resulted in better performance), but now showed how repeated re-study also benefited retention, with ST_N items being better recalled than S_NT_N items. This results highlight the importance of considering higher-order sets of influencing variables in understanding how retrieval impacts memory as whole, going beyond the mere timecourse of specific items that are repeatedly studied or tested, and also assessing it's contextual, procedural and strategic facets.

These aspects fall under the umbrella-term "indirect testing effects" (Roediger & Karpicke, 2006a; Arnold & McDermott, 2013) or the related term "forward testing effects" (Pastötter & Bäuml, 2014). As the central hypothesis on the present work concerns the impact of the specific retrieval requirements on subsequent study and test episodes with similar but new information, we next review some of the literature on some of these effects, *en route* to a clearer definition and distinction of our proposal.

2.2. The indirect effects of retrieval

As already discussed here, several decades before the "testing effect" occupied a prominent place in the spotlights of human memory research, Chizuko Izawa presented a series of experiments testing the hypothesis that while no learning was to occur during retrieval, retrieval attempts would affect subsequent study episodes of the same information, even in the absence of feedback or retrieval success (e.g. Izawa, 1966; 1970; Izawa et al., 2005). She dubbed this phenomenon "test-potentiated learning". But just as indirect/mediated effects were hidden but at play in Karpicke

and Roediger (2008)'s design (Soderstrom et al., 2016), the same can be said about Izawa (1966; 1970; 1971)'s paradigm. In her series of experiments, participants were to learn number-letter paired items in multi-trial cued recall tasks, and the number of re-study and test episodes was manipulated. One of the main findings from this line of research was that, while controlling for the number of preceding study trials, conditions where more interspersed test episodes occurred lead to faster learning (the slope of learning curves was steeper). With the then-widespread assumption that no learning occurred during retrieval episodes, she attributed this increase in learning rates to beneficial effects of retrieval attempts of some items on their subsequent encoding (also, performance across consecutive testing trials didn't improve). As evidence supporting direct effects of testing mounted (i.e., retrieval, apparently, is not a learning-neutral event), a better assessment of test potentiation effects and their disentanglement from direct testing effects was provided by Arnold and McDermott (2013), who analyzed performance on test trials that followed re-study trials with varying previous test trials (1 or 5, manipulated between-participants) conditional on whether the items remained recallable from the previous test trial, or whether they were not recalled in a previous test but after restudy they were so. Arnold and McDermott replicated Izawa's results (now with educationally-relevant word pairs) and further disentangled the testing effect from the test potentiation effect: more preceding test trials lead to better performance after re-study, even when only analyzing items that were newly retrieved after re-study.

While there are still some possible confounds in Arnold and McDermott (2013)'s results (e.g. between-conditions differences related to the spacing effects; Cepeda et al., 2006), the message to researchers on the direct effects of testing is clear: direct effects are hardly alone, and considering the indirect/mediated effects is

crucial for a better understanding of the phenomenon both at basic level as in more applied settings.

Indeed, further results illustrating ways in which retrieval can impact further encoding and retrieval operations also accumulated. For example, Szpunar, McDermott and Roediger (2007, 2008) showed how repeated retrieval (free-recall) after study, and not repeated re-study, protected participants against the hindering effects of the buildup of proactive interference (PI; present learning being negatively affected by the interference of prior learning, usually after prolonged and serial learning of the materials; Postman & Keppel, 1977; Underwood, 1957), resulting in fewer intrusions and better performance both in a surprise test for all participants about the last studied list, as in a final cumulative test. Importantly, these results were obtained using new materials on each study-test cycle. Szpunar et al., (2008) showed that this effect occurred both with unrelated and related (different lists drawn from the same taxonomic category) materials, and Nunes and Weinstein (2012) replicated this effect but using DRM (Roediger & McDermott, 1995) materials, with retrieval resulting in less intrusions from the critical items. In this line, Bäuml and Kliegl (2013) presented evidence that retrieval after study was superior in protecting against the buildup of PI than other known methods, using participant's response latencies as proxies for the release in PI: when experienced retrieval, participants showed shorter response latencies, interpreted by the authors as resulting from reduced memory search sets size (see Wixted and Rohrer, 1993), and thus as less interference from previous materials.

2.3. Indirect effects of retrieval: how tests impact encoding

Some explanations for this effect have been put forward, either focusing on encoding processes or retrieval processes (e.g., Pastötter & Bäuml, 2014) but direct evidence for either has not been put forward, at least in a way that successfully disentangles these memory phases. For example, some authors consider that retrieval after study helps participants segregate lists via a mental context change, which would allow participants, in a subsequent test, to capitalize on the specific contextual cues previously generated to boost performance (e.g. Howard & Kahana, 2002; Bäuml & Kliegl, 2013). While this hypothesis can account for the "release from PI" data (i.e., it fits the results indicating fewer intrusions from a previous learning episode), it is less adequate to cases where PI is not prone to emerge (e.g., Arnold & McDermott, 2013), thus constituting a narrower explanation for what seems to be a more general effect. Additionally, it is somewhat similar to "event boundaries" effects (e.g., Pettijohn, Thompson, Tamplin, Krawietz & Radvansky, 2016; Sargent et al., 2013; *cf*. Pettijohn & Radvansky, 2016), thus placing the role of retrieval in the same level of any other process, event, or behavior that might function as signaling a boundary or reason for segregation in materials by binding it to a specific context. While this factor might be of importance for the forward effects of testing, it surely doesn't tell us the whole story.

Another class of explanations for the beneficial impact of retrieval in the release from PI puts more emphasis on the impact of retrieval on subsequent encoding processes. For example, Wissman, Rawson and Pyc (2011) replicated Szpunar et al., (2008)'s results using sections of complex prose passages, and shown that the superior performance of participants that had to recall the studied information before the next study episode was not merely result of engaging in an intervening activity

(study, interim math tasks and unrelated content practice tests resulted in poorer performance than recall testing; Exps. 3, 4). Relevant for the present thesis is the fact that in Experiment 4, Wissman et al., (2011) included a condition where participants were only to study and recall the final section of the prose passage, a condition where PI was not to emerge. Performance for this group was also worse than for the interim tests group, thus suggesting that the performance boosting effects of retrieval go beyond the mere release from PI. The authors suggest that in this paradigm, modulation of the study strategies by retrieval was very likely to be at play (e.g., generating mnemonic mediators; Pyc & Rawson, 2010).

Some recent findings point to a more adaptive process-based (but still task-dependent) phenomenon. For example, when compared to re-studying, testing one's memory has been found to result in better meta-cognitive calibration (e.g., Karpicke, Butler & Roediger, 2009; Finn & Metcalfe, 2007) allowing one to correct one's previous mistakes in a subsequent test (e.g. Amlund, Kardash & Kulhavy, 1986), make more accurate predictions of future performance (e.g., Koriat & Bjork, 2005), which can result in subsequent differing time allocation in studying information that a previous test showed was not correctly recalled (e.g. Soderstrom & Bjork, 2014). In fact, Soderstrom and Bjork (2014), using a paired-associates learning paradigm where pairs varied in their associative directionality[1], showed how participants became aware of gaps in their knowledge when they were tested, and this lead to longer study-times both overall but especially on items that were not recalled on a previous interim test; this didn't happen when the items were merely re-studied. Also, participants showed increased meta-cognitive accuracy by devoting more time studying *backward*

[1] *Forward* pairs have a strong association from cue to target (e.g. umbrella-rain); *Backward* pairs have a strong association from target to cue (e.g., rain-umbrella); *Forward* pairs tend to result in better performance, but participants are usually unaware of that when studying them (e.g., Nelson, McEvoy & Schreiber, 1998; Koriat & Bjork, 2005).

pairs after an interim test but not after restudy, the same happening to completely unrelated pairs even to a larger degree. Additionally, this effect also occurred for initially studied but non-tested items (Exps. 2-4), suggesting that the testing opportunities allow for the learning of general stimuli characteristics and their intricate relationship with the test requirements, and not only about the specific tested items. Important to the present work is how it appears that participants, by experiencing retrieval, became sensitive to a rather subtle manipulation of the materials, the associative directionality, as proxyed by the further allocation of study time to unrelated and *backward* pairs (overall, study-time allocation grew monotonically with pair-type, while usually there are no differences in participants' judgments of future recallability by pair-type; Koriat & Bjork, 2005, 2006). So not only participants showed sensitivity to the specificities of the materials in their relationship with the task requisites (recall), but they also adapted their encoding strategies towards dealing with a characteristic in the materials that was an obstacle to good performance at test.

Another line of research that points in this direction is Elizabeth Bjork and collaborator's work on how experiencing the *generation effect*'s (Jacoby, 1978; Slamecka & Graf, 1978; for a meta-analysis see Bertsch, Pesta, Wiscott, & McDaniel, 2007) benefits in performance on a test leads participants to adapt their encoding strategies in a subsequent study-test cycle with the same structure (deWinstanley & Bjork, 2004; Bjork, deWinstanley, & Storm, 2007; Bjork & Storm, 2011; Storm, Hickman & Bjork, 2016)[2]. In deWinstanley and Bjork (2004; Exp. 1), participants had

[2]While the *generation effect* and the *testing effect* can seem related or overlapping at face value, some authors contend they refer to different aspects of tasks and instructions (e.g. Carrier & Pashler, 1992; Karpicke & Zaromb, 2010); this distinction is out of the scope of the present work, and we'll review Elizabeth Bjork and collaborator's work on this issue as it is informative of adaptation to task characteristics.

to study two prose paragraphs, phrase by phrase, where to-be-remembered critical words were highlighted and were either intact (*read* items; participants had to read them and write them down) or missing letters (*generate* items; participants had to use the remaining letters and the phrase context to generate the word and write it down) and in subsequent fill-in-the-blank tests they saw several studied phrases with the target words removed, which they were to recall. As expected, performance was superior for *generate* words than for *read* words, the so-called *generation effect*. Interestingly, when they were given another passage to learn – again with the generate/read manipulation on critical items – and the corresponding fill-in-the-blank test, the *generation effect* was eliminated, which lead the authors to speculate that participants altered their encoding strategies in a way that *read* words were processed more effectively. The authors' explanation was that it was the awareness of the relative benefits of generating words that lead participants to encode *read* items the same way they encoded *generate* items on the second passage, namely, that they encoded the words more according to the textual context – a strategy that would help them "solve" the *generate* trials. Accordingly, when the generate/read manipulation was between paragraphs (i.e., the first paragraph, presented phrase-by-phrase, contained only one type of critical item, and a second one the other type; Exp. 2), so that participants were less likely to experience the relative impact of the manipulation on their performance, the *generation* effect persisted on the second study-test trial. The same pattern of results emerged when the generate/read manipulation was between-participants (Exp.3), and when free-recall tests were used instead of fill-in-blank (Bjork & Storm, 2011). In a recent exploration of this phenomenon's boundaries and necessary/sufficient conditions, Storm, Hickman and Bjork (2016) devised a paradigm where, manipulated between participants, (a) one condition was a

direct replication of deWinstanley and Bjork (2004)'s Exp. 1, but in two other conditions the first fill-the-blank test was substituted with a handout asking participants to imagine that they would take a fill-the-blank test about the studied paragraph, complete with examples of the type of questions in such test, and participants were asked to predict whether they would be better at recalling the generated or the read target words (with the majority predicting better performance for generated items). After this, in one condition (b) they were asked to explain why they made that prediction, and in the other (c) they received information about the relative benefits of generating vs. reading for memory performance. The results showed that only participants who actually took the first test managed to eliminate the *generation effect* via better performance with read words, while participants in the other conditions, despite having explicit knowledge of how to better encode read items, were not able to do so. On a second experiment, Storm et al., (2016) selected the items that granted better and worse performance regardless of target condition, and used them to create "advantaged" generate and "advantaged" read conditions, where participant's performance on each type of target was determined to be favored. As expected, performance on the first fill-the-blank test followed this manipulation, with participants in the "advantaged" read condition attaining better performance for read than for generate items. Interestingly, along with a replication condition (as in their Exp. 1), both "advantaged" conditions showed the attenuation of the generation effect on the second passage. This seems to tell us that test experience *per se* is more potent than instructions or explicit information about the characteristics of the stimuli-test relationship. It also points to the importance of test expectancy (e.g., Lundeberg & Fox, 1993; Finley & Benjamin, 2012) in the modulation of subsequent encoding strategies: experiencing the test and expecting for a similar one in the future.

It is the central thesis of the present dissertation that retrieving information from memory in a given context will afford the learner with knowledge regarding the specificities of the retrieval task, allowing for the adaptation of subsequent encoding (and retrieval) strategies. We believe that this framing goes beyond the notion of test expectancy (despite some overlap), as while the test expectancy literature weights more on explicit acknowledgement of the efficacy of encoding strategies to a clearly defined test format and conscious strategic choices leading to better performance (e.g., Finley & Benjamin, 2012; see also Hertzog, Price & Dunlosky, 2008), we are interested in the adaptive character of human memory in responding to retrieval requisites by learning how the studied information's features and the retrieval contexts fit together for good performance. In this regard, two methodological features of the test expectancy research are apt examples of how this line of research differs from the present one: 1) most studies directly manipulate expectancy via explicit instructions (for a review and some exceptions, see Whitten, 2011) and 2) participants are often probed for meta-cognitive judgments, such as judgments of learning (JOLs) at various points across study-test cycles (e.g., Finley & Benjamin, 2012; Hall, Grossman & Elkwood, 1976; Leonard & Whitten, 1983). This distinction will be further discussed in Chapter II.

3. Sensitivity to the relationship between task requisites and stimuli dimensions

3.1. On the relationship between memory and conceptual learning

To better understand what is learned about retrieval requirements via retrieval experience, one has to refer to an unfairly overlooked article by Higham and Brooks (1997). The authors were investigating how participants in memory and

categorization experiments are sensitive to selection rules and contingencies in stimuli sets, using this knowledge to inform responses in memory tasks and using their own episodic status judgments to inform categorization responses regarding the stimuli structure. In Experiment 1, participants were instructed to study a list of words, which were selected using non-salient criteria: word frequency, word class (e.g., being a noun), and number of letters. These words were studied under a regular *levels of processing* (LoP) manipulation: shallow vs. deep encoding instructions (e.g., counting vowels vs. rating their level of understanding; see Craik & Lockhart, 1972; Roediger, Gallo & Geraci, 2002). In the test phase, participants had to complete both a recognition memory test and a classification test, on two lists that contained three types of items each: old (appeared on the study list), new-consistent (distracters consistent with the three selection rules) and new-inconsistent (distracters consistent with one of the selection rules, but inconsistent with the other two). The order of the tests was counterbalanced and in the recognition task participants had to rate a word on how likely it was that it appeared on the study lists, whereas in the classification task participants were instructed that the studied materials had to meet some set criteria to be selected, and that they had to rate the test words' consistency with those rules. While these two type of tests are often treated as two separate worlds, the results point to some similarities between the two, and suggest that people engage on recruiting episodic knowledge in tasks that rely on conceptual knowledge and vice versa. First, discrimination (as measured by A'; see Snodgrass & Corwin, 1988) was above chance for both tasks, showing that participants were both sensitive to the words' episodic status and to the rules employed in the selection study materials. Additionally, the mirror effect (as it applies to this experiment, the LoP manipulation resulting in more 'old' and less 'new' responses in the deep processing condition;

Glanzer & Bowles, 1976) occurred for recognition – as expected – but also for classification, with the LoP manipulation resulting in higher 'old' responses, but lower 'new-consistent' responses, which should not have happened if classification had relied purely on the commonalities activated during study and the items' episodic status did not matter. Moreover, collapsing across LoP conditions, the authors found what they called an "episodic effect" in the classification task ("old" vs. "new-consistent" discrimination; old items were more likely to be deemed consistent with the rule), and a "structural effect" in the recognition task ("new-consistent" vs. "new-inconsistent" discrimination; items consistent with the rule were more likely to be deemed old). This pattern suggests that episodic and conceptual knowledge can, given the chance, interact and even replace one another when responding to a task.

These results are informative to several memory phenomena by illustrating a non-mnesic component of peculiar effects in memory tasks. One of these cases is the broad subject of false memory (e.g., responding "old" to items when they are "new" in recognition tasks, or producing intrusions in free-recall tasks). One example comes from the classic leading-questions task (e.g., Loftus, 1975, 1979; Ceci, Ross & Toglia, 1987) where participants, after seeing a video depicting, for example, car crashes or classroom disruptions, are immediately asked a question that either contains true presuppositions or not. Later in time (e.g., one week), when they are asked a different question regarding the video, participants who were in the experimental (leading-questions) condition falsely report having seen objects or characteristics that were only presupposed (i.e., they were not the focus of the question, nor actually present at the studied scene) by the previous question. While some explanations for this phenomenon state that the leading-questions task alters participants' memory itself (e.g., Loftus, 1979), others highlight the role of the task structure in the critical test as

potent factor driving the effect: in some cases (e.g. Ceci et al., 1987) the final assessment test (after the leading-questions phase) contains both the target materials and the foils that were suggested by the leading-questions. The problem here is that in the leading-questions phase participants in the control condition have an opportunity to correctly re-study the target materials via the leading-question itself (e.g., "Do you remember the girl with the hat?", considering the hat in the video was red) while this doesn't happen for the experimental condition (e.g., "Do you remember the girl with the green hat?", considering the hat was red). If the final test to assess the influence of the leading-questions is composed by the target material, and by the foils related to the experimental condition (e.g., green hat), one interpretation is that, besides the opportunity to re-study target material only being present for the control condition (i.e., asking participants whether they remember the girl with the hat will prompt them to remember the actual red hat), the distracter material appears once in the leading-questions phase only for the experimental condition, and then reappears in the final test. This characteristic of the design allows the control participants to reject the distracter material more easily due to its novelty making it stand out, without the need for correct retrieval of the target information, while this structural feature is absent for the experimental condition (McCloskey & Zaragoza, 1985; Brainerd & Reyna, 1988). This can be considered a form of previous knowledge that control participants acquire and capitalize on, that is not attainable for the experimental participants. This aspect is in line with further results from false memory and prototype memory research were previous knowledge interacts with the test structure in meaningful ways (e.g., Posner & Keele, 1968; Roediger & McDermott, 1995), that is, when the relationship between the materials and the task allows for a classification of deviations from a previously learned concept or rule. If one thinks of classical conceptual learning effects (e.g.,

Bruner et al., 1956; for a review see Murphy, 2002), correctly classifying new instances as members of category due to similarity or overlap with other exemplars or a central rule (for discussions on exemplar and rule-based theories of categorization, see Erickson & Kruschke, 1998; Rouder & Ratcliff, 2006) is a signature of learning. Analogously, in a memory task, one could classify a new instance (e.g., a distracter) as similar to a set of studied exemplars, or its commonalities, and we usually conceive this as a false alarm, but a false alarm that can signal learning.

3.2. On the nature of retrieval-induced strategy adaptation

Our proposal stems from this ideas: repeated retrieval will allow participants to attain knowledge (tacitly, at least) regarding the retrieval requirements, and this will lead to an adaptation of processing strategies in subsequent learning episodes, by capitalizing on that knowledge. We propose that this results in processing similar new stimuli attending to dimensions that will benefit performance on the expected test, and disregarding or inhibiting the dimensions that are deemed relevant. This notion of dimensions derives from Underwood's (1969) and Garner's (1978) conceptualization of memory as a collection of attributes, i.e., stimuli can be processed, stored and represented using different types of information about them. Our position is that experiencing retrieval with specific requisites will modulate which dimensions will be attended to and which will be disregarded.

Importantly, this idea fits in one of the most overarching and successful frameworks of human memory, Bjork and Bjork's (1992) "New Theory of Disuse" (see also Bjork, 2011). This theory states that 1) human memory has a (virtually) infinite storage capacity but a very limited retrieval capacity, and 2) there are two factors of memory "strength": storage strength (SS; how well learned or

interassociated information is) and retrieval strength (RS; how accessible or retrievable information is in a given point in time). This is parallel to the classic distinction between learning and performance (for a recent review see Soderstrom & Bjork, 2015). The ability to recall is considered to be entirely determined by the present retrieval strength, while storage strength is considered a latent variable that may act in delaying forgetting (loss of retrieval strength) and in modulating the gains of retrieval strength. In this view, while not remembering information we know we have in our memories might be frustrating, hypothetically remembering every piece of information we have stored would be unbearable (Bjork, 2011). In other words, it is highly undesirable to have big amounts of information available to recall, as this would be highly costly resource-wise (proactive interference would accumulate to unpractical levels; Bjork, 1972) and also would be counterproductive in terms of keeping memory current and adapted to one's present state and context. Thus, forgetting, or the inability to recall, should be considered not as a cognitive frailty, but as the product of and adaptive and self-updating mechanism (Bjork, 1989; Shiffrin, 1970). In that way, the act of recalling will both be a product of memory selection and production, but also of the inhibition of concurrent memories that are associated and can be elicited by the same set of cues (e.g., Anderson & Bjork, 1994; Anderson, Bjork & Bjork, 1994). So forgetting, broadly speaking, can only be conceived as a relative deficit, as it can be associated with contextual and goal-oriented factors: forgetting (due to inhibitory mechanisms) can and often is the correct and adaptive output in learning situations.

Returning to the parallels between conceptual learning and memory, one can also consider false alarms as adaptive and intelligent responses to current retrieval contexts, and a sign that learning as occurred. For example, in DRM research (Deese,

1959; Roediger & McDermott, 1995) falsely recalling or recognizing the critical word only occurs if participants are able to process the commonalities or common core of the studied items (e.g., Brainerd & Reyna, 1998; Roediger, Balota & Watson, 2001), but these false alarms are reduced in cases where learning is to be impaired, (i.e., the conceptual common core of the lists is less likely to be extracted) as with children tested with adult lists (e.g., Brainerd, Reyna & Forrest, 2002; *cf*. Carneiro, Albuquerque, Fernandez & Esteves, 2007), dementia patients (e.g., Balota, et al., 1999), and young adults under cognitive load (e.g., Seamon, Luo & Gallo, 1998) or tested with lists with weaker association to the critical item (e.g., Roediger, Watson, McDermott & Gallo, 2001). In this sense, not only forgetting or inhibiting aspects of our memories is important to respond to current context, but also memory errors such as congruent or related false alarms can indicate that one remains updated and aware of the invariants and cues in the environment.

The hypothesis that we explore in the present work follows the aforementioned principles, as we consider that by experiencing retrieval one shifts processing strategies towards present and expected requisites, modulating the type of information that is sought after in subsequent learning episodes. This can have positive consequences in terms of memory performance, but also negative ones, as retrieval practice can make us disregard stimuli dimensions that are not useful for performance in the task at hand, but in the long term, this inhibition can mean that we will disregard non-focal information that might be useful, or that can be a potential source of new learning.

CHAPTER II. Adapting to the test structure: Letting testing teach what to learn. Garcia-Marques, Nunes, Marques, Carneiro & Weinstein (2015). *Memory, 23*(3), 365-389

Human memory did not evolve to give optimal performance in our silly laboratory task[s]

Anderson (1990, p. 42)

Introduction

The idea that we will fail to grasp what memory *is* without knowing what memory *is for* is hardly new. Several authors have presented this idea before (Bjork & Bjork, 1992; Neisser, 1978) others have developed specific (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) or even quite general models (Anderson, 1990) around it. In this paper, we join these authors by proposing a narrow version of the same idea. We propose that we encode and store information as a function of the particular ways we have used similar information in the past. More specifically, we contend that the experience of retrieval can serve as a powerful cue to the most effective ways to encode similar information in comparable future learning episodes.

In order to explore these notions, we designed two experiments in which all participants went through multiple study-test cycles of single-category lists with different types of tests. Each test involved recognition, but either included only same-category lures or only different-category lures. From here onwards, we refer to this manipulated feature as "test structure". We expected that the experience of multiple study-test cycles (i.e., repeated testing) would lead to adaptive changes in subsequent

performance as long as test structure remained constant; but this adaptive change would also carry a cost that would be detectable whenever the test structure changed.

As our goals and methods combine ideas that have been explored under the rubrics of test expectancy (e.g, Balota & Neely, 1980; Finley & Benjamin, 2012), the power of testing (e.g., Landauer & Bjork, 1978; Roediger & Karpicke, 2006), and conceptual learning effects (Higham & Brooks, 1997; Kantner & Lindzay, 2010), we will first briefly refer to these literatures, both to review previous findings on the effects of test experiences on subsequent encoding, and to describe how our studies differ from the studies carried out under these rubrics.

Testing expectancy

The first studies identifying test expectancy effects were reported by Meyer (1934; 1936). Meyer's findings were intuitive and compelling: students expecting an essay test always performed better than those expecting a multiple-choice test, regardless of final criterial test format. The take home message appeared to be clear – students should always study as if for an essay test, even when they expect a different kind of test. Further lab research used practice tests to induce expectancy of a final criterial tests, and roughly equated essays with free recall and multiple-choice with recognition. Findings from these studies either converged with the conclusion that expectations of a recall test produced superior performance in all cases (e.g., Balota & Neely, 1980; Hall, Grossman, & Elwood, 1976, exp. 1; Neely & Balota, 1981; Schmidt, 1983; Thiede, 1996) although some studies only found the recall expectancy superiority in recall performance (e.g., Connor, 1977; d'Ydewalle, 1981; Hall, Grossman, & Elwood, 1976, exp. 3; Maisto, DeWaard, & Miller, 1977). Puzzling enough, the picture is even less clear-cut in field studies, where a test expectancy

match benefit is more often found than recall expectancy superiority effects - students expecting an essay do better at essay tests than students expecting multiple-choice test; whereas the converse is true for multiple-choice performance (for a meta-analysis that includes both lab and field studies, see Lundeberg & Fox, 1991)[3]. With some exceptions (e.g., Finley & Benjamin, 2012; Leonard & Whitten, 1983; Tversky, 1973; Whitten & Leonard, 1980), the test expectancy literature has been targeted at exploring and accounting for Meyer's initial findings, either in lab studies that induced recall or recognition expectancies by practice tests or in field studies that used the supposedly more naturalistic test counterparts, namely essay and multiple-choice tests and induced test expectancy by manipulating practice tests, instructions or both. In our case, although we use practice tests to induce test expectancies just as many of the lab studies in this area, we are not interested in how test expectancies affect test performance per se, instead we are more interested in exploring how experience with testing leads to more adaptive encoding of subsequent information for the same type of test. More specifically, we are interested in showing how test expectancy and beliefs about test structure can influence future study episodes and not only performance on tests.

The power of testing

Since its inception (Ebbinghaus, 1885), the scientific study of memory has maintained a sharp contrast between the phases of study and test. Learning would – ideally – occur only during the study phase and testing would serve exclusively to provide a more or less sensitive assessment of learning. It is true that over time there

[3] There were however a few exceptions, even for the more restricted version of Meyer's hypothesis (e.g., McDaniel, Blischak, & Challis, 1994).

have been a few critical voices speaking up against this simplistic view (e.g., Bjork, 1975; Izawa, 1971), but by and large in the memory literature, the investigation of encoding has been separated from the study of retrieval. Moreover, research on encoding processes has predominated over research on retrieval processes. Recently, under the concept of the "testing effect" (for recent reviews, see Roediger & Butler, 2011; Roediger, Putnam, & Smith, 2011), the literature has uncovered a host of processes that occur at testing and that can modify, add, and generally boost learning. For instance, Karpicke and Roediger (2008) have shown that repeated retrieval practice (in the absence of new study opportunities) enhances long-term learning much more than repeated studying of the same information (in the absence of repeated testing).

Our studies, however, are not concerned with the benefits of repeated testing of the same information and long-term retention, but with the changes induced by repeated testing in the encoding of new information. That is, we are interested in test potentiation effects, a much lesser explored subject. Research on test potentiation is usually related to how unsuccessful retrieval attempts enhance learning by improving encoding of the same or related items in subsequent study episodes (Izawa, 1971; Finley & Benjamin, 2012; Karpicke, 2009; Szpunar, McDermott, & Roediger, 2007; Wissman, Rawson, & Pyc, 2011). Our studies depart from this more usual line of research in the sense that we aim to study how the development of strategies during retrieval can change encoding of a different set of items. Despite this difference, we are still interested in a mediated testing benefit and not just in the direct benefits of testing.

Mediated and direct effects of testing have often been treated in an undifferentiated way and only recently have testing and test potentiation effects been

experimentally isolated (see Arnold & McDermott, 2013) and the possible processes by which potentiation occurs clarified (Grimaldi & Karpicke, 2012). Our aim with this study is precisely to enlighten some of the mechanisms that might be responsible for some forms of the test potentiation effect, namely with novel items.

Conceptual Learning

In our studies, although all study lists correspond to exemplars of a given category, in one condition, the recognition test includes only lures from a different category. In this sense, our recognition test becomes very similar to classification tasks used in concept learning (Bruner, Goodnow, & Austin, 1956; for an extensive review, see Murphy, 2002; for a review on implicit learning, see Seger, 1994). In classification tasks, participants have to learn to discriminate items pertaining to different categories (corrective feedback is usually provided during discrimination). Performance is usually assessed by classification accuracy of new exemplars. In recognition memory tasks participants have to discriminate items pertaining to the study list from lures (new items), and corrective feedback is usually not provided during discrimination. Performance is usually assessed by identification accuracy ($P_{(\text{“old” / “old”})} - P_{(\text{“old” / “new”})}$). But classification and recognition tasks can be very similar because participants in classification tasks are sometimes asked to study examples of category A, and only later are they asked to classify old and new items members as members of category A or B (non-A). In classification tasks, the discrimination criteria are feature-based, corrective feedback is usually provided, and the learning episodes are not clearly separated from the testing episodes as usually occurs in memory tasks. In recognition memory tasks, the discrimination criteria are episodic, special care is taken to prevent feature-based learning (i.e., list items should

match foils in every respect other than their encoding status), corrective feedback is not usually provided, and the separation between learning and testing phases is clear-cut, with the tasks for each phase being clearly different. These differences conspire to allow for little concept or feature-based learning to occur, or at least to be ascertainable in recognition tasks. This is unfortunate, because memory and concept-based learning are likely to be closely interwoven in the real world. In fact, the literature has already documented a host of effects of knowledge on memory (e.g., prototype memory effects and false memory effects based on previous associations or knowledge, see Anisfeld & Knapp, 1968; Brainerd & Reyna, 1988; Posner & Keele, 1968; Roediger & McDermott, 1995; Underwood & Freund, 1968) and several theoretical approaches have already attempted to delineate a more meaningful relationship between memory and concept learning (Hintzman, 1986; Brainerd & Reyna, 2001). More specifically, Higham and Brooks (1997) have obtained classical recognition memory effects in a classification task and classical classification effects in a recognition memory task. Thus, in a classification task, participants classified items as better category exemplars when they had a positive episodic status (i.e., they were "old") or when they were deeper processed (see Craik & Lockhart, 1972). Moreover, these results reversed for new items such that a "mirror effect" was obtained (see Glanzer & Adams, 1985). On the other hand, in recognition tests, these authors obtained common classification effects, such as centrality effects – false recognition of new exemplars increases whenever their features are congruent with the features of old items. Given these findings, we believe that inducing test expectancy by retrieval practice will allow us to uncover systematic concept learning effects in recognition and to further characterize the relationship between concept learning and memory.

Experiment 1

In Experiment 1, participants went through four study-test cycles. Each study list was formed from exemplars of a common category, with a different category used for each study-test cycle. For the two between-subjects conditions, the first three cycles differed only in the nature of the recognition test. In the Related Lure condition, all lures were exemplars of the category presented in the study list, matched to the presented items in terms of frequency of production – i.e, mean number of times the item is produced in a category generation task. In the Unrelated Lure condition, all lures were exemplars of a different category (one category per list) from the one presented in the study list. In the last study-test cycle, participants in both conditions received a test like the ones participants had received in the Related Lure condition (i.e., with all lures being exemplars of the same category presented in the study list). Our goal was to examine how participants adapt their encoding strategies to the requirements of the situation in which the learned information is to be retrieved. In this regard, the two groups of participants stood in stark contrast. For the first three study-test cycles, for participants in the Unrelated Lure condition, the identification of the category of the exemplars included in each study list was crucial and sufficient to assure good performance in the subsequent recognition test; whereas for participants in the Related Lure condition, identifying the category to which all of the items belonged was of little value. More specifically, in the Related Lure condition, category knowledge had no diagnostic power when applied to the recognition test because the lures were matched to the presented items in terms of category membership and frequency of production (i.e., centrality). Thus, participants in the Unrelated Lure condition should make their recognition judgments based on the category membership of test items, whereas participants in the Related Lure condition

should base their recognition judgments on the distinctive features of study list members. So, we expect that in test 4, participants in the Unrelated Lure condition fail to adapt to the new test and show more false alarms and possibly less hits than participants in the Related Lure condition, who should already have adapted their encoding strategy to the type of test.

Although we are mostly interested in the performance on test 4, we are confident that recognition performance on the first three tests will be affected by our test structure manipulation, and that participants in the Unrelated Lure condition will greatly outperform participants in the Related Lure condition on these three tests. Also, a more interesting question about performance on the three first tests arises - will the difference in the types of lures included in the recognition test affect criterion in a contrasting way. More specifically, participants in the Unrelated Lure condition should exhibit a highly restrictive criterion in that they should reject all the lures (which are from a different category than the study list items) as well as the least typical studied items because low centrality items are often not recognized as category members (Greenberg & Bjorklund, 1981; Loftus, 1975; McCloskey & Glucksberg, 1978). On the other hand, participants in the Related Lure condition should exhibit the opposite response tendency (i.e., a lenient criterion) because all items on the test are from the study list item category. This leniency bias should be greater for high centrality relative to low centrality items. It is important to note that the stability of the criterion used across the study-test cycles should differ between conditions. Participants in the Related Lure condition should gradually learn to avoid using conceptual knowledge to inform their recognition judgments. Thus, their criterion should become less lenient across the four study-test cycles and the centrality bias should also decrease across cycles. On the contrary, participants in the Unrelated Lure

condition should maintain a stable criterion until the 4th test, which involves a sudden change in the nature of the test (where all lures belong to the study list item category). This change in the nature of the lures should lead to a dramatic change in criterion for participants who had experienced only unrelated lures on previous tests.

Method

Participants. Eighty-eight Lisbon University undergraduates participated in the experiment for payment. Forty-nine participants were assigned to the Unrelated Lure condition (tested with unrelated lures in tests 1-3) and 39 were assigned to the Related Lure condition (tested with related lures in tests 1-3).

Materials. Four categories were selected from the Portuguese category norms (Pinto, 1992) – animals, fruits, occupations, and body parts – to construct the categorical lists and the related lures. These four categories were chosen because they were the categories with the greatest number of exemplars with a category frequency (i.e., probability of being generated as a category exemplar) above 5%. For each category, the 45 most frequent exemplars were selected. Exemplars with more than one word were excluded

For each category, three study lists were built so that, when ordered in terms of decreasing frequency, items in positions 1, and 4; 2, and 5; or 3, and 6 were not presented and could be used as lures in the recognition tests for participants in the Related Lure condition. As list words and lures were counterbalanced between participants, three different study lists per category were created, and each participant studied one of these lists per category. Each study list was composed of 30 exemplars of the same category randomized afresh for each participant. For each study list, two types of recognition tests were created. Both types of test included 15 out of the 30

studied words. But the Related Lure recognition test included the 15 unstudied category members of the studied category, whereas Unrelated Lure recognition test included 15 unstudied and unrelated words, selected from 15 other nonpresented categories. So, in both conditions, each recognition test included a total of 30 words. The unrelated lures used in the three recognition tests of the Unrelated Lure condition were drawn from a set of 45 exemplars of 15 nonpresented categories. For each recognition test this set was divided in three sets, so that 15 lures of 15 different categories were included. Thus, for the Unrelated Lure condition, the lures were different for each one of the three tests, and each lures on a single test came from a different category – but the 15 categories were repeated across the three tests. Test 4 was identical for the two conditions and was composed of the 15 list words and the 15 unstudied members of the same category.

Design. The design was between participants, so participants in the Related Lure condition received tests with lures that were from the same category as studied words in tests 1-3, whereas participants in the Unrelated Lures condition received tests with lures that were from different categories in tests 1-3. Test 4 was identical for all the participants, and included related lures, varying only in terms of category and the specific lures that were used for counterbalancing purposes. The study lists were counterbalanced using a Latin square design, resulting in four different study list orders. The level of correct recognition and false alarms was measured for all tests, so the criterion and sensibility could be calculated for all tests, as well as performance differences (in this case, the comparison of interest was on test 4.

Procedure. Participants were told they were going to study lists of words, and that their memory about those words would be tested afterwards. Words were presented one by one in the center of a computer monitor at the rate of 2,000 ms per

word. After each list (30 words) presentation, all participants solved math problems for 1 minute. After this distractor task, participants received classic recognition instructions, and those in the Related Lure condition received a recognition test composed of 15 presented words and the 15 unpresented category members, while participants in the Unrelated Lure condition received a recognition test composed of 15 presented words and 15 unpresented unrelated words. This cycle of study-math-test was repeated for the four lists. Test 4 was identical for participants in the two conditions, being composed of 15 presented words and 15 related distractors, i.e., same category members. For the recognition tests, each word was presented one by one in the center of the computer monitor, and word order was randomized for each participant. The recognition tests were self-paced and participants responded by using the keyboard. Participants were instructed to press the key "c" if they thought the word had been presented before or the key "m" if they thought the word had not been presented (the key "m" had a red sticker on it and the key "c" had a green sticker on it, and participants were told they should press the red key if they had not seen the word before and the green key if they had seen the word before).

Results and Discussion

We first describe the results in terms of hit and false alarm data, and then turn to discrimination and criterion measures, which were statistically analyzed to address our hypotheses. Mean hits and false alarms from the 4 study-test cycles in the Related and Unrelated Lure conditions are presented in Table 1. The data show a mirror effect. That is, participants in the Unrelated Lure condition produced both more hits and fewer false alarms than participants in the Related Lure condition on tests 1 to 3. On the 4^{th} test, which was the same in the two conditions, participants in the

Unrelated Lures condition produced more hits as on previous lists, but also more false alarms.

TABLE 1
Mean hits and false alarms from the four study–test cycles in the Related and Unrelated Lure conditions for Experiment 1

	Related lures				*Unrelated lures*			
	Cycle 1	*Cycle 2*	*Cycle 3*	*Cycle 4*	*Cycle 1*	*Cycle 2*	*Cycle 3*	*Cycle 4*
Hits	.87 (.10)	.84 (.09)	.81 (.13)	.76 (.17)	.87 (.11)	.90 (.07)	.88 (.10)	.83 (.13)
FAs	.30 (.16)	.27 (.20)	.31 (.22)	.29 (.19)	.06 (.06)	.06 (.03)	.06 (.03)	.41 (.28)

Note: Cycle 4 for Unrelated Lures condition refers to a recognition test containing related lures only. Standard deviations are in parentheses.
FAs, false alarms.

To explore the effect of the lure manipulation on performance throughout the study-test cycles, we subjected both d' (ability to discriminate between studied items and lures) and c (criterion or response bias; reflects the degree to which "yes"/"old" responses dominate or the degree to which "no"/"new" responses are preferred; negative values indicate a bias toward classifying items as "old", and positive values indicate a bias toward classifying items as "new"; Macmillan & Creelman, 2005) to a 2 Lure (Related vs. Unrelated) x 3 Study-Test Cycle (1-3) mixed-model ANOVA and then we separately test the results of our manipulation in the fourth study cycle (in which the participants of both conditions received the same related lure recognition test). As a reminder, note that Lure was manipulated between-subjects. Let us first present the results for the memory discrimination measure, d'. We obtained a main effect of Lure, showing that participants in the Unrelated Lure condition (M = 2.89) outperformed participants in the Related Lure condition (M = 1.73), $F(1, 86) = 149.43$, $p < .001$, MSe = .57, $\eta^2 = .63$. We also obtained a marginal main effect of study-test cycle, $F(2, 86) = 2.90$, $p = .056$, MSe = .26, $\eta^2 = .03$, but most importantly, the two factors interacted showing that the superior performance of participants in the Unrelated Lure condition relative to participants in the Related Lure condition

disappears on the 4th test, F (3, 86) = 29.68, p < .001, MSe = .32, η^2 = .18. These results were expected since the recognition is much easier for Unrelated Lure participants. In the fourth study cycle (in which the participants of both conditions received the same related lure recognition test), the two conditions, however, the two conditions no longer differed (t<1). Figure 1 depicts all the relevant d' scores data.

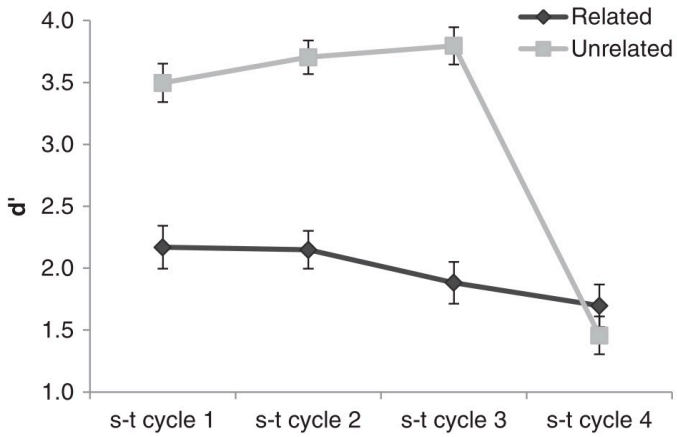

Figure 1. Study–test × lures condition interaction on d′ scores in Experiment 1. Vertical bars denote standard errors of the means.
Note: S-T cycle 4 for Unrelated Lures condition refers to a recognition test containing related lures only.

In the analysis of c (criterion or response bias), we also obtained a main effect of lure, such that participants from the Unrelated Lure condition (M = .15) showed a more biased criterion than participants in the Related Lure condition, who were more liberal (M = - .23), F (1, 86) = 48.65, p = .001, MSe = .19, η^2 = .36. We also obtained the interaction showing that the two factors interacted, F (2, 86) = 3.27, p = .040, MSe = .06, η^2 = .04. To interpret this interaction we contrasted the linear trend between the two conditions and it was indeed different, F (1, 86) = 3.78, p = .055, MSe = .05. And this happened because, whereas Unrelated Lure participants did linearly decrease their bias across study-test cycles, F (1, 86) = 4.33, p = .040, MSe = .05, the same did not

occur for Unrelated Lure participants (F<1). Finally, in the fourth trial (in which the participants of both conditions received the same related lure recognition test), the initial differences between Unrelated and Related Lures participants reversed such that the bias was now stronger and more liberal for Unrelated Lure (M = -.40) relative to Related Lure M = -.08) participants, with a F (86) = 7.08, p = .001, MSe = .33, η^2 = .08. Figure 2 depicts all the relevant c scores data.

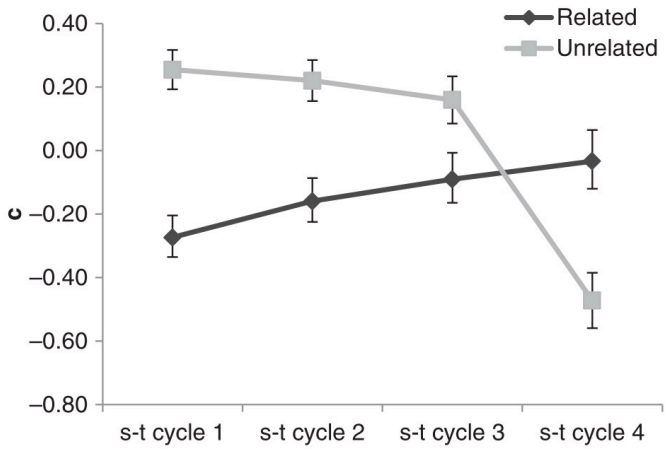

Figure 2. Study–test cycle × lures condition interaction on bias (c) in Experiment 1. Vertical bars denote standard errors of the means.
Note: S-T cycle 4 for Unrelated Lures condition refers to a recognition test containing related lures only.

To further explore the role of conceptual knowledge on criterion placement we performed a median split by centrality of the items and lures used in the Related Lures condition. We had to restrict this analysis to the Related Lures condition because in the Unrelated Lures condition, lures were taken from a different category than the studied items (except for the fourth study cycle). We present hits and false alarm data in Table 2. Data from Table 2 show a centrality effect such that High Centrality targets and lures are more readily accepted as old items than the corresponding Low Centrality items and lures. To analyze these data, we ran two Centrality (High vs. Low) x 4 Study-Test Cycle within-subjects ANOVAs on both d' and c. The results of

the ANOVA for the memory discrimination measure, d', only produced a Centrality main effect, showing that participants' performance was superior for Low Centrality items and lures (M = 2.36) relative to High Centrality Items and Lures (M = 2.13), $F(1, 38) = 488$, $p = .033$, $MSe = .68$, $\eta^2 = .11$.

TABLE 2
Mean hits and false alarms for high and low centrality items and lures (Related Lures condition only) along the study–test cycles for Experiment 1

	High centrality				*Low centrality*			
	Cycle 1	*Cycle 2*	*Cycle 3*	*Cycle 4*	*Cycle 1*	*Cycle 2*	*Cycle 3*	*Cycle 4*
Hits	.90 (.13)	.86 (.12)	.82 (.17)	.77 (.21)	.86 (.14)	.84 (.12)	.80 (.17)	.76 (.20)
FAs	.33 (.25)	.34 (.30)	.27 (.24)	.28 (.25)	.20 (.19)	.18 (.19)	.23 (.22)	.25 (.23)

Note: Standard deviations are in parentheses.
FAs, false alarms.

Figure 3 shows criterion in the Related Lure condition by Centrality and Study-Test Cycle. In the analysis relative to criterion measure c, we also obtained a main effect of Centrality, such that participants more readily accepted High Centrality items and lures (M = -.30) relative to Low Centrality items and lures (M = - .05), $F (1, 38) = 32.92$, $p < .001$, $MSe = .15$, $\eta^2 = .46$. We also obtained a main effect of study-test cycle, $F (3, 114) = 2.66$, $p = .051$, $MSe = .39$, $\eta^2 = .06$, mainly due to the fact that the overall criterion becomes dramatically less and less biased across the study-test cycles. And indeed the linear trend is significant, $F (1, 38) = 7.32$, $p = .010$, $MSe = .41$ and the effect size is considerably larger ($\eta^2 = .16$) than the overall main effect whereas the residuals were non-significant, $F < 1$. Critically, although the interaction was not significant, $F (3, 114) = 2.99$, $p = .119$, $MSe = .23$, $\eta^2 = .05$, the above mentioned linear trend was indeed moderated by Centrality, $F (1, 38) = 6.65$, $p = .015$, $MSe = .17$, $\eta^2 = .15$, whereas the residuals were non-significant, $F < 1$.

Finally, we also compared contrasted High and Low Centrality across Unrelated and Related Lure conditions in the fourth cycle (in which the participants of both conditions received the same related lure recognition test) for both d' and c. Thus

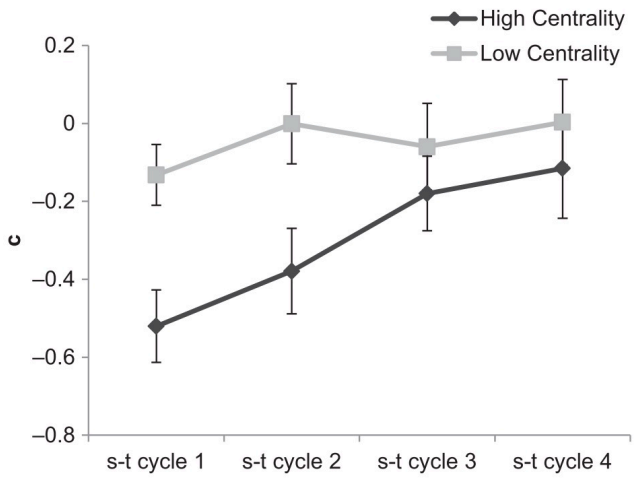

Figure 3. Changes in bias (c) across the four study–test cycles for high and low centrality words (Related Lures condition only) in Experiment 1. Vertical bars denote standard errors of the means.

we first performed a 2 Lure (Unrelated vs. Related) X 2 Centrality (High vs. Low) mixed ANOVA on d' scores, with the last factor being whithin-participants. We obtained a main effect for Centrality, $F(1, 86) = 3.76$, $p = .056$, $MSe = 1.05$, $\eta^2 = .04$, such that participants performed better with Low (M=1.88) than with High Centrality (M=1.56) items and lures. A marginal interaction also emerged, $F (1, 86) = 2.92$, $p = .091$, $MSe = 1.05$, $\eta^2 = .03$ suggesting that the differences between low and high centrality items occurred more pronouncedly in the Unrelated (M= 1.82 vs. M= 1.25) than in the Related Lure (M=1.94 vs. M =1.90) condition. As for the same analysis with c scores, we only replicated the already reported difference between conditions in the last study cycle and this difference was not qualified by centrality ($F < 1$).

We interpreted these data as evidence that, across the study-test cycles, participants in both conditions tried to adapt their encoding and test strategies to the requirements of the specific test they had been faced with. Thus, whereas participants in both conditions probably identified the taxonomic category underlying the first study list, this conceptual knowledge facilitated the performance of participants in the Unrelated Lure condition but it hindered performance of participants in the Related

Lure condition, because the lures were from a different category in the former condition but from the same category in the latter condition. More specifically, participants' conceptual knowledge lead to a liberal criterion in the Related Lure condition (especially in the case of High Centrality) because both targets and lures belonged to the same category; and to a stringent criterion setting in the case of the Unrelated Lure condition because all lures belonged to a different category, which made lure rejection very easy (along with a certain degree of rejection of low centrality items). Across the study-test cycles, participants in the Related Lure condition learned to disregard their knowledge about the category underlying the study lists and their criterion became less and less biased (lenient) overall and in especially for High Centrality targets and lures. However, the same did not occur in terms of memory discrimination. Apparently, the constant change of categories across the study-test cycles may have limited the possibility of developing a strategy that would improve memory performance, while at the same time may have facilitated insight regarding the low diagnosticity of category membership.

Participants in the Unrelated Lure condition, on the contrary, kept relying on their conceptual knowledge to perform on the recognition test, such that when on the last study-test cycle the structure of the test was changed to include only lures from the same category as the studied items, participants' criterion placement showed a greater degree of bias than the initial bias shown by the participants in the Related condition.

Thus, participants in both conditions initially relied on their conceptual knowledge, showing the centrality effect, in other words being more likely to endorse High Centrality items as old. However, across the first three study-test cycles, participants in the Related Lure condition learned to avoid this conceptually-based

response because it promoted limited performance. The same did not happen to participants in the Unrelated Lure condition, for whom the conceptually-based response strategy allowed them to achieve very high levels of performance until the last critical study-test cycle. Thus, these participants in the Unrelated Lure condition were never given the opportunity to learn how to avoid this response strategy and when they were confronted with a recognition test in which all the lures were from the same category as the studied items (the 4th test), they exhibited an even more liberal response criterion than they would probably have shown in the absence of previous tests (i.e., a criterion probably similar to the criterion shown by participants in the Related condition when they were tested for the first study-test cycle).

Experiment 2

We contend that the requirements of experienced tests shape the way participants encode similar new information or new information in similar contexts. These test-driven differences of encoding should affect performance on similar tests but they can also affect performance on different tests if participants are surprised by an unwarned change in test format. In Experiment 2, the new test we use is free recall. We relied on the same Related/Unrelated Lure manipulation as in Experiment 1, but on the fourth and last test we asked our participants to perform a free recall test, and we did not warn them of this change in test format. We reasoned that participants in the Related Lure condition would learn a strategy based on the individual features of the studied items (a distinctive feature-based strategy), since a category-based strategy would be quite inefficient given the structure of the test lists they had experienced, whereas participants in the Unrelated Lure condition would learn to rely precisely on the categorical or thematic nature of studied items (a relational, conceptual-based

strategy). Since feature-based encoding strategies enhance free recall of lists of related items (Einstein & Hunt, 1980; see also, Hunt & McDaniel, 1993; Nairne, 2006) we expect that despite the Unrelated Lure participants outperforming the Related Lure participants across the first three study-test cycle, this performance difference would be reversed in the final study-test cycle, in which the test requires free recall.

Method

Participants. Ninety-four Lisbon University undergraduates participated in this experiment for payment. Forty-six participants were assigned to the unrelated condition (tested with unrelated distractors in tests 1-3) and forty-eight were assigned to the related condition (tested with related distractors in tests 1-3).

Materials, Design and Procedure. The same materials, design, and procedure used in Experiment 1 were also used in this Experiment, with two exceptions: first, the final test was a free recall test rather than a recognition test – after studying the fourth list and solving math problems, the participants were instructed to write on a blank sheet as many words as they could recall from the last list of words; and second, in the first three recognition tests the participants' responses were given on a sheet of paper, where all the test items were listed, rather than on the computer. In the study phase, the words were presented one by one on the computer screen (randomized afresh for each participant), but after each list presentation, followed by the math distractor task, the participants were instructed to read the words written on the sheet of paper and to decide whether each word had been presented on the previous list. For each word, participants were asked to circle "yes" if they thought that the word had been presented and "no" if they thought it had not. The recognition test was administered in a self-paced manner and the test words were presented in a random

order for each version (without the items from positions 1 and 4, 2 and 5 and 3 and 6) and list presentation order (as determined by a Latin square design), resulting in a total of 12 test word presentation orders for each condition. The randomization of the words on each test was constant because the tests were administered on paper.

Results and Discussion

Hits and false alarms from the first 3 study-test cycles in both conditions are presented in Table 3. The data show that participants in the Unrelated Lure condition produced about the same level of hits but noticeably fewer false alarms relative to participants in the Related Lure condition.

TABLE 3
Mean hits and false alarms from the three study–test cycles in the Related and Unrelated Lure conditions for Experiment 2

	Related lures			*Unrelated lures*		
	Cycle 1	*Cycle 2*	*Cycle 3*	*Cycle 1*	*Cycle 2*	*Cycle 3*
Hits	0.89 (.10)	0.84 (.14)	0.84 (.15)	0.84 (.18)	0.88 (.17)	0.84 (.09)
FAs	0.20 (.13)	0.20 (.15)	0.18 (.13)	0.05 (.13)	0.02 (.11)	0.01 (.05)

Note: Standard deviations are in parentheses.
FAs, false alarms.

To explore the effect of the lure manipulation on recognition performance throughout the first three study-test cycles, we performed 2 Lure (Related vs. Unrelated) x 3 Study-Test Cycle (1-3) mixed-model ANOVAs on both d' and c. Regarding the memory discrimination measure, d', we only obtained a main effect of Lure, showing that participants in the Unrelated Lure condition (M = 3.51) outperformed participants in the Related condition (M = 2.45), $F(1, 90) = 64.89$, $p < .001$, MSe = 1.70, $\eta^2 = .42$. In the analysis relative on the criterion measure, c, we again obtained a main effect of Lure, such that participants in the Unrelated Lure condition (M = .45) showed a more stringent criterion than participants in the Related Lure condition (M = - .14), $F(1, 90) = 82.63$, $p < .001$, MSe = .29, $\eta^2 = .48$. We also

obtained a main effect of Study-Test Cycle, F (3, 90) = 4.48, p = .013, MSe = .09, η^2 = .05, mainly due to the fact that the overall criterion became more stringent across the three study-test cycles. The two factors did not interact significantly, F (2, 180) = 2.37, p = .100, MSe = .09, η^2 = .02. Thus, the difference in bias between Lure conditions remained reliable across the three study-test cycles. However, across the three study-test cycles, participants in the Related Lure condition exhibited a positive linear trend, F (1, 90) = 5.99, p = .016, MSe = .12, η^2 = .06 (with a non-significant residuals, F < 1), whereas this linear trend was absent from the data of participants in the Unrelated Lure condition, F (1, 90) < 1. Figure 4 depicts these data.

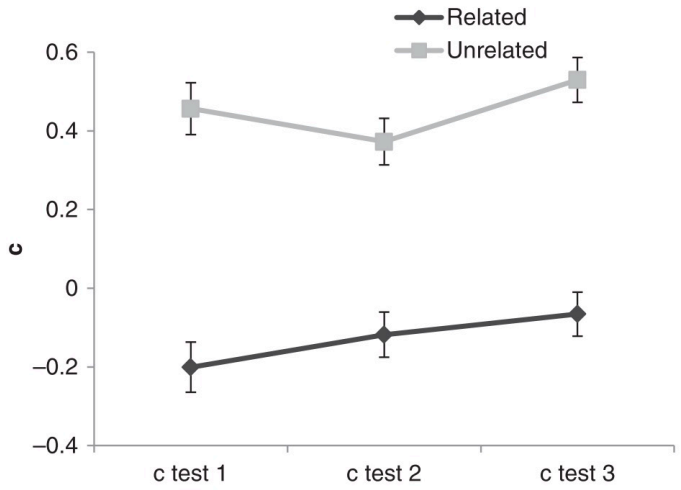

Figure 4. Changes in bias (c) for the first three study–test cycles for Related and Unrelated Lures participants in Experiment 2. Vertical bars represent standard errors of the means.

Again, we conclude that across the first three similar study-test cycles, participants in both conditions tried to adapt their encoding and test strategies to the requirements they were confronted with. More specifically, as in Experiment 1, participants' conceptual knowledge lead to a liberal criterion setting for participants in the Related condition and to a stringent criterion setting for participants in the Unrelated Lure condition. Across the study-test cycles, participants in the Related Lure learned to disregard their knowledge about the categories underlying the study

lists and their criterion became less liberal. However, and like in Experiment1, this did not occur in terms of memory discrimination. Participants in the Unrelated Lure condition, on the contrary, maintained their encoding and response strategy based on conceptual knowledge to perform in the recognition test.

Like in Experiment 1, to further explore the role of conceptual knowledge on criterion placement, we split both items and lures used in the Related Lure condition into Low and High Centrality lures. We present the hits and false alarm data for these two types of items in Table 4. Data from Table 4 shows a centrality effect such that High Centrality lures are more readily falsely recognized than Low Centrality lures.

TABLE 4
Mean hits and false alarms for high and low centrality items and lures (Related Lures condition only) along the first three study–test cycles for Experiment 2

	High centrality			*Low centrality*		
	Cycle 1	*Cycle 2*	*Cycle 3*	*Cycle 1*	*Cycle 2*	*Cycle 3*
Hits	0.88 (.19)	0.84 (.16)	0.81 (.18)	0.86 (.19)	0.86 (.14)	0.87 (.15)
FAs	0.27 (.23)	0.23 (.18)	0.24 (.19)	0.16 (.18)	0.17 (.18)	0.13 (.14)

Note: Standard deviations are in parentheses.
FAs, false alarms.

To follow-up in the analyses of these data, we computed two 2 Centrality (High vs. Low) X 3 Study-Test Cycle mixed-model ANOVAs (the last factor being within-participants) on both d' and c. Two main effects emerged from the results for the memory discrimination measure, d'. Thus, a Centrality main effect emerged, indicating that participants' performance was superior for Low Centrality items and lures (M = 2.80) relative to High Centrality items and lures (M = 2.29), $F(1, 46) = 20.28$, $p < .000$, MSe = .91, $\eta^2 = .31$. A Study-Test Cycles main effect was also significant, due to the better performance on the first study-test (M = 2.76) relative to the remaining cycles (M = 2.39 and M = 2.48), $F(2, 92) = 3.25$, $p = .043$, MSe = 1.09, $\eta^2 = .07$.

In the analysis relative to criterion measure c, we also obtained a main effect of Centrality. The main effect of Centrality indicated that participants more readily accepted High Centrality targets and lures (M = -.23) relative to Low Centrality targets and lures (M = - .04), F (1, 46) = 6.50, p < .014, MSe = .39, η^2 = .12. Critically, although the interaction was not significant, F (2, 92) = 1.41, p = .249, MSe = .26, η^2 = .03, a difference in the acceptance of high centrality targets and lures (M = -.40) and low centrality targets and lures (M = .-.07) emerged at the first study-test cycle, F (1, 46) = 6.89, p = .012, MSe = .37, η^2 = .13, but failed short from significance for the remaining cycles (both Fs < 1). Figure 5 depicts these data.

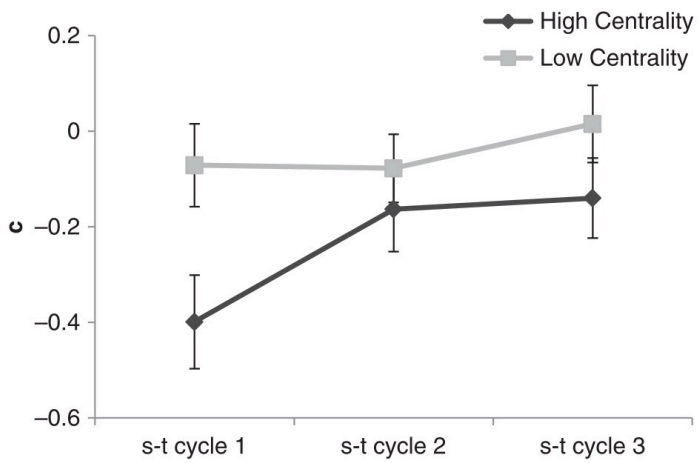

Figure 5. Changes in bias (c) across the first three study–test cycles for high and low centrality words (Related Lures condition only) in Experiment 2. Vertical bars denote standard errors of the means.

We interpret these data as replicating results from Experiment 1, in that although participants' conceptual knowledge lead to a liberal criterion setting in the Related Lure condition, across the study-test cycles, participants learned to disregard their knowledge about the category underlying the study lists and their criterion became less biased. This effect could be documented in both the overall gradual "de-

liberalizing" of the criterion and in the gradual reduction of difference in criterion between High and Low Centrality items across the study-test cycles.

Finally, we compared free recall in the fourth study-test cycle between the Related and Unrelated Lure conditions. The number of intrusions was small and not significantly different between the two conditions, so we ignored them in subsequent analyses and just focused on correct recall. Performance differed by Lure condition, $t(90) = 2.70$, $p = .001$, $SD = 5.60$, $d = .56$, such that participants in the Related Lure condition ($M = 13.55$) outperformed participants in the Unrelated Lure condition ($M = 10.40$). To check whether the Centrality effect found across the previous study-test cycles in recognition would also be obtained in free recall, we split the recalled items in High and Low Centrality and computed a 2 Lure condition (Related vs. Unrelated) X 2 Centrality (High vs. Low) mixed-model ANOVA, the last factor being within-participants. The analysis revealed two main effects, a Lure main effect that off course, reproduced the previously obtained difference, and a Centrality main effect, such that High Centrality targets ($M = 6.70$) were better recalled than Low Centrality targets ($M = 5.50$) , $F(1, 91) = 14.89$, $p < .001$, $MSe = .22$, $\eta^2 = .14$. The interaction was not significant, $F < 1$. Thus, although participants in the Related condition learned how to overcome a knowledge-based bias in their responses to recognition tests, such that the Centrality effect as measured by criterion was gradually overcome, a Centrality effect reappeared for these participants when the retrieval context was no longer detrimental to such a response strategy (i.e., at the free recall test).

As we have argued that test requirements lead to an adaptive change in encoding and test strategies across the study-test cycles, and that it was these strategy changes that accounted for the differences obtained in the last study-test cycle (in the free recall test), we measured the correlation between d'/c and the number of items

recalled. Because test requirements varied as a function of condition, we computed these correlations separately for the Related and Unrelated Lure conditions. These correlations are depicted in Table 5.

TABLE 5
Correlations between the d′ and c scores obtained in the first three study–test cycles and the number of items recalled in the final cycle (computed by Lures condition) in Experiment 2

	d′			c		
	Cycle 1	Cycle 2	Cycle 3	Cycle 1	Cycle 2	Cycle 3
Free recall (*Related lures*)	.61*	.59*	.51*	−0.09	−0.03	0.04
Free recall (*Unrelated lures*)	0.13	0.27	0.17	−0.03	−0.17	−.31*

*$p < .05$.

Data from Table 5 suggest that, for participants in the Related Lure condition, the encoding and response strategies that allowed them to discriminate between list items and lures in the first three study-test cycles was similar to the strategy they adopted in the last study-test cycle, as the correlations between d' scores across the study-test and the number of recalled items were all statistically significant and strong. No correlation was found between c scores and the number of items recalled. On the contrary, participants in the Unrelated Lure condition showed a completely different performance pattern. The encoding and response strategies that led them to successfully discriminate between items and different category lures seems to be quite different from the strategy these participants used in recall. This is, for participants in the Unrelated Lure condition, the correlations obtained between d' scores and the number of items recalled were small and non-significant. On the other hand, for these participants, response criterion on the last recognition test did predict the number of items recalled on the free recall test, such that the more stringent the criterion, the fewer items were recalled in the last study-test cycle (free recall test). As we

interpreted this stringency in criterion as evidence of knowledge-based encoding and response strategy, this result concurs with our interpretation.

General Discussion

In this paper, we argued that specific requirements of retrieval contexts in which previous learning is assessed affect future encoding of similar information or information acquired in similar contexts. In two studies, in which the instructional sets were always identical across conditions (i.e., memorize categorized study lists), we simply altered the nature of the recognition tests by including related (same-category) or unrelated (different-category) lures at test. This difference guaranteed a higher performance from participants in the Unrelated Lure condition and promoted the setting of a conceptual-based stringent response criterion that remained stable until the retrieval requirements were changed. On the other hand, for participants in the Related Lure condition, the initial conceptual-based criterion would be very liberal and the level of performance attainable by persisting in that criterion would be very poor (due to high levels of false alarms). As such, these participants learned to avoid this conceptual-based criterion and supposedly relied on a more distinctive or feature-based encoding and response strategy. When, in Experiment 1, the nature of the recognition test was changed for participants in the Unrelated Lure condition, such that all lures were from the same category as the study list, performance deteriorated and the response criterion became as liberal as the initial response bias exhibited by participants in the Related Lure condition. Although this difference in criterion from the first three to the last study-test cycle (in Experiment 1) was quite dramatic, we account for it in terms of the same conceptual-based encoding and response strategy that gave rise to very different consequences when the test requirements were

changed. Moreover, we found more direct evidence for the hypothesized learning process of avoiding a conceptual-based strategy, which occurred for participants in the Related Lure condition, as suggested by the linear of gradually reducing bias for High Centrality items. It is also noteworthy that this effect emerged only for participants in the Related Lure condition (in both Experiment 1 and 2). Also, we were able to show that the avoidance of a knowledge or conceptual-based strategy and the development of an item or distinctive-based strategy gave an advantage on the free recall test to participants in the Related Lure condition, compared to participants in the Unrelated Lure condition, who supposedly did not learn to avoid the thematic or conceptual-based strategy (Hunt & McDaniel, 2003; Nairne, 2006). Finally, correlations between d' and c scores from the three first study-test cycles and the number of items recalled at the last study-test cycle (free recall test) provided evidence that converged with the idea that the nature of the lures included in the successive recognition tests induced different encoding and processing strategies that benefited or hindered performance on the free recall test. The same-category included in the recognition tests given to participants in Related Lure condition induced them to adopt strategies based on item distinctive features. These feature-based strategies apparently enhanced performance at free recall because the d' scores and the number of items recalled were always strongly associated for these participants. On the contrary, only the c score from the third study-test cycle predicted the level of free recall for participants in the Unrelated Lure condition, showing a negative association. Thus, the more stringent was the criterion adopted by participants in the Unrelated Lure condition at the recognition test that preceded the last study-test cycle, the worse their performance was on the free recall test. As we explained before, we took stringency of criterion as an indication of knowledge or relational encoding and

response strategy. Thus this inverse relationship corroborates the idea that the adoption of a relational strategy hinders free recall of related lists (Hunt & McDaniel, 2003; Nairne, 2006). Taken together these results point to a mediated test benefit – the ability to adapt encoding strategies to the nature of previous materials and retrieval tasks. This retrieval induced strategy adaptation seems to be a valuable process to potentiate learning of related but different sets of items. Since we explored how testing potentiated the learning of new items and not how testing potentiated the learning of previously unrecalled but studied items (e.g. Karpicke, 2009), our proposed explanation of potentiation is not incompatible with other explanations such as the ones based on the activation of a specific search-set during retrieval that would lead to a better subsequent encoding (Grimaldi & Karpicke, 2012).

However, we think we must address three caveats before we conclude. The first is related to the suggestion made by Zaromb and Roediger (2010) that the power of repeated testing derives from the increased use of an organizational or conceptual-based strategy. In fact, Zaromb and Roediger showed that repeated testing led to better recall and to higher recall output organization than repeated studying. Thus, if repeated testing leads to reliance on a more relational, conceptual-based or organizational strategy, how is it possible that our participants in the Related Lure condition became less reliant on it across study-test cycles? One possible answer is that the organizational benefits found by Zaromb and Roediger occurred from repeated testing of the same list, whereas in our case, a new study list was presented in each of the study-test cycles. This constant change of list may have made a specific conceptual-based strategy difficult to adopt and the poor diagnosticity of general conceptual-based strategies particularly easy to detect. Future research will be needed to clarify whether in study-test cycles with the same category conceptual-based

strategies are avoided or used by participants that have studied the lists under conditions similar to our Related Lure condition.

The second caveat has to do with our preference for talking about an encoding and retrieval strategy. Why don't we simply argue that our participants learned different response strategies during the performance of recognition tests with different test structures, while the encoding strategy remained invariant across conditions? We concur that this account can easily fit the results from Experiment 1. However, in Experiment 2, we changed the fourth test for both conditions, such that the specific requirement learned from previous recognition tests could no longer be used by any participant, independently of their condition, and we still obtained the predicted difference across conditions. Such a difference is typically obtained when the encoding set favors either relational conceptual-based or distinctive feature-based encoding (Hunt & McDaniel, 2003; Nairne, 2006). We agree that our paradigm cannot completely disentangle encoding and retrieval processes, but neither is it our aim (which is more how the interplay between these two processes can promote adaptation to a context with a given structure). Thus we contend that our proposal of an adaptive change of both encoding and retrieval strategies as a function of requirements of the relevant retrieval context fits better with the data from the two experiments.

The final and last caveat has to do with an alternative but related possible account, the *desirable difficulties in learning* framework from Bjork (2013; 1994a; 1994b). According to Bjork (2011), manipulations that make successful learning more difficult during study (introducing interference, spacing learning, or using testing instead of restudy) often enhance long-term retention and transfer. In our case, although we always used similar study lists and standard recognition tests for both

conditions, tests for the Unrelated Lure condition were patently easier, and thus the poorer performance of participants in the Unrelated Lure condition at free recall (Experiment 2) may simply indicate that the difficulties felt by participants in the Related Lure condition were indeed desirable. Although we do concur with this assessment, we believe that every encoding strategy has inevitable associated costs that become noticeable only under certain circumstances, that is, when the retrieval requirements change in appropriate ways. Note that whereas it is certainly true that participants who faced desirable difficulties (i.e., participants in the Related Lure condition) did perform better at the final study-test cycle in both experiments, it is very likely that participants in the Unrelated Lure condition would continue to outperform their colleagues if the final test had remained unchanged. Thus, we believe that although some difficulties in learning can be desirable in most long-term retention or transfer cases, it should always be possible to find a specific retrieval context in which the desirable difficulties cease to be desirable. Only future research can provide further insight into these questions.

We began by quoting John Anderson who argued that human memory did not evolve to give optimal performance in our silly laboratory tasks. And we agree. But we would add that even our silly laboratory tasks might lead our memories to "evolve" in adaptive ways to these tasks if we allow them to.

CHAPTER III. Adapting to the retrieval requirements: when testing word location hinders semantic activation. Marques, Garcia-Marques & Orghian (*invited for resubmission*) *Memory & Cognition*.

Introduction

The more testing resembles study, the better memory performance gets. This is indeed a very popular principle in the study of memory. Both, the encoding specificity principle (Tulving & Thompson, 1973) and its conceptual close relative, the transfer of appropriate processing approach (Morris, Bransford & Frank, 1977), emphasize the importance of the similarity between encoding and retrieval contexts or between the operations carried out at study and test. However, these principles have been often conveyed as one-way processes, running from encoding to retrieval. This makes sense because most often than not, as memory experiments include only a single study and a single testing phase, thus making it difficult to understand how the nature and the requirements of a specific test affects the way the information is later on encoded[4]. But, recent research has demonstrated that in successive study-test cycles, retrieval can affect encoding. In a paper that directly addressed the dynamic changes in encoding strategies induced by repeated testing, Garcia-Marques, Nunes, Marques, Carneiro and Weinstein (2015) showed that when participants went through successive study-test cycles with tests that either encouraged or discouraged conceptual processing (by manipulating the relationship between the lures in recognition test and the words studied), they adapted their encoding strategies towards

[4] It is certainly possible to induce test expectancies by instructions (e.g. Meyer, 1934, 1936) but it seems a relative weak way to let the nature of the test to act on encoding relative to actual test practice.

the relevant dimensions of the stimuli, and away from the irrelevant ones. More specifically, participants tested with related lures (from the same category as the studied items) seemed to encode the words attending to their distinctive features and ignoring category membership, while the opposite occurred with the participants tested with unrelated lures. This was observed when, on a final test, the structure changed without notice after encoding: whereas participants tested with unrelated lures shown a patently superior performance on a series of three study-test cycles, their performance dropped drastically when a final test contained only related lures (Exp. 1) or was a free-recall test (Exp. 2). In addition, two more pieces of evidence for the idea that participants were adapting their strategies across study-test cycles were found in this paper. First, there was a gradual decrease of bias for high typicality words for participants tested with related lures, suggesting that they were learning to rely less on conceptual information at study. And second, there was an increasing correlation obtained between recognition performance (over earlier tests) and final recall (Exp. 2) in participants that had to discriminate between presented and non-presented instances of the same category. This same correlation was not verified when the participants had to discriminate between presented instances of a category and non-presented unrelated words: instead, the recall for these participants was correlated with recognition bias but not with discrimination performance. These results fall into the host of effects recently dubbed "forward testing effects" (Pastötter & Bäuml, 2014).

In this paper, we follow up this line of research, by testing whether making different dimensions of the same stimuli more or less relevant to test in repeated study-test cycles, leads to corresponding differences at encoding. Something like a

"reversed" transfer of appropriate processing, flowing, this time, from retrieval to encoding.

The spatial dimension meets semantics

The spatial dimension of stimuli is a central attribute that can be encoded in memory, allowing for discrimination between memories and serving as retrieval cue, when relevant (e.g. Rajaram, 1998; Kornblum, 1992; Underwood, 1969). Moreover, and of interest for the present studies, several results point to a more automatic nature of encoding spatial information of stimuli (e.g. Hasher & Zacks, 1979; Logan, 1998; McCormack, 1982; Tsal & Lavie, 1988, 1993; Schulman, 1973), and although there are some examples of controlled influences of spatial processing (see Naveh-Benjamin, 1987, 1988; Park & Mason, 1982), performance on incidental spatial tasks is usually above change level (Caldwell & Masson, 2001). In a further illustration of this pattern, using the process dissociation procedure (PDP; Jacoby, 1991) on a memory for objects location task with young and old adults, Caldwell and Masson (2001) found an advantage of controlled processes on performance for younger participants, but no difference in automatic processes by age. Taken together, these data suggest that while the encoding of some spatial aspects is largely automatic, it can still improve with practice and specific strategies or goals.

If the phenomena described by Garcia-Marques et al., (2015) also occur for the spatial dimension of stimuli, repeatedly testing memory for words' location should lead participants to employ strategies focusing more on relative positions of the items, and less focused on their meaning, when compared with testing memory for the words themselves. Additionally, this would only become apparent if the nature of the test changed after encoding, and against the participants' expectations. In the present

studies, this was done by switching unexpectedly the tests between two test conditions (a semantic cue or a location cue test) and also by using a new task. The new task was the modified free association task (Hourihan & MacLeod, 2007) that, while not explicitly requiring retrieval of previously experienced information, would benefit more from previous conceptual encoding (semantic) over non-conceptual (spatial) encoding.

Present studies

In the present experiments we test these hypotheses using a modified version of the game "Concentration" (also known as "Pairs", "Memory" or "Pelmanism"). The rules of the game are usually as follows: a deck of cards containing symbols, images, or words (matching in content) is placed on an array facing down, and the player must turn up any two cards at a time with the goal of finding a matching pair. If a pair is a matching one, they are kept facing up on the array, whereas if the player fails to find a matching pair the cards are turned down again. The game ends when all pairs are revealed. This game was adapted as our experimental paradigm due to the fact that the coexistence of the spatial and semantic dimensions of items (the cards) is at the very core of the game's nature.

In our version of the game, 16 cards are laid out facing down in a 4 x 4 array, each containing a word from the 16 most typical items of a common category (Pinto, 1992). The nature of the pairs is revealed when, in a first phase, two cards are randomly flipped at a time revealing two of the category members that should be treated as a pair, and flipped back after few seconds until all the cards were presented, thus, defining the to-be-remembered pairs. After this phase, the goal of the game is disclosed to participants who will have to either, given a randomly flipped card, recall

the word with which it was paired (semantic condition; the location of the target word is disclosed) or recall it's location (spatial condition; the target word itself is disclosed).

Experiment 1

In Experiment 1, participants went through four study-test cycles of the memory game. We manipulated the test requirements (semantic or location-based) for the first three cycles and the nature of the last test, that could match or not the requirements in the previous tests. This resulted in four between-participants conditions: two experimental conditions (Semantic cue for the first three tests and Location cue for the fourth test - SL; Location cue for the first three test and Semantic cue for the fourth test - LS) and two control conditions (Semantic cue for the four tests - SS; Location cues for the four tests - LL). Note that the four cycles regarded different material, such that in each cycle a different category was studied and tested. If, while studying the same materials, being tested repeatedly with location or semantic cue tests impacts the way participants encode the words in the study phase, we should observe a performance drop whenever the final test changes abruptly. We also hypothesize that this drop in performance will be more accentuated in the case where participants learned to respond to the location requirements and are given a semantic cue test in cycle four (LS condition) when compared with the LL condition. As the nature of spatial encoding is predominately automatic, a final location test should be less affected by the requirements in previous tests than a final semantic test would be.

Method

Participants

A total of 88 undergraduates of the University of Lisbon participated in the experiment in exchange for course credits. The experiment was ran in groups of five to eight participants, and the assignment to the experimental conditions was randomized, leaving 23 participants in the SS condition, 22 in the SL condition, 23 in the LS condition and 20 in the LL condition. All participants reported normal or corrected normal vision.

Materials and Apparatus

The words featured on the cards were drawn from the Portuguese category norms (Pinto, 1992). The 16 most typical exemplars from four categories (occupations, mammals, fruits, and body parts) were chosen, excluding exemplars with more than one word. The experiment was built and ran on the python-based software *OpenSesame* (Mathôt, Schreij & Theeuwes, 2012).

Design

The design was 2 (Tests 1-3: Semantic vs Location) X 2 (Test 4: Match vs Mismatch), both factors being manipulated between participants. The dependent variable was the proportion of correctly recalled words or identified cards.

Procedure

Participants were told they were going to play an adapted version of the classic Memory/Concentration game, and were informed of the differences in the version they were playing. The instructions at this study phase were to pay attention

to the flipped word pairs, because their memory for them would subsequently be tested, without disclosing any information about the nature of the memory test. Participants would see 16 cards facing down on the screen, concealing words on the other side, one word per card. The cards were represented by 16 evenly spaced magenta rectangles on a black background. After the participant pressed a key, two cards at a time would then be randomly flipped, exposing two words that should be treated as a to-be-remembered pair, at a rate of 3000ms/pair until all (eight) pairs were presented. Each pair was only presented once.

After studying the 8 pairs, participants performed a filler task (*sudoku*) for 2 minutes, followed by the test instructions, according to the condition they were assigned to. When the test was semantic (S), participants were told that one card would be flipped, revealing one of the previously presented words, and the card that contained the other corresponding pair member was highlighted. Their task was to recall which word was it, by clicking on the location of the highlighted card and typing down the word. To go to the next pair participants pressed an 'OK' button. When the test was location (L), participants were told that one card would be flipped, revealing one word, and the corresponding pair member would appear at the bottom of the screen. Their task was to click on the card that they thought contained the pair member, i.e., it's location. For the control conditions (SS and LL) this study-test cycles were repeated four times. For the experimental conditions (SL and LS), the same type of study-test cycles were repeated three times, and before the fourth test, the instructions highlighted the change in the nature of task, and the corresponding test was administered. No feedback was provided. A different category was used on each cycle. Category order, word pairing, and pair testing order were randomized afresh for all participants.

Results and discussion

We first examined whether participants exhibited performance gains across study-test cycles, which would be an evidence of the benefit of adapting encoding strategies across cycles. This analysis will also tell us which type of test benefits more from successive cycles. Only the two "match" conditions (SS and LL) were used in this analysis, so as to observe performance across four identical study-test cycles instead of only three (as in the "mismatch" conditions). We conducted a 2 (Condition: SS vs LL) X 4 (Study-test cycles: 1-4) mixed-model ANOVA with the dependent variable being the proportion of correct responses. No main effects emerged, but the two factors interacted, $F(3, 123) = 3.19$, $p = .03$, *Cohen's d* = .56. As it's shown on Figure 1, this interaction is mainly due to the fact that on test 1 the SS participants started off worse ($M = .31$) than their LL counterparts ($M = .46$) , $t(41) = 2.38$, $p = .02$, *Cohen's d* = .74, showing a boost in performance from cycle 1 to cycle 2. From test 2 on, their performance was identical. This general pattern was expected, given the aforementioned automatic nature of the encoding of spatial attributes, and to the fact that comparably, a semantic test would be harder but more prone to benefits of

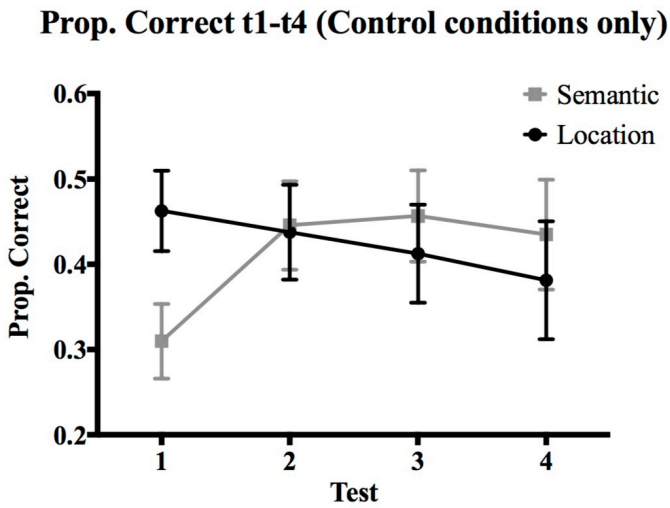

Fig. 1: Proportion of correct responses across study-test cycles for SS and LL (control) conditions in Exp. 1
Vertical bars denote the standard errors of the mean

practice. Moreover, the fact that the 16 positions repeat across cycles (only the configuration of the pairs and the words change across cycles) might be leading to proactive interference buildup in the LL condition, and to release from it in the SS condition (e.g., Goggin & Wickens, 1971; but see Chun & Jiang, 1998) and this can very plausibly explain the relatively stable performance and lack of improvement in LL condition.

Next, in order to explore the effects of the prior tests (1-3) on participants' performance on the final test (T4), we averaged the performance on the first three tests in each condition (semantic vs location) and conducted a 2 (Type 1-3 test: S vs L) X 2 (Type test 4: S vs L) ANOVA, with both factors between participants, and the dependent variable being the proportion of correct responses on T4. The results are depicted on Figure 2. Again, no main effects emerged, but the interaction was significant, $F(1, 84) = 10.22$, $p = .002$, *Cohen's d* = .70, such that, while both control conditions (SS and LL) were numerically superior in terms of performance on test 4 than the experimental conditions (LS and SL), the difference between the SS (M= .44) and the LS (M= .18) conditions was patently larger than between LL (M= .38) and SL (M= .29). In other words, and as expected, the performance drop was more accentuated when the test was of semantic nature, and previous testing was location-based.

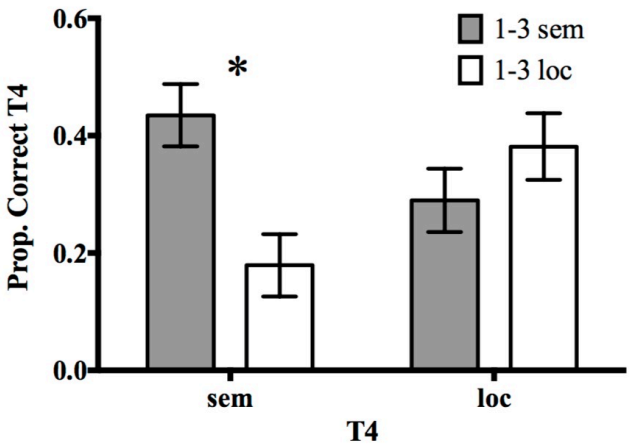

Figure 2. Proportion of correct responses on test 4 (semantic or location) given previous semantic or location tests 1 to 3.
Note: Vertical bars denote standard errors of the mean; * $p < .005$

The fact that performance dropped when the test changed unexpectedly after encoding isn't completely surprising, if one draws the parallels with the test expectancy (e.g. Lundeberg & Fox, 1991) and task switching (e.g. Monsell, 2003) literatures. But this result is novel and of interest to our proposal because i) the 'test expectancy' or 'task-set' (in the aforementioned literatures nomenclature) was induced by repeated testing, with the instructions – which were short and general depictions of the task – only being disclosed at the moment of each test, and ii) performance suffered more from a switch between previous spatial processing to a final semantic test than from a switch between previous semantic processing and a final location-based test. According to our proposal, participants are learning to encode the pairs by ignoring the stimuli attributes deemed irrelevant and focusing on the ones deemed relevant. The automatic nature of location encoding can explain why the performance drop was steeper in the LS than in the SL condition, as even if participants were disregarding the location at the study phase in cycle 4 in SL condition, this information was encoded nonetheless.

In sum, the results in Experiment 1 support our hypothesis, but it's possible to argue that the mere effects of specific task practice, the unexpected task switch, and the related switch cost can account for the performance drops, without meaningful qualitative changes on encoding strategies. We believe this would be a poor explanation, because adapting ones' encoding strategies based on test experience should be one of the inherent benefits of practice, and only when the task targets change after encoding can this effect be noticeable. Nonetheless, we concur that a final novel task that is sensitive to differential encoding of item's attributes would constitute better evidence to our proposal. We address these issues on Experiment 2.

Experiment 2

Experiment 2 was designed to provide converging evidence of adaptation of encoding strategies to the specific retrieval requirements, and specifically, the consequences of repeatedly testing location on the encoding of semantic attributes. If participants who are repeatedly tested with location cues learn to ignore semantic and conceptual aspects of stimuli, then the stimuli's semantic network should be less activated than if they were repeatedly tested with semantic cues. To test this, an explicit memory task could confound the results with the influence of learnt retrieval strategies, and not encoding strategies, so a novel task that wouldn't strongly rely on conscious and intentional retrieval would be more adequate (see MacLeod, 2008; Roediger, 1990). Moreover, this task should be differentially sensitive to conceptual encoding of the items, so that attending to meaning and to relational properties of the word pairs should result in a larger benefit in the performance on this task, compared to attending to shallower and spatial properties. Such a task is the one advanced by Hourihan and MacLeod (2007), which was devised to "capture" implicit conceptual memory. The authors reasoned that words that were encoded conceptually (that is, in terms of their meaning and taxonomic relations) would as a result have their semantic network more activated (see, e.g., Nelson & Goodmon, 2002; Nelson, Kitto, Galea, McEvoy & Bruza, 2013) than words that were encoded non-conceptually (that is, in terms of superficial or perceptual characteristics). Participants in Hourihan and MacLeod's (2007) experiments studied words under (Exp. 1) *generation effect* (Slamecka & Graf, 1978) or (Exp. 2) *levels of processing* (Craik & Lockhart, 1972) conditions, and afterwards performed free-association task, where they were asked to silently read a word and then, without any constraints other than answering as fast as possible, were instructed to say the first word that came to their mind. The words at

this phase were conceptually and non-conceptually encoded earlier and new words. The response was made using a voice key, and response times were recorded. Results shown that words that were encoded conceptually (i.e., generated, or processed semantically) were associated with faster responses than new or non-conceptually encoded words, a sign that the encoding manipulations affected the activation of item's semantic networks and thus the ease with each participants would produce an associate. In the present experiment we used an adapted word association task to investigate the activation of item's semantic network after being successively tested with location or semantic cues. Being tested with semantic cues should lead to facilitated semantic network activation for the studied items (as found by Hourihan & McLeod, 2007). However, if test requirements do actually guide subsequent encoding, then being repeatedly tested with location cues should lead the participants to increasingly ignore semantic information and therefore the facilitated access to the corresponding network activation should decrease or fail to occur, because it isn't relevant for the task. We used high and low typicality words to both assess the task's sensitivity to the activation of semantic network and to test the hypothesis that highly typical instances of the study categories (e.g. *dog* as a typical instance of category animal) should exhibit high levels of activation in both conditions because the study list is entirely composed by instances of the same category. The high level of activation for high typical words should facilitate the assembling of the respective lexical network very easily, both for old and new words, making this comparison useless as a test for the episodic status of the free association word. On the contrary, low typicality words should elicit faster associative responses only when they were previously encountered at study, and especially if they were processed conceptually rather than non-conceptually. In other words, low typicality items would benefit more

from previous conceptual processing than highly typical words. We used few target words intermixed with filler words to decrease the probability of strategic and recollective responding due to increased salience of the relationship with the previous tasks. Note that while in Hourihan & MacLeod's (2007) paradigm conceptual processing is induced by instructions, in our case it is elicited by test experience. Participants went through four study-test cycles before the final free association task.

Method

Participants

A total of 96 undergraduates from the University of Lisbon participated in the experiment in exchange for course credits. The data from two participants was removed from analysis, due to poor performance (0 correct responses in the 4 tests). Each participant was assigned to the 'semantic' (S; 46 participants) or the 'location' (L; 48 participants) condition in a randomized manner. Each session was run in individual cubicles.

Materials & Apparatus

The materials were the same as in Experiment 1, but now with added 16 most typical items of a fifth category (trees) from the Portuguese norms (Pinto, 1992), and 20 unrelated words that fell into the word length and frequency in tongue intervals of the categorized words. Vocal RT's in the free association task were collected via microphone and a voice key. The stimuli on the free association task were selected from the category present at the last study-test cycle and the new category that was never experienced, by ordering them by typicality and selecting the 3 most typical and 3 least typical of the set of 16. The decision to use new category words instead of just the unrelated words was taken because it constitutes a better comparison term with the

studied words, given the categorical nature of the stimuli. The unrelated words served as fillers and were included in order to diminish the chances of explicitly relating the task with the previous study-test cycles. Also, the decision to use new items from a new category instead of new items from the category encountered on cycle 4 was taken because conceptual processing of these items, especially the high typicality ones, (e.g., Barsalou, 1985; Mervis & Rosch, 1981; Neely, 1977; Rosch, Simpson & Miller, 1976) should also activate non-presented exemplars' semantic network, which would attenuate the diagnosticity of the task to assess differential encoding. Importantly, using the highest and lowest typical items of categories allowed us to check whether our version of Hourihan and MacLeod's (2007) task was sensitive to semantic network strength (as would be expressed by faster response times to highly typical items). Thus, participants had to respond to 3 old high typicality words, 3 old low typicality words, 3 new high typicality words and 3 new low typicality words, along with the 15 unrelated items.

Procedure

Procedure on cycles 1 to 4 was the same as in Exp. 1, now with participants in the semantic (S) and location (L) conditions performing the same kind of test across all cycles. As in Exp. 1, category order was randomized afresh for each participant and no feedback was provided. Participants performed a filler task (*Sudoku*) during 2 minutes after each study trial, and after the test on cycle 4, before the free association task. After the last filler task, the free association instructions were presented, in a different font than previous instructions. Participants were told they were going to perform a new task, where they had to use a microphone to respond, and were instructed how to use it (maintaining the distance between their mouth and the

microphone, avoiding emitting any other sound than the response, and to respond in a loud projected voice). They were instructed that they would see words on the center of the screen, which they should read in silence and quickly say out loud the first word that came to their mind, not worrying whether it is bizarre, made sense or has been seen or used recently. Although participants were not informed that they were being timed, participants were encouraged to respond as fast as they could. At the beginning of each trial the word 'Ready?' appears at the center of the screen. To start the trial participants pressed the spacebar key, which leads to a blank of 300ms, followed by the cue word. The words appeared at the center of the screen one by one at random order, and remained visible until a verbal response was detected. When all 27 items were responded to, the experiment ended and participants were quickly debriefed.

Results and Discussion

Again, we first looked at performance across study-test cycles to assess whether there were increments in performance. We conducted a 2 (Condition: Semantic vs Location) X 4 (Test: 1 – 4) mixed model ANOVA on the proportion of correct responses. This time, a test main effect emerged, $F(3, 276) = 4.19$, $p = .006$, *Cohen's d* = .43, indicating that participants performance got better across cycles. Performance on test 4 ($M = .47$) was significantly superior than performance on the first test ($M =. 38$), $t(186) = 2.83$, $p = .001$, *Cohen's d* = .38. Because, as in Exp. 1, each cycle involves a new list, one can consider this effect as a forward testing effect (Pastötter & Bäuml, 2014) or a special case of test potentiation effects (Arnold & McDermott, 2013), in line with our hypothesis that participants will adapt to the test targets and learn to encode the relevant dimensions of the stimuli, while ignoring the

irrelevant. Importantly, the two factors interacted, indicating that the performance increase across cycles was mainly due to participants in the Semantic condition,

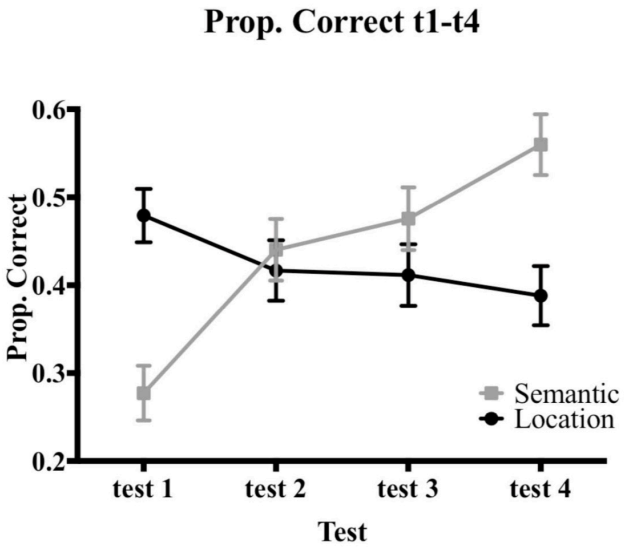

Fig. 3: Proportion of correct responses across study-test cycles Location and Semantic conditions in Exp. 2.
Vertical bars denote the standard errors of the mean

$F(3,276) = 16.21$, $p = .017$, *Cohen's d* = .48, replicating the results of Exp. 1, and supporting our hypothesis that performance gains are more evident in the Semantic condition. These data are depicted in Figure 3.

As in Hourihan and MacLeod (2007), the RTs on the free-association task were submitted to a trimming procedure, namely, Thompson (2006)'s implementation of Van Selst and Jolicœur (1994)'s with a shifting z-score criterion procedure, which excluded outliers using a cutoff criterion for each participant and condition, adjusting for the respective sample sizes and SDs. The procedure removed 14.1% of the trials. We first compared RTs for the high and low typicality category members, collapsing for episodic status of the items in order to test our hypothesis that generating free associates to high typicality items is faster than for low typicality ones, in order to check the task's sensitivity. We ran a 2 (Condition: Location vs Semantic) X 2

(Typicality: High vs Low) mixed-model ANOVA on the RTs, with the last factor being within-participants. The only significant effect was a Typicality main effect, $F(1,79) = 12.77$, $p < .001$, *Cohen's d* = .80, with neither condition, $F(1,79) = 1.66$, $p = .202$, *Cohen's d* = .29, nor the interaction, ($F<1$), reaching statistical significance. As expected, RTs were faster for high typicality items (M = 1625 ms) than for low typicality ones (M = 1762 ms), indicating that the task is sensitive to the items' typicality, and thus to pre-existent strengthened semantic networks[5]. Given the generally faster responses for high typicality words, we performed 2 (Condition: Semantic vs Location) X 2 (Episodic Status: Old vs New) mixed model ANOVAs, with the last factor being within-participants, separately for high typicality and low typicality items. For the high typicality items, as we expected, neither condition [$F(1,88) = 2.21$, $p = .14$, *Cohen's d* = .32] nor episodic status or the interaction ($F<1$) reached statistical significance. This indicates that for these items free-association responses were always fast, regardless of prior encounters with the word in any way. For the low typicality words, our comparison of interest, we obtained a main effect of episodic status with old items exhibiting faster response times (M = 1639 ms) than new items (M = 1734 ms), $F(1,90) = 7.1$, $p = .009$, *Cohen's d* = .56. But more importantly, the two factors interacted, indicating that the old/new difference in RTs is mainly due to faster responses to old items from participants in the Semantic condition, $F(1,90) = 6.65$, $p = .012$, *Cohen's d* = .54. For these participants average response times were 188ms faster for old than for new items, while this difference for participants in the Location condition this difference was virtually non-existent (3ms). These data are depicted on Figure 4. Mean response times for all item conditions are reported on Table 1. The fact that this old-new difference only appeared for Semantic

[5] This claim applies specially to the categories we used, which were common categories used very often in daily life.

participants is a strong indication that experiencing different test targets guided encoding strategies.

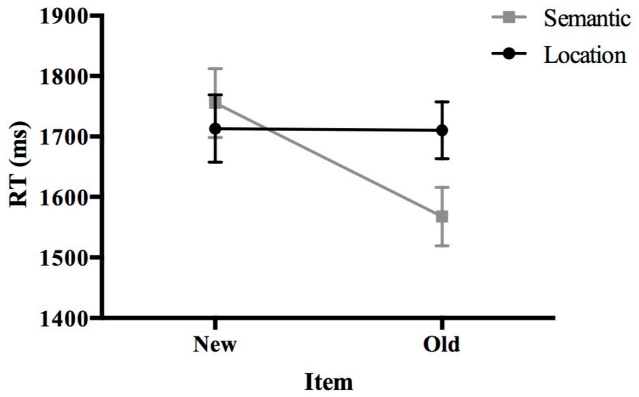

Figure 4. Mean RTs (in miliseconds) of free-association responses for old and new low typicality items in Exp. 2.
Note: vertical bars denote the standard errors of the means.

Table 1

Mean response times (ms) as a function of previous tests for free association in Exp. 2

Condition	Item				
	New High Typicality	Old High Typicality	New Low Typicality	Old Low Typicality	Fillers
Location	1670 (61)	1692 (61)	1795 (59)	1781 (50)	1723 (37)
Semantic	1621 (62)	1646 (61)	1752 (59)	1638 (50)	1697 (37)

Note: standard errors in parenthesis

General Discussion

In this paper we extended and gathered additional evidence for the claim that test experience can impact the way we subsequently encode new but similar information. Our results converge with Garcia-Marques' et al., (2015) in that, while maintaining the same encoding context and instructions, participants who experienced tests that either favored (Semantic tests) or disfavored (Location tests) conceptual processing exhibited performance drops when the test targets changed abruptly after

encoding. Especially, after a series of semantic tests, when the test suddenly changed to an non-conceptual one, the performance decreased drastically (Exp. 1) , whereas a facilitation was observed in the generation of free-associates to previously experienced words (Exp. 2), a task that benefits from a conceptual processing. Whereas in Garcia-Marques et al., (2015) the nature of the test was manipulated by changing the nature of the lures, in the present studies this was achieved by changing the requirements of task, what encouraged the participants to attend and encode different dimensions (Lockhead, 1972; Underwood, 1969) of the stimuli.

On Exp. 1 participants went through four study-test cycles in a modified Memory game, where test targets had either a semantic or a location nature, and were either kept constant or changed abruptly after encoding in the fourth test. Evidence for adaptation of encoding strategies was found in the performance drops that occurred when test targets changed after encoding, especially when the change was from Location practice to final Semantic test. This suggests that being tested with Location tests encouraged the neglecting of semantic attributes of stimuli, whereas when being tested with Semantic test the location information is harder to neglect (due to its more automatic character), even if it is not useful for the task at hand. On Exp. 2 a similar paradigm was used, with participants being tested with either Semantic or Location tests across four cycles, and performing a final free-association task to new and old (experienced in the last cycle) items. On this task, facilitation on old items when compared with new (as assessed by faster response times in generating associates) only occurred for participants previously tested with Semantic cues, which was expected since these item's semantic networks were more activated. This result indicates that for participants in Location condition, the items encountered on the fourth cycle were treated less conceptually, with the accompanying neglect of its

semantic attributes what explains the attenuating access to studied words' associates. We interpret these results as further evidence for the phenomenon described by Garcia-Marques et al., (2015): retrieval experience can guide subsequent processing strategies towards the dimensions of stimuli that better suit the retrieval targets, in new but similar study-test episodes. In other words, the participants learned the structure of the test they were presented with and adjusted their encoding strategies accordingly. The present studies extend the Garcia-Marques and colleagues' idea to another stimuli dimension, location, and using a novel paradigm. Thus, our results can be seen as a conceptual replication of these authors' studies, suggesting that this learning phenomenon is pervasive in many facets of human memory. One could argue that it is possible to explain the current results in terms of an adaptation of response strategies and not encoding strategies (i.e., encoding strategies remained stable, but participants learned different strategies to respond to the test requirements). While this could be true for Exp. 1 in which the same type of test was continuously used, this explanation doesn't fit so well with the results of Exp. 2, where participants had to respond to a completely novel – and deemed unrelated – task. If test experience only impacted response strategies there should have been a condition main effect with Semantic participants being faster than Location participants on all items, which didn't occur. Moreover, post-hoc analysis revealed that different levels of typicality (an inherently semantic variable) of the target items in the interim tests differently affected the performance in the Semantic condition, whereas for the participants in the Location condition the performance levels were not affected by the target word's typicality. Also notice that in the Semantic condition, facilitation in generating associate only occurred for lower typicality items, but performance on the previous test was better for higher typicality items, again showing the special role this variable

has for the Semantic condition. These results further support our claim that participants in Location condition neglected the semantic aspects and taxonomic structure of the stimuli, and weakens the possibility that the facilitation in free-association found for Semantic participants was due to overlapping response strategies. Nonetheless, we concur that the present paradigm can't completely disentangle encoding strategies from retrieval strategies and further work should address this issue.

The pattern of results found in Experiment 2 fits well with an inhibition account. In fact, in Experiment 2, we were able to show that lexical network access to low typicality words was inhibited in location relatively to semantic cue conditions. This result was however obtained in a test that was completely different from the tests used in the previous study-test cycles. This change of test can be considered the setting of an independent test criterion for inhibition, similar to the probe independent criteria set forth by Anderson and Spellman (1995).

We do recognize however that this experiment was not conceived as a test of an inhibition account and thus we are not able to disentangle inhibition from a number of alternative accounts, but we will draw consequences for an inhibition account of the present results in the hope that it may inspire future research.

An inhibition-based account of the present phenomenon would be as follows: participants tested with Location cue tests adapted their processing strategies across study-test cycles by focusing on the spatial dimension of stimuli, and inhibiting semantic processing of items, because this knowledge wasn't useful to the type of test they were being submitted to. For participants in both conditions, during the first study episode, encoding the words should activate their semantic dimension (as well as their location counterpart), but for participants repeatedly tested with Location cues

this would mean creating representations of the words that would be cluttered with irrelevant (semantic, in this case) and relevant information, which would then have to be selected for retrieval, given the specific retrieval requirements [see Hasher (2007), Hasher, Lustig & Zacks (2007) for a similar framework on inhibition both at encoding and at retrieval]. Inhibitory processes (especially) at encoding would hinder semantic activation of the words, because the specific retrieval goals of the task at hand required so, i.e., activating meaning and semantic mediators couldn't help to fulfill the goal and could actually impair performance on the task. Here, retrieval will serve as an opportunity to learn which processing strategies would be more fitted or unfitted to employ in subsequent similar episodes. It is important to note, though, that this inhibition account of our results differs in one important way from current inhibition explanations of effects in memory research [directed forgetting (e.g., MacLeod, 1998) or retrieval inhibition (e.g., Anderson & Green, 2001; Anderson & Spellman, 1995) effects are apt examples]: while in the mentioned examples inhibition acts on the same stimuli, in our case it acts on subsequent processing of new but similar stimuli. In the directed forgetting case, it acts on the to-be-forgotten studied items, as assessed by a later test on those items, and in the retrieval inhibition case, it acts on the non-practiced studied stimuli. In our case, inhibition acted on subsequent study-test cycles, on completely new sets of stimuli, thus describing inhibition in a higher level of analysis, as part of an iterative dynamic process that would grant memory an adaptive character. In our case, the inhibition would be attribute and stimulus-based, something that would seem crucial for grasping the invariants of one's environment. Further research should provide more informative tests to this phenomenon and clarify its potential limits and boundaries.

CHAPTER IV. Losing conceptual focus: how new learning is impaired during retrieval. Marques & Garcia-Marques (*submitted*)

Introduction

Memory research suggesting that repeated retrieval or testing benefits long-term retention has accumulated at an impressive rate over the last decade. Repeated retrieval has been shown to improve long-term retention and learning in many ways, both with typical lab materials and with educationally relevant ones. For instance, Karpicke and Roediger (2008) had college students learn a list of 40 Swahili-English word pairs, manipulating whether the pairs remained in the list (and were repeatedly practiced) or were dropped off the study or off the test list after the first time they were recalled. Students returned for the final test a week later. Repeated studying after learning had no effect on delayed recall, but repeated testing produced a large positive effect. In another example, using a more applied setting, Roediger and Karpicke (2006) showed that taking a memory test enhances later retention of educationally relevant materials relative to restudy of the same material. In two experiments, students studied prose passages and took one or three immediate free-recall tests, without feedback, or restudied the material the same number of times as the students who received tests. Participants were tested 5 min, 2 days, or 1 week later in a final test. When this test was given after 5 min, repeated studying improved recall relative to repeated testing. However, on the delayed tests, prior testing produced substantially greater retention than studying the same material. The beneficial effects of repeated retrieval and memory testing for long term retention and learning represent a very robust finding, obtained in hundreds of studies, across different materials, memory

tests designs, populations, both in the lab and applied settings and with different schedules and designs – for reviews see Roediger and Butler (2011) or Nunes and Karpicke (2015).

Interestingly, research showing detrimental effects of repeated retrieval has also been put forth. Notably, Finn and Roediger (2013) showed that learning of new information presented at test is sometimes impaired. In fact, Finn and Roediger's study was initially developed to test the hypothesis that repeated retrieval would improve retention on novel related information when compared to restudy. In this study, participants learned 3 pieces of information: a person's face, name, and profession. In phase 1, participants in all conditions learned faces and names. In phase 2, participants either restudied the face–name pairs (the restudy condition) or tried to retrieve the names with the faces as cues, and received correct answer feedback (the test condition). In addition, in both conditions the corresponding profession was presented for study just after each photo was restudied or tested. On either an immediate or in a delayed final test, participants were cued with each photo and had to recall both the corresponding name and profession. Contradicting the original hypothesis, results showed that names were better recalled in the test condition (i.e., a testing effect was found), but professions were better recalled in the restudy condition. Davis and Chan (2015) replicated these results and tried to further understand the conditions under which the effect is reversed, i.e., conditions where test-enhanced new learning is more likely to happen. Interestingly, they only found a slight testing advantage when the name and profession were presented in different blocks, a case where the "new learning" is not to happen while retrieving other related information. At the moment the reasons for this learning hindrance are not clear.

Further evidence for a similar deficit comes from Kantner and Lindsay

(2010)'s work on the effect of feedback on recognition judgments. In Exp. 4 of their paper, participants were assigned to three conditions: 1) a category-rule recognition condition, where participants had to study a list of words and perform a recognition test devised so that the episodic status of items overlapped with a category rule (all targets were names of large objects, all distracters were names of small objects), 2) a simple categorization condition, where participants had to respond to the same test words as their category-rule recognition counterparts, but did not study a word list and were instead instructed to classify the words in categories A and B, and 3) a simple recognition condition, identical to the category-rule condition, except there was no overlap between the episodic and conceptual statuses of items. Corrective feedback at test was manipulated between participants. In the categorization and conceptual learning literature, corrective feedback has long been shown to be a crucial aspect of the process of learning category boundaries (e.g. Bruner, Goodnow & Austin, 1956), so one could expect it to enhance performance whenever it helps attain conceptual knowledge. Curiously, Kantner and Lindsay (2010)'s results showed that feedback only impacted performance in the simple categorization condition, whereas it had no effect neither on the simple recognition nor on the category-rule recognition conditions. This lack of effect on the latter condition, in which participants could directly point to the item selection rule that by itself guarantees maximal performance, suggests that the aforementioned learning impairment represents a deficit in conceptual learning that occurs while the participants retrieve related information and in this respect, it is a result akin to Finn and Roediger (2013).

In this article we continue to explore the possible learning costs of repeated testing. In particular, we examined the possibility of a conceptual learning cost, in terms of a hindrance in learning of new information during test. Before we develop

our own framework, we refer to previous important research that have interweaved the study of memory with conceptual learning and, in this way, have helped clarify the similarities between recognition memory tasks and conceptual learning tasks (e.g. Bruner et al., 1956).

On the parallels between conceptual learning and memory

In an ingenious but sadly overlooked paper, Higham and Brooks (1997) demonstrated how participants are sensitive to selection rules and contingencies in stimuli sets, using this knowledge, even if tacitly, to inform memory responses and using their own episodic status judgments to inform categorization responses regarding the stimuli structure. In one of their studies, participants were instructed to study a list of words, which were compiled using non-salient criteria such as being: (a) low frequency (b) nouns (c) with seven or eight letters (Higham & Brooks, 1997; Exp. 1). These words were studied under a regular levels of processing (LoP) manipulation (i.e., shallow vs. deep encoding instructions; see Craik & Lockhart, 1972). In the test phase, participants had to complete both a recognition memory test and a classification test, on two lists that contained each three types of items: old (appeared on the study list), new-consistent (new words consistent with the three selection rules) and new-inconsistent (new words consistent with one of the selection rules, but inconsistent with the other two). The order of the tests was counterbalanced and in the recognition task participants had to rate a word on how likely it was that it appeared on the study lists, whereas in the classification task participants were instructed that the studied materials had to meet some set criteria to be selected, and that they had to rate the test words' consistency with those rules. While these two types of tests are often treated as two separate worlds, the results point to some

similarities between the two, and suggest that people engage on recruiting episodic knowledge in tasks that rely on conceptual knowledge and vice versa. First, discrimination (as measured by A'; see Snodgrass & Corwin, 1988) was above chance for both tasks, showing that participants were both sensitive to the words' episodic status and to the rules employed in the selection study materials. Additionally, the mirror effect (as it applies to this experiment, the LoP manipulation resulting in more 'old' and less 'new' responses in the deep processing condition; Glanzer & Bowles, 1976) occurred for recognition - as expected - but also for classification, with the LoP manipulation resulting in higher 'old' responses, but lower 'new-consistent' responses, which should not have happened if classification had relied purely on the commonalities activated during study. Moreover, collapsing across LoP conditions, the authors found what they called an "episodic effect" in the classification task ("old" vs. "new-consistent" discrimination; old items were more likely to be deemed consistent with the rule), and a "structural effect" in the recognition task ("new-consistent" vs. "new-inconsistent" discrimination; items consistent with the rule were more likely to be deemed old). This pattern suggests that episodic and conceptual (or structural, in the author's nomenclature) knowledge can, given the chance, interact and even replace one another when responding to a task. Another way to put this is that under some circumstances one can consider recognition tasks as a valid outlet for conceptual learning and classification tasks as a valid outlet for recognition memory.

The impact of retrieval on (new) learning

Our experiments assessed the hypothesis that repeatedly experiencing similar memory tasks enables people to learn the task structure and guide further encoding in comparable situations by drawing on acquired task knowledge to respond. Following

Finn and Roediger (2013) and Davis and Chan (2015)'s results we hypothesize that when this knowledge is only available in the moment of retrieval its conceptual processing will be impaired, as opposed to when it is present at study. We assessed new learning by looking at the false alarms deriving from category knowledge present on the memory tasks.

Experiment 1

In a paradigm adapted from Garcia-Marques, Nunes, Marques, Carneiro, and Weinstein (2015), participants in a thematic encoding condition or a thematic retrieval condition went through a series of study and recognition test cycles. In cycles 1-3 of the thematic encoding condition, study lists contained exemplars of common categories, with a different category in each cycle. Recognition tests in cycles 1-3 of the thematic encoding condition contained the studied items along with foil items taken from a pool of unrelated words, except for two list-related distracters (i.e., members of the studied categories). For example, a participant in the thematic encoding condition might study the names of 30 fish in cycle 1, then receive a recognition test with 15 old fish, 13 new unrelated words, and 2 new fish. In cycles 1-3 of the thematic retrieval condition, study lists contained unrelated words. Recognition tests in cycles 1-3 of the thematic retrieval condition contained the studied items along with two list-related distracters (i.e., unrelated words) and the new categorized words. For example, a participant in the thematic retrieval condition might study the names of 30 unrelated words in cycle 1, then receive a recognition test with 15 old unrelated words, 13 new categorized words, and 2 new unrelated words. After the three study-test cycles, participants in both conditions took a final recognition test in which they were instructed to accept items presented in any part of

the experiment (either in study lists, or as distracters on tests 1-3), and reject completely new items. This procedure is similar to that used by Jacoby, Shimizu, Daniels, and Rhodes (2005) and Shimizu and Jacoby (2005) with the difference being that in the final recognition test the authors used only items that have served as distracters in a previous test and new items, whereas we used both items from previous study lists and distracters from previous tests. This final recognition test contained old items (targets from lists and distracters from tests) and new items (unrelated words and non-presented members of each category). As the recognition tests were very easy in both conditions because study items and most lures were thematically divergent, we expected very good performance across the study-test cycles. The minority of lures that are similar to the study lists were nevertheless expected to promote a higher level of false alarms, if participants noticed the systematic difference between items and lures across the study-test cycles. Critically, however, in the final recognition test we expected the thematic encoding condition to show better recognition of the themes presented during the study-test cycles than participants in the thematic retrieval condition. To assess thematic extraction and learning we employed a criterion commonly used in conceptual learning – the difference in false recognition for thematic-congruent versus thematic-incongruent lures (akin to the difference in false recognition of rule-consistent versus rule-inconsistent items used by Higham and Brooks, 1997).

Method

Participants

Sixty-four undergraduates of Universidade de Lisboa took part in the study in exchange for course credit. They were randomly assigned to one of two experimental

conditions. The minimum sample size to ensure power to detect an effect was defined a priori using estimates of effect size obtained in previous studies in which test expectancies were manipulated by test experience (Garcia-Marques et al., 2015). The G*Power 3.1 software was used to calculate the minimum sample size for and effect size f = .40, and power = .80, and indicated that at least 52 participants would have to be tested. It slightly exceeded the sample size suggested by the software because the experiment was run in groups of approximately eight people.

Design

We used a between-participants design, so that participants assigned to the thematic encoding condition studied categorized word lists, while participants assigned to the thematic retrieval condition studied unrelated word lists. The critical manipulation was the source of thematic information: Recognition tests 1-3 were similar in the two conditions, except for the two list-related distracters (category words for the thematic encoding condition vs. unrelated words for the thematic retrieval condition), and the episodic status of the categorized and unrelated words (see Table 1 for a depiction of the structure of the task). Hits, list-related false alarms, and list-unrelated false alarms were measured for tests 1-3. Recognition test 4 was the same for all participants (varying only in list version, for counterbalancing purposes) and thematic and athematic hits and false alarms were measured on this test.

Table 1
Types of study lists and foils for Thematic Encoding and Thematic Retrieval conditions in Experiment 1.

Task Structure		
	Thematic Encoding	Thematic Retrieval
Study Lists	Thematic	Athematic
List-Unrelated Foils	Athematic	Thematic
List-Related Foils	Thematic	Athematic

Materials

To construct the study and test lists, we used three categories from the Portuguese word association norms (Pinto, 1992) – occupations, mammal animals and plants – and a selection of 297 unrelated words (excluding items related to the chosen categories, and items formed by more than one word) from the Multifunctional Computational Lexicon of Contemporary Portuguese (Nascimento, Casteleiro, Marques, Barreto, & Amaro, 2000). These three categories were chosen because they are the categories with the greatest number of exemplars with category typicality higher than 5%. For each category we selected the 45 more typical exemplars and ordered them in terms of decreasing category typicality. We then created three list versions per category composed of 30 words each, by removing every third exemplar to be in a list (i.e., the first list did not contain exemplars in positions 1, 4, etc.; the second list did not contain exemplars in positions 2, 5, etc.; and the third list did not contain exemplars in positions 3, 6, etc.). Thus we obtained a total of nine lists (three per category). This procedure allowed us to use the 45 (15 per list) removed items as category-related distracters, and as new thematic items in test 4. From the unrelated word pool, we created 9 lists with 30 words each, matching their categorical counterparts in terms of lexical frequency.

Each study list in the thematic encoding condition was composed of 30 items from the same category, randomized afresh for each participant, whilst recognition tests 1 to 3 were composed of 15 studied words and 15 distracters. The distracters in the thematic encoding condition were 13 words from the unrelated lists and 2 study-list related distracters, which were the items removed from the original category list according to list version.

Each study list in the thematic retrieval condition was composed of 30 unrelated words, randomized afresh for each participant, and recognition tests 1 to 3 comprised 15 studied words and 15 distracters. In the thematic retrieval condition, the distracters were 13 words from one category and 2 new words from the unrelated pool. Note that, except for the 2 study-list related distracters, the 1-3 tests were similar in the two conditions. All tests were administered on a single sheet of paper, with item order randomized per list version, category, and condition.

The final test (test 4) was composed of 18 old words: 9 thematic, and 9 athematic words (having appeared during study-test cycles 1 to 3 as targets or distracters depending on condition) and 18 new words: 9 thematic (belonging to the studied categories) and 9 athematic (from the unrelated word pool). The 9 thematic old words were the 3 most frequent items in each presented or tested category, and the 9 athematic old words were the corresponding unrelated items that appeared in the previous study-test cycles. The 9 thematic new words were the ones removed from the category lists according to list version, 3 per study-test cycle, and the 9 athematic new words were taken from the unrelated words pool, controlling for lexical frequency (see Table 2). There were 3 versions of the final test, according to the category list version (depending on which items were removed from the lists at study), with item order randomized for each version, and this test was also administered on a single sheet of paper.

Table 2
Provinience of old and new items on the final test for Thematic Encoding and Thematic Retrieval conditions in Experiment 1.

Final Test Structure		
	Thematic Encoding Condition	Thematic Retrieval Condition
9 Old Thematic Items	From Study Lists	From Test Lists
9 Old Athematic Items	From Test Lists	From Study Lists
9 New Thematic Items	Related to Study Lists	Related to Test Lists
9 New Athematic Items	Related to Test Lists	Related to Study Lists

Procedure

Participants were told they were participating in a memory experiment and were instructed to pay attention to a list of words because they would be tested on them later, without any additional information regarding the task or the list. Participants were randomly assigned to one of the experimental conditions. The words were presented at the center of a computer screen at the rate of 2000ms/word and the order was randomized afresh for each participant. After the study phase, participants solved arithmetic problems for 1 minute to prevent them from rehearsing the studied words. After this task, the experimenter handed the participant the corresponding sheet with the recognition test, and instructed them to tick "Old" if they had seen the word and "New" if they had not seen the word. When participants finished the test, the experimenter collected the test sheet and instructed them that they would see another list, thus beginning the next study-test cycle. This procedure was carried out for three study-test cycles, with the same instructions. The order in which each category (or the corresponding unrelated list) was studied or tested was counterbalanced between participants. After the third recognition test, participants performed the arithmetic task for 1 minute, and were then given the instructions for the final recognition test: their task was to respond "Old" if they had seen the word anywhere during the experiment (study or test lists), and "New" if they had not. All of the recognition tests were self-paced. After the final test participants were thanked and debriefed.

Results and Discussion

As four participants across the 3 study-test cycles systematically obtained a higher proportion of list-unrelated false alarms than hits, we excluded them from the

analyses (it is likely that they mistakenly switched yes and no responses on the sheets), thus leaving 27 participants in the Thematic Encoding condition and 33 in the Thematic Retrieval condition.

Study-test cycles

Hits, list-unrelated and list-related false alarms across the three study-test cycles are presented in Figure 1. Participants performed very well in the task, with a high level of hits and a low level of list-unrelated false alarms. Thus, participants appeared to easily grasp the structure of the recognition tests and use it to their benefit for accomplishing high performance. List-related false alarms, however, were much higher than list-unrelated false alarms, a signature of conceptual learning according to Higham and Brooks (1997), and another indication that participants grasped of the structure of the recognition tests. To test whether conceptual learning was constant across conditions, we performed a 2 condition (Thematic Encoding vs. Thematic Retrieval) X 3 study-test cycle (1st vs. 2nd vs. 3rd) X 2 type of false alarms (List-Unrelated vs. List-Related) mixed ANOVA with the last two factors being within-participants.

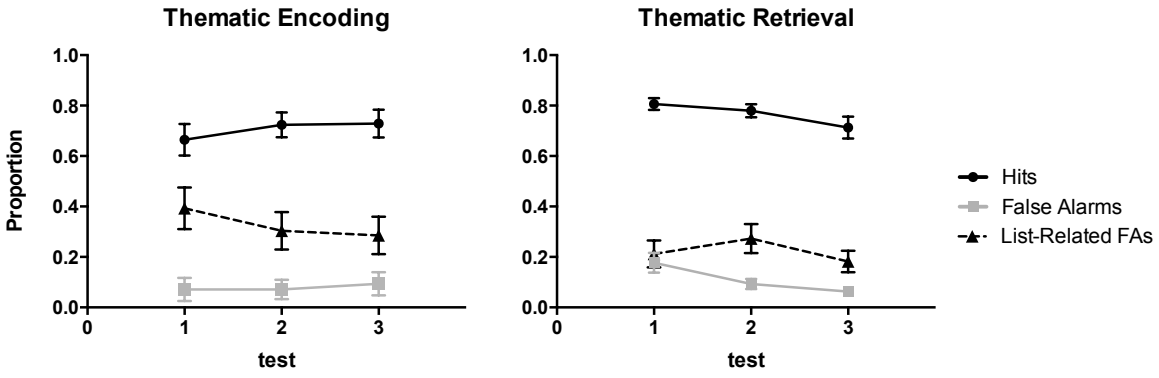

Fig. 1. Mean proportion of hits, false alarms and list-related false alarms across tests for Exp. 1.
Note: Error bars denote standard errors of the mean

A main effect of type of false alarm emerged, with the difference between list-related and list-unrelated false alarms reaching statistical significance, $F(1, 58) = 29.17$, $p = .001$, mean difference = .21, *Cohen's d* = 1.42[6]. A significant interaction between thematic condition and type of false alarms was also found, $F(1, 58) = 8.35$, $p = .005$, *Cohen's d* = .76. The mean difference between Thematic and Athematic False Alarms for the Thematic Encoding condition was .31, whereas for the Thematic Retrieval condition was .10, showing that the magnitude of the effect was higher in Thematic Encoding than in the Thematic Retrieval. Under Thematic Encoding conditions, this difference between list-related and list-unrelated false alarms is reminiscent of false memory effects obtained with categorized lists (e.g. Shiffrin, Huber, & Marinelli, 1995; Buchanan, Brown, Cabeza, & Maitson, 1999). However, under thematic retrieval conditions, this result is novel because it represents an instance in which false memories stem from categorized lures and not from categorized lists. That is, theme activation occurred only at test and not at encoding. Thus, recognition memory can be seen as akin to learning to discriminate between two lists, and the presence of a theme in one of these lists benefits performance. The finding of a significant difference between list-related and list-unrelated false memories, even in a condition in which only the lures are categorized, is illustrative of the strong relationship between conceptual learning and memory.

Final recognition test

For a more strict comparison between thematic encoding and retrieval conditions, we compared thematic and athematic false memories (i.e., false memories belonging to the categories presented at study or test, or false memories belonging to

[6] All *Cohen's d's* for the ANOVAs were calculated from the F values, according to Friedman (1982)

the unrelated word pool, respectively). Note that in the thematic encoding condition, thematic information appeared in the categorized study lists, whereas in the thematic retrieval condition, thematic information appeared in the categorized lures, across the three recognition tests. However, the same final recognition test was presented to participants of both conditions. We conducted a 2 condition (Thematic Encoding vs. Thematic Retrieval) X 2 type of false memory (Thematic vs. Athematic False Memories) mixed model ANOVA with the second factor being within-participants. Both main effects were statistically significant: Thematic Encoding (M = .33) vs. Retrieval (M = .20), $F(1, 58) = 8.61$, p = .002, *Cohen's d* = .77; Thematic (M = .38) vs. Athematic False Memories (M = .15), $F(1, 58) = 35.14$, $p < .001$, *Cohen's d* = 1.56. More importantly, however, an interaction qualified these effects, $F(1, 58) = 23.12$, $p < .001$, *Cohen's d* = 1.26, with the mean difference between Thematic and Athematic False Alarms for the Thematic Encoding condition being .42, and .07 for the Thematic Retrieval condition. These data are depicted on Figure 2. Recall that on this final test, participants were supposed to accept all previously presented items (i.e., both list items and lures), and thus the conditions are directly comparable. The

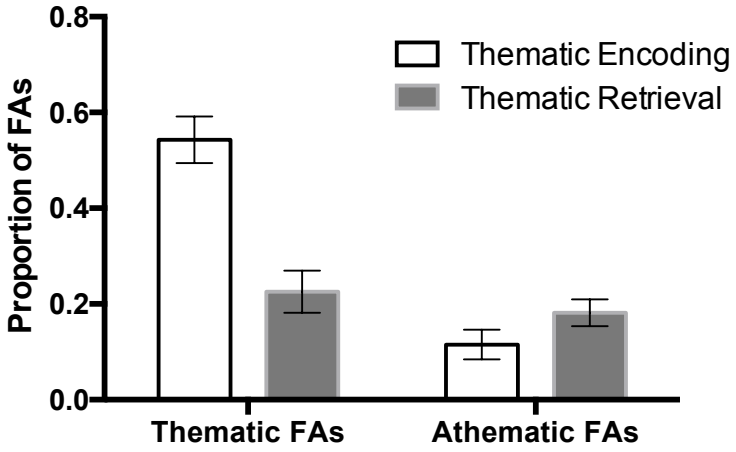

Fig. 2
Proportion of thematic and athematic false alarms (FAs) for Thematic Encoding and Thematic Retrieval conditions
Note: Vertical bars denote the standard errors of the means

difference that emerged points to a conceptual learning deficit for participants in the thematic retrieval condition. These were the participants for whom the thematic information was conveyed by the lures during the recognition tests.

Before we make further claims about these results, there are at least two caveats to deal with. First, in our paradigm, participants in the thematic encoding condition saw the thematic item twice in each study-cycle (i.e., once at study and another time during the test, through the "old" items), whereas participants in the thematic retrieval condition saw the thematic item only once in each study-test cycle (i.e., during the test, through the lures). Second, the thematic encoding participants received the thematic information blocked in each study list, whereas thematic retrieval participants received the thematic information interspersed through each recognition test. Both repetition (Benjamin, 2001) and blocking of categorized sub-lists (Jacoby, 1972; McDermott, 1996; Mulligan, Guyer, & Beland, 1999; *cf.* Kornell & Bjork, 2008) enhance relational and conceptual processing and thus the so-called relative deficit for participants in the thematic retrieval condition may simply result from these uninteresting factors. In the next study, we addressed these alternative accounts.

Experiment 2

In order to address the possible sources of confounding variables in Experiment 1 we used a within-subjects design to explore the occurrence and peculiarities of a relational processing deficit at test. Moreover, we changed the nature of the final recognition test such that the relative conceptual deficit found in Experiment 1 would no longer become detrimental for performance. Participants had

to study word lists composed of exemplars of two common categories (e.g., categories A and B), and perform recognition tests where all the targets were from only one of the studied categories (A, for example) and all the distracters were drawn from a new category (C, for example). As in Experiment 1, this was done across three study-test cycles with new categories on each cycle, and a final recognition test on the items presented along the different cycles of the experiment (study lists and test lists). This design allows us to compare thematic false memories arising during encoding and retrieval (i.e., comparing category B and category C false memories in the final test, following the aforementioned example), while discarding the alternative explanations of the results in Experiment 1.

Method

Participants

Ninety undergraduates from Universidade de Lisboa participated in the experiment in exchange for course credit or payment. This sample size was determined in order to have at least 10 participants per cell in our between-participants factor (category order for counterbalancing purposes). The data from one participant was removed due to patently low performance (100% hits on test 1, but 0 on tests 2 and 3).

Design

The design was mixed 9 Category-order Versions (V1 to V9) X 2 Category Locus (Encoding vs Retrieval), the first factor being between-participants and the latter within-participants. All participants went through 3 study-test cycles, being exposed to 3 categories per cycle (with only one appearing both at study and as target

at test, according to the category-order condition), and performed a final cumulative recognition test of memory for distracters and targets, which was the same for all participants. We used a partial latin-square design to define the category order conditions, ensuring that all categories were experienced in all study-test cycles (1 to 3) and both at encoding and at retrieval (this between-participants factor only served counterbalancing purposes; see Table 3). The final test was composed by exemplars that appeared either at study or at test, and by new exemplars of these categories that were never experienced. Using this design we are able to directly compare, in the final recognition test, conceptual processing of the category items by looking at the false alarm rates on categories that appeared either during encoding or during retrieval.

Table 3
Depiction of the distribution of categories through study-test cycles for the nine between-participants factors (versions) used for counterbalancing purposes in Exp. 2. Each letter corresponds to a category. The categories of interest on the final test (test 4) are the ones that appeared only once on each cycle (e.g., in version 1: B, C, E, F, H and I).

version	study categories 1	test categories 1	study categories 2	test categories 2	study categories 3	test categories 3
1	A and B	A and C	D and E	D and F	G and H	G and I
2	B and C	B and D	E and F	E and G	H and I	H and A
3	C and D	C and E	F and G	F and H	I and A	I and B
4	D and E	D and F	G and H	G and I	A and B	A and C
5	E and F	E and G	H and I	H and A	B and C	B and D
6	F and G	F and H	I and A	I and B	C and D	C and E
7	G and H	G and I	A and B	A and C	D and E	D and F
8	H and I	H and A	B and C	B and D	E and F	E and G
9	I and A	I and B	C and D	C and E	F and G	F and H

Materials

Some of the materials in the study and test lists were the same as in Experiment 1, but because now each of the three study-test cycles involves three categories, exemplars from other categories from the Portuguese norms (Pinto, 1992) were used, thus totaling nine categories: sports, fish, atmospheric phenomena, fruits, birds, clothing, occupations, mammal animals, and body parts. To create the lists we

ordered the category exemplars in terms of typicality, selected the twenty most typical exemplars, and removed the items from the positions 3, 6, 9, 12, and 15. Therefore, there were nine categorized lists of fifteen items each, which could be used as part of the study lists or part of the test lists. This way, two lists of 15 words each from two different categories composed each study list. Each test comprised one of the presented lists (15 targets), and a new category list (15 distracters).

For the final recognition test, the previously removed words from positions 3, 6, 9, 12, and 15 were used as distracters, while words from positions 8, 10, 11, 13, and 14 (from the original category list) were used as targets (they were part of the study or test lists, depending on the category order condition). As participants experienced nine categories throughout the experiment, this final test had 90 items to respond to (5 targets and 5 distracters per category). Remember that on this final test an item is deemed a target if participants previously experienced it either at study or at test, and is deemed a foil when it is an exemplar of the experienced categories that was never experienced itself.

Procedure

The procedure was similar to the one in Experiment 1: participants were randomly assigned to one of the version conditions, and received general memory instructions ("pay attention the following list because your memory about it will be tested"). After the onset of the experiment the words were presented at the center of screen, at a rate of 2000ms/word. After the study phase participants performed arithmetic problems for 1min to prevent rehearsal. When they finished this filler task, participants received the test instructions. They were told that they were going to see some words on the screen, an their task was to press the 'A' key if they thought the

word appeared on the study phase (old) and to press the 'L' key if they thought it did not appear (new). During the test, the words appeared one by one at the center of the screen, with participants' responses making the new word appear. After the test, a screen indicated that a new list was going to be presented, to which they should pay attention because their memory for it would be subsequently tested, thus initiating a new study-test cycle. This was repeated two more times, totaling three cycles. Before the final test, a new instruction screen informed the participants of the different nature of the task, so that they should respond "Old" or "New" considering the whole experiment, i.e., all items that appeared on any study list or test.

Results and Discussion

Study-test cycles

To analyze performance on tests 1 to 3, we first looked at the pattern of hits and false alarms. As illustrated on Table 4, overall hits were at ceiling level (*M*=.87) and false alarms were at floor level (*M*=.05), indicating that the participants understood the test structure, and used category knowledge to boost their performance. Globally, there were no significant differences across study-test cycles.

Table 4
Mean Hit and False Alarm Rates for tests 1-3 in Exp. 2

	test 1	test 2	test 3
Hits	0.89 (0.01)	0.87 (0.02)	0.85 (0.01)
FAs	0.04 (0.01)	0.06 (0.01)	0.06 (0.01)

Note: standard errors in parentheses

However, in our previous work (Garcia-Marques et al., 2015) we found that typicality tends to moderate performance (d') and response bias (c). With this exploratory goal in mind, we split the test items between high and low typicality (dropping one mid-typicality item from the targets and one from the distracters) and included typicality

as a variable in the following analysis. We first compared hits for high and low typicality items across tests, running a 2 Typicality (High vs Low) X 3 Test (1 vs 2 vs 3) within-subjects ANOVA on the proportion of hits. Only a typicality main effect emerged, $F(1, 88) = 12{,}06$, $p<.001$, *Cohen's d* = .74, with greater proportion of hits for high typicality (M=.90) than low typicality (M=.86) items, suggesting that participants were using category knowledge to inform their responses, akin to the classification effects in recognition in Higham and Brooks (1997) and to our own findings (Garcia-Marques et al., 2015).

Next, in order to explore whether participants showed sensitivity to category knowledge in the type of false alarms responses, we ran a 2 Typicality (High vs Low) X 3 Test (1 vs 2 vs 3) within-subjects ANOVA on the proportion of false alarms. In line with our original hypothesis, no main effects (F<1) nor interaction, $F(2, 176)=1.10$, $p>.1$, were significant, showing that participants were not differentially producing false alarms relying on typicality, nor there were changes across study-test cycles, even when identifying the distracter's category was a pre-condition for better performance. Note that differences in false alarms for high and low typicality items across test cycles would be a signature of conceptual processing at test, as these categories only appear during the test phase and their conceptual processing (i.e., figuring out the category) is encouraged from test 1 onwards.

Final recognition test

We next analyzed results of the final recognition test to investigate whether participants exhibited dissimilar conceptual processing of items presented either at test or at study. As this final test contained old and new items of the experienced categories, and the counterbalancing assured that all categories were experienced in

all positions (at study or at test, at each study-test cycle), comparing the participants' ability to discriminate old from new words within each category will give us a measure of conceptual processing occurring at encoding or retrieval. We also explored how participant's performance is affected by the item's category typicality.

First, for all participants, we collapsed responses in terms of the locus of the categories during study-test cycles 1-3, so that we could compare hits and false alarms on the categories that appeared once either during the study phase or the test phase as distracters. A hit consisted in responding 'old' to a word that appeared either as a study item or as a distracter, and a false alarm consisted in responding 'old' to a new word from the category that appeared either in the study or the test phase. We conducted a 2 Typicality (High vs Low) X 2 Locus (Study vs Retrieval) X 2 Response (Hits vs False Alarms) within-subjects ANOVA on the responses proportions. Overall, performance was high, with the mean proportion of hits (M= .75) being larger than the proportion of false alarms (M= .43), as evidenced by a significative main effect of response, $F(1, 88) = 329.04$, $p < .001$, *Cohen's d* = 3.87. A main effect of typicality was also obtained, with a higher proportion of 'old' responses for more typical (M = .62) than less typical (M = .57) items, $F(1, 88) = 13.15$, $p < .001$, *Cohen's d* = .77. Also, these factors interacted, $F(1, 88) = 9.23$, $p = .003$, *Cohen's d* = .65, showing that while the proportion of hits was the same for typical and less typical items, the proportion of false alarms was higher for high typicality (M = .48) than for low typical items (M = .39). The locus of the categories main effect failed to reach conventional statistical significance, $F(1, 88) = 3.02$, $p = .09$, *Cohen's d* = .37, but showed that the proportion of 'old' responses produced to items from categories that appeared at study (M = .62) was numerically superior to that of items from categories that appeared at retrieval (M = .57). The interaction with typicality reached

conventional statistical significance, F(1, 88) = 4.04, p = .05, *Cohen's d* = .43, indicating that more 'old' responses were produced for high than for low typicality items only for categories presented at study (M = .65 vs. M = .58), while for the categories presented at test this did not happen (M = .58 vs M = .56). The category locus by response interaction and the three-way interaction failed to reach statistical significance (F<1). As the manipulation of the locus of knowledge on this experiment was within participants and only thematic false alarms were possible to produce (contrary to Experiment 1), we calculated discrimination (d') and bias measures (c) using the Signal Detection Theory framework (e.g., Macmillan & Creelman, 2005; Green & Swets, 1966). This enabled us to examine the impact of conceptual knowledge locus on both the discriminability performance and on the amount and quality of information required to make a memory decision. The sensitivity (or discriminability) measure, d', allowed us to check whether participant's discrimination performance differs for items from categories presented at study or at test; the bias measure, c, allows us to check whether participants will need less information to accept an item presented either at study or at test as old. We also analyzed the influence of the item's typicality on these measures as a proxy to conceptual processing (differing sensitivity or bias for high and low typicality items is interpreted as the items having been encoded conceptually, i.e., according to their taxonomic structure). We first compared d' scores for categories that appeared either at study or at test. Performance was superior when the category locus was at test (M = 1.17) than when it was at study (M = .95), t(174) = 2.03, p = .04, *Cohen's d* = .31. We interpret these results as a deficit in processing items conceptually at test, even if this was encouraged by the task structure. Interestingly, this is a case where a deficit actually leads to better performance: inhibited conceptual encoding of items at test led

to greater discrimination memory, given that distractors on the final recognition test were members of the experienced categories. We then compared c scores for categories presented at study or at test. While criterion scores for categories presented at study were lower (M = -.37) than for those presented at retrieval (M = -.33), suggesting a slightly more lenient criterion, this difference was not significant, t(174) = .44, p = .66, *Cohen's d* = .07.

To further examine this pattern, we ran the same exploratory analysis we did for study-test cycles, taking into account the item's typicality, by performing a median split on targets and distracters, dropping the mid typicality items from each group (one item from the targets, and one from the distracters). We first conducted a 2 Typicality (High vs Low) X 2 Locus (Study vs Retrieval) within-subjects ANOVA on d' scores. We obtained a main effect of typicality, with more typical items resulting in worse performance (M = 1.16) than their lower typicality counterparts (M = 1.49), F(1,88) = 5.64, p = .02, *Cohen's d* = .46, with neither category locus nor the interaction reaching significance (F<1). This superiority of low typical items is, again, similar to the false memory effects with categorized lists (Shiffrin et al., 1995; Buchanan et al., 1999), and was expected to arise given our specific task demands (the categorical nature of the lists was salient, and both targets and distracters were exemplars of the same category), confirming that participants grasped the task structure. Note that the absence of the category locus main effect (and interaction) is not comparable to the previous analysis because of the dropping of the two mid typicality items. Next, we conducted a 2 Typicality (High vs Low) X 2 Locus (Study vs Retrieval) within-subjects ANOVA on the c scores. In terms of main effects, category locus was not significant, F(1,88) = 2.35, p = .129, *Cohen's d* = .33, but typicality was, F(1,88) = 7.58, p = .007, *Cohen's d* = .59. Participants were overall

more lenient towards accepting higher typicality (M = -.44) than lower typicality items (M = -.31) as old, again reminiscent of Higham and Brooks' (1997) classification effects on recognition. Interestingly, the interaction between typicality and category locus was significant, $F(1,88) = 4.83$, $p = .03$, *Cohen's d* = .47, showing that this leniency towards accepting highly typical exemplars was greater for categories experienced at study, with those that appeared at test not showing differences in terms of typicality. These results are depicted on Figure 3**.** We interpret this pattern as further evidence that only the categories presented at study were subjected to conceptual processing, as leniency towards accepting highly typical items is a sign that these items were processed in terms of their taxonomic structure. The absence of difference in criterion for high and low typical exemplars experienced at test is, thus, interpreted as a deficit in conceptual processing of these items.

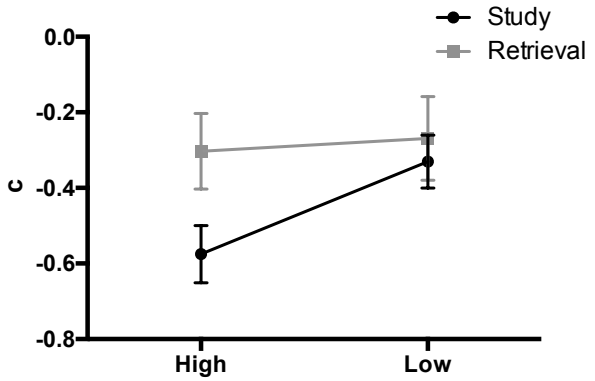

Fig. 3 Mean criterion (c) scores for high and low typicality items experienced at the study and retrieval phases in Exp. 2. *Note*: vertical bars denote the standard errors of the means.

General Discussion

A consistent pattern emerged from the present studies, where even when task structure strongly encourages conceptual processing of information at test participants exhibit a deficit in doing so. This hindrance in new learning at retrieval was shown to occur in two tasks with different structures, but that shared the usefulness of

conceptual knowledge (categorical, in our case) in maximizing performance. On Experiment 1 we examined this deficit in terms of implicit conceptual learning, by comparing false alarms that were consistent or inconsistent with the categorical lists experienced in interim study-test cycles, and shown that the consistent-inconsistent difference was lower when the participants processed the category at test, whereas when participants processed the category at encoding the difference was larger, with a significantly higher rate of produced related false alarms. On Experiment 2 we show that this deficit also occurs in discriminative memory, by comparing participant's ability to discriminate old and new items from categories encountered in interim study-test cycles. In this case, this hindrance in conceptual processing benefited performance, as targets and distracters were exemplars from the same category. If participants had processed the interim study lists nonconceptually, discrimination between old and new items on the final recognition task would have been facilitated, as their memory would be less cluttered with irrelevant category knowledge, and/or activated exemplars from the encountered categories.

We interpret this phenomenon as a deficit, albeit a relative one, as only information processed during the study phase shows signatures of conceptual processing, namely differential responding according to conceptual characteristics of the items such as semantic relatedness or category typicality. Despite the fact that the task structure itself encouraged participants to process items at test conceptually (as in Garcia-Marques et al., 2015; Higham & Brooks, 1997) their ability to do so was hindered. We consider this to be a relative deficit because there are always contexts where not processing conveyed information conceptually, and thus neglecting knowledge, will be beneficial for performance, as in Experiment 2 (a notorious case where something similar happens, i.e., more knowledge leading to a hindrance in

performance, is the fan effect; e.g., Anderson, 1974; Anderson & Reder, 1999). Moreover, these data are in line with the studies showing impairment of new learning at retrieval (Finn & Roediger, 2013; Davis & Chan, 2015), where the new learning was relational (e.g., Einstein & Hunt, 1980) in nature.

The specific causes of this deficit are not clear yet, but we speculate that it may well be related to the characteristics of one's mental state when at retrieval, what Tulving (1983) has dubbed "Retrieval Mode". According to Tulving and his collaborators (Tulving, 1983; Lepage, Ghaffar, Nyberg, & Tulving, 2000; Nyberg et al., 1995), during the retrieval mode one mentally holds in the background of focal attention a segment of one's past, treating incoming and on-line information as ''retrieval cues'' for that particular segment, and during the retrieval mode one also refrains from retrieval-irrelevant processing. We believe that this decoupling of attention from perception and its corresponding shift to an internal attentional focus can carry important consequences; namely, that new information present while one is engaged in the retrieval mode will tend not to be learned or retained. Further research is needed to test this hypothesis, as in our experiments we did not directly manipulate the retrieval mode. This typically involves instructions (e.g., Karpicke & Zaromb, 2010) whereas in the present experiments we guided encoding strategies via test structure, leaving the instructions as broadest as possible.

In addition, we speculate these effects will have important practical consequences. Namely, in the form of costs that can co-exist but also limit the straightforward real-world application of the so-called testing effect, both in it's direct and indirect facets (Nunes & Karpicke, 2015; Roediger & Butler, 2011). More specifically, we wonder whether the improvement of performance achieved by

repeated testing of the same information is associated with a cost in terms of ability to learn new information.

CHAPTER V. General Discussion

1. Summary

The central goal of this dissertation was to put forward and explore the role of repeated retrieval in guiding further processing of new but similar information via learning of the retrieval requirements. Although many aspects regarding the direct effects of repeated retrieval of the same information have been thoroughly explored (e.g., Roediger & Butler, 2011) the same is not true for the so-called indirect (e.g. Arnold & McDermott, 2013) or forward (e.g., Pasttotter & Bäuml, 2014) effects of testing. We present evidence showing that we process information as a function of the particular ways we retrieved similar information in comparable learning situations, and also extend this notion using further dimensions of the stimuli (thus suggesting that the proposed phenomenon is a general aspect of our memory functioning), while accommodating this evidence with results showing impairments on the learning of new information during retrieval.

In Chapter II (Garcia-Marques et al., 2015), we propose that repeated study-test episodes allow for the learning of the test structure and its requisites, which will guide further encoding of similar, but new, information in subsequent comparable study-test episodes. Specifically, we show how changing the nature of the distracters in recognition tests impacts the way we approach further information. By manipulating whether distracters were related to the studied lists (i.e., from the same taxonomic categories) we were able to observe how participants adapted their encoding strategies across study-test cycles towards the dimensions that the previous tests favored, namely, relational or item-specific approaches to the items; we also

argue that these differences in encoding strategies only become apparent when the structure of the tests changes abruptly from what is expected.

In chapter III (Marques, Garcia-Marques & Orghian, 2016) we provide converging evidence for this phenomenon using a different paradigm where the nature of the task is manipulated to encourage processing of either semantic or spatial aspects of the items, while keeping the encoding conditions stable. We show that when the test requisites are location-based and word meaning is less likely to be useful for good performance, participants tend to ignore the semantic aspects of words, as was evident in the performance drop when the requisites change abruptly to semantic (Exp. 1) or when participants had to generate free-associates to low typicality studied items (Exp. 2).

In Chapter IV (Marques & Garcia-Marques, 2016) we show that even when the task structure strongly encourages conceptual encoding of items, such as when identifying the categories to which targets and foils on a test belong, this processing is hindered at the moment of retrieval, in line with recent research showing that during retrieval new learning is impaired.

In the present chapter, we will further discuss the implications of each of the empirical chapters and some of its limitations. Also, we will suggest overall limitations of the present work, and discuss the importance this phenomenon for memory theories, the specificities, benefits and costs of repeated retrieval, and how it fits on current theories of memory and learning.

2. Further discussion on the learning of test requisites and its impact on encoding

As already stated along the introduction, for several decades the experiments devised to study human memory functioning generally equated learning with the

encoding phase, with retrieval being a mere vehicle for assessment of the information that was successfully/unsuccessfully encoded and had suffered the effects of decay in varying degrees over time (cf. Roediger, Knight & Kantowitz, 1977; Lewandowski & Oberauer, 2009). This vision of retrieval processes evolved, now retrieval being considered a 'memory modifier' (Bjork, 1975; also see Raaijmakers & Shiffrin, 1981) and one of the central processes in an adaptive and self-actualizing memory system, and lead to an increased interest on the role of retrieval both in theory development (e.g., Anderson, 1990; Bjork & Bjork, 1994) as in applied settings (e.g., Karpicke & Roediger, 2006; Nunes & Karpicke, 2015). Actually, it was the use-inspired characteristic of the general theme of retrieval (Stokes, 1997; see Karpicke & Grimaldi, 2012) that greatly resulted in an increase in funding (e.g., Fitzpatrick & Dolezalek, 2013) and thus in the scrutinizing of retrieval processes. With the focus on how retrieval processes – to a larger extent than encoding processes – can benefit long-term retention, as evidenced by increased performance in delayed memory tests the literature has accumulated myriad effects depicting memorial benefits of repeated retrieval.

The work presented on Chapter II (Garcia-Marques et al., 2015) differs from this approach both in theoretical purposes and methods. As we were not interested in investigating the conditions that lead to better long-term retention of specific materials, but on dynamic changes in processing strategies that accompany our repeated encounters with similar study-test situations, each study-test cycle was about new information that shared an underlying structure (i.e., graded category structure; e.g., Barsalou, 1985). Moreover, our critical test (the forth, in both experiments) was administered shortly after the last study phase. This way, changes in encoding strategies induced by the structure of the first three tests could be assessed by looking

at how participants approached the materials at the fourth study phase as evidenced by their performance patterns in the final test where the requisites changed without notice.

Our results illustrate how participants' strategies evolve across study-test cycles where the encoding context remains stable, but the retrieval context was manipulated. The data relative to the progressive unbiasing of criterion for highly typical items in participants tested with related distracters suggests that they were abandoning an alluring relational processing strategy (as the nature of the materials, common categories, was easily discernible), and moving towards processing a more adequate dimension of the materials, the distinctive aspects of individual items (Einstein & Hunt, 1980). By including an abrupt change in the nature of the last test, we could observe the type of information (i.e., the appropriate stimuli dimension) that was being attended to by participants in both groups (especially when this final test structure was novel for all participants, as in Experiment 2).

The important message from this chapter is threefold. First, we advanced in establishing the phenomenon of "retrieval-induced strategy adaptation" as a further *indirect testing effect*. This proposed phenomenon differs from previously described indirect effects of retrieval in two important ways: a) it refers to the study and retrieval of new information on each cycle, and not of the same information (cf. Arnold & McDermott, 2013; Soderstrom & Bjork, 2014) and b) it refers to adaptive and intelligent strategy shifts without specific instructions on how to better process the stimuli for the tasks at hand as instructions only broadly indicated that participant's memories would be tested, with no reference to strategies or operations that could enhance performance (cf. Pyc & Rawson, 2010, 2012). Secondly, we described a set of conditions where memory and concept learning-like effects are to

be intertwined at two levels of explanation: 1) when there is a possible overlap between item's episodic status and conceptual features, recognition tasks can benefit from the use of rules or exemplar similarity to ensure good performance (e.g. Higham & Brooks, 1997), and 2) repeated (and sequential) experience with a given test requisite appears to lead participants to disregard stimuli dimensions that are deemed irrelevant and to focus on those deemed relevant, akin to what happens in category learning tasks (e.g., Blair, Watson, Walshe & Maj, 2009; Deng & Sloutsky, 2015; Rehder & Hoffman, 2005). Our results regarding "Related Lures" participants' criterion scores across test cycles and the subsequent better performance in the final tests are illustrative of this. Thirdly, it is informative of the relative status of specific encoding or retrieval strategies as optimal for memory performance (cf. Zaromb & Roediger, 2010), as it is possible to encounter contexts where a previously deemed non-efficient strategy will be the most adequate one, depending on the structure of critical tests. This aspect is a notable example of the learning vs. performance duality (e.g., Soderstrom & Bjork, 2015): sometimes learning occurs in the absence of performance increments, and sometimes performance increases but little learning is taking place. While one can usually consider that the latter situation is to be avoided and the former encouraged, it is important to note that this is mostly adequate in educational or training contexts, where the main goal is to create conditions that lead to long-term and permanent information retention, skills change, and comprehension, as assessed by clearly defined requisites in critical tests. The phenomenon described in Garcia-Marques et al., (2015), besides being applicable to training and educational contexts (students appear to be sensitive to the particular structures of tests on their memory, and capitalize that knowledge; e.g., Attali & Bar-Hillel, 2003), can also account for daily-life uses of memory, whenever the same retrieval requirements are

repeatedly experienced. Another way to state this is that whenever retrieval is experienced, learning about the relation between studied items and retrieval requirements will occur, and it should always be possible to assess this learning, if we think of critical tests in a relative light. These implications will be further examined ahead on this Discussion.

2.1. Limitations and future studies

One possible limitation of the experiments presented in Chapter II is that it is unresolved whether the pattern of results in final tests really stems from learning to use item-specific or relational processing, as defined by Einstein and Hunt (1980; Hunt & Einstein, 1981). This was not the central goal of our investigation, which was focused on adaptation of encoding and response strategies to task requisites on a broader level (as is explored on Chapter III), i.e., how participants learned to focus processing on differences or similarities/common features (see Hunt & McDaniel, 1993). However, a better assessment of these two specific processing strategies could provide even stronger evidence for our claim that Related Lures participants learned to avoid a relational processing strategy via experiencing test structure, and better situate our results in the relational/item-specific literature. This is also important as it could give us a clearer picture of the procedural aspect of the described phenomenon, and to better distinguish it from a more general (but not unrelated; see Chapter II) explanation of our results, such as merely being the outcome of experiencing "desirable difficulties" (for reviews see Bjork, 2011; McDaniel & Butler, 2011; for an example of recent research interpreting these data as such, see Cho & Neely, 2016). Although the progressive unbiasing of criterion moderated by typicality and the final

superior free-recall performance for Related Lures participants are indeed characteristics more associated with less reliance on relational strategies, it can be possible to further define these contrasting strategies by employing some modifications in the paradigm.

First, across the initial three study-test cycles, assessing participants' study-time in the study phases could give us a proxy of participants' processing strategy. We employed experimenter-paced study settings, in which participants would see the word on the screen at a rate of 2000 ms/word but, as N. Soderstrom stated, "all study is self-paced" (personal communication, August 14th, 2016), i.e., even in experimenter-paced paradigms we can't be sure about the attention (both qualitatively and quantitatively) that participants devote to stimuli. Thus, letting participants study at their pace and measuring it can be highly informative on their attention allocation procedures, and their goals towards the stimuli (e.g., Dunlosky & Ariel, 2011; Soderstrom & Bjork, 2014). Our hypothesis regarding study-time would be as follows: on the first study episode study-time should be evenly distributed on both conditions, moderated by typicality[7] (when unconstrained, participants tend to allocate more time in deemed-easier items, if we take typicality as a proxy for item difficulty; e.g., Metcalfe & Kornell, 2003; Son & Metcalfe, 2000); after experiencing the recognition tests with unrelated lures, we expect participants to devote more time on the items on the beginning of the list until they attain the respective category, and then the study-time should drop, as their goal of concept attainment is already

[7] While it is intuitive to hypothesize about how typicality would impact study-time allocation, further research on the sensitivity of time measures on this variable is needed. On a further note, the proposed manipulation could also result in no differences between conditions: for example, Benjamin (2003) showed that participants usually predict better future recall for common words, but postdict (after being tested) better future recall of uncommon words; this dynamic could affect results in the proposed paradigm as multiple tests are employed and a general time-allocation bias toward less typical words could occur.

satisfied (also, the less typical items should be allotted more study-time on the beginning of the list); when participants experience a recognition test with related lures, we expect study-time to be evenly distributed along the study lists, while moderation by typicality should become less accentuated across cycles. This would be a strong indication that participants in the Unrelated Lures condition are indeed relying on conceptual knowledge extraction and use, while Related Lures participants avoid doing so.

Another way to examine the processing strategies could be employed in the final free-recall test, by assessing the time course of cumulative-recall scores (for a review see Wixted & Rohrer, 1994), that is, how many items were recalled by participants over time during the test. Research by Daniel Burns and collaborators (e.g., Burns, 2006; Burns & Hebert, 2005; Burns & Schoff, 1998) has shown that relational and item-specific processing of words result in markedly different cumulative-recall curves: relational processing is associated with very high recall scores in the early phase of the test, while item-specific processing is associated with poorer initial performance, but overall resulting in better final performance. Measuring and analyzing these data patterns would provide us further evidence of ways participants approached the final study list, especially since in our experiment the intrusions rate was very low and thus could not be informative.

3. Further discussion on the effect of testing word location in semantic activation

Departing from the phenomenon described in Chapter II, we went on to extend it to further stimuli dimensions. If the adaptation of encoding (and retrieval) strategies to the experienced retrieval requirements is an ubiquitous characteristic of human

memory, it should occur along several attainable dimensions of stimuli. Quoting Wendell Garner, "There is nothing to prevent the human information processor from asking a series of yes-no questions such as: 'Is the stimuli red, is the stimuli free, is the stimuli square, etc'." (Garner, 1978, p. 105); our hypothesis is that experiencing retrieval requirements in comparable learning situations will guide "which questions are asked" about stimuli, i.e., according to which dimensions should the stimuli be processed, following the classical characterization of stimuli as a multidimensional collection of attributes and features (e.g., Tulving & Thompson, 1973; Underwood, 1969; Wickens, 1970). With the goal of gathering converging evidence on retrieval-induced strategy adaptation, we devised a task where the specific retrieval requirements were manipulated, as to favor either meaning-based (or semantic) or spatial-based (or location) processing. These dimensions were chosen as they both are central for representation and retrieval (Hasher & Zachs, 1979; Underwood, 1969), and because they naturally co-occur at encoding. As in Chapter II, we expected that experience with a given requisite would guide processing strategies towards relevant dimensions and away from the irrelevant ones, thus leading to performance drops when the task requisites change abruptly (Exp. 1), and to performance patterns illustrating disregard or inhibition of the deemed-irrelevant dimension (Exp. 2).

The research presented at Chapter III is informative and important in several ways. First, it is a conceptual replication of Garcia-Marques et al., (2015), again showing how repeated experience with specific test requirements tunes our encoding (and retrieval) strategies, and also showing that that becomes apparent when the test requirements change abruptly and accordingly. Also, it shows the asymmetric nature of stimuli dimensions, in terms of their memorial impact and position in the automatic-deliberate continuum (e.g., Caldwell & Masson, 2001; Hasher & Zachs,

1979), as learning to disregard the semantic dimension of the stimuli resulted in steeper performance drops when the requisites changed to location than the other way around. Importantly, it also provides a clearer picture of the central goal of this dissertation, by illustrating the pervasiveness of retrieval-induced strategy adaptation, and hinting at the influence of inhibitory and facilitatory processes as accounts of the effect.

Given our proposed inhibition explanation of the effect, some aspects related to our claims and methods, and also related to the inhibition concept in cognitive psychology itself (see Macleod, 2007) should be further clarified. First, as we state, the experiments that we present were not devised as tests of an inhibitory account (we were interested in gathering further evidence for retrieval-based adaptation of strategies), although our Experiment 2, by using a novel and deemed unrelated task to assess semantic network activation via free-association response time, can be considered an apt criterion for inhibition-related phenomenon[8]. Another important aspect is our definition of inhibition (researchers should always state the employed definition of inhibition, due to some ambiguity in the use of this concept; see Macleod, 2007). In the discussion of Chapter III we put forward explanation of our effect that encompasses inhibition in the general processing of stimuli dimensions (as opposed to item, category, or location-specific; e.g., Anderson & Green, 2001; Anderson & Spellman, 1994; Hasher & Zacks, 1988; Tipper, 2001), moderated by retrieval-induced test expectancies, and acting on new but comparable sets of stimuli. Thus, we are referring to what Hasher and Zacks (1988) defined as attentional control

[8] In a similar vein, Experiment 2 of Garcia-Marques et al., (2015), Chapter II, can also imply inhibitory mecanisms acting on the Related Lures participants, hindering conceptual processing and focusing on item-specific information, thus resulting in a free-recall advantage. Despite not constituting a truly independent test (Anderson & Spellman, 1995), as the study-phase preceding free-recall was presented as just another study episode much alike the previous ones, we consider the results an indication of inhibition, as the nature of the materials (common categories) should encourage conceptual processing.

of what information enters working memory, embedded in the more general and parsimonious definition of inhibition as "the stopping or overriding of a mental process, in whole or in part, with or without intention" (Macleod, 2007). In our paradigm, as study lists for all conditions are always members of the same common category, a relational and meaning-based processing strategy is expected to be developed early on (as is evident in Chapter II; Garcia-Marques et al., 2015), only to be overridden or tuned by the subsequent experience with specific retrieval requirements. That way, the larger performance drop at requisites switch in Exp. 1, and the null effect of episodic status on low typicality items in Exp. 2 when participants were previously tested with location cues constitute signs of inhibitory processes underlying the adaptation of strategies to test requisites.

3.1. Limitations and future studies

There are two main caveats in the research presented on Chapter III that might limit our conclusions in a related way. One is the mentioned asymmetrical proneness to the buildup of proactive interference (PI) across study-test cycles. As we mentioned, while for semantic tests on each cycle a new category is presented and is focal to the task, for location tests, where the focal aspect is not meaning-based and participants are encouraged to attend to the relative positions of the words, the same set of locations is repeated. It is reasonable to think that in the semantic condition, the previously attended contents would not hinder the processing subsequent ones in a significant manner, while for the location condition that is more prone to happen, as inherently semantic stimuli dimensions have been shown to be more effective in producing release from PI, when compared to more physical dimensions (Wickens,

1970) - on related examples, release from PI is more prone to occur when categories change between trials, or with bilinguals when language changes between lists (e.g., Goggin & Wickens, 1971; Dillon, McCormack, Petrusic, Cook & Lafleur, 1973). This aspect could weaken our results' diagnosticity, as it could obscure potential increments in performance due to practice (and adaptation to the requirements), in conditions with location tests (which can not be completely disentangled from an explanation based on the more automatic nature of spatial processing), and also leading to the second caveat by creating a disadvantage in analyzing the RT data on Exp. 2: as location (LL) condition participants exhibit significantly poorer performance than their semantic counterparts, and we were assessing their response times to free-associate to words tested on cycle 4 (and new words). This last caveat is also connected with our decision to use six old and six new items (three high and three low typicality) in the free-association task, as presenting participants with a higher number of old items could lead to contamination of the implicit test with deliberate processes and strategic responding (Hourihan & Macleod, 2007). As such, a conditionalized analysis of the free-association RTs on the correctly recalled items turns unfeasible, and such analysis could discard an explanation based on poorer performance[9].

Future studies that contemplate these caveats would have the following characteristics: the cards disposition should change across cycles (e.g., different polygons on each cycle, instead of the 4 x 4 quadrilateral; this would also resemble how some people actually play the Concentration game); the final free-association test could encompass more items, both our focal items (old and new, high and low

[9] Nonetheless, the current data constitutes a good depiction of qualitative differences in processing between conditions. Notice that requisites condition only impacted RTs for the low typicality items, and that post-hoc analysis shown that typicality impacted participants' performance across study-test cycles only in the semantic condition.

typicality) as fillers (so to attenuate explicit contamination); record participant's free-association verbal responses, so to check whether they are merely repeating the word that was previously paired with the target, and to assess whether participants in either condition were producing more related or more idiosyncratic responses.

4. Further discussion on the relative conceptual processing deficit at retrieval

In Chapter IV we further explore the parallels between memory and conceptual learning, and the role of retrieval. We aimed at integrating our results that show how test requisites by themselves can impact subsequent encoding and responding strategies, potentially guiding participants to approach memory tasks as conceptual learning tasks, with data showing how new learning (of relational type) is hindered during retrieval (e.g., Davis & Chan, 2015; Finn & Roediger, 2013; Kantner & Lindsay, 2010). This aspect is of importance, because in our research (Chapter II) we created conditions that lead participants to process the study items conceptually, as attaining the category would grant optimal performance in a recognition test, by classifying items that were members of the studied category as old. We wanted to investigate whether this "retrieval-induced strategy adaptation" would also apply to processing items at test, i.e., if conceptual information was only available at test, in order to evaluate some boundary conditions and to provide a clearer explanatory standing point: the notion that we are describing an higher-order effect that is not directly dependent on the specific stimuli at hand, but on the dynamic relationship between stimuli structure and test structure. Said in another way, repeated retrieval, besides its direct effects on specific items, appears to impact subsequent cognition by guiding focus on the relevant dimensions and disregard on the irrelevant ones; while

the moment of retrieval is central in guiding this effect, the learning processes implied are to be continuously at play in series of study-test cycles and not isolatedly at retrieval, as specific conceptual or relational processing during memory tests seems to be hindered.

To achieve this we developed adaptations of the paradigm presented in Chapter II (Garcia-Marques et al., 2015), manipulating the locus of conceptual information: at the study list or at the test, as distracters. We did that having two goals in mind: a) directly pitting encoding against retrieval, i.e., creating conditions where participants were encouraged (in the sense of Garcia-Marques et al., 2015) to process items conceptually either as study-items/targets or as distracters, and b) to further illustrate the relative status of processing strategies efficacy, as it is always conditional on task requisites.

On Experiment 1 of Chapter IV, we assessed the "conceptual deficit at retrieval" focusing on an implicit conceptual learning measure, the difference between false alarms that are congruent with the presented information and false alarms that are incongruent with it (Higham & Brooks, 1997), after participants went through a series of study-test cycles where the locus of thematic/conceptual knowledge was manipulated. On the encoding condition, study lists were exemplars of a category and at test all distracters were unrelated words, except for two new exemplars of the studied category; on the retrieval condition study lists were unrelated words and at test all distracters were exemplars of a category, except for two new unrelated words. In the final memory-for-foils test[10], participants in the retrieval condition did not

[10] In this test, we also included items that were experienced as targets because they were also experienced as distracters by participants in the two conditions. Moreover, we believe that this also focused participant's attention on the experiment as a whole, and not specifically on the distracters. Thus, our final test on Exp. 1 is not a pure "memory-for-foils" test in strict sense, even though we were not interested on performance on targets.

exhibit a significant difference in the production of thematic and athematic false alarms, contrary to their encoding condition counterparts who produced a high level of thematic false alarms. Despite the confounding variables described in Chapter IV (encoding condition participants were exposed to a higher number of thematic items overall, which were presented in blocked format; Benjamin, 2001; Jacoby, 1972; McDermott, 1996; Mulligan, Guyer, & Beland, 1999), that can obscure direct conclusions from these results, it is important to note that with this design we strongly encouraged participants in the retrieval condition to process distracters conceptually, as attaining the corresponding category would grant them superior performance. Note that performance was equally high in both conditions, and that participants in the retrieval condition consistently produced false alarms to unrelated distracters, a sign that they grasped the test structure (akin to Chapter II results), and were using category knowledge to inform their responses (in this case, to reject thematic items). Moreover, this experiment depicts a case where participants can treat a set of unrelated items as a category itself (even lists of unrelated stimuli are subject to organization; e.g., Tulving, 1962), thus enabling them to use a thematic-athematic stimuli strategy to inform their responses. It is not reasonable to think that participants in the retrieval condition attained high performance solely because their memory for unrelated lists was very accurate, at least to the level of their encoding condition counterparts, as semantic relatedness usually helps performance[11] (e.g., Epstein, Philips & Johnson, 1975; Poirier & Saint-Aubin, 1994). Thus, despite not constituting a strong test to our hypothesis that conceptual processing is impaired at retrieval even if strongly encouraged (due to the mentioned confounds), Experiment 1 of Chapter IV constitutes moderate evidence to our hypothesis, while also providing a clear

[11] While also increasing false memories (see Gallo, 2006).

illustration of the central topic of this dissertation: the phenomenon of retrieval-induced strategy adaptation.

In Experiment 2, we aimed at solving the aforementioned confounds, by using a within-participants design that allowed for the equal exposure to thematic information at encoding and at retrieval. This way all participants experienced a total of nine categories across three study-test cycles. The study-lists were members of two categories, while at test the targets were always exemplars of just one, while the distracters were exemplars of a new category. By using this design we were able to directly compare discriminatory memory on categories that appeared at encoding (part of study lists) or at retrieval (as the distracters) in a final test. Via counterbalancing the category presentation order, it was possible to assess whether participants were better at discriminating old from new exemplars of categories that were experienced only once, either at encoding or at retrieval.

Again, the results in the three study-test cycles mirror results presented at Chapter II, Unrelated Distracters condition, with hit rates nearing ceiling levels, and false alarm rates nearing floor levels, a sign that participants grasped the test structure and were using category knowledge to inform their responses, and a signature of conceptual processing of the items. Importantly, exploratory analysis showed that the hits during the study-test cycles were influenced by item typicality (participants were more likely to produce a hit when it was more typical of the category) while false alarms were not. We interpret this pattern both as further evidence of retrieval induced strategy adaptation, especially at encoding, and also as preliminary evidence of the central hypothesis of Chapter IV, as false alarms ("old" responses to items that only appeared during retrieval and never during encoding, across the interim study-test

cycles) were immune to a patently conceptual variable, category typicality (Barsalou, 1985).

On the final test, by analyzing participant's performance pattern on old and new items from the experienced categories, we presented evidence for the hypothesized conceptual processing at retrieval deficit. On a first note it is important to clarify that a conceptual processing deficit will always be of relative nature, as the structure of the assessing task will determine whether conceptual processing will aid or hinder performance. In the case of the final test of Experiment 2 of Chapter IV, such a deficit would grant better performance, as we were comparing discriminatory performance on items from the same category, and a more superficial and item-specific processing approach would lead to fewer thematic-based false alarms, as compared to relational/conceptual processing which would result in theme-based responding (e.g., Buchanan et al., 1999; Smith, Ward, Tindell, Sifonis & Wilkenfeld, 2001). The results indeed supported our hypothesis, with participants exhibiting better discrimination performance for categories presented at test, and also showing a bias towards accepting high typicality items only for categories presented at study (this bias was absent for categories experienced at test, indicating that participants were not sensitive to a conceptual variable when they were retrieving).

4.1. Limitations and future studies

As already stated, Experiment 2 was motivated by the problematic confounding variables in Experiment 1, whose main virtue was to create a condition where conceptual processing at test was not only highly encouraged but it was almost obligatory for maintaining good performance across study-test cycles. We concur that

the problems with Experiment 1 turned the results less conclusive, but in Experiment 2 these problems were overcame and the results further supported our hypothesis.

Nonetheless, we will address two important limitations and their tentative solutions in order to strengthen our argument. The first one is the absence of new distracters from new categories in the final test of Experiment 2. We didn't include these distracters as the final test would require responding to ninety items, after a series of three study-test cycles, which could lead to participant fatigue or boredom (see D'Angiulli & LeBeau, 2002) if we were to include further items. But including these new distracters would allow us to assess not only intra-category discrimination memory moderated by category locus, but also discrimination between experienced categories and completely new items, i.e., compare thematic false alarms from categories presented at study or test with athematic false alarms (akin to the comparison in Experiment 1). Also, we could assess veridical memory by comparing hits from categories at encoding or retrieval with unrelated false alarms. This would give us a good baseline to compare conceptual processing to, and better estimates of the magnitude of the proposed deficit.

Another possible limitation is related with our proposed explanation of the deficit as stemming from the consciousness and attention byproducts of entering the retrieval mode (REMO; Lepage et al., 2000; Nyberg et al., 1995; Tulving, 1983). As our experiments were designed with the objective of extending Finn and Roediger (2013)'s and Davis and Chan (2015)'s results to situations where conceptual processing at test was encouraged by retrieval-induced task adaptation, we did not directly manipulate REMO, which is usually done via instructions (e.g., Karpicke & Zaromb, 2010; Lepage et al., 2000), and instead manipulated the locus of conceptual knowledge in study lists and recognition tests to let the nature of the task itself guide

participant's strategies. Directly manipulating REMO, i.e., instructing participants to retrieve information from memory compared to a non-mnesic instruction would allow to test whether the conceptual processing deficit at test is a result of the consciousness peculiarities of REMO. As this manipulation is usually done by instructing for and providing recall tests (e.g., recall vs generation; Karpicke & Zaromb, 2010), it would fall out of the scope of Chapter IV and the present dissertation, because the task requirements of free recall tests lie on the studied materials. Thus, an experiment where recognition-type tests are administered but instructions are manipulated in a way that requires access to the past vs. access to world knowledge, for example, could address the present issues. One solution could be using well know symbols (e.g., the apple logo, traffic signs, playing cards suits, etc.) as to-be-studied materials; after studying a list of symbols under general instructions (e.g., "pay attention to these symbols as the following task will be about them"), participants would have to perform a recognition test that would contain the targets and distracters that were new/made-up symbols. The manipulation of interest would be the type of instructions provided at the beginning of the tests: intentional retrieval (e.g., "were the following symbols present on the list you just saw?") or unintentional retrieval (e.g., "have you came across this symbol in your daily life?". After a series of study-test cycles, they would receive a surprise final memory-for-foils test (akin to our test 4 in Chapter IV) where they had to respond whether they saw the symbols anytime during the experiment. This last test would contain symbols used as distracters and completely new/made-up symbols, and if it is indeed the intentionality of retrieval (one of the core assumptions of REMO theorizing; Nyberg et al., 1995; Tulving, 1983) that hinders our capability of conceptual processing at test, we would expect better performance for the unintentional condition. Moreover, if the stimuli were pre-tested

for frequency (e.g., how frequent are the symbols in daily life) we could expect that participants who received intentional retrieval instructions would gradually abandon a bias towards high frequency symbols, while this should not happen in the unintentional retrieval instructions participants.

5. Final Considerations

The research here presented delineated a peculiar phenomenon of retrieval-induced strategy adaptation, as a characteristic of a dynamic and self-actualizing memory system. This phenomenon is to be at play whenever there is a situation where repeated study-test cycles are to be experienced, which is central in current memory research due to the increased interest on the testing effect this last decade, while also constituting an important aspect of most educational scenarios, and even in our daily lives (outside the lab and the classroom) whenever one has to repeatedly encode and retrieve information about the world. In this final section we will draw some initial theoretical implications, and explore how the present research can inform best practices in applied settings.

First, it is important to reiterate how our proposal differs from similar recent accounts of forward testing effects and task adaptation, in order to better frame our results theoretically. As already mentioned, one of the key differences of our studies to similar literature that addresses the impact of practice and/or test expectations on subsequent cognition (e.g., Balota & Neely, 1980; Cho & Neely, 2016; Finley & Benjamin, 2012; Tversky, 1973) is our goal of exploring how experiencing testing events affords the human information processer with knowledge that can impact how new information is subsequently encoded for an expected test with the same requisites. The lack of directing instructions (in all studies from chapters II, III and

IV, pre-encoding instructions were general as "pay attention to the following words because your memory will be tested") and the fact that we used final tests in order to assess qualitative differences in encoding and representation of the stimuli, and not to ascertain which conditions resulted in better long-term retention and comprehension (as already reviewed here, that goal has successfully garnered interest from researchers for more than a decade, resulting in myriad robust results on the direct effects of retrieval; Nunes & Karpicke, 2015; Roediger & Butler, 2011) mirror this goal. A comparison with recent research that also used test practice to induce test expectancies with the goal of finding whether test expectancy impacts encoding strategies will help to further clarify our claims.

In a recent publication, Cho and Neely (2016) sought to test whether test expectancy would result in qualitative changes in encoding strategies. In their Experiment 1, participants were to engage in four (practice) study-test cycles before a final critical one. In cycles 1-4, the type of test was manipulated so that one group received cued-recall tests with semantic cues (e.g., "LEG", to elicit the correct response "ARM") and other group received cued-recall tests with orthographic cues (e.g., "A_M", to elicit the correct response "ARM"), so to elicit expectations of receiving that specific test before study-phases. On the last cycle the critical test contained mixed cues (semantic and orthographic) and the authors found a disordinal test expectancy effect (the "gold standard" of test expectancy effects; Finley & Benjamin, 2012; Lundeberg & Fox, 1991), so participants expecting semantic cues would perform better in semantic trials than orthographic trials and vice-versa. As the authors noted, by itself this effect doesn't comprise robust evidence for qualitative changes in encoding, as this could be a byproduct of specific test practice and reflect learnt response processes. To better test the encoding changes hypothesis, on their

next experiment they assured that both test expectancy groups experienced both kinds of tests equally, while now inducing expectancy through instructions before study. To achieve this, both groups now received three semantic cues (S) tests and three orthographic cues tests (O) in two distinct orders (SOOSSO vs OSSOOS) with the final critical test being the same as in their Experiment 1. The authors reasoned that if a test expectancy effect akin to the one in the previous experiment emerged, it would constitute evidence that test expectancy induced changes in encoding, and if not, the previous results could just be interpreted as mere practice effects. As the disordinal test expectancy effect now was not obtained, the authors concluded that test expectancy was not causing qualitative changes in encoding. While this conclusion might seem at odds with our proposal, we believe this data is consistent with our results and hypothesis. We concur that data from their Experiment 1 does not constitute robust evidence of retrieval-induced strategy adaptation, and that's why in our studies we also took in account word typicality when assessing possible impacts of practice in encoding (in Chapter II we show that participants tested with related distracters linearly unbias their criterion for highly typical words, disregarding a stimuli dimension not deemed useful; also their d' scores in interim cycles correlated with performance in the final free-recall test, while for unrelated distracters participants only the c scores of the third test correlated – negatively – with free-recall performance; in Chapter III we show in a different test how participants tested with location cues show no facilitation in eliciting free-associates for experienced low typicality words, while participants tested with semantic cues show a clear facilitation; also, in Chapter IV typicality only impacted responses for information present during encoding). Moreover, the design in Experiment 2 of Cho and Neely (2016) could elicit learning in the same vein of E. Bjork and collaborator's results

(e.g., Bjork & Storm, 2011; deWinstanley & Bjork, 2004; Storm et al., 2016) where the *generation effect* ceased to emerge when participants experienced retrieving information in generate and read trials, i.e., by employing both types of tests in different schedules (while also giving instructions) could have afforded participants with the opportunity to develop strategies that would go beyond mere test expectancy and practice effects[12], as different test formats could elicit an overall deeper or semantic encoding strategy (the development of optimal strategies should occur over time, with repeated experience on tests with specific requisites; when test requisites change often and instructions are provided, this could result in the adoption of a strategy that best fits both test's requirements). This alternative interpretation of Cho and Neely's (2016) results is an apt example of the issues explored in this dissertation.

Besides informing testing effects research by describing an indirect effect of retrieval practice, we believe our hypothesis and results fit well in contemporary memory and learning theories. For example, R. Bjork and E. Bjork's (1992) New Theory of Disuse (see also Bjork, 2011) states that in order to keep our memories current (e.g., remembering my present, and not former, phone number) some memories are adaptly forgotten (retrieval induced forgetting and retrieval inhibition effects aptly illustrate this; e.g., Anderson, Bjork & Bjork, 1994; Anderson & Spellman, 1995) and in some contexts some aspects (and not whole memory traces) are forgotten or inhibited if it helps attain a specific goal, such as when learning a new language inhibits access and causes forgetting of your native tongue (e.g., Levy, McVeigh, Marful & Anderson, 2007). Our proposal parallels these notions, as encoding information as a function of the specific ways we have used similar information in the past will help maintaining our attentional focus current, by

[12] In that line, we hypothesize that if word typicality was analysed across cycles in their Exp. 1, it should only impact participants in the semantic cues condition.

disregarding aspects of information that are deemed unuseful (this knowledge is acquired through past similar tests) to allow for additional focus on the relevant/useful aspects of stimuli in an uncluttered fashion. We also propose that our results fit an inhibitory account, i.e., that our proposed learning of test structure and its impact of encoding are mainly caused by inhibitory processes. While indeed the present research does not constitute a test on an inhibitory account, several signatures of such influences can be identified in our results (e.g., gradually reducing bias for high typicality items when that stimuli dimension is not useful, as shown in Chapter II; mitigated semantic activation with repeated experience with location cue tests, as evidenced by no facilitation in free associating "old" words, as shown in Chapter III; hindered conceptual processing of distracters, even if encouraged, as shown in Chapter IV). These results are in line with Lynn Hasher and collaborator's framework on inhibition as an important factor in regulating cognition throughout the lifespan (e.g., Hasher, 2007; Hasher, Lustig & Zacks, 2007; Lustig, Hasher & Zacks, 2007). These authors view inhibition as acting first at encoding - by limiting the amount and array of incoming information, so to facilitate the subsequent retrieval process and prevent that it becomes a difficult sorting of cluttered relevant and irrelevant information - and also at retrieval - by suppressing concurrent activations that might come to mind, with the goal of focusing memory search. The present empirical chapters II and III provide a good a description of learned inhibition at encoding via test experience, while Chapter IV can reflect a facet of inhibiting deemed-unuseful activations at retrieval[13].

[13] In our case, this impediment to conceptual processing at test lead participants to better discriminate old and new items from the same category, and to produce few thematic false alarms (Chapter IV). But these deficit-derived benefits are always to be considered as relative in nature; if a subsequent task was one of implicit memory, such as a fill-in-the-gap, this deficit would lead to poorer memory (but for older adults, who have been shown to exhibit a

We also believe that our proposal helps in the understanding of the daily uses of memory, as attaining knowledge about how to encode information in ways that maximize its adequacy to retrieval requirements should be central in regulating and updating memory functioning. One very apt example is social categorization (i.e., classifying people into groups according to shared characteristics), as encoding and retrieval of information about others is pervasive in our social functioning. To address social categories attainment and use the "Who said what?" paradigm (Taylor, Fiske, Etcoff & Ruderman, 1978; see also Klauer & Wegener, 1998) has been one the most successfully used, providing stable results with a wide array of variations in social categories and contexts, thus constituting a good proxy for everyday social information use (cf. Sani, Bennett & Soutar, 2005). In typical experiments, participants read or listen to statements by a number of people as part of a conversation or discussion, accompanied with each speaker's photo; the speakers differ, for example, in one social category such as race; in a subsequent test, participants are given the studied sentences and photos from the previous conversation agents for them to match. The classic result from this task is more within-category errors (e.g., misattributing a statement from one white person to another white person) than between-category errors (e.g., misattributing a statement from one black person to a white person), a strong indication of bias. However, we believe that with practice at retrieving this person memory information, one can adapt our processing strategies towards the specific discriminations that we encounter at test. For example, if two social categories are present in the stimuli at the same time (e.g., race and sex) but only one of them differs at interim tests, participants should gradually disregard the deemed unuseful category across study-test cycles thus

[13] *(cont.)* deficit at inhibiting information both at study and at test, good performance should be expected; see Rowe, Valderrama, Hasher & Lenartowicz, 2006).

boosting performance; if a surprise final test (in the same vein of the ones in Chapters II and III) required for the retrieval of both social categories, performance is expected to significantly drop. This would be a further indication that we are able to adapt our encoding (and retrieval) strategies to the nature of the requisites that we ordinarily face, and also a sign of the pervasiveness of the phenomenon.

Another context where the present research has implications is Educational Practice. As research suggesting the innumerous benefits of repeated testing for long-term retention, comprehension and transfer mounts (for a recent review see Nunes & Karpicke, 2015), tests are now seen as important components of educational policies and not as mere knowledge assessment tools (e.g., Benjamin & Pashler, 2015). We believe that the present research will help better characterize how repeated testing will impact learning in three important ways. First, it constitutes lab proxies of the phenomenon of students adapting their study methods to the experienced and expected tests, suggesting metacognitive self-regulation (e.g., Atalli & Bar-Hilel, 2003; Ross, Green, Salisbury-Glennon & Tollefson, 2006). Secondly, as we have noted, this retrieval-induced strategy adaptation can benefit performance, but in some cases it can hijack future performance on a test with requisites that are at odds with previous ones. This point especially applies to cases where multiple-choice tests are repeatedly used (another problem that arises from the use of this tests is the learning and maintaining of incorrect information present on lures, when students don't know the correct answer; see Marsh, Roediger, Bjork & Bjork, 2007), as this can encourage recognition-based, and not thorough and productive processes to reach the answer. However, recent educational research has suggested that with some adaptations, multiple-choice testing can be a very efficient teaching tool (e.g., Little, Bjork, Bjork & Angello, 2012; Spark, Bjork & Bjork, 2016). For example, using competitive

incorrect alternatives (e.g., Little & Bjork, 2015) and confidence-weighted multiple choice (i.e., selecting answers by deliberatively assessing their confidence in a given answer relative to the other alternatives; Bruno, 1989) should elicit more elaborative processes at responding, thus benefiting long-term retention and comprehension (Sparck, Bjork & Bjork, 2016); this application, according to our research, should encourage students to use more elaborative study strategies, instead of studying superficial aspects that could aid them in rejecting wrong answers[14]. Thirdly, the present research could inspire educators to use test structure/requirements *per se* as tool, especially when these requirements can overlap with the "real world" goals to be attained, and attention to the specificities of test structure could allow for the development of more optimal encoding and responding strategies. On a more speculative note, the use of different test formats along academic years in various formative levels should lead to both the development of different strategies that can be useful in different contexts, while also aligning with the direct benefits of testing (i.e., strengthening of the memory traces, as assessed in delayed critical tests) and the advantages of interleaving both materials as study and retrieval practices in retention and transfer of knowledge (e.g., Birnbaum, Kornell, Bjork & Bjork, 2013; Pan, Pashler, Potter & Rickard, 2015; Richman, Bjork, Finley & Linn, 2005).

[14] In a more extreme case, even this kind of test could lead to the development of more mechanized strategies of elaborative studying and fact-relating.

REFERENCES

Abbott, E. E. (1908). On the analysis of the memory function in orthography. (Bachelor's Dissertation). Retrieved from www.ideals.illinois.edu.

Abbott, E.E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs*, *11*, 159–177.

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *8*(4), 463-470.

Amlund, J. T., Kardash, C. A. M., & Kulhavy, R. W. (1986). Repetitive reading and recall of expository text. *Reading Research Quarterly*, *21*, 49-58.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451-474.

Anderson, J. R. (1985). Cognitive Psychology and its Implications (2nd Ed.). New York: Freeman.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*(2), 186-197

Anderson, M. C., & Bjork, R. A. (1994). Mechanisms of inhibition in long-term memory: A new taxonomy. In D. Dagenbach & T. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 265-325). New York: Academic Press.

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, *410*(6826), 366-369.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: memory retrieval as a model case. *Psychological Review*, *102*(1), 68-100.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063-1087.

Anisfeld, M., & Knapp, M. (1968). Association, synonymity and directionality in false recognition. *Journal of Experimental Psychology*, *77*(2), 171–179.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940-945.

Balota, D. a., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in

recall and recognition. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(5), 576–587.

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.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629-654.

Bäuml, K. H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language*, *68*(1), 39-53.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(4)*,* 941-947.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*(2), 297-305.

Benjamin, A. S., & Pashler, H. (2015). The Value of Standardized Testing: A Perspective From Cognitive Psychology. *Policy Insights From the Behavioral and Brain Sciences*, *2*(1), 13-23.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & cognition*, *35*(2), 201-210.

Birnbaum, I. M., & Eichner, J. T. (1971). Study versus test trials and long-term retention in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 516-521.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*(3), 392-402.

Bjork, E. L., & Storm, B. C. (2011). Retrieval experience as a modifier of future encoding: Another test effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1113-1124

Bjork, E. L., deWinstanley, P. A., & Storm, B. C. (2007). Learning how to learn: Can experiencing the outcome of differential encoding strategies enhance subsequent learning? *Psychonomic Bulletin & Review, 14*(2), 207-211.

Bjork, R. A. (1968). All-or-none subprocesses in the learning of complex sequences. *Journal of Mathematical Psychology*, *5*(1), 182-195.

Bjork, R. A. (1972). Theoretical implications of directed forgetting. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 217-235).

Washington, D.C.: Winston.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger and F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309-330). Hillsdale, NJ: Erlbaum.

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

Bjork, R. A. (1994b). Institutional impediments to effective training. In D. Druckman and R. A.Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance* (pp. 295-306). Washington, DC: National Academy Press.

Bjork, R. A. (2011). On the symbiosis of learning, remembering, and forgetting. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: a Festschrift in honor of Robert A. Bjork* (pp. 1-22). London, UK: Psychology Press.

Bjork, R. A. (2013). Desirable difficulties perspective on learning. In H. Pashler (Ed.), *Encyclopedia of the mind.* Thousand Oaks: Sage Reference.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes*, (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum.

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1196-1206.

Brainerd, C. J., & Reyna, V. F. (1988). Memory loci of suggestibility development: Comment on Ceci, Ross, and Toglia. *Journal of Experimental Psychology: General*, *117*(2), 197-200.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. In H. W. Reese & R. Kail (Eds.), *Advances in child development and behavior* (Vol. 6, pp. 41-100). San Diego: Academic Press.

Brainerd, C. J., Reyna, V. F., & Forrest, T. J. (2002). Are Young Children Susceptible to the False–Memory Illusion?. *Child development*, *73*(5), 1363-1377.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12-21.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York, NY: Wiley

Bruno, J. E. (1989). Using MCW–APM Test Scoring to Evaluate Economics Curricula. *The Journal of Economic Education*, *20*(1), 5-22.

Buchanan, L., Brown, N. R., Cabeza, R., & Maitson, C. (1999). False memories and semantic lexicon arrangement. *Brain and Language*, *68*(1), 172-177.

Burns, D. J. (2006). Assessing distinctiveness: Measures of item-specific and relational processing. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 109–130). Oxford, NY: Oxford University Press.

Burns, D. J., & Schoff, K. M. (1998). Slow and steady often ties the race: Effects of item-specific and relational processing on cumulative recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 1041-1051.

Burns, D., & Hebert, T. (2005). Using cumulative-recall curves to assess the extent of relational and item-specific processing. *Memory*, *13*(2), 189-199.

Caldwell, J. I., & Masson, M. E. (2001). Conscious and unconscious influences of memory for object location. *Memory & Cognition*, *29*(2), 285-295.

Carneiro, P., Albuquerque, P., Fernandez, A., & Esteves, F. (2007). Analyzing false memories in children with associative lists specific for their age. *Child Development*, *78*(4), 1171-1185.

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

Ceci, S. J., Ross, D. F., & Toglia, M. P. (1987). Suggestibility of children's memory: Psycholegal implications. *Journal of Experimental Psychology: General*, *116*(1), 38-49.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354-380.

Cho, K. W., & Neely, J. H. (2016). The roles of encoding strategies and retrieval practice in test-expectancy effects. *Memory*, 1-10. Advance online publication.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*(1), 28-71.

Connor, J. M. (1977). Effects of organization and expectancy on recall and recognition. *Memory & Cognition*, *5*(3), 315–318.

Craik F. I. (2010). Levels of processing in human memory. In M. A. Gernsbacher, R.

W.Pew, L. M. Hough, J. R. Pomerantz (Eds.). *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 76–82). New York: Worth Publishers.

Craik, F. I. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior*, *9*(2), 143-148.

Craik, F. I. (2002). Levels of processing: Past, present... and future?. *Memory*, *10*(5-6), 305-318.

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

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

D'Angiulli, A., & LeBeau, L. S. (2002). On boredom and experimentation in humans. *Ethics & Behavior*, *12*(2), 167-176.

d'Ydewalle, G. (1981). Test-expectancy effects in free recall and recognition. *Journal of General Psychology*, *105*(2), 173-195.

Davis, S. D., & Chan, J. C. (2015). Studying on borrowed time: How does testing impair new learning?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1741-1754.

Deese, J. (1958). *The psychology of learning*. New York: McGraw-Hill.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17-22.

Deng, W. S., & Sloutsky, V. M. (2015). The development of categorization: Effects of classification and inference training on category representation. *Developmental Psychology*, *51*(3), 392-405.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*(6), 945-955.

Dillon, R. F., McCormack, P. D., Petrusic, W. M., Cook, G. M., & Lafleur, L. (1973). Release from proactive interference in compound and coordinate bilinguals. *Bulletin of the Psychonomic Society*, *2*(5), 293-294.

Donaldson, W. (1971). Output effects in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 577-585.

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. Ross (Ed.). *Psychology of learning and motivation* (Vol. 54, pp. 103–140). San Diego, CA US: Elsevier Academic Press.

Ebbinghaus, H. (1885). *Über das Gedächtnis* [Memory: A contribution to

experimental psychology]. Leipzig: Duncker & Humblot.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 588-598.

Epstein, M. L., Phillips, W. D., & Johnson, S. J. (1975). Recall of related and unrelated word pairs as a function of processing level. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(2), 149-152.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107-140.

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: evidence from the test-expectancy paradigm. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(3), 632–52.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 238-244.

Finn, B., & Roediger, H. L. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1665-1681.

Fitzpatrick, S.M. & Dolezalek, M. (2013). Diversifying your funding portfolio: The role of private funders. In R.J. Sternberg (Ed.). *Writing Successful Grant Proposals from the Top Down and Bottom Up,* (pp. 255-276). Los Angeles: Sage Publications.

Forlano, G. (1936). *School learning with various methods of practice and rewards* (Teachers College Contributions to Education No. 688). New York: Teachers College, Columbia University, Bureau of Publications.

Friedman, H. (1982). Simplified determinations of statistical power, magnitude of effect and research sample sizes. *Educational and Psychological Measurement*, *42*(2), 521-526.

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

Gallo, D.A. (2006). *Associative illusions of memory*. New York: Taylor & Francis.

Garcia-Marques, L., Nunes, L. D., Marques, P., Carneiro, P., & Weinstein, Y. (2015). Adapting to test structure: Letting testing teach what to learn. *Memory*, *23*(3), 365–380.

Garner, W . R. (1978). Aspects of a stimulus: Features, dimensions, and configurations. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 99-133). Hillsdale, NJ: Erlbaum.

Gates, A.I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40).

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*(1), 8-20.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 21-31.

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

Goggin, J., & Wickens, D. D. (1971). Proactive interference and language change in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *10*(4), 453-458.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greenberg, M. S., & Bjorklund, D. F. (1981). Category centrality in free recall: Effects of feature overlap or differential category encoding? *Journal of Experimental Psychology: Human Learning & Memory, 7*(2), 145-147.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding?, *Memory & Cognition, 40*(4), 505–513.

Hall, J. W., Grossman, L. R., & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & cognition*, *4*(5), 507–13.

Hasher, L. (2007). Inhibition: Attentional Regulation in Cognition. In H. L. Roediger, Y. Dudai, & S. M. Fitzpatrick (Eds.). Science of Memory: Concepts. (pp. 291–294). New York: Oxford University Press.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, *108*(3), 356-388.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G.H. Bower (Ed.), *The psychology of learning and motivation, Vol. 22* (pp. 193-225). New York: Academic Press.

Hasher, L., Lustig, C., & Zacks, R. T. (2007). Inhibitory mechanisms and the control of attention. In A. Conway, C. Jarrold, M. Kane, A. Miyake, A., & J. Towse (Eds.), *Variation in working memory*. (pp. 227-249). New York: Oxford University Press.

Hertzog, C., Price, J., & Dunlosky, J. (2008). How is knowledge generated about memory encoding strategy effectiveness?. *Learning and Individual Differences, 18*(4), 430-445.

Higham, P. A., & Brooks, L. R. (1997). Learning the Experimenter's Design: Tacit Sensitivity to the Structure of Memory Lists. *The Quarterly Journal of Experimental Psychology: Section A*, *50*(1), 199-215.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411–428.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562-567.

Hourihan, K. L., & MacLeod, C. M. (2007). Capturing conceptual implicit memory: The time it takes to produce an association. *Memory & cognition*, *35*(6), 1187-1196.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269-299.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 497-514.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language*, *32*(4), 421-445.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, *83*(2, Pt. 1), 340-344.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, *8*(2), 200–224.

Izawa, C., Maxwell, S., Hayden, R.G., Matrana, M., & Izawa-Hayden, A.J.E.K. (2005). Optimal foreign language learning and retention: Theoretical and applied investigations on the effects of presentation repetition programs. In C. Izawa & N. Ohta (Eds.), *Human learning and memory: Advances in theory and application: The 4th Tsukuba International Conference on Memory* (pp. 107–134). Mahwah, NJ: Erlbaum.

Jacoby, L. L. (1972). Effects of organization on recognition memory. *Journal of Experimental Psychology*,*92*(3), 325–331.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513-541.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, *12*(5), 852-857.

Jacoby, L.L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649–667.

Jones, H.E. (1923–1924). The effects of examination on the performance of learning. *Archives of Psychology, 10*, 1–70.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition*, *38*(4), 389–406.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469–86.

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, *24*(3), 401-418.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966-968.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*(3), 227-239.

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

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the" Who said what?" paradigm. *Journal of Personality and Social Psychology*, *75*(5), 1155-1178.

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition*, 1(3), 347-355.

Kolers, P. A. (1979). Reading and knowing. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *33*(2), 106-117.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 187-194.

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: a comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133-1145.

Kornblum, S. (1992). Dimensional overlap and dimensional relevance in stimulus-response and stimulus-stimulus compatibility. In G.E. Stelmach & J. Requin (Eds.), *Tutorials in Motor Behavior II* (pp. 743-777). North-Holland, Amsterdam.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories is spacing the "enemy of induction"?. *Psychological Science*, *19*(6), 585-592.

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology*, *109*(3), 451-464.

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

Leonard, J. M., & Whitten, W. B. (1983). Information stored when expecting recall or recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(3), 440–455.

Lepage, M., Ghaffar, O., Nyberg, L., & Tulving, E. (2000). Prefrontal cortex and episodic memory retrieval mode. *Proceedings of the National Academy of Sciences*, *97*(1), 506-511.

Levy, B. J., McVeigh, N. D., Marful, A., & Anderson, M. C. (2007). Inhibiting your native language the role of Retrieval-induced forgetting during Second-language acquisition. *Psychological Science*, *18*(1), 29-34.

Lewandowsky, S., & Oberauer, K. (2008). The word-length effect provides no evidence for decay in short-term memory. *Psychonomic Bulletin & Review*,

15(5), 875-888.

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & cognition*, *43*(1), 14-26.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337-1344.

Lockhead, G. R. (1972). Processing dimensional stimuli: a note. *Psychological Review*, *79*(5), 410-419.

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

Loftus, E. F. (1979). The malleability of human memory: Information introduced after we view an incident can transform memory. *American Scientist, 67*(3), 312-320.

Logan, G. D. (1998). What is learned during automatization? II. Obligatory encoding of spatial location. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(6), 1720-1736.

Lundeberg, M. A., & Fox, P. W. (1991). Do laboratory findings on test expectancy generalize to classroom outcomes?. *Review of Educational Research*, *61*(1),

94-106.

Lustig, C., Hasher, L., & Zacks, R. (2007). Inhibitory deficit theory: Recent developments in a "new view". In C. M. MacLeod & D. S. Gorfein (Eds.), *Inhibition in cognition* (pp. 145–162). Washington, DC: American Psychological Association.

MacLeod, C. M. (2007). The concept of inhibition in cognition. In D. S. Gorfein & C. M. MacLeod (Eds.), *Inhibition in cognition* (pp. 3–23). Washington, DC: American Psychological Association.

MacLeod, C.M. (1998). Directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1-57). Mahwah, NJ: Lawrence Erlbaum Associates.

MacLeod, C.M. (2008). Implicit memory tests: Techniques for reducing conscious intrusion. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 245-263). New York: Psychology Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates. New York.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*(2), 194-199.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods*, *44*(2), 314-324.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724–760.

Mccloskey, M. E., & Glucksberg, S. A. M. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472.

McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, *114*(1), 1-16.

McCormack, P. D. (1982). Coding of spatial information by young and elderly adults. *Journal of Gerontology*, *37*(1), 80-86.

McDaniel M. A., Butler A. C. (2011). A contextual framework for understanding when difficulties are desirable. In Benjamin A. S. (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). New York, NY: Psychology Press.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through

retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(2), 371-385.

McDaniel, M. A., Blischak, D. M., & Challis, B. (1994). The effects of test expectancy on processing and memory of prose. *Contemporary Educational Psychology, 19*(2), 230-248.

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, *35*(2), 212-230.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*(1), 89-115.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*(4), 530-542.

Meyer, G. (1934). An experimental study of the old and new types of examination: I. The effect of the examination set on memory. *Journal of Educational Psychology*, *25*(9), 641-661.

Meyer, G. (1936). The effect of recall and recognition on the examination set in classroom situations. *Journal of Educational Psychology*, *27*(2), 81-99.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134-140.

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

Mulligan, N. W., Guyer, P. S., & Beland, A. (1999). The effects of levels-of-processing and organization on conceptual implicit memory in the category exemplar production test. *Memory & Cognition*, *27*(4), 633-647.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 27-46). New York: Oxford University Press.

Nascimento, M. F. B., Casteleiro, J. M., Marques, M. L. G., Barreto, F., & Amaro, R. (2000). Léxico multifuncional computorizado do português contemporâneo [Multifunctional computational lexicon of contemporary Portuguese](data file). *Available from Centro de Linguística da Universidade de Lisboa website: http://www.clul.ul.pt*.

Naveh-Benjamin, M. (1987). Coding of spatial location information: An automatic process?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 595-605.

Naveh-Benjamin, M. (1988). Recognition memory of spatial location information: Another failure to support automaticity. *Memory & Cognition*, *16*(5), 437-445.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226-254.

Neely, J. H., & Balota, D. A. (1981). Test-expectancy and semantic-organization effects in recall and recognition. *Memory & Cognition*, *9*(3), 283–300.

Neisser U. (1978). Memory: What are the important questions?. In M. M. Grunenberg, P. E. Morris, & R. E. Sykes (Eds.), *Practical Aspects of Memory* (pp. 3–24). Academic Press: London.

Neisser, U. (1985). Toward an ecologically oriented cognitive science. In T.M. Schlecter, & M.P. Toglia (Eds.) *New Directions in Cognitive Science* (pp. 17-32), Ablex Publishing Corp, Norwood, N.J.

Nelson, D. I., & Goodmon, L. B. (2002). Experiencing a word can prime its accessibility and its associative connections to related words. *Memory & Cognition*, *30*(3), 380-398.

Nelson, D. L., Kitto, K., Galea, D., McEvoy, C. L., & Bruza, P. D. (2013). How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & Cognition*, *41*(6), 797-819.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Available at www.usf.edu/FreeAssociation

Nunes, L. D., & Karpicke, J. D. (2015). Retrieval-based learning: Research at the interface between cognitive science and education. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (pp. 1-16). John Wiley & Sons, Inc.

Nunes, L. D., & Weinstein, Y. (2012). Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory*, *20*(2), 138-154.

Nyberg, L., Tulving, E., Habib, R., Nilsson, L. G., Kapur, S., Houle, S., ... & McIntosh, A. R. (1995). Functional brain maps of retrieval mode and recovery of episodic information. *NeuroReport*, *7*(1), 249-252.

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language*, *83*, 53-61.

Park, D. C., & Mason, D. A. (1982). Is there evidence for automatic processing of spatial and color attributes present in pictures and words?. *Memory & Cognition*, *10*(1), 76-81.

Pastötter, B., & Bäuml, K. H. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology: Cognitive Science, 5,* Article 286.

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology*, *58*(3), 193-198.

Pettijohn, K. A., & Radvansky, G. A. (2016). Walking through doorways causes forgetting: Event structure or updating disruption?. *The Quarterly Journal of Experimental Psychology*, *69*(11) 2119-2129.

Pettijohn, K. A., Thompson, A. N., Tamplin, A. K., Krawietz, S. A., & Radvansky, G. A. (2016). Event boundaries and memory improvement. *Cognition*, *148*, 136-144.

Pinto, A. C. (1992). Categorização de itens verbais: Medidas de frequência de produção e de tipicidade [Categorization of verbal items: Measures of production frequency and typicality]. Porto: Relato Técnico de Centro de Psicologia Cognitiva.

Poirier, M., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity and item position. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *50*(4), 408.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of*

Experimental Psychology, 77(3 Pt. 1), 353–363.

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal of Experimental Psychology: General*, *106*(4), 376-403.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335-335.

Pyc, M. A., & Rawson, K. A. (2012). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 737-746.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93-134.

Rajaram, S. (1998). The effects of conceptual salience and perceptual distinctiveness on conscious recollection. *Psychonomic Bulletin & Review*, *5*(1), 71-78.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1-41.

Reyna, V. F., & Brainerd, C. J. (1998). Fuzzy-trace theory and false memory: New frontiers. *Journal of Experimental Child Psychology*, *71*(2), 194-209.

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive

science to education: Generation and interleaving effects. In *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 1850-1855). Lawrence Erlbaum Mahwah, NJ.

Rock, I. (1957). The role of repetition in associative learning. *The American Journal of Psychology*, *70*(2), 186-193.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*(9), 1043-1056.

Roediger, H. L., & Arnold, K. M. (2012). The one-trial learning controversy and its aftermath: Remembering Rock (1957). *The American Journal of Psychology*, *125*(2), 127-143.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27.

Roediger, H. L., & Gallo, D. A. (2002). Levels of processing: Some unanswered questions. In M. Naveh-Benjamin, M. Moscovitch, & H. L. Roediger (Eds.), *Perspectives on human memory and cognitive aging: Essays in honour of Fergus I. M. Craik* (pp. 28-47) Philadelphia: Psychology Press

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210

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

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), 803–814.

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 Press.

Roediger, H. L., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels of processing framework. *Memory, 10*(5-6), 319-332.

Roediger, H. L., Knight, J. L., & Kantowitz, B. H. (1977). Inferring decay in short-term memory: The issue of capacity. *Memory & Cognition, 5*(2), 167-176.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre, & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford: Elsevier.

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*(3), 385-407.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, *2*(4), 491-502.

Rosner, S. R. (1970). The effects of presentation and recall trials on organization in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, *9*(1), 69-74.

Ross, M. E., Green, S. B., Salisbury-Glennon, J. D., & Tollefson, N. (2006). College students' study strategies as a function of testing: An investigation into metacognitive self-regulation. *Innovative Higher Education*, *30*(5), 361-375.

Rowe, G., Valderrama, S., Hasher, L., & Lenartowicz, A. (2006). Attentional disregulation: a benefit for implicit memory. *Psychology and Aging*, *21*(4), 826-830.

Runquist, W. (1986). Changes in the rate of forgetting produced by recall tests. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *40*(3), 282-289.

Sani, F., Bennett, M., & Soutar, A. U. (2005). The ecological validity of the "who said what?" technique: An examination of the role of self-involvement, cognitive interference and acquaintanceship. *Scandinavian Journal of Psychology*, *46*(1), 83-90.

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., ... & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition*, *129*(2), 241-255.

Schmidt, S. R. (1983). The effects of recall and recognition test expectancies on the retention of prose. *Memory & cognition*, *11*(2), 172–80.

Schulman, A. I. (1973). Recognition memory and the recall of spatial location. *Memory & Cognition*, *1*(3), 256-260.

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.

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*(2), 163–96.

Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure?. *Science*, *168*(3939), 1601-1603.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM -

Retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4, 145–166.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 267.

Shimizu, Y., & Jacoby, L. L. (2005). Similarity-guided depth of retrieval: Constraining at the front end. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *59*(1), 17-21.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, *4*(6), 592-604.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language*, *26*(6), 589-607.

Smith, S. M., Ward, T. B., Tindell, D. R., Sifonis, C. M., & Wilkenfeld, M. J. (2000). Category structure and created memories. *Memory & Cognition*, *28*(3), 386-395.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology:*

General, *117*(1), 34-50.

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99-115.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance an integrative review. *Perspectives on Psychological Science*, *10*(2), 176-199.

Soderstrom, N. C., Kerr, T. K., & Bjork, R. A. (2016). The Critical Importance of Retrieval - and Spacing - for Learning. *Psychological Science*, *27*(2), 223-230.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204-221.

Sones, A. M., & Stroud, J. B. (1940). Review, with special reference to temporal position. *Journal of Educational Psychology*, *31*(9), 665-676.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications*, *1*(3), 1-10.

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

Stokes, D. E. (1997). Pasteur's Quadrant: Basic science and technological innovation. Washington D.C.: Brookings Institution Press.

Storm, B. C., Hickman, M. L., & Bjork, E. L. (2016). Improving encoding strategies as a function of test knowledge and experience. *Memory & cognition*, *44*(4), 660-670.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, *35*(5), 1007-1013.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392-1399.

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*(7), 778-793.

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology*, *49A*(4), 901–918.

Thompson, G. L. (2006). An SPSS implementation of the nonrecursive outlier

deletion procedure with shifting z-score criterion (Van Selst & Jolicoeur, 1994). *Behavior Research Methods*, *38*(2), 344-352.

Tsal, Y., & Lavie, N. (1988). Attending to color and shape: The special role of location in selective visual processing. *Perception & Psychophysics*, *44*(1), 15-21.

Tsal, Y., & Lavie, N. (1993). Location dominance in attending to color and shape. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(1), 131-139.

Tulving, E. (1962). The effect of alphabetical subjective organization on memorizing unrelated words. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *16*(3), 185-191.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*(2), 175-184.

Tulving, E. (1983). *Elements of episodic memory.* Oxford: Clarendon.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381-391.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes

in episodic memory. *Psychological review*, *80*(5), 352-373.

Tversky, B. (1973). Encoding processes in recognition and recall. *Cognitive Psychology, 5*(3), 275–287.

Underwood, B. J. (1957). Interference and forgetting. *Psychological review*, *64*(1), 49-60.

Underwood, B. J. (1969). Attributes of memory. *Psychological Review*, *76*(6), 559-573.

Underwood, B. J., & Freund, J. S. (1968). Errors in recognition learning and retention. *Journal of Experimental Psychology*, *78*(1), 55–63.

Van Selst, M. & Jolicœur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology*, *47*(3), 631-650.

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

Whitten, W. B. (1974). *Retrieval "depth" and retrieval component processes: A levels-of-processing interpretation of learning during retrieval*. Technical Report No. 54, Human Performance Center, University of Michigan, Ann

Arbor, Michigan.

Whitten, W. B. (2011). Learning from and for tests. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: a Festschrift in honor of Robert A. Bjork* (pp. 217-234). London, UK: Psychology Press.

Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 127–134.

Wickens D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review, 77*, 1–15.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*(6), 1140-1147.

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1024-1039.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*(1), 89-106.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*(8), 995–1008.