Universidade de Brasília

Institute of Psychology

Department of Basic Psychological Processes

Graduate Program in Behavioral Sciences

Doctoral Dissertation

Direct and Indirect Effects of Fluid Intelligence on the Retrieval Practice Effect: Experiment and Simulations

Efeitos Diretos e Indiretos da Inteligência Fluida Sobre o Efeito de Prática de Lembrar: Experimento e Simulações

by

Marcos Felipe Rodrigues de Lima

Brasília, 11th December 2023

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Direct and Indirect Effects of Fluid Intelligence on the Retrieval Practice Effect: Experiment and Simulations

Efeitos Diretos e Indiretos da Inteligência Fluida Sobre o Efeito de Prática de Lembrar: Experimento e Simulações

by

Marcos Felipe Rodrigues de Lima

Doctoral dissertation submitted in partial fulfillment of the requirements for the degree of Doctor in Behavioral Sciences in the Graduate Program in Behavioral Sciences at the University of Brasília (Research area: Cognition and Behavioral Neurosciences).

Supervisor: Luciano Grüdtner Buratto, PhD

Brasília, 11th December 2023

Examining Committee:

Luciano Grüdtner Buratto, PhD (President)

Graduate Program in Behavioral Sciences

University of Brasília

Antônio Jaeger, PhD (External member)

Graduate Program in Psychology: Cognition and Behavior

Federal University of Minas Gerais

Roberta Ekuni de Souza, PhD (External member)

Department of Social and Institutional Psychology, Center of Biological Sciences

State University of Londrina

Goiara Mendonça de Castilho, PhD (Internal member)

Graduate Program in Behavioral Sciences

University of Brasília

Ricardo José de Moura, PhD (Substitute member)

Graduate Program in Behavioral Sciences

University of Brasília

Brasília, 11th December 2023

Acknowledgments

Ao professor Luciano Grüdtner Buratto, por sua orientação sempre respeitosa, encorajadora e cuidadosa ao longo de sete anos de convivência desde a graduação. Aprendi com ele por meio de instruções diretas, mas também de maneira observacional, através do seu exemplo como professor e pesquisador. Sou grato por ter sido aluno do Luciano e por tê-lo como uma grande fonte de inspiração.

Aos professores Antônio Jaeger, Goiara Mendonça de Castilho, Josemberg Moura de Andrade, Ricardo José de Moura, Roberta Ekuni de Souza e Rui de Moraes Jr., por gentilmente aceitarem participar de minha banca examinadora de qualificação e/ou de defesa da tese, contribuindo para que eu pudesse entregar um trabalho final mais satisfatório.

Aos professores, servidores, estagiários e terceirizados que atuam no Instituto de Psicologia da Universidade de Brasília, por ouvirem minhas queixas e pedidos, e por tentarem viabilizar soluções para os problemas do cotidiano universitário. A pesquisa científica só é possível graças à contribuição direta ou indireta de muitos atores. Muito obrigado a todos vocês.

Aos membros e ex-membros do Laboratório de Processos Cognitivos (LPC), por suas contribuições nos diferentes projetos em que estive envolvido ao longo dos anos. Em especial, agradeço ao Cadu Klier, por nossas conversas nas reuniões do LPC; e à Larissa Grizza, por ter aceitado encarar a missão de coletar dados comigo.

Aos participantes, por cederem generosamente um espaço em suas agendas para participarem da pesquisa.

À CAPES, por me apoiar financeiramente ao longo do meu doutorado.

À Alexandra Elbakyan e aos criadores de conteúdo *on-line*, por produzirem e compartilharem conhecimento gratuito, aberto e de qualidade para todos.

Aos meus colegas e amigos espalhados pelo Distrito Federal, Brasil e mundo afora, que fazem parte da minha memória autobiográfica, repleta de boas lembranças que compartilhamos. Também agradeço à família Sales, por seu apoio desde minha chegada a Brasília.

Por fim, à minha família, que torceu por mim durante a aventura em Brasília que decidi viver nos últimos 12 anos. Em especial, agradeço aos meus pais, Miryam e David, e à minha irmã, Mirella, por sempre acreditarem em mim e por serem meus maiores apoiadores.

Todos vocês algum dia vão se formar, eu espero. Quando vocês se formarem, por favor, não se esqueçam das pessoas, da sociedade. Procurem sempre desenvolver algum tipo de atividade social […], para que vocês possam auxiliar pessoas que não vão ter como te pagar […]. Todas as profissões têm alguma serventia junto à sociedade.

—Nerckie, durante apresentação de um curso gratuito de matemática para vestibulandos (https://www.youtube.com/watch?v=M4bDd2otryI)

Table of Contents

List of Tables

List of Figures

List of Abbreviations, Acronyms, and Symbols

AC	Attentional control
ANOVA	Analysis of variance
BF_{10}	Bayes Factor
$BF_{Inclusion}$	Model-average Bayes Factors
CI	Confidence interval
Cog	Cognitive ability
d	Cohen's *d*
D	Kolmogorov–Smirnov statistic
EM	Episodic memory
Exp.	Experiment
F	*F*-statistic
gC	Crystallized intelligence
gF	Fluid intelligence
hr	Hour
ID	Individual-difference
KR-20	Kuder and Richardson's internal consistency index
M	Mean
ms	Millisecond
N	Sample size
n	Subsample size
NFC	Need for cognition
p	*P*-value

PC_R	Observed proportion correct in the rereading condition (actual participant)
$\widehat{PC}_T$	Predicted proportion correct in the retrieval practice condition (actual participant)
P_R	Probability of successful recall in the final test for a randomly chosen item in the rereading condition (ideal participant)
P_T	Probability of successful recall in the final test for a randomly chosen item in the retrieval practice condition (ideal participant)
P_{T-s}	Probability of successful recall in the final test for a randomly chosen item in the retrieval practice condition (supported by study memory; ideal participant)
P_{T-t}	Probability of successful recall in the final test for a randomly chosen item in the retrieval practice condition (supported by test memory; ideal participant)
r	Pearson correlation coefficient
RE	Rereading
RP	Retrieval practice
s	Second
S_R	Study memory strength of an item assigned to the rereading condition
S_{T-s}	Study memory strength of an item assigned to the retrieval practice condition
S_{T-t}	Test memory strength of an item assigned to the retrieval practice condition
SD	Standard deviation
t	*t*-test statistic or response threshold
$\widehat{TE}$	Predicted retrieval practice effect in the dual-memory framework
WMC	Working memory capacity

z_{crit}	Critical point that cuts off an area under the standard normal distribution
η_p^2	Partial eta-squared
χ_{SB}^2	Satorra–Bentler scaled (mean-adjusted) chi-square

General Abstract

Retrieving information from memory (i.e., retrieval practice) is a learning technique that, on average, enhances long-term retention—a phenomenon known as the *retrieval practice effect*. However, retrieval practice does not benefit all learners equally. The overall goal of this dissertation is to investigate direct and indirect effects of fluid intelligence (gF) on the magnitude of the retrieval practice effect (Chapter 1). The dissertation is divided into three main manuscripts. The first manuscript (Chapter 2) comprises a narrative literature review on individual differences in the retrieval practice effect. While studies have examined personality traits and cognitive abilities, consistent links between individual differences and the retrieval practice effect remain elusive. The analysis is complicated by the heterogeneous procedures employed in these studies. Some findings indicate that the impact of an individual-difference variable on the magnitude of the retrieval practice effect might depend on other individual differences or contextual factors. The second manuscript (Chapter 3) presents the results of an experiment ($N = 144$) designed to extend the findings of Minear et al. (2018) and to present a novel set of analyses. We found that the retrieval practice effect and performance during the practice phase were contingent on gF and item difficulty. Moreover, we observed positive correlations between gF and the amount of new items participant recalled during the practice phase in Cycles 1–3. Additionally, we found an indirect effect of gF on the retrieval practice effect mediated by performance during the practice phase. The third manuscript (Chapter 4) explores the dual-memory framework (Rickard & Pan, 2018), which was used to derive two simple models: the fixed-threshold model and the random-threshold model. These models were tested against the data collected in the experiment described in the second manuscript. The random-threshold model yielded a point estimate closer to the empirical value we obtained than the fixed-threshold model, although the empirical confidence interval

overlapped with estimates from both models. Drawing from these three manuscripts, we propose a research agenda (Chapter 5), which includes, but is not limited to: exploring whether the impact of individual differences on the retrieval practice effect depends on other individual differences (e.g., participants' spontaneous encoding strategy use) and contextual factors (e.g., extended practice); testing whether individual differences impact the retrieval practice effect indirectly through the learners' ability to generate and retrieve mediators, and to monitor and shift mediators after unsuccessful retrieval attempts; and testing different models derived from the dual-memory framework with various data points.

Keywords: retrieval practice, testing effect, fluid intelligence, individual differences, quantitative model

Resumo Geral

Recuperar informações da memória (i.e., prática de lembrar) é uma estratégia de aprendizagem que, em média, melhora a retenção a longo prazo—um fenômeno conhecido como *efeito de prática de lembrar*. Contudo, a prática de lembrar não beneficia todos os aprendizes igualmente. O objetivo geral desta tese é investigar os efeitos diretos e indiretos da inteligência fluida (gF) sobre a magnitude do efeito de prática de lembrar (Capítulo 1). A tese está dividida em três manuscritos principais. O primeiro manuscrito (Capítulo 2) consiste em uma revisão de literatura narrativa sobre as diferenças individuais no efeito de prática de lembrar. Embora estudos tenham examinado traços de personalidade e habilidades cognitivas, vínculos consistentes entre diferenças individuais e o efeito da prática de lembrar permanecem incertos. A análise é complicada pelos procedimentos heterogêneos utilizados nesses estudos. Alguns resultados indicam que o impacto de uma variável de diferenças individuais na magnitude do efeito da prática de lembrar pode depender de outras diferenças individuais ou fatores contextuais. O segundo manuscrito (Capítulo 3) apresenta os resultados de um experimento ($N = 144$) delineado para expandir os achados de Minear et al. (2018) e para apresentar um novo conjunto de análises. Observamos que o efeito da prática de lembrar e o desempenho durante a fase de prática dependeram da gF e da dificuldade do item. Além disso, observamos correlações positivas entre gF e a quantidade de novos itens que os participantes se recordaram nos Ciclos 1–3 da fase de prática. Adicionalmente, encontramos um efeito indireto da gF sobre o efeito da prática de lembrar, mediado pelo desempenho durante a fase de prática. O terceiro manuscrito (Capítulo 4) explora o arcabouço teórico de memória dual (Rickard & Pan, 2018), que foi utilizado para derivar dois modelos simples: o modelo de limiar fixo e o modelo de limiar aleatório. Esses modelos foram testados com os dados coletados no experimento descrito no segundo manuscrito. O modelo de limiar aleatório produziu uma

estimativa pontual mais próxima do valor empírico que obtivemos do que o modelo de limiar fixo, embora o intervalo de confiança empírico tenha se sobreposto às estimativas de ambos os modelos. Com base nesses três manuscritos, propomos uma agenda de pesquisa (Capítulo 5), que inclui, mas não se limita a: explorar se o impacto das diferenças individuais no efeito da prática de lembrar depende de outras diferenças individuais (p.ex., o uso espontâneo de estratégias de codificação) e fatores contextuais (p.ex., prática estendida); testar se as diferenças individuais afetam o efeito da prática de lembrar de forma indireta por meio da habilidade dos aprendizes de gerar e recuperar mediadores, e de monitorar e modificar mediadores após tentativas de lembrar malsucedidas; e testar diferentes modelos derivados do arcabouço teórico de memória dual com diversos pontos de dados.

Palavras-chave: prática de lembrar, efeito de testagem, inteligência fluida, diferenças individuais, modelo quantitativo

Resumo Expandido

Efeitos Diretos e Indiretos da Inteligência Fluida Sobre o Efeito de Prática de Lembrar: Experimento e Simulações

Marcos Felipe Rodrigues de Lima

Orientador: Luciano Grüdtner Buratto, PhD

Recuperar informações da memória (i.e., prática de lembrar) melhora, em média, a retenção a longo prazo. Contudo, a prática de lembrar não beneficia todos os aprendizes igualmente. Por isso, estudos de diferenças individuais têm buscado identificar se características dos aprendizes moderam o efeito de prática de lembrar. Nesta tese, focamo-nos na inteligência fluida (gF)—a habilidade de resolver novos problemas, engajar-se em raciocínio indutivo, sequencial e quantitativo, e que é tipicamente avaliada por meio de tarefas não-verbais e supostamente não influenciadas pela cultura.

Embora alguns estudos tenham previamente explorado a relação entre gF e a magnitude do efeito de prática de lembrar, duas lacunas podem ser destacadas. A primeira delas é a falta de estudos investigando se a gF exerce um efeito indireto sobre a prática de lembrar, mediado pelo desempenho durante a fase de prática. A segunda é a ausência de uma clara orientação teórica nos estudos, possivelmente em decorrência de hipóteses contemporâneas do campo predizerem primariamente padrões grupais. Com base nessas lacunas, o objetivo geral da tese é investigar os efeitos diretos e indiretos da gF sobre a magnitude do efeito de prática de lembrar. Cinco objetivos específicos foram almejados ao longo de três manuscritos.

O primeiro manuscrito (Capítulo 2) busca revisar a literatura sobre as diferenças individuais no efeito de prática de lembrar. Variáveis de diferenças individuais moderam o efeito de prática de lembrar? Embora estudos tenham examinado traços de personalidade e habilidades

cognitivas, vínculos consistentes entre diferenças individuais e o efeito da prática de lembrar permanecem incertos. A análise é complicada pelos procedimentos heterogêneos utilizados nesses estudos. Alguns resultados indicam que o impacto de uma variável de diferenças individuais na magnitude do efeito da prática de lembrar pode depender de outras diferenças individuais ou de fatores contextuais.

O segundo manuscrito (Capítulo 3) consiste na descrição de um experimento. Participantes ($N = 144$) primeiramente estudaram 40 pares de palavras suaíli–português, releram metade dos pares e se engajaram em prática de lembrar da outra metade. Em sessões separadas, os participantes completaram um teste de recordação com pistas e um teste de gF. Esse delineamento permitiu atingir outros três objetivos específicos da tese. Primeiro, buscamos generalizar duas interações triplas encontradas por Minear et al. (2018) em análises restritas a participantes que se beneficiaram da prática de lembrar. Em geral, estendemos com sucesso resultados prévios. Segundo, investigamos se a gF está positivamente relacionada à quantidade de novos itens que os participantes se recordam durante a fase de prática. Correlações positivas consistentes foram observadas nos Ciclos 1–3. Terceiro, testamos e observamos um efeito indireto da gF sobre o efeito de prática de lembrar, mediado pelo desempenho durante a fase de prática.

O terceiro manuscrito (Capítulo 4) apresenta o arcabouço teórico de memória dual como um candidato viável para instanciar modelos quantitativos específicos capazes de predizer diferentes relações entre variáveis de diferenças individuais e a magnitude do efeito da prática de lembrar. Nós introduzimos uma abordagem para simular dados, que leva em consideração a proporção média recordada na condição releitura e a correlação de uma variável de diferenças individuais e a proporção recordada na condição releitura. Por meio dessa abordagem, derivamos dois modelos simples, o modelo de limiar fixo e o modelo de limiar aleatório. Usando os dados

coletados no experimento descrito no segundo manuscrito, o modelo de limiar aleatório produziu uma estimativa pontual mais próxima do valor empírico que obtivemos do que o modelo de limiar fixo, embora o intervalo de confiança empírico tenha se sobreposto às estimativas de ambos os modelos.

Com base na revisão de literatura, no estudo experimental e na tentativa inicial de derivar modelos a partir do arcabouço dual de memória, nós propomos a seguinte agenda de pesquisa (Capítulo 5):

1. Explorar se o impacto das diferenças individuais no efeito da prática de lembrar depende de outras diferenças individuais (p.ex., o uso espontâneo de estratégias de codificação e de lembrar) e fatores contextuais (p.ex., espaçamento entre itens, prática estendida).
2. Investigar a fidedignidade do efeito de prática de lembrar (p.ex., teste–reteste; entre tarefas, materiais e intervalos de retenção).
3. Adotar procedimentos alternativos para reduzir diferenças entre aprendizes durante a prática (p.ex., usando procedimentos baseado em critério).
4. Replicar nossos resultados usando análise de mediação.
5. Testar se diferenças individuais impactam o efeito de prática de lembrar indiretamente por meio da habilidade dos aprendizes de gerar e recuperar mediadores.
6. Testar se diferenças individuais impactam o efeito de prática de lembrar indiretamente por meio da habilidade dos aprendizes de monitorar e mudar mediadores após tentativas de recuperação malsucedidas.

7. Testar diferentes modelos derivados do arcabouço teórico dual de memória com vários pontos de dados.
8. Investigar por que, e sob quais condições, a descrição conceitual e a implementação matemática do arcabouço teórico dual de memória levam a predições distintas.
9. Propor experimentos críticos contrastando predições distintas de diferentes arcabouços teóricos (memória dual vs. bifurcação; Halamish & Bjork, 2011), bem como explorar modelos híbridos.
10. Explorar se predições baseadas nos modelos de limiar fixo e aleatório são sensíveis à escolha da distribuição de probabilidade usada para modelar as forças de memória.

Palavras-chave: prática de lembrar, efeito de testagem, inteligência fluida, diferenças individuais, modelo quantitativo

Chapter 1 – General Introduction

General Introduction

Several scholars have argued that studying retrieval processes is essential for understanding human memory (e.g., Gazzaniga, 1991; Moscovitch, 2007; Rajaram & Barber, 2008; Roediger, 2000). Prominent examples highlighting the importance of retrieval include the role of retrieval cues in memory tests (Tulving & Pearlstone, 1966; Tulving & Thomson, 1973), the involvement of inhibitory processes during selective retrieval (Anderson, 2003; Anderson et al., 1994), and the "sin" of blocking in the tip-of-the-tongue state (R. Brown & McNeill, 1966; Schacter, 2021). Another crucial example involves the fact that retrieval practice—the focus of the present dissertation—alters memory representations (Bjork, 1975), making retrieved items subsequently more retrievable (Halamish & Bjork, 2011; Whiffen & Karpicke, 2017).

In recent decades, it has been argued that practicing retrieval is a key activity for long-term retention (Karpicke & Roediger, 2008; Roediger & Butler, 2011). This claim is supported by hundreds of experiments showing that retrieval practice, compared with different control conditions, enhances long-term retention (Carpenter et al., 2008; Cavendish et al., 2022; Coane, 2013; Karpicke & Blunt, 2011; Klier & Buratto, 2023; Rawson & Zamary, 2019; Roediger & Karpicke, 2006b). This *retrieval practice effect*—also known as the *testing effect*—has been demonstrated since the early decades of the 20th century (Abbot, 1909; Gates, 1917; Spitzer, 1939), with sporadic investigations throughout the subsequent decades (Carrier & Pashler, 1992; Glover, 1989; Hogan & Kintsch, 1971; McDaniel & Masson, 1985; Melton, 1967; Tulving, 1967; Wheeler & Roediger, 1992). However, the resurgence of interest in this phenomenon occurred only in the early years of the 21st century (Roediger & Karpicke, 2006a, 2006b).

In this chapter, we provide a brief introduction to the retrieval practice literature. How has retrieval practice been investigated? What are the benefits of retrieval practice for memory? We

begin by describing a procedure typically employed in contemporary research, as well as some benefits of retrieval practice. Next, we introduce two caveats on terminology we used throughout the dissertation. We then highlight important gaps in the literature on individual differences in the retrieval practice effect. Finally, we outline the objectives and the general structure of the dissertation.

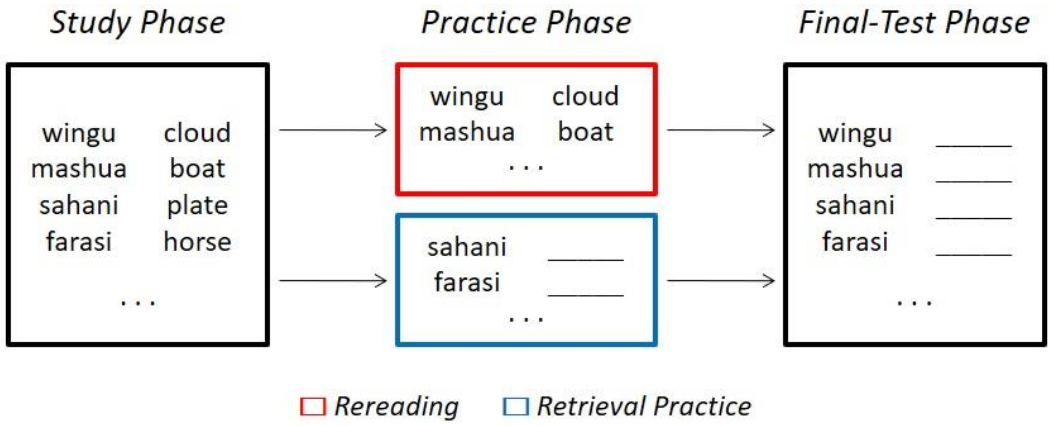
Direct and Indirect Benefits of Retrieval Practice

Contemporary research on the retrieval practice employs a variety of experimental paradigms, but a typical three-phase procedure is often used (see Figure 1.1). In the *study phase*, participants are initially exposed to the to-be-learned material. Then, in the *practice phase*, they reread a portion of that material (rereading condition) and attempt to retrieve the remaining portion of the material (retrieval practice condition). Figure 1.1 illustrates a situation in which retrieval practice is induced through a cued-recall test, wherein participants are presented with a Swahili word and asked to recall its English translation. However, this induction can be accomplished with other tasks, such as free-recall tests, fill-in-the-blank questions, and multiple-choice questions (Cavendish et al., 2022; Little & Bjork, 2015; Moreira, Pinto, Justi, & Jaeger, 2019). Finally, in the final-test phase, participants are tested on all the previously studied material. The goal is to assess whether memory performance in this test differs based on the learning strategy adopted in the practice phase, namely, rereading and retrieval practice.

Retrieval during the practice phase enhances subsequent retention of the items when compared with various control conditions (Carpenter & Yeung, 2017; Roediger & Karpicke, 2006b). Some researchers refer to this as a direct benefit of retrieval practice (Karpicke, 2017; Roediger & Karpicke, 2006a), namely, the strengthening of the memory traces (Halamish & Bjork, 2011) resulting from the successful act of retrieving items from memory.

Figure 1.1

Experimental Procedure Commonly Used in Retrieval Practice Studies



Retrieval practice also offers indirect benefits, which encompass the advantage provided to learners through events occurring after the retrieval attempt. For instance, experiments suggest that retrieval practice enhances the subsequent encoding of material (test-potentiated learning; Arnold & McDermott, 2013a, 2013b) and facilitates the learning of new information (forward testing effect; Pastötter & Bäuml, 2019; Szpunar et al., 2008). Moreover, there is evidence that retrieval practice improves later retrieval organization (Arnold & McDermott, 2013a; Cavendish et al., 2022; Congleton & Rajaram, 2012; Rawson & Zamary, 2019; Zaromb & Roediger, 2010), narrows the retrieval search set (Hopper & Huber, 2018; Lehman et al., 2014; Racsmány et al., 2018), and protects memory from the detrimental effects of acute stress (Smith et al., 2016; for other benefits, see Roediger et al., 2011).

It is important to note that many studies provide corrective feedback after the retrieval attempt during the practice phase. Although not always necessary (e.g., Roediger & Karpicke, 2006b; Storm et al., 2014), retrieval practice followed by corrective feedback may be important for subsequent memory performance (e.g., Finley et al., 2011; Silva et al., 2023; Tse et al., 2010),

especially in situations where the initial recall during retrieval practice is low (Rowland, 2014; but see Alamri & Higham, 2022, for detrimental effects of feedback). The value of feedback possibly lies in its diagnostic value, enabling learners to be exposed again to the correct information, monitor their own knowledge, and allocate greater attentional resources to that content during subsequent study opportunities. Studies incorporating feedback after retrieval practice allow for the assessment of the combined direct and indirect benefits of retrieval practice (Karpicke, 2017).

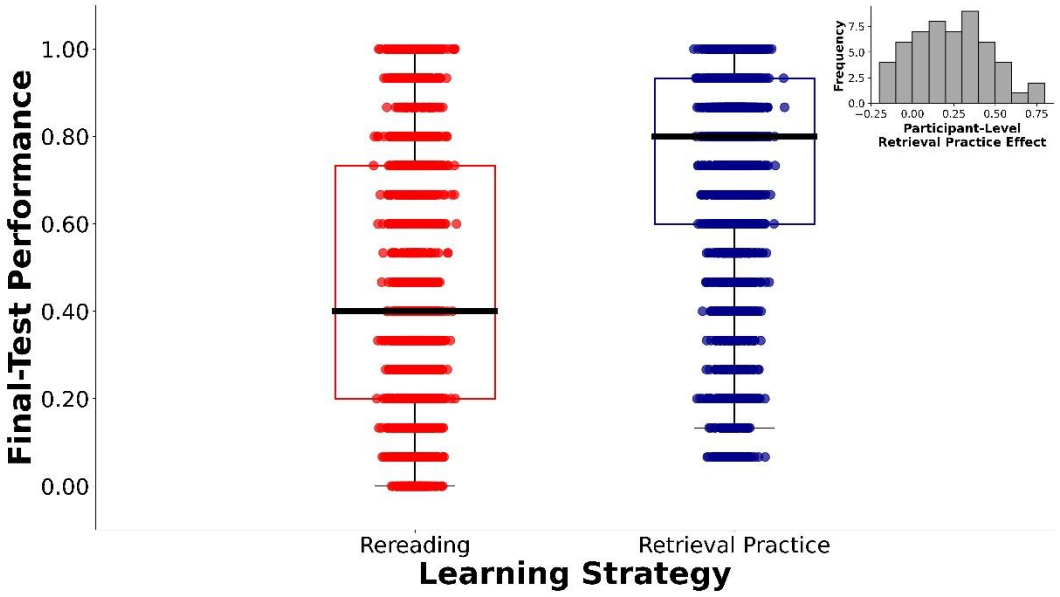
Caveats on Terminology Used in the Dissertation

There is no single definition of the retrieval practice effect (McDermott, 2021). Firstly, as mentioned in the previous section, there are direct and indirect benefits of retrieval practice. Secondly, retrieval practice has been compared to various control conditions, including rereading (Bertilsson et al., 2021; Roediger & Karpicke, 2006b), pleasantness-rating tasks (Cavendish et al., 2022; Congleton & Rajaram, 2012), semantic elaboration tasks (Coane, 2013; Karpicke & Smith, 2012), and even a no-exposure condition (Glover, 1989; Shaffer & McDermott, 2020). Thirdly, the study and practice phases illustrated in Figure 1.1 depict a duration-based procedure, where the researcher determines in advance the number of practice trials per item (Pyc & Rawson, 2009). However, several studies employ criterion-based procedures, where additional practice manipulation occurs after each item has been successfully recalled a predetermined number of times (Friedman et al., 2017; Karpicke & Roediger, 2008). Given these considerations, the existence, and the magnitude, of the retrieval practice effect in an experiment will depend on factors such as the procedure, the number of practice opportunities, and the control condition against which retrieval practice is compared (Kornell et al., 2012; Moreira, Pinto, Starling, & Jaeger, 2019). This aspect will be briefly revisited in the General Discussion (Chapter 5).

In this dissertation, the term *retrieval practice effect* will be used to convey two distinct meanings, both of which are illustrated in Figure 1.2. The first meaning—the most traditional one—refers to the group-level retrieval practice effect: On average, recall in the final-test phase is higher in the retrieval practice condition than in the rereading (or in other control) condition. In Figure 1.2, this phenomenon is evidenced by a greater concentration of superior performances in the retrieval practice condition (blue dots) as opposed to rereading condition (red dots). Importantly, our interest is on both direct, indirect, and combined benefits of retrieval practice.

Figure 1.2

Group- and Participant-Level Retrieval Practice Effects



Note. Inset graph shows the distribution of participant-level retrieval practice effects scores. Data based on Lima and Buratto (2023b), Session 2.

In individual-difference research, the group-level retrieval practice effect can be broken down into multiple scores, as long as studies manipulate learning strategy within subjects. We will

refer to these scores as participant-level retrieval practice effects: They represent the proportion recalled in the retrieval practice condition minus the proportion recalled in the rereading condition, essentially a difference score (Agarwal et al., 2017; Pan et al., 2015). These scores quantify how much each participant benefits—or does not benefit at all—from practicing retrieval, compared to the rereading condition. Considering this definition, the negative scores in the inset graph of Figure 1.2 indicate that some participants did not benefit from practicing retrieval (Brewer & Unsworth, 2012; Minear et al., 2018; Pan et al., 2015).

Gaps in the Individual-Difference Literature on the Retrieval Practice Effects

The replicability of the group-level retrieval practice effect has been well-documented in the extant literature, as evident in meta-analytic reviews (Pan & Rickard, 2018; Rowland, 2014; Yang et al., 2021). Crucially, this effect is not confined to laboratory settings but extends to practical applications as well. For instance, classroom studies employing educationally relevant materials have replicated the retrieval practice effect (e.g., Carpenter et al., 2009; Jaeger et al., 2015; for reviews, see Moreira, Pinto, Starling, & Jaeger, 2019; Schwieren et al., 2017; Yang et al., 2021). Moreover, retrieval practice has demonstrated its benefits in improving memory retention among patients with conditions like multiple sclerosis (e.g., Sumowski et al., 2013) and traumatic brain injury (e.g., Sumowski et al., 2010), and has enhanced the oral naming of familiar words in patients with aphasia (e.g., Middleton et al., 2015). These findings underline the potential utility of retrieval practice in educational and cognitive rehabilitation contexts (Dunlosky et al., 2013; Lima, Cavendish, et al., 2020; Moreira, Pinto, Starling, & Jaeger, 2019).

Considering the widespread endorsement of retrieval practice in applied contexts, it becomes imperative to assess its utility across different learners. However, as depicted in the inset graph of Figure 1.2, not all learners benefit from retrieval practice. Do individual-difference

variables moderate the retrieval practice effect? This question holds both empirical and theoretical significance. Empirically, identifying moderator variables informs researchers and educators about the boundary conditions of the retrieval practice effect (Roediger et al., 2010). Theoretically, the presence of such moderator variables can refine contemporary hypotheses in the field.

In this dissertation, we focus on fluid intelligence (gF)—the ability to solve novel problems, engage in inductive, sequential, and quantitative reasoning, ability typically assessed through nonverbal and supposedly culture-free tasks (Engle et al., 1999; Walrath et al., 2020). While some studies have explored the relationship between gF and the magnitude of the retrieval practice effect (Brewer & Unsworth, 2012; Minear et al., 2018; Robey, 2019; Wenzel & Reinhard, 2019), none have investigated whether gF exerts an indirect effect on the retrieval practice effect mediated by performance during the practice phase.

Another gap in literature is that studies on individual differences predominantly lack a clear theoretical orientation. One reason for this gap might be that contemporary hypotheses in the field primarily predict group-level patterns (Carpenter, 2009; Lehman et al., 2014; Pyc & Rawson, 2009, 2010, 2012). Recently, the dual-memory framework (Rickard & Pan, 2018) has been introduced to account for the magnitude of the retrieval practice effect in cued-recall tests. While its primary purpose was not to account for individual differences in the retrieval practice effect, we contend that its quantitative implementation could predict positive, negative, or null relationships between the retrieval practice effect and individual-difference variables under certain scenarios (cf. Rickard, 2020). This dissertation outlines a modeling approach that can be used to derive models and predictions from the dual-memory framework.

Objectives and General Structure of the Dissertation

The overall goal of this dissertation is to investigate direct and indirect effects of gF on the magnitude of the retrieval practice effect. We pursued five specific objectives: (a) reviewing the literature on individual differences in the retrieval practice effect; (b) extending some findings of Minear et al. (2018), which investigated the relationship between gF and the retrieval practice effect, while also taking into account item difficulty; (c) assessing whether gF correlates with the amount of new items participants recall in each cycle during the practice phase; (d) examining whether there is an indirect effect of gF on the retrieval practice effect mediated by performance during the practice phase; and (e) outlining the dual-memory framework as a feasible candidate to instantiate specific quantitative models able to predict different relationships between individual-difference variables and the magnitude of the retrieval practice effect.

The dissertation is organized into three main manuscripts. The first manuscript (Chapter 2) comprises a narrative literature review on individual differences in the retrieval practice effect. This chapter presents the current state of the science on the topic, highlights inconsistencies in the literature, and outlines future research directions (objective *a*). The second manuscript (Chapter 3) presents the results of an experiment designed to extend the findings of Minear et al. (2018) and to present a novel set of analyses (objectives *b*, *c* and *d*). The third manuscript (Chapter 4) outlines the dual-memory framework and how it can be used to derive models and generate predictions for studies on individual differences (objective *e*). The General Discussion section (Chapter 5) provides an assessment of the main contributions of this dissertation, presents limitations, and proposes a research agenda for further studies exploring individual differences in the retrieval practice effect.

Chapter 2 – Retrieval Practice Effect and Individual Differences: Current Status and Future Directions

Manuscript 1

Abstract

Retrieval practice improves retention more effectively than other learning strategies on average. However, retrieval practice may not benefit all learners equally. If this holds true, recommendations regarding the use of retrieval practice in educational settings need to be nuanced. Do individual-difference variables moderate the retrieval practice effect? This article comprises a narrative review on the relevant literature exploring this question. While studies have examined personality traits and cognitive abilities, consistent links between individual differences and the retrieval practice effect remain elusive. Heterogeneous procedures in these studies complicate the analysis. Some findings indicate that the impact of an individual-difference variable on the magnitude of the retrieval practice effect might depend on other individual differences or contextual factors. Additionally, studies on individual difference predominantly lack a clear theoretical orientation, as contemporary accounts typically predict only group-level patterns. We argue that the dual-memory framework can be used for generating quantitative models and predictions for individual-difference studies. Our suggestions for future research includes testing quantitative predictions derived from the dual-memory framework, investigating whether participants' spontaneous encoding and retrieval strategies mediate the relationship between individual differences and the magnitude of the retrieval practice effect, and exploring third variables—such as lag and the number of retrieval practice opportunities—that could impact the relationship between individual differences and the magnitude of retrieval practice effect.

Keywords: retrieval practice, testing effect, test-enhanced learning, memory, individual differences

Retrieval Practice Effect and Individual Differences: Current Status and Future Directions

A ubiquitous claim made by cognitive scientists is that retrieval practice improves learning and long-term retention (Carpenter, 2009; Pastötter et al., 2011; Roediger & Karpicke, 2006b). Some methods of practicing retrieval include self-testing, explaining recently learned concepts to a colleague, and summarizing previously read texts in a closed-book format. Indeed, the retrieval practice effect has been demonstrated in both laboratory and classroom experiments (e.g., Carpenter et al., 2009; Jaeger et al., 2015; Lima, Venâncio, et al., 2020). Based on the findings from these experiments, researchers have advocated for the incorporation of this learning technique into educational contexts (Dunlosky et al., 2013; Trumbo et al., 2021; Yang et al., 2021).

While retrieval practice improves retention more than other learning strategies on average, it may not benefit all learners equally. If this holds true, recommendations regarding the use of retrieval practice in educational settings need to be nuanced. Do individual-difference variables moderate the retrieval practice effect? In four main sections, this narrative review focuses on the relevant literature exploring this question. Initially, we characterize the phenomenon and delineate a justifiable recommendation based on the extant literature. Subsequently, we summarize the findings of individual-difference research on the retrieval practice effect, pointing out important methodological issues that need to be considered by scholars. We then introduce the dual-memory framework (Rickard & Pan, 2018) as a viable theoretical framework for deriving models and generating quantitative predictions regarding the relationship between the retrieval practice effect and individual-difference variables. Finally, in our final considerations, we discuss two important concepts for research on individual differences, namely, reliability and portability. Throughout this chapter, we present a research agenda, which is later revisited in Chapter 5.

The Retrieval Practice Effect: Experimental Evidence

During the Sensation and Perception lecture, the instructor introduced concepts related to the biological basis of vision (Goldstein & Cacciamani, 2022). In the final part of the lecture, the students were given sheets containing a series of statements about the topic, some of which were complete (e.g., "The fovea, a region with the highest density of cone photoreceptors in the retina, is related to focal vision.") and others with blank spaces (e.g., "The _____ can change its shape to adjust the focus of objects located at different distances." [lens]). The students were instructed to restudy the complete statements either by rereading it or by attempting to fill in the blanks for the incomplete ones, practicing retrieval of factual knowledge. After 15 min, the instructor collected the sheets. Two days later, at the beginning of the next lecture, the students received an unexpected closed-book test, where they were asked to write down everything they remembered from the previous lecture.

Experiments investigating the retrieval practice effect mirror certain aspects of this description. In these experiments, participants encode a to-be-learned material (e.g., content about the biological basis of vision), practice it through either rereading or retrieval practice, and, after a given retention interval (e.g., 2 days), take a memory test. In the given example, the retrieval practice effect would be evident if, on average, students recalled more information on the memory test for material they practiced through retrieval compared to the material they reread.

Using variations of this general procedure, experiments have demonstrated that retrieval practice, compared with control conditions, enhances long-term retention (Klier & Buratto, 2023; Pyc & Rawson, 2010). This enhancement occurs even when retrieval practice is compared with conditions involving semantic elaboration of the material (e.g., Coane, 2013; Karpicke & Blunt, 2011; but see Yang et al., 2021). Researchers have found the retrieval practice effect in studies

with longer retention intervals (e.g., 16 weeks; Carpenter et al., 2009) and with educationally relevant materials (e.g., Ekuni & Pompeia, 2020; McDaniel et al., 2011), indicating the ecological validity of the effect.

Is this evidence sufficient to support the recommendations for using retrieval practice in educational contexts? It is important to note that most of the available evidence is experimental. Experimental psychologists design tightly controlled environments—through random assignment, counterbalancing, and holding other variables constant (Roediger & Yamashiro, 2020)—to make cause-and-effect claims possible (Cronbach, 1957). They assume that the mechanisms underlying experimental effects are homogeneous across participants (see, e.g., Healey & Kahana, 2014) and that the individuals in their sample are interchangeable instances with any other instances from the same population (Borsboom et al., 2009). In other words, under this reasoning, it would make no difference whether, in a given experiment, Mary took or did not take part in the study along with other n participants. This is because process inference—rather than population inference—often is the goal in experimental psychology (cf. Hayes, 2022, pp. 65–67).

The situation changes slightly when the goal is the application of a psychological principle. For instance, from Mary's perspective, what really matters is whether retrieval practice is useful for her, not for instances supposedly interchangeable with her. Therefore, returning to the question from the previous paragraph, there is sufficient evidence to legitimately recommend retrieval practice in educational contexts if the objective is to improve students' learning and retention, compared with no/filler activity, testing with fewer questions, and rereading (Yang et al., 2021). Teachers who suggest that their students engage in retrieval practice can anticipate an overall improvement in retention across students. However, concluding that retrieval practice will improve retention for *each* individual would be an ecological fallacy (McDermott, 2021). The assumption

of homogeneity of the retrieval practice effect, often only implicit in experimental research, is likely unrealistic. For example, while experiments demonstrate the retrieval practice effect at the group level, there are also studies reporting that a sizeable number of participants did not benefit from retrieval practice (e.g., Minear et al., 2018; Pan et al., 2015; Robey, 2019).

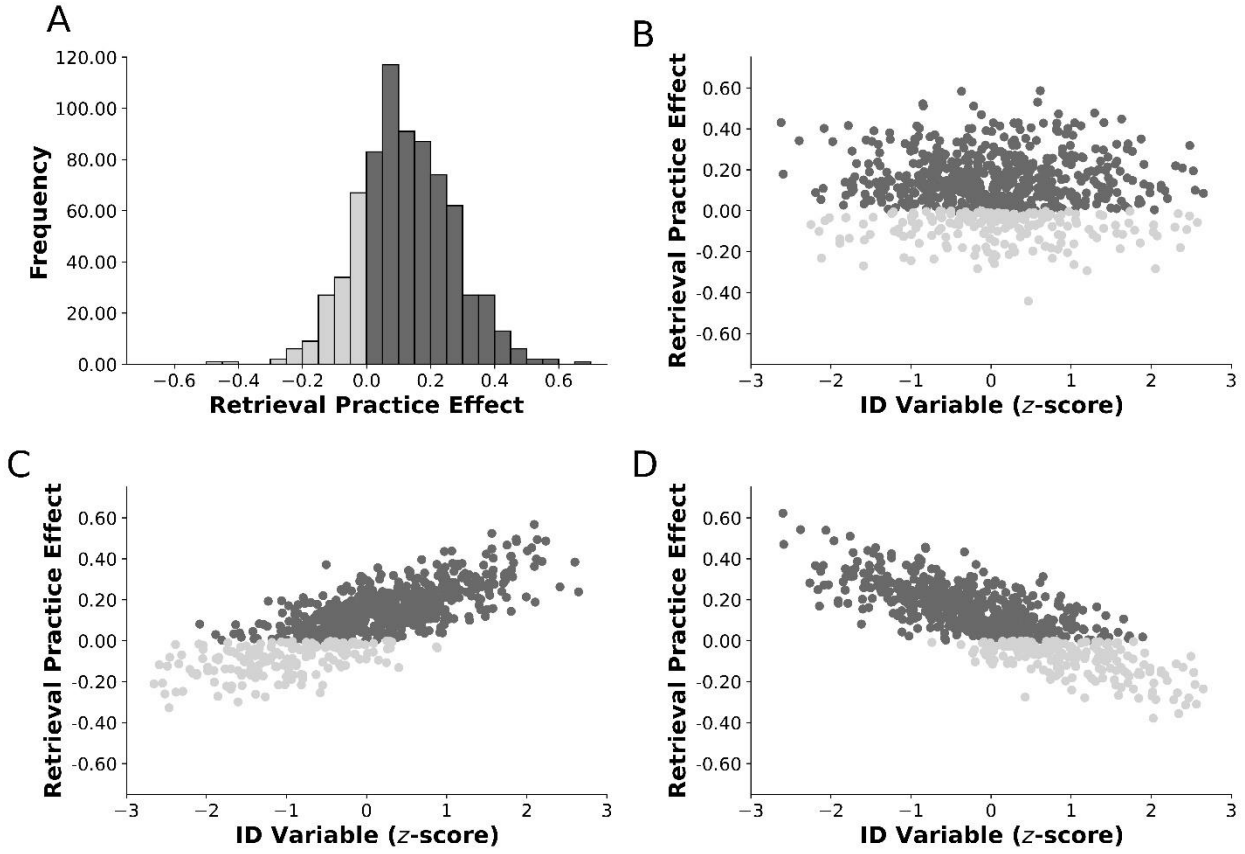
The panel A of Figure 2.1 displays the frequency distribution of the retrieval practice effect at the participant level—recall performance on retrieval practice condition minus recall performance on rereading condition—for five combined experiments (Brewer & Unsworth, 2012; Pan et al., 2015, Experiments 1 and 2; Robey, 2019, Experiments 1 and 2; $N_{combined} = 739$). The overall retrieval practice effect for these experiments was greater than zero, $M_{Difference} = .11, 95\%\ \text{CI}\ [.09, .12]$. This is represented by the majority of participants benefiting from retrieval practice (indicated by the dark gray bars). Importantly for our discussion, 28.55% of participants did not benefit from retrieval practice (indicated by the light gray bars). A natural question arises: Do individual-difference variables moderate the retrieval practice effect?

A decade ago, researchers had already questioned whether retrieval practice would benefit different learners equally (Brewer & Unsworth, 2012). This question is both empirically and theoretically relevant. Empirically, experimental psychologists aim not only to establish that a phenomenon is indeed replicable but also identify its boundary conditions (Roediger & Yamashiro, 2020). For instance, if researchers identify one characteristic of learners that predict whether they will fall more to the left or to the right of the frequency distribution in Figure 2.1, panel A, this will certainly have practical implications. Such finding would suggest that the recommendation for using retrieval practice needs to be qualified in terms of this characteristic. Theoretically, identifying profiles of learners who consistently do not benefit from retrieval practice would inform contemporary accounts in the field. For example, if individuals with lower gF benefit more

from retrieval practice than those with higher gF, researchers would need to propose or revise hypotheses that account for this moderator variable.

Figure 2.1

Frequency Distribution and Hypothetical Relationships of Participant-Level Retrieval Practice Effect Scores with an Individual-Difference Variable



Note. ID = individual-difference. Data combined in panel A were described by Brewer et al. (2021). The remaining panels display simulated data showing a null correlation (B), a positive correlation (C), and a negative correlation (D) between participant-level retrieval practice effect and an individual-difference variable. Magnitudes of correlations in panels B, C, and D were chosen only for illustrative purposes. Dark gray bars and datapoints indicate positive retrieval practice effects, whereas light gray bars and datapoints indicate null and negative retrieval practice effects.

In summary, the retrieval practice effect has been well established in experimental research (for meta-analytic reviews, see Rowland, 2014; Schwieren et al., 2017; Yang et al., 2021). The phenomenon has been replicated in various learning conditions, materials, and criterion tests, making retrieval practice a highly useful learning technique in educational contexts (Dunlosky et al., 2013). However, this utility is expected to be evident at the group level, which may obscure the fact that retrieval practice may not work for some learners. It is this aspect that individual-difference research aims to explore.

Individual-Difference Variables as Potential Moderators of the Retrieval Practice Effect

Brewer and Unsworth (2012) outline three potential relationships between individual-difference variables the retrieval practice effect, focusing on cognitive abilities. Here, we extend their description by including the term *trait* to encompass personality characteristics as another potential set of individual-difference variables that could moderate the retrieval practice effect, given their exploration by some researchers. These potential relationships are illustrated in Figure 2.1, panels B–D. These scenarios are not exhaustive (they could be nonlinear, for example) and they represent empirical patterns that are not necessarily predicted by contemporary accounts in the field.

The first scenario supports the universal recommendation of retrieval practice for all learners: The participant-level retrieval practice effect is not consistently related to individual-difference variables (Figure 2.1, panel B). Learners across the spectrum of these variables are equally likely to benefit from engaging in retrieval practice. The second scenario suggests that retrieval practice is more beneficial for those learners who already possess a higher level of latent ability or trait. In this case, learners who employ their cognitive resources suboptimally might benefit less from retrieval practice (Figure 2.1, panel C). If this case holds true, teachers and

educators would need to explore alternative methods to enhance the learning of students with lower latent abilities or traits. Lastly, the third scenario suggests that learners with higher latent abilities or traits are already employing their cognitive resources optimally and, consequently, benefit less—or do not benefit at all—from retrieval practice. Conversely, learners with lower latent abilities or traits, who are presumed to use suboptimal encoding strategies, could reap greater benefits from retrieval practice (Figure 2.1, panel D).

Individual Differences in the Retrieval Practice Effect

Tables 2.1 and 2.2 summarize characteristics and results, respectively, of studies on individual differences in the retrieval practice effect. To clarify, our interest lies in individual-difference variables as potential moderators of the retrieval practice effect in healthy individuals. Therefore, we did not include studies that investigated whether the retrieval practice effect would emerge in different populations, such as comparing a clinical group with an age-matched healthy control group (e.g., Minear et al., 2023; Sumowski et al., 2010). Similarly, studies involving between-subject manipulations of learning strategy (i.e., retrieval practice vs. control condition) were not included in these tables since they did not measure the retrieval practice effect at the participant level (e.g., Jaeger et al., 2015; Stenlund et al., 2017; Wiklund-Hörnqvist et al., 2014).

As seen in Table 2.1, most studies used word pairs as the to-be-learned material and cued-recall tests during both the practice and the final-test phases. Regarding the number of practice cycles and the provision of feedback, the scenario was more heterogeneous. It is important to note that studies incorporating corrective feedback or multiple practice cycles introduce indirect benefits of retrieval practice (Karpicke, 2017). Given this heterogeneity, it should be emphasized that there is no single retrieval practice effect, but rather multiple different effects, depending on the experimental paradigm adopted. We will briefly return to this issue in the Final Considerations.

Table 2.1

Characteristics of Studies on Individual Differences in the Retrieval Practice Effect

Study	Sample	Material	Retrieval practice task	Cycles	Feedback	Retention interval	Final test
Agarwal et al. (2017)	College students	110 general knowledge facts	CR test	1	Yes and no	10 min or 2 days	CR test
Bertilsson et al. (2021)	Upper secondary-level students	60 Swahili–Swedish word pairs	CR test	6	Yes	5 min, 1 week, or 4 weeks	CR test
Bertilsson et al. (2017), Exp. 2	Upper secondary-level students	60 Swahili–Swedish word pairs	CR test	6	Yes	5 min, 1 week, or 4 weeks	CR test
Brewer & Unsworth (2012)	College students	40 weakly associated English–English word pairs	CR test	1	Yes	1 day	CR test

Study	Sample	Material	Retrieval practice task	Cycles	Feedback	Retention interval	Final test
Jonsson et al. (2020), Exp. 1	Upper secondary-level students	60 Swahili–Swedish word pairs	CR test	6	Yes	5 min, 1 week, or 4 weeks	CR test
Jonsson et al. (2020), Exp. 2	Upper secondary-level students	60 Swahili–Swedish word pairs	CR test	6	Yes	5 min, 1 week, or 4 weeks	CR test
Minear et al. (2018)	College students	48 Swahili–English word pairs	CR test	4	Yes	2 days	CR test
Moreira, Pinto, Justi, & Jaeger (2019), Exp. 1	6th grade students	One encyclopedic text	Fill-in-the-blank test	1	No	1 week	Fill-in-the-blank and MC tests
Moreira, Pinto, Justi, & Jaeger (2019), Exp. 2	4th grade students	One encyclopedic text	Fill-in-the-blank test	2	No	1 week	Fill-in-the-blank and MC tests

Study	Sample	Material	Retrieval practice task	Cycles	Feedback	Retention interval	Final test
Pan et al. (2015), Exp. 1	Adults from the Amazon Mturk worker pool	40 weakly associated English–English word pairs	CR test	1	Yes	1 day	CR test
Pan et al. (2015), Exp. 2	College students	40 weakly associated English–English word pairs	CR test	1	Yes	1 day	CR test
Robey (2019), Exp. 1	College students	40 English–English word pairs from five categories[a]	CR test	2	Yes	30 min	CR test
Robey (2019), Exp. 2	College students	40 English–English word pairs from five categories[a]	CR test	2	Yes	15 min	CR test

Study	Sample	Material	Retrieval practice task	Cycles	Feedback	Retention interval	Final test
Tse et al. (2019), Exp. 1	College students	80 general knowledge facts	CR test	2	Yes	Immediate and 2 days	CR test
Tse et al. (2019), Exp. 2	College students	80 general knowledge facts	CR test	2	Yes	Immediate and 2 days	CR test
Tse & Pu (2012)	College students	40 Swahili–English word pairs	CR test	12	No[b]	1 week	CR test
Wenzel & Reinhard (2019), Exp. 2	College students	Textbook chapter on the brain's lateralization	MC and open-ended questions	1	Yes	1 week	MC and open-ended questions
Wiklund-Hörnqvist et al. (2022)	Upper secondary-level students	60 Swahili–Swedish word pairs	CR test	6	Yes	1 week	CR test

Note. All studies used rereading as the control condition. CR = cued-recall. MC = multiple-choice.

Study	Sample	Material	Retrieval practice task	Cycles	Feedback	Retention interval	Final test

[a] Related – high imageability nouns, related – low imageability nouns, unrelated – high imageability nouns, unrelated – low imageability nouns, and nonsense words. [b] Although there was no feedback after retrieval attempts, the repeated retrieval practice condition (S-T-S-T-S-T-S-T-S-T-S-T) was compared with the repeated study condition (S-S-S-T-S-S-S-T-S-S-S-T), where S stands for study blocks and T stands for test (retrieval practice) blocks.

Table 2.2

Summary of Studies on Individual Differences in the Retrieval Practice Effect

Study	Personality			Cognitive Ability						
	Grit	NFC	TA	AC	Cog	EM	gC	gF	Reading	WMC
Agarwal et al. (2017)										_[b]
Bertilsson et al. (2021)	✗	✗								✗
Bertilsson et al. (2017), Exp. 2	✗	✗								✗
Brewer & Unsworth (2012)				✗		–		–		✗
Jonsson et al. (2020), Exp. 1					✗					
Jonsson et al. (2020), Exp. 2					✗					
Minear et al. (2018)[a]							✗	✗		✗
Moreira, Pinto, Justi, & Jaeger (2019), Exp. 1									✗	
Moreira, Pinto, Justi, & Jaeger (2019), Exp. 2								✗	✗	✗

Study	Personality			Cognitive Ability						
	Grit	NFC	TA	AC	Cog	EM	gC	gF	Reading	WMC
Pan et al. (2015), Exp. 1						✗				
Pan et al. (2015), Exp. 2						✗				
Robey (2019), Exp. 1						✗		✗		
Robey (2019), Exp. 2						✗		✗		
Tse et al. (2019), Exp. 1			✗							✗
Tse et al. (2019), Exp. 2			✗							–[c]
Tse & Pu (2012)			–							✗
Wenzel & Reinhard (2019), Exp. 2								+		
Wiklund-Hörnqvist et al. (2022)		✗								

Study	Personality			Cognitive Ability						
	Grit	NFC	TA	AC	Cog	EM	gC	gF	Reading	WMC

Note. Exp. = experiment. NFC = need for cognition. TA = trait anxiety for test taking. AC = attentional control. Cog = cognitive ability (composite index including measures of gF, WMC, EM, visuospatial short-term memory, and updating). EM = episodic memory. gC = crystalized intelligence. gF = fluid intelligence. Reading = reading ability. WMC = working memory capacity. The ✗ symbol denotes no effect, the plus symbol denotes a positive effect, and the minus symbol denotes a negative effect.

[a] Results for the full dataset. [b] Agarwal et al. (2017) observed a negative significant correlation and three nonsignificant ones. [c] Tse et al. (2019), Experiment 2 observed two negative regression coefficients (one significant and one marginally significant) and two nonsignificant ones.

Individual-difference variables can be grouped into two broader categories, namely, personality traits (i.e., grit, need for cognition, and trait anxiety for test taking) and cognitive abilities (e.g., gF and working memory capacity; WMC; see Table 2.2). Studies focusing on personality traits were predominantly conducted by the same research team, and they did not report any associations, a pattern similar to that shown in Figure 2.1, panel B (Bertilsson et al., 2021; Bertilsson et al., 2017; Wiklund-Hörnqvist et al., 2022).[1] A notable exception was an experiment that observed that participants with lower trait anxiety for test taking (test anxiety) benefited more from retrieval practice (Tse & Pu, 2012). However, this finding was not replicated (Tse et al., 2019). Tse and Pu's focus was on a three-way interaction, which we will discuss below.

The majority of studies focused on cognitive abilities. Studies investigating the effects of attentional control, crystallized intelligence, reading skills, and general cognitive ability (i.e., a latent factor combining different abilities) indicated that the retrieval practice effect is independent of these variables (Brewer & Unsworth, 2012; Jonsson et al., 2020; Minear et al., 2018; Moreira, Pinto, Justi, et al., 2019). It is important to note, however, that at most, that these constructs have been explored by few studies (i.e., no more than two).

Five studies found no relationship between WMC and the retrieval practice effect (Bertilsson et al., 2021; Bertilsson et al., 2017; Brewer & Unsworth, 2012; Minear et al., 2018; Moreira, Pinto, Justi, et al., 2019, Experiment 2). Tse et al. (2019) tested eight hierarchical regression models, including various predictors such as measures of test anxiety, WMC, mood, and interactions between these variables. In their Experiment 2, Tse et al. found WMC as a negative predictor in only two models—after immediate tests for both interesting and boring facts.

[1] Minear et al. (2018) also measured need for cognition and grit, along with measures of the Big Five constructs, academic entitlement, academic self-efficacy, test anxiety, and stress (see their Footnote 1). However, Minear et al. did not report the relationships observed between these variables and the participant-level retrieval practice effect.

In a previous study, however, the focus was on examining the combined effect of test anxiety and WMC: Test anxiety had a negative correlation with the retrieval practice effect for participants with lower WMC but not for those with higher WMC (Tse & Pu, 2012).

Similarly, Agarwal et al. (2017) assessed the effect of an individual-difference variable contingent on other design features. Specifically, Agarwal et al. found that participants with lower WMC benefited more from retrieval practice with feedback in a 2-day retention interval—but not in other three conditions, retrieval practice with feedback in a 5-min retention interval, retrieval practice without feedback in a 5-min retention interval, and retrieval practice without feedback in a 2-day retention interval. Taken together, these studies suggest that the impact of an individual-difference variable on the magnitude of the retrieval practice effect may depend on other individual differences or contextual factors (see Roediger, 2008, for a contextualist approach to memory phenomena).

Studies examining the relationship between the retrieval practice effect and episodic memory abilities, as well as gF, have yielded mixed results. Brewer and Unsworth (2012) observed that participants with lower episodic memory abilities and with lower gF scores, as opposed to those with higher abilities and scores, benefited the most from retrieval practice (similar to Figure 2.1, panel D). However, in two replication attempts, Pan et al. (2015) did not observe the same pattern of results concerning the episodic memory measure (similar to Figure 2.1, panel B).

Other studies also failed to find significant correlations between the retrieval practice effect and measures of episodic memory or gF (Moreira, Pinto, Justi, & Jaeger, 2019; Robey, 2019). Wenzel and Reinhard (2019, Experiment 2) reported a result pattern that was the opposite of that reported by Brewer and Unsworth (2012): Retrieval practice benefited participants with average and above-average gF, but did not benefit participants with below-average gF (similar to Figure

2.1, panel C). Finally, Minear et al. (2018), who also considered item difficulty, identified a three-way interaction: The retrieval practice effect for easy items was greater for lower gF participants, while the retrieval practice effect for difficult items was greater for higher gF participants. This result, however, only emerged when the analyses were restricted to individuals who benefited from retrieval practice.

To date, studies have predominantly examined the direct relationship between individual-difference variables and the retrieval practice effect. However, a few studies have examined whether encoding and retrieval strategies spontaneously employed by participants during retrieval practice experiments are related to the retrieval practice effect at the participant level (for exceptions using self-reported strategy use, see Minear et al., 2018; Robey, 2019). For example, it is known that better semantic organization is associated with higher recall in the final-test phase (Cavendish et al., 2022; Rawson & Zamary, 2019), although investigations relating this type of strategy to studies on individual differences is lacking. It is possible, for instance, that the impact of individual-difference variables on the retrieval practice effect is mediated by difference in participants' spontaneous encoding and retrieval strategies. This appears to be an interesting avenue for future research.

In summary, studies have not identified individual-difference variables that consistently relate to the retrieval practice effect. However, this literature is much less extensive than the experimental literature demonstrating the retrieval practice effect at the group level (Pan & Rickard, 2018; Rowland, 2014; Yang et al., 2021). Furthermore, given the heterogeneity of procedures in studies on individual differences, it is important to further examine the heterogeneity of these studies.

Methodological Heterogeneity Across Studies

In human memory research, the interplay between material characteristics, task complexity, and individual differences can lead to the recruitment of different cognitive processes (Healey & Kahana, 2014), thus contributing to inconsistent results (Roediger, 2008). If the relationship between an individual-difference variable and the retrieval practice effect can be influenced by other variables (i.e., higher-order interactions), these variables need to be carefully considered in future investigations. The following four illustrative examples aim to point out important aspects of methodological heterogeneity that need to be considered by researchers.

First, Agarwal et al. (2017) and Brewer and Unsworth (2012) examined whether WMC moderates the retrieval practice effect. While their experiments varied in several dimensions, we focus on one here: The number of intervening items between successive presentations of the same item (i.e., lag), which ranged from 0 to 9 in the Agarwal et al. study. These lags imposed different demands on learners' capacity to keep the information active in the face of distractions, such as the interference of subsequent items (Conway et al., 2005). In contrast, the Brewer and Unsworth study had a lag of at least 20 items, thus exceeding learners' capacity, and rendering WMC irrelevant. During the practice phase, participants in the Agarwal et al. study performed better (approximately 80%) than those in the Brewer and Unsworth study (46%). Of course, differences in material difficulty could also account for the different results. Nonetheless, exploring this three-way interaction (Learning Strategy × WMC × Lag) could indeed open up an intriguing research avenue.

Second, the number of retrieval opportunities varies across studies (see Table 2.1). In experiments using word pairs, participants might transition from mediated to direct retrieval with extended practice (Crutcher & Ericsson, 2000; Dikmans et al., 2020; Kole & Healy, 2013).

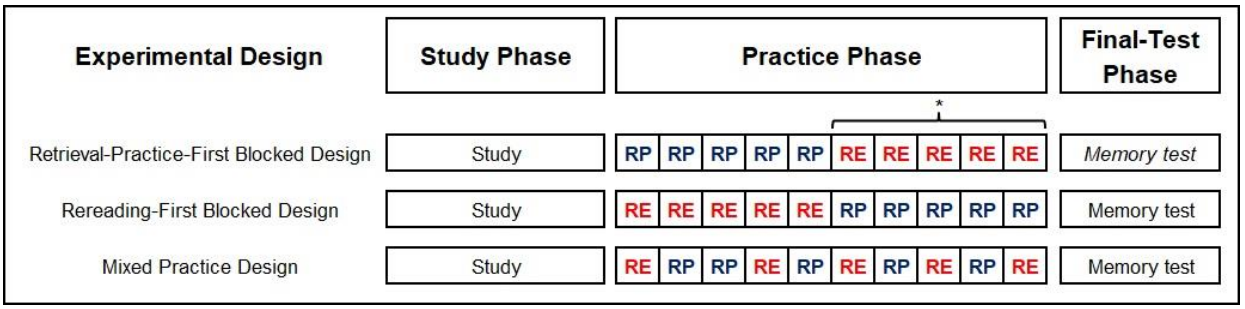
Consequently, the varying number of retrieval practice opportunities could engage different cognitive mechanisms. If different learners have distinct learning trajectories, the relationship between individual-difference variables and the retrieval practice effect might change over time. The key point here is that researchers should convert between-study differences into independent variables to identify the so-called "hidden moderators" (Klein et al., 2018).

The final examples pertain to how items are presented during the practice phase (Abel & Roediger, 2017). Figure 2.2 illustrates three different designs (cf. Gupta et al., 2024). Items assigned to one condition can be temporally separated from those in another condition. This order—retrieval-practice-first blocked or rereading-first blocked designs—might (e.g., Robey, 2019) or might not be (e.g., Brewer & Unsworth, 2012) counterbalanced across participants. Alternatively, rereading and retrieval practice items can be randomly intermixed during the practice phase (mixed practice design; e.g., Wiklund-Hörnqvist et al., 2022).

Crucially, a study involving a 5-list learning demonstrated that retrieving information from episodic, semantic, or short-term memory after each one of the initial four lists, as opposed to rereading them, enhanced the encoding of the fifth list, as measured by a later memory test (Pastötter et al., 2011). This is the forward testing effect (Pastötter et al., 2011) or the test-potentiated new learning (Yang et al., 2021). Likewise, Gupta et al. (2024) found that studies using retrieval-practice-first blocked designs are partially confounded by the forward testing effect (see Figure 2.2), that is, the retrieval practice effect appears smaller, possibly because the blocked presentation benefits the subsequent encoding of the material in the rereading condition (for an independent, but similar argument, see Mulligan et al., 2022).

Figure 2.2

The Potential Confounding Role of the Forward Testing Effect in Studies on the Retrieval Practice Effect



Note. RE = rereading. RP = retrieval practice. Asterisk indicates which set of items is expected to benefit from the forward testing effect. Italics denote the memory test in which is expected a better performance in the rereading condition (in retrieval-practice-first blocked design), compared with this same condition in the rereading-first blocked design. Based on Gupta et al. (2024).

The significance of these findings lies in their indication that the experimental design could impact the retrieval practice effect at the group level, thereby influencing the functional relationship between this effect and individual-difference variables. Notably, studies using blocked designs observed smaller retrieval practice effects (Brewer & Unsworth, 2012; Minear et al., 2018; Robey, 2019) compared to those using mixed practice designs (Bertilsson et al., 2017; Pan et al., 2015). Importantly, Gupta et al. (2024) claimed that higher ability learners are likely to benefit more from the confounding forward testing effect. If this claim holds true, it is possible that such methodological characteristic adds noise in the retrieval practice effect at the participant level (e.g., by changing the rank order of participants in terms of benefits from retrieval practice). Additionally, other findings suggest that even unrelated tasks, if administered *before* the memory task, might also lead to this confounding effect, improving encoding in the rereading condition (Pastötter et al., 2011).

How can these potential interpretive problems be mitigated? Researchers must be explicit about the specific effect they intend to investigate. If the focus is on the direct effect of retrieval practice, it is essential that the experiment adopts a mixed practice design with only one practice cycle and without feedback after retrieval practice. Alternatively, for those interested in a combination of direct and indirect benefits of retrieval practice, a mixed practice design (with one or more practice cycles with feedback) should be used. Finally, to prevent interpretive issues arising from the confounding forward testing effect, researchers should avoid blocked practice designs and the application of cognitive tasks (e.g., gF tasks) before the main experiment.

Introducing the Dual-Memory Framework

Cognitive scientists have proposed a number of contemporary accounts of the retrieval practice effect. These accounts differ in scope, with some describing empirical patterns and others positing cognitive mechanisms underlying the phenomenon. They include, but are not restricted to, the elaborative retrieval hypothesis (Carpenter, 2009), the mediator-effectiveness hypothesis (Pyc & Rawson, 2010), and the episodic context account (Lehman et al., 2014). However, contemporary accounts typically predict group-level patterns. They are, at least in their original formulations, silent about potential individual differences. One consequence of this state of affairs is that studies on individual differences—which seek to answer questions such as "Do individual-difference variable *X* relate to the magnitude of the retrieval practice effect?"—may not necessarily advance predictions based on extant accounts.

The dual-memory framework (Rickard & Pan, 2018) provides a viable theoretical framework for generating quantitative models and predictions regarding the relationship between individual-difference variables and the magnitude of the retrieval practice effect. This descriptive framework relies on the idea of strength of memory traces. Initially developed to account for

findings from experiments using cued-recall tasks, the dual-memory framework posits that rereading and retrieval practice items are recalled if their memory strengths are above a fixed response threshold. Importantly, while rereading items are modeled by a single memory strength dimension, retrieval practice items are modeled by two distinct and independent memory strength dimensions (Rickard & Pan, 2018).

The dual-memory framework predicts the retrieval practice effect based on the probability of correct responses in the rereading condition (rereading proportion correct), PC_R, using a quadratic function: $Retrieval\ practice\ effect = PC_R - PC_R^2$ (Rickard & Pan, 2018). This function, illustrated as the solid line in Figure 2.3, suggests that larger retrieval practice effects are expected when $PC_R = .50$, decreasing as PC_R approaches 0 or 1. Now, consider the case where an individual-difference variable (e.g., gF) has a positive correlation with memory performance (for evidence for latent correlations between gF and long-term memory, see Unsworth, 2019). But what will be the relationship between gF and the retrieval practice effect? The dual-memory framework proposes that the correlation between gF and the retrieval practice effect will be "(in part) a joint consequence of (1) the relation between the [individual-difference] variable and [rereading] proportion correct, and (2) the relation between [rereading] proportion correct and the [retrieval practice effect]" (Rickard, 2020, p. 789).

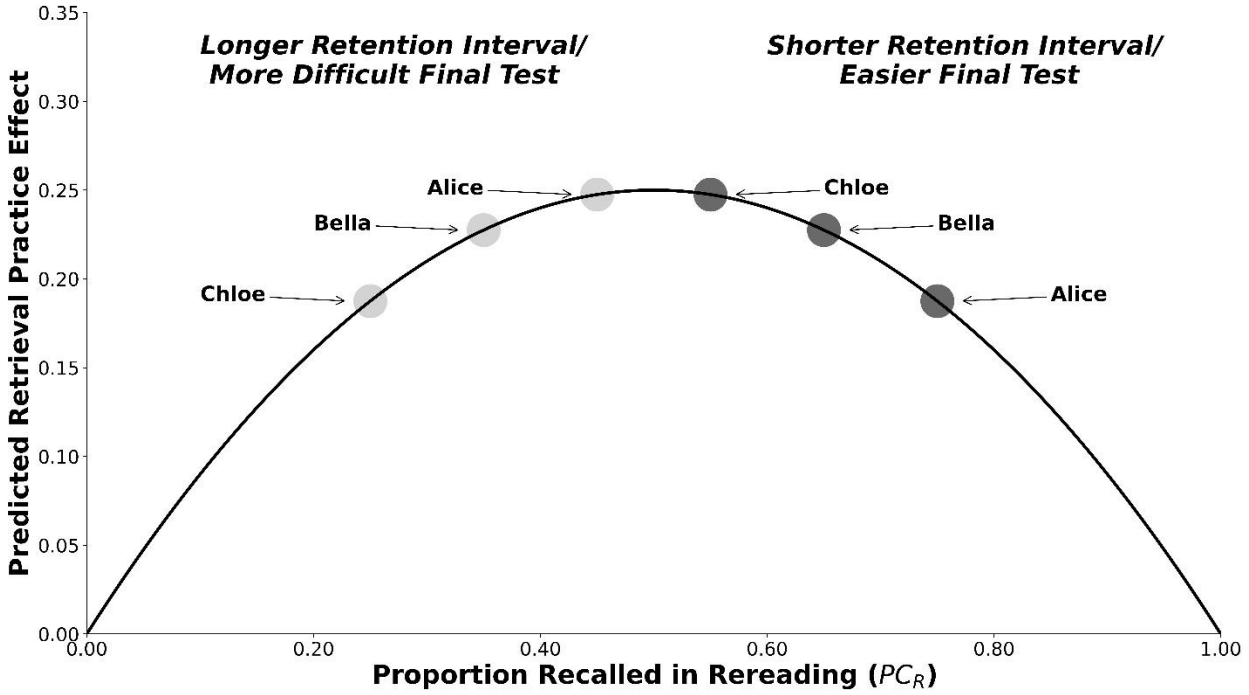
An illustrative example is provided below.[2] Suppose Alice, Bella, and Chloe recall .75, .65, and .55 of the rereading items, respectively, in a hypothetical experiment with a shorter retention interval or an easier final test. Further, assume that their true gF scores were ranked as follows: $Alice > Bella > Chloe$. In this scenario, the dual-memory framework predicts retrieval practice effects of .19, .23, and .25 for Alice, Bella, and Chloe, respectively (rounding to two decimal

[2] The mathematical details and a modeling approach are presented in Chapter 4.

places; see the dark gray points in Figure 2.3). In this example, the average $PC_R > .50$, and gF scores and the retrieval practice effects are negatively correlated, a consequence of the quadratic relationship between PC_R and the predicted retrieval practice effect described earlier.

Figure 2.3

Illustrative Example of the Dual-Memory Framework in Two Hypothetical Experiments with Different Retention Intervals and Difficulties in the Final Test



Note. Solid line represents the predicted retrieval practice effect (*y*-axis) as a function of the proportion recalled in rereading (*x*-axis).

What if this hypothetical experiment were instead one with a longer retention interval or a more difficult final test? In this new scenario, some forgetting would be expected. For instance, let us assume that Alice, Bella, and Chloe recall .45, .35, and .25 of the rereading items, respectively. In this hypothetical situation, the participants' rank-order in rereading performance was preserved,

consistent with studies indicating a high correlation between immediate and delayed recall (Gates, 1918; Jonsson et al., 2014). The dual-memory framework now predicts retrieval practice effects of .25, .23, and .19 for Alice, Bella, and Chloe, respectively, reversing the order of the magnitude of the retrieval practice effects (see the light gray points in Figure 2.3). Now, with an average $PC_R < .50$, assuming that the true gF scores remain the same, the gF scores and the retrieval practice effects are positively correlated. Of course, this pattern will change depending on the strength of the correlation between PC_R and an individual-difference variable (see Rickard, 2020).

Our take-home message here is that the dual-memory framework has the advantage of being, at least in principle, able to describe how different empirical patterns would emerge from the complex interaction between characteristics of learners, materials, and tasks. A potential avenue for future studies is to randomly assign participants to tasks that induce an average PC_R above or below .50 (e.g., easier and more difficult tests, respectively) and explore whether the correlations between an individual-difference variable and the retrieval practice effect align with the predictions of the dual-memory framework.

Final Considerations

In methodology, certain concepts exhibit polysemic characteristics. In experimental research, effects are said to be *reliable* when they are replicable across participants or situations (e.g., Brewer & Unsworth, 2012, p. 408; Carpenter, 2009, p. 1563) or even when they are statistically significant (e.g., Nickerson, 2000, p. 256 and p. 288; Roediger & Karpicke, 2006b, p. 252; Rowland, 2014, p. 1446). In individual-difference research, scores of a measure are deemed reliable, in a psychometric sense, if they consistently yield error-free scores (Mair, 2018; Nunally & Bernstein, 1994). This equates to repeating tasks with the same participants, showing that those benefiting most from retrieval practice at time i tend to do so at time $i + 1$. Some researchers

emphasize the importance of demonstrating psychometric reliability in the retrieval practice literature (Lima & Buratto, 2023b; McDermott, 2021).

The retrieval practice effect is reliable in the experimental sense (Pan & Rickard, 2018; Rowland, 2014; Yang et al., 2021). However, there is only preliminary evidence regarding its psychometric reliability (Lima & Buratto, 2023b). More studies are needed to investigate whether the retrieval practice effect exhibits psychometric reliability across various experimental paradigms, including different retention intervals, tasks, and materials (Brewer & Unsworth, 2012; McDermott, 2021).

Closely related, the retrieval practice effect is not a singular phenomenon but rather comprises multiple distinct effects. In other words, the retrieval practice lacks *portability*, meaning that due to the diverse procedures used in individual-difference studies, retrieval practice effects estimated by these tasks do not represent a fixed unit (Rouder & Haaf, 2019)—they can be thought as random effects, akin to the meta-analysis literature (Hedges & Vevea, 1998; Lima & Buratto, 2023a). This variability in effects across different procedures raises the possibility that these effects may have different relationships with individual-difference variables. For example, holding other variables constant, retrieval practice effects from two experiments, one with a single practice cycle and another with 20 cycles, might have varying correlations with individual-difference variables.

The main challenge in interpreting divergent results in the retrieval practice literature is that these studies often vary in multiple factors simultaneously. Additionally, there are few attempts at close replication (Pan et al., 2015). In cases of conflicting results of studies using heterogeneous procedures, efforts have not been made to follow up investigating the source of the discrepant results. We believe that close replications, where only one factor is manipulated at a

time, are essential to potentially identify “hidden moderators” in the relationship between individual-difference variables and the retrieval practice effect (Klein et al., 2018).

In his influential article, Cronbach argued that a united discipline is warranted to address important problems in psychology (Cronbach, 1957). We believe this is particularly true for research on the retrieval practice effect. We hope that, in the future, experimental and individual-difference approaches will make joint efforts to address the fascinating problems from human memory research.

Chapter 3 – Direct and Indirect Effects of Fluid Intelligence on the Retrieval Practice Effect

Manuscript 2

Abstract

The main aim of this study was to investigate the relationship between fluid intelligence (gF) and the retrieval-practice effect. Participants first studied 40 Swahili–Brazilian-Portuguese word pairs, then reread half of the word pairs and retrieval-practiced with feedback the other half. In separate sessions, they then completed cued-recall and gF tests. Three key questions were addressed. First, we attempted to generalize the two 3-way interactions found by Minear et al. (2018) in their analyses restricted to participants benefiting from retrieval practice. Overall, we successfully extended their results. Second, we investigated whether gF is related to the amount of new items participants recall during the practice phase. Consistent positive relationships were found in Cycles 1–3 (*r*s between .30 and .42). Third, we tested and found an indirect effect of gF on the retrieval practice effect mediated by performance during the practice phase. Learners with higher gF may be particularly skilled at generating effective mediators and at monitoring and replacing less-effective ones after retrieval failures during practice. To test this hypothesis, further studies should measure mediator production, shift, and retrieval, and correlate them with gF. In addition, we employed a duration-based procedure, in which the researcher determines in advance the number of practice trials per item. We propose future studies employing criterion-based procedures, where additional practice manipulation occurs after each item has been successfully recalled a predetermined number of times. This research agenda has the potential to sharpen our understanding of the conditions and cognitive mechanisms underlying individual differences in the retrieval practice effect.

Keywords: retrieval practice, testing effect, test-enhanced learning, memory, fluid intelligence

Direct and Indirect Effects of Fluid Intelligence on the Retrieval Practice Effect

What are the four Piagetian stages of cognitive development? What does the concept of object permanence entail? What is the difference between assimilation and accommodation? Correctly answering these questions necessitates prior exposure to these topics and retrieving information from memory. Cognitive scientists refer to attempts to retrieve information from memory as *retrieval practice* (Karpicke, 2017; Mulligan et al., 2022), a learning technique that enhances long-term retention (Karpicke & Roediger, 2008; Ludowicy et al., 2023; Minear et al., 2023; Roediger & Karpicke, 2006b). This phenomenon is known as the *retrieval practice effect.*

In this chapter, we focus on the relationship between gF and the retrieval practice effect. Does gF relate to the magnitude of the retrieval practice effect? Does gF predict performance in the practice phase? Is there an indirect effect between gF and the retrieval practice effect mediated by performance during the practice phase? We begin by describing studies that have investigated the relationship between gF and the retrieval practice effect. Next, we highlight a gap in the literature on individual differences. We then report the results of an experiment aimed at extending findings from a previous study (Minear et al., 2018). Finally, we discuss our results and propose a research agenda for studies on individual differences.

Studies on Individual Differences in gF and the Retrieval Practice Effect

An experiment on the retrieval practice effect is straightforward. In the study phase, participants are initially exposed to the to-be-learned material. Then, in the practice phase, they reread half of that material and engage in retrieval practice for the other half. Rereading and retrieval practice may occur one (e.g., Buchin & Mulligan, 2017, 2019; Roediger & Karpicke, 2006b, Experiment 1) or multiple times (e.g., Lima & Buratto, 2023b; Minear et al., 2023; Racsmány et al., 2018) for each item. Finally, in the final-test phase, participants take a memory

test. The retrieval practice effect is defined as better memory performance in the final-test phase in the retrieval practice condition than in the rereading condition. This effect has been repeatedly observed in both laboratory (Carpenter & Yeung, 2017; Klier & Buratto, 2023; Pyc & Rawson, 2010; Racsmány et al., 2018; Roediger & Karpicke, 2006b; for reviews, see Adesope et al., 2017; Rowland, 2014) and classroom studies (Agarwal, 2019; Batsell et al., 2017; Ekuni & Pompeia, 2020; Kenney & Bailey, 2021; Leeming, 2005; for reviews, see Moreira, Pinto, Starling, & Jaeger, 2019; Yang et al., 2021). Furthermore, the effect also emerges across the lifespan (Guran et al., 2020; Jaeger et al., 2015; Karpicke et al., 2016; Meyer & Logan, 2013) and with memory- and language-impaired populations (Friedman et al., 2017; Middleton et al., 2015; Sumowski et al., 2010).

The retrieval practice effect is conceptualized as a group-level phenomenon. However, when examining, for each participant, the difference in performance between retrieval practice and rereading in the final-test phase (referred to as the participant-level retrieval practice effect), it becomes apparent that some participants exhibit better memory performance after rereading than after retrieval practice (Brewer & Unsworth, 2012; Lima & Buratto, 2023b; Minear et al., 2018; Sumowski et al., 2013). Thus, a decade ago, researchers began to ask whether the retrieval practice effect is moderated by individual differences (Brewer & Unsworth, 2012).

Studies on individual differences in the retrieval practice effect include a memory task and one or multiple individual-difference tests. The memory task is similar to the experimental procedure previously described. Individual-difference tests consist of administering personality scales (e.g., Bertilsson et al., 2021; Bertilsson et al., 2017) or cognitive abilities tasks (e.g., Brewer & Unsworth, 2012; Robey, 2019) with the purpose of measuring constructs and relating these measures with the retrieval practice effect. Here we are interested in gF—the ability to solve novel

problems, engage in inductive, sequential, and quantitative reasoning, which is typically measured by nonverbal and supposedly culture-free tasks (Engle et al., 1999; Walrath et al., 2020).

So far, only seven studies have investigated the direct effect of gF on the retrieval practice effect (Brewer & Unsworth, 2012; Minear et al., 2018; Moreira, Pinto, Justi, & Jaeger, 2019, Experiment 2; Robey, 2019, Experiments 1 and 2; Starling et al., 2019; Wenzel & Reinhard, 2019, Experiment 2). Four of them failed to observe a moderating role of gF in the retrieval practice effect (Moreira, Pinto, Justi, & Jaeger, 2019, Experiment 2; Robey, 2019, Experiments 1 and 2; Starling et al., 2019). This was the case in the Robey (2019) study, even when structural equation modeling was employed to analyze the combined data from the two experiments, making it challenging to attribute the null results to low statistical power.

Wenzel and Reinhard (2019, Experiment 2) did observe a retrieval practice effect for participants with average and above-average gF, but not for those with below-average gF. This was consistent with a "rich-gets-richer" effect, that is, participants with high ability benefit most from retrieval practice. Brewer and Unsworth (2012), however, found a moderating effect in the opposite direction: Participants with lower gF benefited more from retrieval practice than those with higher gF. Yet, interpreting these conflicting results is complicated due to several differences across studies. These variations included the types of materials used (such as an introductory chapter on brain lateralization or English–English word pairs), the number of practice opportunities (ranging from one to four), retention intervals (ranging from 15 min to 1 week), and even analytical methods (including regression-based moderation, structural equation modeling, correlations, and quartile-based analysis of variance [ANOVA]). Therefore, it is possible that one or more of these differences contributed to the diverse outcomes across the studies.

The seventh study investigating the relationship between gF and the retrieval practice effect also took into account item difficulty (Minear et al., 2018). This study will be discussed in more detail for the purposes of this chapter. Notably, four key findings emerged. First, Minear et al. split their sample into participants who showed a positive, a negative, or a null retrieval practice effect (hereafter, *positive*, *negative*, and *null testers*, respectively). While positive and negative testers did not differ in Raven's Advanced Progressive Matrices (Raven et al., 1998; hereafter, *Raven*) scores or recall during the practice phase, they differed in the final-test phase. Specifically, negative testers outperformed positive testers in the final-test phase. Second, negative testers outperformed positive testers in the final-test phase for the rereading condition, while positive testers outperformed negative testers in the final-test phase for the retrieval practice condition.

Third, Minear et al. (2018) used quartile analyses, splitting positive testers into low and high gF group based on their Raven scores. They found a Learning Strategy (rereading vs. retrieval practice) × Group (low gF vs. high gF) × Difficulty (easy vs. difficult items) interaction: The low gF group showed a larger retrieval practice effect for easy items than for difficult ones, whereas the high gF group exhibited a larger retrieval practice effect for difficult items than for easy ones. Fourth, during the practice phase, although the high gF group outperformed the low gF group in the retrieval practice condition, Minear et al. found that similar patterns for the low gF group/easy items and the high gF group/difficult items. These results underscore the importance of considering both participants' abilities and task difficulty in studies on individual differences.

An Important Gap in the Literature on Individual Differences

The previously described studies explored the potential *direct* effect of gF on the retrieval practice effect. In other words, these studies have examined whether gF correlates with the magnitude of the retrieval practice effect. While this is important, we argue that it is equally crucial

to explore whether gF exerts an *indirect* effect on the retrieval practice effect. Below, we provide justifications for the importance of this topic.

So far, all existing studies on individual differences in the retrieval practice effect have employed duration-based procedures (Agarwal et al., 2017; Bertilsson et al., 2021; Bertilsson et al., 2017; Brewer & Unsworth, 2012; Minear et al., 2018; Pan et al., 2015; Robey, 2019; Wenzel & Reinhard, 2019; Wiklund-Hörnqvist et al., 2014; Wiklund-Hörnqvist et al., 2022). In these procedures, the researcher predetermines the number of practice trials per item (Pyc & Rawson, 2009; Vaughn & Rawson, 2011). This implies that the nominal exposure time to the material is fixed across participants, but not necessarily the effectiveness of this exposure, defined here as the performance achieved during the practice phase.

Several findings suggest that equating the total exposure time does not ensure equal learning and retention. Recent studies indicate individual differences in learning efficiency, that is, how quick learners master items to a criterion of one successful recall (Zerr et al., 2018; Zerr et al., 2021). Notably, quick learners often demonstrate better retention in a subsequent memory test (Zerr et al., 2018). Additionally, a meta-analytic review identified a positive relationship between performance during the practice phase (when feedback was withheld) and the magnitude of the retrieval practice effect (Rowland, 2014). Although this result was aggregated for potential individual differences, it aligns with the notion that duration-based procedures might not always facilitates learners with different characteristics in mastering the material.

An illustrative case supporting this interpretation involves the *negative testing effect* (Peterson & Mulligan, 2013), the finding that retrieval practice impairs retention when initial retrieval heavily relies on intraitem relational processing (e.g., cue–target relationship), but the final-test phase heavily relies on interitem relational processing (i.e., rhyme relationship shared by

targets from different items). Researchers from the University of North Carolina at Chapel Hill initially described and replicated the effect (Mulligan & Peterson, 2015; Peterson & Mulligan, 2013), while researchers from Kent State University reported five unsuccessful attempts to replicate it (Rawson et al., 2015). An across-site, close replication experiment demonstrated that the disparate results were linked to across-site differences in WMC, gF, and performance during the practice phase (Mulligan et al., 2018).

Finally, in some studies, conditional analyses indicated that the likelihood of recalling items in a memory test increased as they were successfully retrieved more times during the practice phase (Ariel & Karpicke, 2018; Finley et al., 2011; Lima, Venâncio, et al., 2020). This kind of analyses may be plagued by idiosyncratic item selection effects (Horton, 1987; Slamecka & Graf, 1978)—easier items are more likely to be recalled in both the practice and the final-test phases. However, the key point here is that if participants with higher gF have a larger pool of items with high retrieval success during the practice phase than participants with lower gF, they might be more likely to benefit from retrieval practice. In other words, gF can exert an indirect effect on the retrieval practice effect mediated by performance during the practice phase.

In summary, we posit that if individual-difference variables are linked to the effectiveness of exposure during the practice phase, it will be challenging to disentangle the extent to which differences in the magnitude of the retrieval practice effect across participants are due to individual differences in encoding, retention, or a combination of both (akin to the idea of stage analysis; Crowder, 1976/2015). Although Minear et al. (2018) demonstrated that, for positive testers, participants with high gF exhibited superior performance in the practice phase and better retention in the final-test phase, their experiment did not explore the mediating role of performance in the

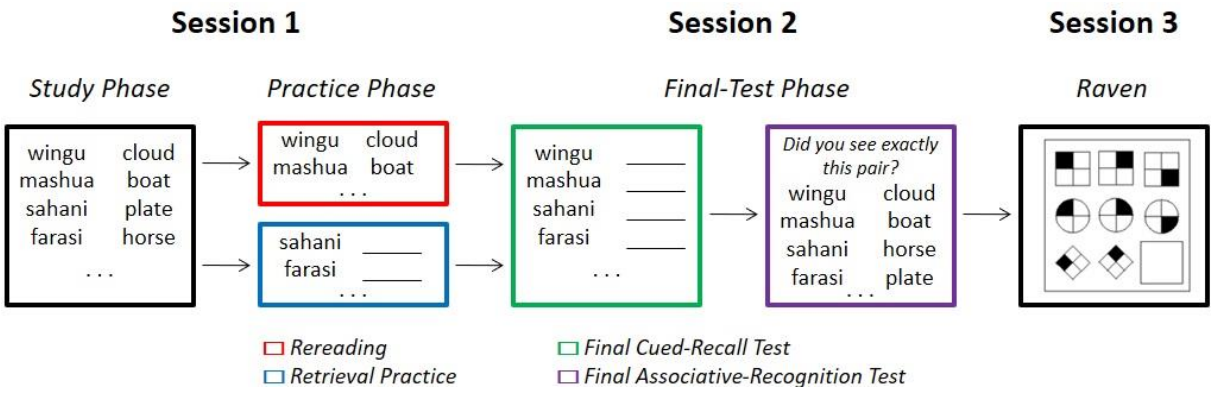
practice phase in the relationship between gF and the retrieval practice effect. One of the objectives of this study was to fill this gap.

The Current Experiment

Our general procedure is depicted in Figure 3.1. During Session 1, participants studied Swahili–Brazilian-Portuguese word pairs. They then reread half of the word pairs and engaged in retrieval practice (with feedback) the other half. For all items, rereading and retrieval practice trials was repeated in four cycles. Participants were instructed to return to the laboratory for Session 2 after 1 or 2 days, depending on their assigned condition. In Session 2, participants completed a final cued-recall test, followed by a final associative-recognition test. In Session 3, scheduled for one of the following weeks, participants took the Raven.

Figure 3.1

General Procedure



Note. During practice phase, corrective feedback was provided after all trials. Retention interval (i.e., between Sessions 1 and 2) was either 1 or 2 days, whereas intersession interval (i.e., between Sessions 2 and 3) was approximately 1 week.

This design enabled us to pursue three main objectives. Firstly, in light of the recent emphasis on the importance of replication (LeBel et al., 2019; Pashler & Wagenmakers, 2012), our aim was to generalize the three-way interactions found by Minear et al. (2018) in their analyses restricted to positive testers. Following the recommendations of LeBel et al., we referred to our study as an *extension* of Minear et al.'s work—rather than a *replication*, acknowledging the departures in our design from theirs. For transparency, it is crucial to note that, unlike Minear et al., we intended to operationalize difficulty in the final test by manipulating retention interval—the final-test phase was expected to be more difficult for participants assigned to a 2-day retention interval and easier for those assigned to a 1-day retention interval. To foreshadow, our retention interval manipulation failed to produce the intended effect, leading us to categorize our items as easy and difficult in a *post hoc* manner (Minear et al., 2018 also categorized their items as easy and difficult, although they adopted an *a priori* categorization).[3] In addition, owing to the lack of an effect of retention interval, we collapsed the data from the 1-day and 2-day intervals for most of the analyses.

Secondly, we explored whether gF is related to the amount of new items participants recall in each cycle during the practice phase. Although Minear et al. (2018) observed that participants with high gF had higher recall than participants with low gF during the practice phase, this analysis was restricted to positive testers. Here, we departed from Minear et al. and analyzed the data for the overall sample. Additionally, we aimed to assess whether individual differences in gF are linked to the post-retrieval re-encoding effects of feedback (Liu et al., 2018). Specifically, we examined the extent to which feedback after retrieval practice in cycle *i* contributes to participants

[3] In our experiment, a three-way interaction in the practice phase would not be expected, given that the manipulation of the retention interval occurred *after* the practice phase. However, this same interaction would make sense considering the difficulty of the items.

correctly recalling novel items in cycle $i + 1$. For this purpose, each analysis included only items not recalled in previous practice cycles.

Thirdly, we examined whether there is an indirect effect of gF on the retrieval practice effect. To address this, we posed three key questions: Does gF predict the average performance in the practice cycle? Does performance in the practice phase predict the magnitude of the retrieval practice effect? Most importantly, is there an indirect effect of gF on the retrieval practice effect mediated by performance during the practice phase? A simple mediation model was used to address these three questions simultaneously.

Method

Participants

This study is part of a larger project that originally aimed to investigate whether the correlation between gF and the retrieval practice effect differs for participants assigned to retention intervals of 1 and 2 days (see Chapter 4). Therefore, this comparison guided the sample size calculation.

The minimal required sample size ($N = 120$) was calculated with G*Power (v. 3.1.9.2; Faul et al., 2007), with an alpha level set at .05 (one tailed) and power set at .80 to detect a difference between two independent Pearson's *r*s coefficients (i.e., comparing correlations between the participant-level retrieval practice effect and gF scores across 1- and 2-day retention intervals). The allocation ratio across groups was set at 1. Based on simulations (see Chapter 4), the target *r*s were set at –.23 (1-day group) and .23 (2-day group). In practice, we oversampled participants to increase statistical power.

Participants were undergraduates or graduates recruited through ads posted on university bulletin boards, on social media posts, and Introduction to Psychology courses. The initial sample

size comprised 152 participants, of whom six were excluded for failing to attend Sessions 2 or 3, and two for not following the instructions in the memory task in Session 1. The final sample size consisted of 144 participants (96 cisgender women, 46 cisgender men, 1 transgender man, and 1 nonbinary person; $M_{age} = 21.61$ years, $SD = 4.15$). The research was approved by the Research Ethics Committee from University of Brasília (Appendix A). Participants provided informed consent before starting the tasks (Appendix B). Materials, data, and codes will be made openly available on the Open Science Framework.

Design

A 2 × 2 mixed-factorial design was employed, where the learning strategy (rereading, retrieval practice) and the retention interval (1 day, 2 days) were manipulated within- and between-subjects, respectively. Seventy-four participants were assigned to return to the laboratory 1 day after the Session 1 (range: 16–32 hr) and 70 participants were assigned to return to the laboratory 2 days after the Session 1 (range: 41–55 hr).

Materials

Word Pairs

Forty Swahili–Brazilian-Portuguese word pairs were selected from a normative database (Lima & Buratto, 2021). Swahili words are suitable for memory research because they (a) are based on the Latin alphabet, (b) exhibit adequate wordlikeness (i.e., they resemble Brazilian-Portuguese words), and (c) are unlikely to be familiar for Brazilian participants. These 40 word pairs were divided into two sets of 20 pairs each for counterbalancing purposes (Appendix C). For each participant, one set was assigned to the rereading condition and the other set to the retrieval practice condition. Based on Lima and Buratto's norms, difficulty was matched across sets

(average recall accuracies *M*s = .41). Unless otherwise stated, word pairs were presented in white color on a black screen.

Raven Advanced Progressive Matrices

The Raven test assesses gF and comprises two sets of items (Raven et al., 1998). We used the first two items from Set I for training and the 18 odd-numbered items from Set II for testing. Each item consists of a 3×3 matrix of geometric patterns, with the pattern in the bottom right corner missing. The task required participants to select, from eight alternatives, the one that properly completes the missing corner, considering both the horizontal and vertical patterns of each matrix. We used the Brazilian version of Raven test (Nunes & Nunes, 2015). In the present study, considering only the 18 odd-numbered items from Set II, a confirmatory factor analysis using the weighted least squares means and variance adjusted estimator indicated a satisfactory fit of a one-factor model to the data, $\chi^2_{SB}(135) = 143.60$, p = .29, Comparative Fit Index = .95, Tucker–Lewis Index = .94, Root Mean Square Error of Approximation = .02, 90% CI [.00, .05] (T. A. Brown, 2015), with a good internal consistency index, KR-20 = .71, 95% CI [.63, .77].

Procedure

The experiment took place over three sessions (see Figure 3.1). Tasks in Sessions 1 and 2 were administered on a computer using PsychoPy (Peirce et al., 2019), while Session 3 employed a paper-and-pencil format. All sessions were conducted individually.

Session 1

In Session 1, participants were instructed to study a series of word pairs, trying to learn the Brazilian Portuguese translation of Swahili words. During the study phase, participants saw 40 word pairs in random order. Each trial started with a fixation cross at the center of screen for 500 ms, followed by a word pair displayed for 7 s (e.g., *wingu–cloud*; *mashua–boat*). After the

presentation of the last word pair, participants engaged in a 1-min distractor task, involving simple mathematical operations (e.g., 12 × 8).

Next, participants were informed that all word pairs would be practiced in one of two possible ways. On rereading trials, participants saw a word pair on the screen for 7 s and were instructed to reread the pair and to type in the Brazilian Portuguese word. On retrieval practice trials, participants saw a Swahili word on the screen for 7 s and were instructed to recall and to type in the Brazilian Portuguese word translation. On both rereading and retrieval practice trials, after 7 s, regardless of participants' response, the correct response was displayed below the typed response in orange color for 2 s. This color was chosen to draw participants' attention to the correct response (feedback).[4] Rereading and retrieval practice trials were intermixed, each starting with a 500-ms fixation cross. Participants completed four practice cycles, with word pairs order randomized anew in each cycle. Each cycle was followed by a 1-min distractor task. After the final distractor task period, participants were informed to return to the laboratory after 1 or 2 days, based on their assigned condition.

Session 2

Session 2 began with the final cued-recall test. Each trial started with the presentation of a Swahili word. Participants were instructed to type in its Brazilian Portuguese translation and press Enter to proceed to the next trial. Regardless of the response, a trial concluded after 15 s. The 40 Swahili cues were presented randomly, and no feedback was given during the final cued-recall test. Following this, participants completed a final associative-recognition test. They were

[4] Feedback likely served distinct roles during the practice phase. Feedback after retrieval practice is believed to enhance subsequent memory performance, particularly when the initial recall is low (e.g., Finley et al., 2011; Silva et al., 2023; Tse et al., 2010). During rereading, on some trials, participants mistakenly copied the Swahili word instead of the Brazilian-Portuguese word. We posit that feedback after rereading likely helped capture participants' attention, prompting them to resume the task accurately.

informed that only words studied in Session 1 would be presented in this task. Some words formed intact pairs (e.g., *wingu–cloud* and *mashua–boat*), while others were rearranged pairs (e.g., *wingu–boat* and *mashua–cloud*). Rearranged pairs were pseudorandomly created for each participant, adhering to two constraints. Firstly, they involved a simple exchange between two pairs practiced using the same learning strategy—rereading pairs were rearranged only among themselves, and the same was done with retrieval practice pairs. Secondly, for each participant, there were 10 pairs per category (rereading/intact, rereading/rearranged, retrieval practice/intact, retrieval practice/rearranged). Participants were instructed to distinguish between old, intact pairs and new, rearranged pairs. The task was self-paced. "Old" and "new" responses mapped to the left and right keys, respectively, on the keyboard. After completing the final associative-recognition test, participants scheduled their Session 3 for one of the following weeks, according to their availability.

Session 3

In Session 3, participants completed the Raven test. During training, participants were instructed about the task and were allowed to ask questions while completing the first two items from Set I. The training was self-paced. Subsequently, during testing, participants were informed that they had up to 10 min to complete as many items as possible from the 18-odd numbered items from Set II. They had the option to skip items (and return to them later), but were cautioned that the items became progressively more challenging—so skipping items would mean advancing to more difficult ones. The researcher verbally notified participants when 5 and 10 min had passed. Upon completing the Raven test, all participants were debriefed, thanked, and dismissed.

Statistical Analyses

In frequentist analyses, the alpha level was set at .05, except when adjusted for post hoc multiple comparisons. Whenever the sphericity assumption was violated in ANOVAs, the Greenhouse–Geisser correction was applied to adjust the degrees of freedom. Since frequentist tests cannot provide support for their null hypotheses (Dienes, 2014), Bayesian analyses estimated the average strength of evidence relative to two sets of competing models. We report the model-average Bayes Factors ($BF_{Inclusion}$) of models including an effect compared with models excluding it (van den Bergh et al., 2022). For example, in a 2 (learning strategy) × 2 (retention interval) × 4 (cycle) mixed ANOVA, the $BF_{Inclusion}$ for the Learning Strategy × Cycle interaction quantifies the strength of evidence for models including this interaction term compared with models excluding it, with higher-order interaction terms excluded. For *t* tests, we report BF_{10}, using a Cauchy distribution width as prior (i.e., 0.707). Both the $BF_{Inclusion}$ and BF_{10} are continuous measures, but we also used verbal labels for qualify them, based on categories suggested by Lee and Wagenmakers (2013). Frequentist and Bayesian analyses were conducted in R (R Core Team, 2023) and JASP (Version 0.17.1; JASP Team, 2018), respectively.

Results

Memory Task

Practice Phase

Table 3.1 presents the proportion of correctly typed targets during the practice phase. Several noteworthy patterns emerge from Table 3.1. First, rereading performance remained stable and consistently high across cycles for both retention-interval groups (except for Cycle 1). This was expected since participants were typing in the target words already displayed on the screen, and since they were getting used with the procedure during Cycle 1. Second, participants

demonstrated an improvement in learning the word pair associations in the retrieval practice condition, as evidenced by the progressively increasing correct proportions across Cycles 1–4. Third, as expected, the performance in the rereading condition was approximately equivalent across the retention-interval groups, and the same was true in the retrieval practice condition.

Table 3.1

Practice Phase Proportion Correct, M (SD)

Group and Learning Strategy	Cycle			
	1	2	3	4
1-Day Group				
Rereading	.94 (.09)	.96 (.05)	.97 (.04)	.98 (.04)
Retrieval practice	.14 (.09)	.29 (.17)	.46 (.21)	.59 (.23)
2-Day Group				
Rereading	.92 (.14)	.95 (.12)	.97 (.05)	.97 (.06)
Retrieval practice	.12 (.10)	.30 (.18)	.46 (.24)	.58 (.25)

A 2 (learning strategy) × 2 (retention interval) × 4 (cycle) mixed ANOVA supported these observations. We found extreme evidence for a learning strategy effect, $F(1, 142) = 1{,}699.60$, $p < .001$, $\eta_p^2 = .92$, $BF_{inclusion} = 1.10 \times 10^{78}$, indicating that performance was markedly better in rereading trials ($M = .96$, $SD = .05$) compared to retrieval practice trials ($M = .37$, $SD = .17$). Additionally, there was extreme evidence for a cycle effect, $F(2.05, 291.10) = 440.31$, $p < .001$, $\eta_p^2 = .76$, $BF_{inclusion} = 3.52 \times 10^{92}$, suggesting a consistent performance improvement across Cycles 1, 2, 3, and 4 (Ms = .53, .63, .72, and .78, SDs = .08, .10, .12, and .13, respectively), all Bonferroni-corrected $ps < .001$. However, these two main effects were qualified by extreme

evidence for a Learning Strategy × Cycle interaction, $F(2.15, 304.70) = 292.19$, $p < .001$, $\eta_p^2 = .67$, $BF_{inclusion} = 4.20 \times 10^{132}$.

Given the distinct task demands of rereading and retrieval practice trials, we partitioned the data by learning strategy and examined this interaction further using two repeated-measures ANOVAs. For rereading trials, the extreme evidence for a cycle effect, $F(2.03, 290.30) = 12.78$, $p < .001$, $\eta_p^2 = .08$, $BF_{inclusion} = 2.11 \times 10^5$, was primarily due to lower performance in Cycle 1 ($M = .93$, $SD = .12$) compared to the Cycles 2, 3, and 4 (Ms = .96, .97, and .98, SDs = .09, .05, and .05, respectively), all $ps_{adj} \leq .02$. Cycles 2, 3, and 4 did not differ from each other, $ps_{adj} \geq .26$. For retrieval practice trials, the extreme evidence for a cycle effect, $F(1.96, 280.60) = 552.11$, $p < .001$, $\eta_p^2 = .79$, $BF_{inclusion} = 5.92 \times 10^{142}$, indicated differences between Cycles 1, 2, 3, and 4 (Ms = .13, .30, .47, and .59, SDs = .09, .17, .23, and .24, respectively), all $ps_{adj} < .001$. Importantly, in the initial ANOVA, both retention-interval groups performed comparably, $F < 1$, $p = .66$, indicating moderate evidence against an effect, $BF_{inclusion} = 0.21$. This result was expected, given that the retention-interval manipulation occurred after the practice phase. No other interactions reached significance, Fs < 1, ps ≥ .55, $BF_{inclusion} \leq 0.21$.

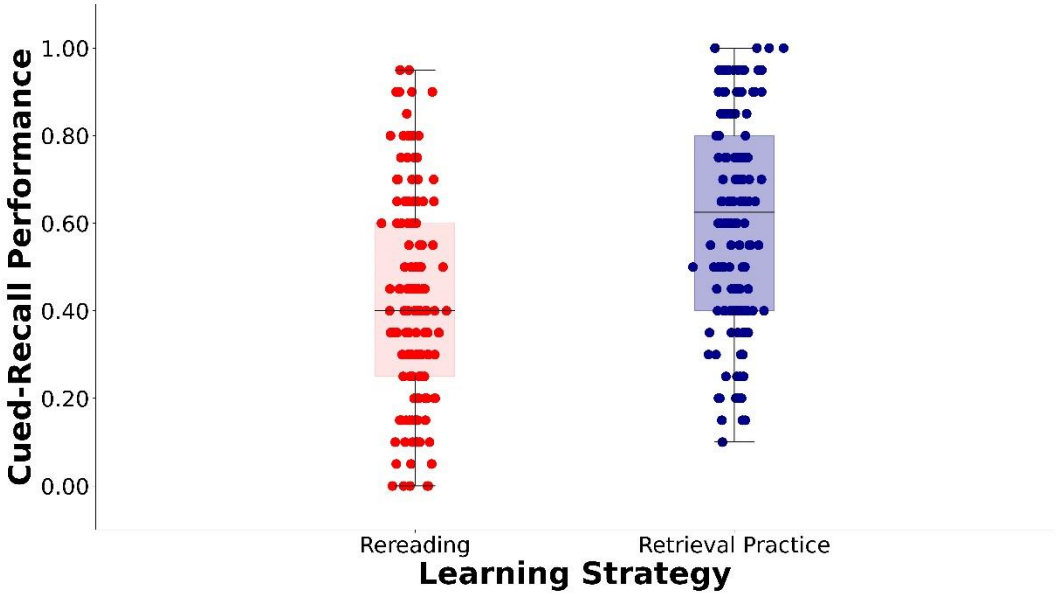
Final Test Phase

Final Cued-Recall Test. Figure 3.2 illustrates the proportion of correctly recalled targets during the final cued-recall test. As depicted in Figure 3.2, when collapsed across retention intervals, the final cued-recall performance was superior for the retrieval practice condition ($M = .60$, $SD = .24$) compared to the rereading condition ($M = .42$, $SD = .24$). A 2 (learning strategy) × 2 (retention interval) mixed ANOVA supported this group-level retrieval practice effect, $F(1, 142) = 156.79$, $p < .001$, $\eta_p^2 = .52$. Bayesian analysis indicated extreme evidence for models including the learning strategy, compared with models excluding it, $BF_{Inclusion} = 2.71 \times 10^{21}$.

Surprisingly, the 1-day group exhibited a small (*M* = .54, *SD* = .23), but nonsignificant advantage over the 2-day group (*M* = .48, *SD* = .21) in final cued-recall performance, $F(1, 142) = 2.57$, $p = .11$, $\eta_p^2 = .02$. Bayesian analysis indicated anecdotal evidence for models excluding retention interval as a predictor, compared with those including it, $BF_{Inclusion} = 0.83$. The interaction term was nonsignificant, $F < 1$, $p = .46$, and showed moderate evidence for a model without the interaction term compared with a model with the interaction term, $BF_{Inclusion} = 0.22$. These findings collectively indicate that the retrieval practice effect remained consistent regardless of the retention interval. The reader interested in the final cued-recall test performance broken down by retention intervals should refer to Figure D1 in the Appendix D.

Figure 3.2

Final Cued-Recall Test Performance as a Function of Learning Strategy



Final Associative-Recognition Test. During the research planning phase, we anticipated the possibility of a floor effect on the final cued-recall test for some participants. Consequently,

our primary rationale for including a final associative-recognition test was to assess whether participants, facing a test thought to rely on a recall-to-reject process (Malmberg, 2008), could demonstrate that specific items were in a latent state close to benefiting from retrieval practice. However, because the final associative-recognition test was administered after the final cued-recall test, it is conceivable that performance on the recognition test might have been influenced by the preceding final cued-recall test (cf. Knouse et al., 2016). To address this concern partially, the analyses presented in this section are segregated for items recalled and not recalled in the final-cued recall test.[5]

Figure 3.3 illustrates the proportion of word pairs correctly answered during the final associative-recognition test, calculated as $(hits + correct\ rejections)/number\ of\ items$.[6] Although we did not include an "I don't know" alternative in the final associative-recognition test, which might have potentially increased guessing responses, the average performance in all conditions was above chance (.50), all *p*s < .001. For the analysis restricted to items recalled in the final cued-recall test (Figure 3.3, panel A), the proportion of word pairs correctly answered exhibited a ceiling effect, remaining consistently similar across different learning strategies. A 2 (learning strategy) × 2 (retention interval) mixed ANOVA failed to reveal any significant effects, *F*s(1, 137) < 1, *p*s ≥ .74, with moderate evidence supporting models excluding each effect of interest, $BF_{Inclusion} \leq 0.19$. The reader interested in performance on the final associative-

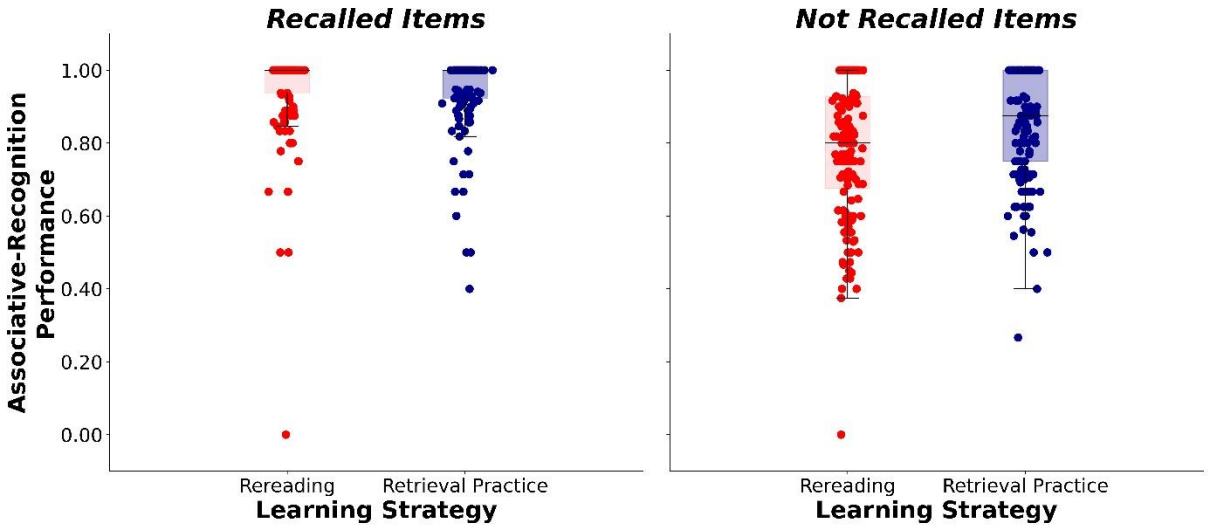
[5] In the conditional analysis for items *recalled* in the final cued-recall test, five participants were excluded for not recalling any items in the rereading condition (three in the 1-day group, and two in the 2-day group). Similarly, in the conditional analysis for items *not recalled* in the final cued-recall test, five participants were excluded for recalling all items in the retrieval practice condition (one in the 1-day group, and four in the 2-day group). Different participants were excluded from each of these conditional analyses. The results presented in this section are therefore based on 139 cases.

[6] The denominator is not two due to the conditional character of these analyses, in which different numbers of intact and rearranged pairs were possible for each participant.

recognition (for both recalled and not recalled items) test broken down by retention intervals should refer to Figure D2 in the Appendix D.

Figure 3.3

Final Associative-Recognition Test Performance as a Function of Learning Strategy



When considering only the items not recalled in the final cued-recall test, the pattern is different. As depicted in Figure 3.3 (panel B), the final associative-recognition test performance was superior for the retrieval practice condition (M = .86, SD = .16) than for the rereading condition (M =.78, SD = .19). A 2 (learning strategy) × 2 (retention interval) mixed ANOVA supported this group-level retrieval practice effect, $F(1, 137) = 18.00$, $p < .001$, $\eta_p^2 = .12$. Bayesian analysis indicated extreme evidence for models including the learning strategy, compared to those excluding it, $BF_{Inclusion} = 684.22$. The remaining effects were not significant, Fs(1, 137) < 1,

$ps \geq .56$, with moderate evidence supporting models excluding each effect of interest, $BF_{Inclusion} \leq 0.20$.[7]

Raven

This study operationalized gF as performance on the Raven test. On average, participants answered slightly over half of the items correctly (*M* = 9.85, *SD* = 2.73). Performance was numerically higher for the 2-day group (*M* = 10.14, *SD* = 3.09) compared to the 1-day group (*M* = 9.58, *SD* = 2.33), Welch's $t(128.13) = -1.23$, $p = .22$. As the retention interval had no effect on memory or gF measures, subsequent analyses pooled the two groups together. For analytical purposes, the number of correct answers in Raven was transformed into a *z*-score for the overall sample. Frequencies for the Raven test are displayed in Table 3.2.

Minear et al.'s (2018) Quartile Analyses

In the following sections, we outline our efforts to extend Minear et al.'s (2018) findings. Initially, we categorized our participants into positive, negative, and null testers. Subsequently, we compared positive and nonpositive testers across different measures. We then describe our *post hoc* subgrouping of positive testers into low and high gF groups, as well as the categorization of word pairs into easy and difficult items. Finally, we present our results following Minear et al.'s quartile analyses.

[7] Two participants answered all trials incorrectly in the rereading condition (please note the two red dots at the bottom of Figure 3.3). These same participants answered all items correctly in the retrieval practice condition. It is possible, therefore, that a lack of understanding of the task is not the reason for the low performance in the rereading condition. Nevertheless, we reran the ANOVAs reported in this section, excluding these two cases. The conclusions remained the same in these sensitivity analyses.

Table 3.2

Frequencies for the Raven Test

Score	*Z*-score	%
1	–3.24	1.39
4	–2.14	0.69
5	–1.78	4.17
6	–1.41	7.64
7	–1.05	2.78
8	–0.68	11.11
9	–0.31	15.28
10	0.05	14.58
11	0.42	11.81
12	0.79	14.58
13	1.15	9.03
14	1.52	3.47
15	1.88	3.47

Positive and Nonpositive Testers

Following Minear et al. (2018), we categorized participants as positive testers ($n = 115$), negative testers ($n = 14$), and null testers ($n = 15$). However, due to the small group sizes, we combined the latter two groups into a single category, *nonpositive testers* ($n = 29$). Subsequently, we compared positive and nonpositive testers based on their performance during the practice phase, final-test phase, and Raven scores. These findings are summarized in Table 3.3. There were

nonsignificant differences between positive and nonpositive testers in rereading performance during the practice phase, overall performance in the final cued-recall test, and Raven scores. However, differences emerged in retrieval practice performance during the practice phase and final cued-recall test performance after both rereading and retrieval practice.

Table 3.3

Means (SDs) of Practice, Final-Test and Raven Performance for Positive and Nonpositive Testers

Measure	Positive testers	Nonpositive testers	*t*	*p*	*d*	Minear et al. (2018)
Practice, rereading	.96 (.06)	.97 (.04)	–1.22	.23	–0.20	—
Practice, retrieval practice	**.38 (.17)**	**.31 (.17)**	**2.11**	**.04**	**0.44**	✗
Final test, rereading	**.38 (.22)**	**.55 (.24)**	**–3.26**	**.002**	**–0.72**	✓
Final test, retrieval practice	**.64 (.23)**	**.48 (.24)**	**3.08**	**.004**	**0.66**	✓
Final test, overall	.51 (.22)	.52 (.24)	–0.11	.92	–0.02	✗
z-Raven	–0.01 (1.02)	0.05 (0.95)	–0.33	.74	–0.07	✓

Note. The Minear et al. (2018) column indicates whether our result, in each row, agrees (✓) or disagrees (✗) with Minear et al.'s result. Significant differences are shown in bold.

Post Hoc Subgrouping

Although the caveats against *post hoc* subgrouping have been long recognized in the methodology literature (cf. Preacher et al., 2005), our objective was to adhere as closely as possible to Minear et al.'s (2018) analytical procedures. In line with this goal, among positive testers, we used quartile analyses similar to previous studies (Brewer & Unsworth, 2012; Minear et al., 2018): Participants in the first (low gF group; $n = 33$) and in the fourth (high gF group; $n = 18$) quartiles were selected based on their Raven scores. Subsequently, based on the final cued-recall test, items were median-split into easy ($M_{recall} = .63, SD = .12$) and difficult ($M_{recall} = .40, SD = .06$) ones. Mean recall represents the average proportion of participants who recalled items on the final-cued recall test, regardless of item assignment to rereading or retrieval practice. Notably, values from the present item difficulty definition exhibited a strong correlation with those based on Lima and Buratto's (2021) norms, r = .90, 95% CI [.81, .95], reassuring the reliability of the current definition.

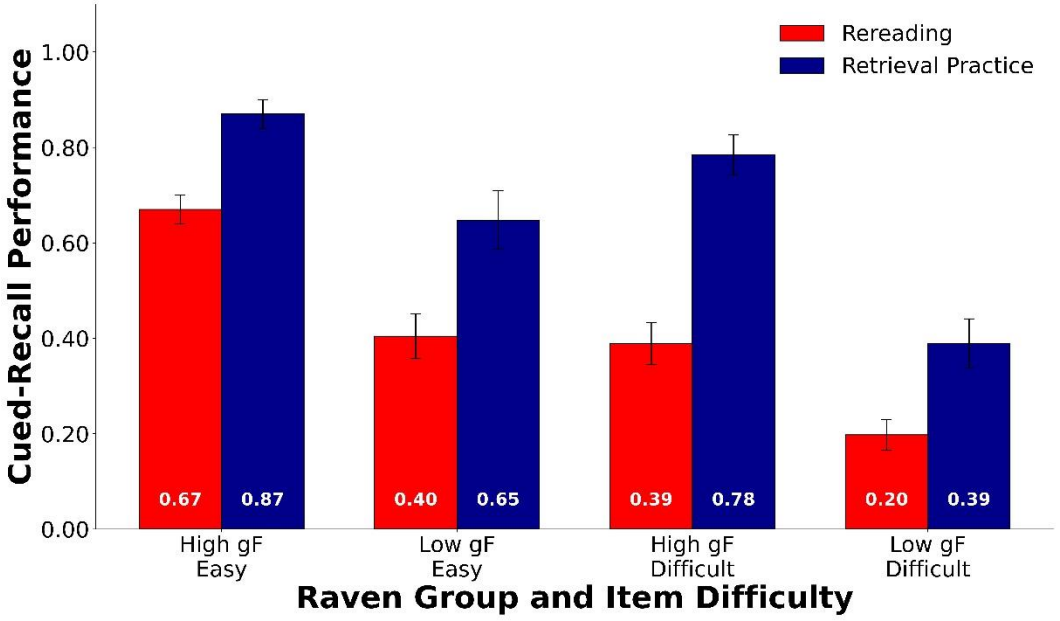
Retrieval Practice Effect, gF, and Item Difficulty

The first quartile analysis focused on the final cued-recall test and involved a 2 (learning strategy) × 2 (Raven group) × 2 (difficulty) mixed ANOVA, with the Raven group as the between-subjects factor. Unsurprisingly, there was moderate-to-extreme evidence for the three main effects: learning strategy, $F(1, 49) = 186.49$, $p < .001$, $\eta_p^2 = .79$, $BF_{Inclusion} = 6.25$; Raven group, $F(1, 49) = 23.59$, $p < .001$, $\eta_p^2 = .33$, $BF_{Inclusion} = 51.57$; and difficulty, $F(1, 49) = 104.66$, $p < .001$, $\eta_p^2 = .68$, $BF_{Inclusion} = 4.36 \times 10^{13}$. As illustrated in Figure 3.4, these main effects indicate superior performance (a) following retrieval practice (i.e., a group-level retrieval practice effect), (b) for the high gF group, and (c) for easy items. Importantly, these results were qualified by moderate evidence for a Learning Strategy × Raven Group × Difficulty interaction, $F(1, 49) =$

10.79, $p = .002$, $\eta_p^2 = .18$, $BF_{Inclusion} = 8.83$. For the low gF group, we observed a numerically—although not significantly—greater retrieval practice effect for easy items ($M_{Difference} = .24$, $SD = .21$) compared to difficult ones ($M_{Difference} = .19$, $SD = .17$), paired $t < 1$, $p = .33$, $d = 0.17$. According to a Bayesian criterion, this provided anecdotal evidence against an effect, $BF_{10} = 0.29$. Conversely, for the high gF group, we observed a greater retrieval practice effect for difficult items ($M_{Difference} = .40$, $SD = .17$) compared to easy ones ($M_{Difference} = .20$, $SD = .14$), paired $t(17) = -5.29$, $p < .001$, $d = -1.25$, providing extreme evidence for an effect, $BF_{10} = 446.19$.

Figure 3.4

Final Cued-Recall Test Performance as a Function of Learning Strategy, Raven Group, and Item Difficulty



Note. Means are presented within each corresponding bar. Error bars represent ±1 standard error of the mean.

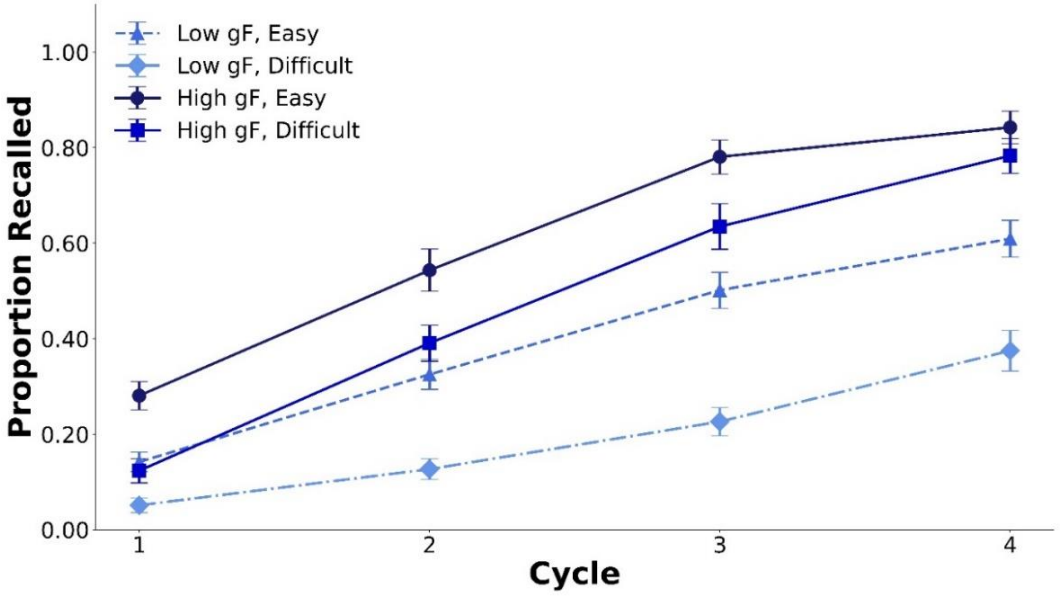
Performance During the Practice Phase, Raven, and Item Difficulty

The second quartile analysis was restricted to retrieval practice trials in the practice phase (Session 1), and encompassed a 2 (Raven group) × 2 (difficulty) × 4 (cycle) mixed ANOVA, with the Raven group as the between-subjects factor. During the practice phase, there was extreme evidence for the three main effects: Raven group, $F(1, 49) = 45.51$, $p < .001$, $\eta_p^2 = .48$, $BF_{Inclusion} = 5.24 \times 10^5$; difficulty, $F(1, 49) = 98.18$, $p < .001$, $\eta_p^2 = .67$, $BF_{Inclusion} = 3.69 \times 10^{11}$; and cycle, $F(2.33, 113.94) = 270.30$, $p < .001$, $\eta_p^2 = .85$, $BF_{Inclusion} = 4.94 \times 10^{48}$. As depicted in Figure 3.5, these main effects indicate (a) superior performance for the high gF group, (b) enhanced performance for easy items, and (c) increased performance across practice cycles. Importantly, these findings were qualified by extreme evidence for a Raven Group × Difficulty × Cycle interaction, $F(3, 147) = 7.53$, $p < .001$, $\eta_p^2 = .13$, $BF_{Inclusion} = 413.18$.

Two 2 (Raven group) × 4 (cycle) mixed ANOVAs probed the three-way interaction. As illustrated in Figure 3.5, the performance advantage for the high gF group over the low gF group was greater for difficult items, $F(1, 49) = 52.98$, $p < .001$, $\eta_p^2 = .52$, $BF_{Inclusion} = 3.87 \times 10^6$, than for easy items, $F(3, 49) = 26.40$, $p < .001$, $\eta_p^2 = .35$, $BF_{Inclusion} = 3.29 \times 10^3$. Notably, the cycle effect sizes were nearly identical for both the high gF group, $F(3, 147) = 180.65$, $p < .001$, $\eta_p^2 = .79$, $BF_{Inclusion} = 1.87 \times 10^{34}$, and the low gF group, $F(3, 147) = 178.10$, $p < .001$, $\eta_p^2 = .78$, $BF_{Inclusion} = 1.11 \times 10^{46}$, indicating that, collapsing for item difficulty, their improvement was consistent across cycles—although they did not start from similar points. Most important, for difficult items, the performance advantage of the high gF group over the low gF group substantially increased across cycles, $F(3, 147) = 25.32$, $p < .001$, $\eta_p^2 = .34$, $BF_{Inclusion} = 9.53 \times 10^9$, whereas this advantage increment over cycles was anecdotal for easy items, $F(3, 147) = 2.87$, $p = .04$, $\eta_p^2 = .06$, $BF_{Inclusion} = 1.10$.

Figure 3.5

Practice Phase Proportion Correct in the Retrieval Practice Trials as a Function of Raven Group, Item Difficulty, and Cycle



Note. Error bars represent ±1 standard error of the mean.

Novel Analyses

Learning During Retrieval Practice Trials

During the practice phase, retrieval practice trials lasted 7 s and were followed by a 2-s feedback. This setup allowed participants to view the correct response, raising the possibility that this feedback influenced participants in correctly recalling new items in subsequent cycles. Do participants' Raven scores relate to the amount of new items they recall in each cycle? To address this question, we employed an analysis similar to one previously conducted by Arnold and McDermott (2013a). They explored whether the organization of free recall, measured using the adjusted-ratio-of-clustering scores (Roenker et al., 1971), in one test correlates with the proportion of new items recalled in a subsequent test—indicating learning during the interim rereading block

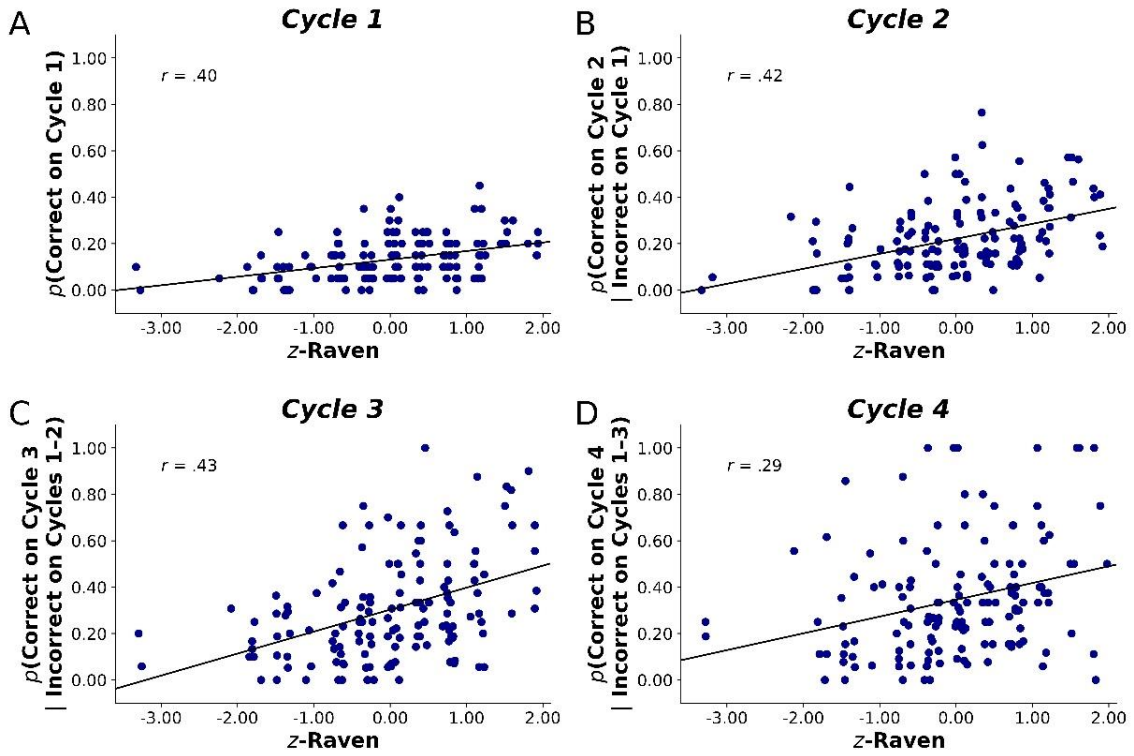
between the two tests. Here, adjusted-ratio-of-clustering scores were replaced with Raven scores, and these scores were correlated with the proportion of new items recalled in each retrieval practice cycle—specifically, focusing on items recalled in a cycle that were not recalled in any previous retrieval practice trials.

Figure 3.6 illustrates the results of these analyses. While the Cycle 1 indexes the quality of encoding during the study phase, the subsequent cycles partially capture post-retrieval re-encoding effects of feedback. As depicted in Figure 3.6, the correlations between Raven scores and the proportion of items recalled across cycles ranged from .29 to .43. This suggests that participants with higher gF not only recalled more items across cycles (as implied by the quartile analysis in the previous section; see Figure 3.5), but also continued to learn more new items compared to participants with lower gF.

These correlation analyses, however, present an artifact, namely, that the observed correlations in Cycles 2–4 might stem from a smaller pool of items not yet recalled in any previous retrieval practice trials. In particular, holding constant the number of items recalled in a given cycle, participants with a smaller pool of items not yet recalled will have a higher proportion of recall in that cycle. Given that the proportion of items recalled in Cycle 1 correlated with Raven scores, it is possible that this effect might have "spilled over" into subsequent cycles, providing an unfair advantage to participants with higher gF. To explore this possibility, we conducted analyses akin to the previous ones but considered the absolute number of items recalled in a cycle.

Figure 3.6

Proportion of Items Recalled in a Cycle, Considering That They Were Not Recalled in Any Previous Retrieval Practice Trials, as a Function of Raven Scores

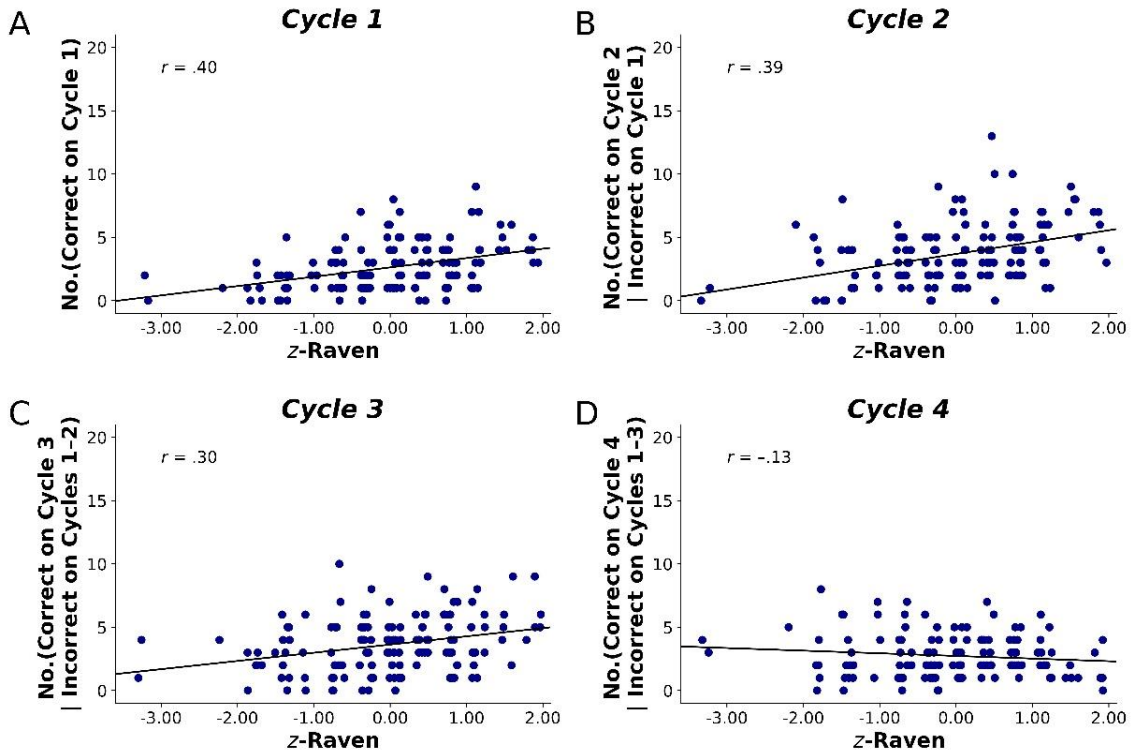


Note. Best-fit line and Pearson's *r* are shown in each panel. Datapoints were jittered to improve visualization.

As depicted in Figure 3.7, the correlations between Raven scores and the number of items recalled across Cycles 1–3 ranged from .30 to .40, providing converging evidence for the previous analyses. However, the correlation in Cycle 4 was numerically negative and nonsignificant. We will revisit these seemingly contradictory findings in the Discussion section.

Figure 3.7

Number of Items Recalled in a Cycle, Considering That They Were Not Recalled in Any Previous Retrieval Practice Trials, as a Function of Raven Scores



Note. Best-fit line and Pearson's *r* are shown in each panel. Datapoints were jittered to improve visualization.

In summary, both sets of analyses suggest that participants with higher gF exhibit superior initial encoding of items during the study phase, as evidenced by their performance in Cycle 1. Moreover, these participants also seem to benefit more from the 2-s feedback (at least in Cycles 1 and 2), as indicated by their performance in Cycles 2 and 3.

Mediation Analysis

To recap, our quartile analyses suggested that the influence of gF on the retrieval practice effect depends on item difficulty. Furthermore, these analyses indicated that the high gF group outperformed the low gF group across the practice cycles. A convergent result emerged in the

analyses of new item learning during the practice phase, as presented in the previous section. However, to the best of our knowledge, no prior study on individual differences has explored a potential indirect effect between gF and the retrieval practice effect mediated by performance during the practice phase. The following analysis aimed to fill this gap.

First, for each participant, we computed the mean proportion recalled across the four practice cycles (i.e., practice recall performance in retrieval practice trials). We then tested the following simple mediation model:

$$\widehat{Practice\ recall} = i_{Practice\ recall} + a \times z_{Raven}, \tag{3.1}$$

$$\widehat{Retrieval\ practice\ effect} = i_{Retrieval\ practice\ effect} + c' \times z_{Raven} + b \times Practice\ recall. \tag{3.2}$$

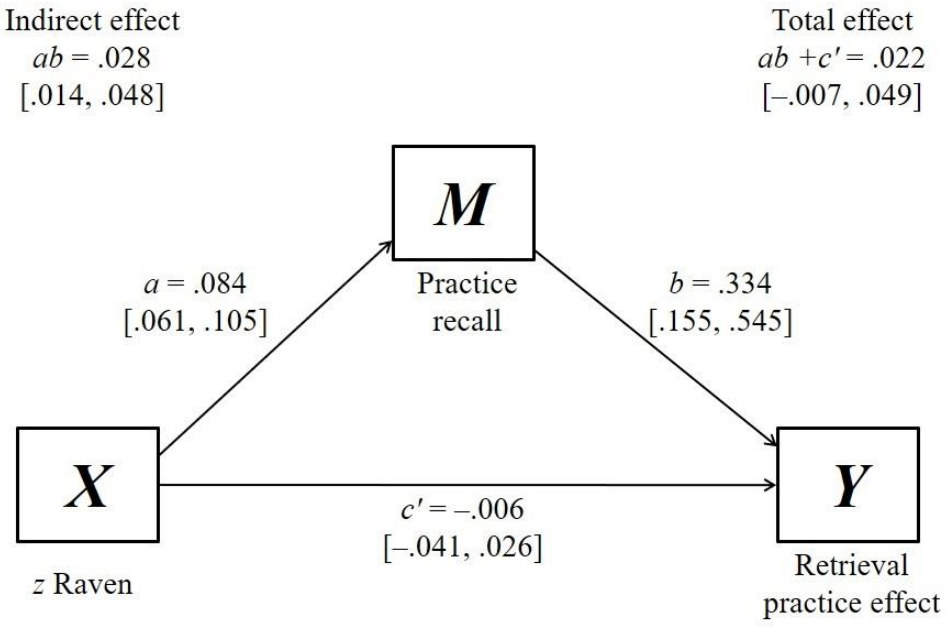
This mediation model allowed us to examine the direct effects of Raven on the practice recall performance (coefficient *a* in Equation 3.1) and on the retrieval practice effect (coefficient c' in Equation 3.2). It also allowed us to explore the direct effect of practice recall performance on the retrieval practice effect (coefficient *b* in Equation 3.2). Crucially, the mediation model quantifies the indirect effect of Raven on the retrieval practice effect mediated by performance during the practice phase (Hayes, 2022). Our inferences were based on 95% bias corrected and accelerated confidence intervals, estimated using 5,000 bootstrap samples, as the normal theory method assumes the normality of the sampling distribution of the product *ab*, even when it tends to deviates from normality (Hayes, 2009, 2022).

As illustrated in Figure 3.8, a one-unit difference on Raven (*z*-scores) corresponded to a .08 difference in the mean proportion recalled across the four practice cycles (path *a*). Moreover, controlling for Raven's effects, a one-unit difference in the mean proportion recalled across the four practice cycles corresponded to a .33 difference in the magnitude of the retrieval practice

effect (path *b*). Most notably, we identified an indirect effect of Raven on the retrieval practice effect: A one-unit difference on Raven (*z*-scores) led to a .03 difference in the magnitude of the retrieval practice effect, as a result of the effect of Raven on the mean proportion recalled across the four practice cycles, which in turn affects the magnitude of the retrieval practice effect (indirect effect *ab*). Lastly, controlling for the effects of practice recall performance, Raven yielded a nonsignificant direct effect on the retrieval practice effect.

Figure 3.8

Simple Mediation Model Results



Note. 95% bias corrected and accelerated confidence intervals, estimated using 5,000 bootstrap samples, are presented in brackets.

Discussion

Our objectives were threefold. First, we attempted to generalize the two 3-way interactions found by Minear et al. (2018) in their analyses restricted to positive testers—that is, the magnitude of the retrieval practice effect depends on gF and item difficulty, and the advantage of high over low gF participants during learning increased across cycles for difficult items, but remained constant for easy items. Overall, we successfully extended their results. Second, we explored whether gF is related to the amount of new items participants recall in each cycle during the practice phase. Consistent positive relationships were found in Cycles 1–3. Third, we examined and found an indirect effect of gF on the retrieval practice effect mediated by performance during the practice phase. The discussion is structured around these three objectives.

Minear et al.'s (2018) Quartile Analyses

We successfully replicated the group-level retrieval practice effect. Additionally, we found that, among items not recalled in the final cued-recall test, more retrieval practice items were correctly recognized in the final associative-recognition test. This finding suggests that a higher proportion of retrieval practice items were just below the threshold (in the final cued-recall test), but in a latent state close to benefiting from retrieval practice, as evidenced by their correct recognition in the final associative-recognition test, which is thought to rely on a recall-to-reject process (Malmberg, 2008). Subsequently, we attempted to extend Minear et al.'s (2018) findings.

Three of our comparisons yielded qualitatively similar results to those obtained by Minear et al. (2018): There were no differences between positive and nonpositive testers in Raven scores, positive testers exhibited an advantage in the final cued-recall test after retrieval practice, and negative testers exhibited an advantage in the final cued-recall test after rereading (see Table 3.2). Despite these similarities, differences in rereading and retrieval practice were more symmetrical

in our study than in Minear et al.'s, where the difference in the rereading condition was more than twice as large (23%) as the difference in the retrieval practice condition (9%). Notably, our positive testers outperformed nonpositive testers during the practice phase, while Minear et al. reported no differences between positive and negative testers. This suggests that the positive and negative testers in Minear et al.'s study primarily differed in retention (at least after retrieval practice), whereas our positive and nonpositive testers diverged in learning and potentially in retention as well.

In the analyses restricted to positive testers, during the final-test phase (i.e., the first quartile analysis), the high gF group exhibited a greater retrieval practice effect for difficult items compared to easy ones, while the opposite pattern was observed for the low gF group—although, in the latter case, the differences were only numerical (Figure 3.4). Nominal labels denoting easy and difficult items are only relative to each other. It is plausible, for instance, that items nominally labeled as easy and difficult correspond to different regions on the continuum of difficulty for learners with distinct ability levels.

Our results from the second quartile analysis converged with those of Minear et al. (2018): For difficult items, the advantage of the high gF group over the low gF group increased across cycles, but this increment was only anecdotal for easy items (Figure 3.5). Perhaps the main discrepancy in results between studies lies more at the descriptive level. In our study, the high gF group demonstrated enhanced learning of difficult items compared to the low gF group with easy items; in Minear et al.'s study, these performances seemed indistinguishable.

Lima and Buratto (2023b) recently provided evidence for the test–retest reliability of the retrieval practice effect (30 word pairs, six practice cycles, and a 5-min retention interval). However, they observed relatively low estimates (intraclass correlation coefficients between .33.

and .35). These low reliabilities could imply a high degree of misclassification of positive and negative—or nonpositive—testers. In other words, a participant classified as a positive tester at one point could be classified as a negative tester at another. In light of this low reliabilities, the generalization of Minear et al.'s (2018) quartile analyses is important, especially given the recent emphasis on replication (LeBel et al., 2019; Pashler & Wagenmakers, 2012). One plausible explanation for the similarity between our study and Minear et al.'s could be the potentially higher reliability of the retrieval practice effect in procedures involving 40 word pairs, four practice cycles, and 1- or 2-day retention intervals—although this hypothesis requires further exploration in future studies.

Relationship Between gF and Post-Retrieval Re-Encoding Effects of Feedback

Participants with higher Raven scores exhibited superior initial encoding for retrieval practice items in the study phase (as evident in performance on Cycle 1), and benefited more from the 2-s feedback in the practice phase, at least in Cycles 1 and 2 (as evident in performances on Cycles 2 and 3). Why would gF be linked to improved encoding and the post-retrieval re-encoding effects of feedback? The correlation observed in Cycle 1 seems to align with the literature demonstrating correlations between latent variables of long-term memory and gF (for a review, see Unsworth, 2019). It is possible that participants with higher gF employ more effective encoding strategies. While Minear et al. (2018) found that high gF participants were more likely to report using deep strategies and the keyword strategy, Robey (2019) observed weak correlations between gF measures and two self-reported strategy use measures (*r*s between –.08 and .10, across Experiments 1 and 2). However, these two studies differed in how they assessed self-reported strategy use—global assessment in Minear et al. (2018) versus item-by-item assessment in Robey

(2019). More research linking cognitive abilities, self-reported strategy use, and retrieval practice is needed.

Regarding the correlations observed in Cycles 2 and 3, two theoretical accounts can help interpret these findings. The mediator-effectiveness hypothesis posits that retrieval practice, compared to rereading, supports learners for using more effective mediators during encoding (Pyc & Rawson, 2010). For instance, the mediator *salami* is said to be effective if it is retrievable (due to its phonological similarity to the cue *sahani*) and decodifiable (due to its semantic relationship to the target *plate*). The mediator-shift hypothesis posits that retrieval failures allow learners to switch from less-effective to more-effective mediators (Pyc & Rawson, 2012). For instance, if the mediator *wind* fails to help associate *wingu* with *cloud*, feedback after a retrieval failure might enable the learner to replace *wind* with *wing* as a mediator.

The key idea here is that learners with higher gF might be especially skilled at generating effective mediators and at monitoring and replacing less-effective ones after retrieval failures. Minear et al. (2018) found that the high gF group were four times more likely (28%) to report using the keyword method than the low gF group (7%). However, our study did not explicitly measure the production, shift, and retrieval of mediators. An interesting avenue of research involves assessing whether higher gF learners are more likely to retrieve mediators and shift mediators after retrieval failures (Pyc & Rawson, 2010, 2012; but see Karpicke & Smith, 2012; Lehman & Karpicke, 2016, for criticisms and evidence against mediator-based accounts).

In Cycle 4, Raven scores positively correlated with the *proportion* of new items recalled (r = .29) but negatively correlated with the *number* of new items recalled (r = –.13). It is well-known that discrepant results can arise depending on the analytical tool applied to the same data (e.g., Carpenter et al., 2008; Silberzahn et al., 2018). In our case, we used different definitions of the

amount of new items recalled. The proportion measure ranges from 0 to 1 across participants and cycles, whereas the absolute measure ranges from 0 to the total number of items not yet recalled, which itself varies across participants and cycles. Consequently, the proportion measure biases correlations upward, while the absolute measure biases them downward. (Note that the datapoints are distributed along almost the entire range of the *y*-axis in Cycles 3–4 of Figure 3.6, but they have a restricted range in Cycles 3–4 of Figure 3.7.) In Cycle 4, for instance, two participants might recall the same number of items, yet the different denominators in the proportion measure cause their scores to differ. This bias interpretation is supported by an analysis showing negative correlations between gF and the number of items yet not recalled on Cycles 2, 3, and 4 (*r*s = –.40, –.47, and –.50, respectively). In other words, participants with higher gF tend to have fewer items not yet recalled on Cycle 4, which results in a small denominator and a high proportion.

So, which measure should we trust? We take the conservative position that measures leading to converging results are trustworthy (i.e., in Cycles 2 and 3). When results diverge, as in Cycle 4, it is essential to explore the source of the conflicting results. The key lesson is that conditional analyses, while informative, can present challenges in situations where there is greater heterogeneity in the pool of items across participants. We anticipate this could occur in latter cycles of the practice phase or even in initial cycles when the materials are easier than ours. Researchers should consider these factors when interpreting such analyses.

The Indirect Effect of gF on the Retrieval Practice Effect

In studies on individual differences employing duration-based procedures (Pyc & Rawson, 2009), the presence of variability in the practice phase entails different levels of learning across participants. Our study revealed that the Raven (*z*-scores) predicted the mean proportion recalled

across the four practice cycles. In simpler terms, this means that gF seems to play a significant role in differences observed in learning in duration-based procedures.

Minear et al. (2018) had similar findings in their quartile analyses, but their results were restricted to positive testers, whereas our result was the first to include the entire sample. We acknowledge that it is not entirely clear why the best analytical approach would be breaking down the data two consecutive times—first by positive and negative (or nonpositive) testers; and then by low and high gF groups—rather than using the original continuous measures. Methodologists have long been discouraged *post hoc* subgrouping due to reduced statistical power and potential impact on interaction tests, especially when removing midrange values (cf. Preacher et al., 2005). Despite our initial skepticism about this approach, we followed Minear et al.'s methods closely in this chapter, only deviating from them in our novel analyses, which used the full sample and avoided dichotomizations. Despite these concerns, we find it reassuring that most of Minear et al.'s findings were generalized even with a much smaller sample in our study.

Taking the statistical diagram in the Figure 3.8 as a reference, prior studies on individual differences has primarily explored path *c* (*z* Raven → Retrieval practice effect; Brewer & Unsworth, 2012; Robey, 2019), while certain experimental studies have presented correlational analyses involving the path *b* (Practice recall → Retrieval practice effect; Ariel & Karpicke, 2018; Finley et al., 2011; Lima, Venâncio, et al., 2020).[8] Minear et al. (2018) was the sole study, to our knowledge, that examined paths *a* (*z* Raven → Practice recall) and *c* separately. Our study related an individual-difference variable with both performances in the practice and in the final-test phases

[8] Figure 3.8 depicts path *c*′ instead of path *c*. The main difference between them is that path *c*′ takes into account the presence of other predictor variables in the model, whereas path *c* does not include other predictor variables, akin to a simple regression model (or to the bivariate correlation, more frequently presented in studies on individual differences), taking the individual-difference variable as the predictor variable and the retrieval practice effect as the criterion variable. Our notation is consistent with Hayes (2022). We will turn to path *c* itself in Chapter 4.

simultaneously (i.e., considering all paths in Figure 3.8). This approach allowed us to demonstrate that gF might influence the retrieval practice effect indirectly through practice phase performance. Importantly, once practice phase performance was controlled for, gF itself did not directly impact the retrieval practice effect. Further studies on this topic are needed, replacing variables in each path or even exploring more complex mediation models.

In the previous section, we suggested that learners with higher gF are especially skilled at generating effective mediators and at monitoring and replacing less-effective mediators after retrieval failures. This notion gains significance, especially in studies with limited practice cycles. However, some studies suggest that participants may shift from mediated retrieval to direct access after extended (≥ 10 cycles) retrieval practice (Crutcher & Ericsson, 2000; Dikmans et al., 2020; Kole & Healy, 2013, Experiment 2). If this holds true, the ability to generate and monitor the effectiveness of mediators could become less relevant after extended retrieval practice. Future studies could explore this potential scenario. Additionally, just as the magnitude of the retrieval practice effect between low and high gF groups seems to depend on item difficulty (Minear et al., 2018), it might also be influenced by the number of practice cycles. For example, items practiced more frequently are strengthened more than less frequently practiced items, although the benefits diminish over time (Pyc & Rawson, 2009; Vaughn & Rawson, 2011). A conceptual replication of Minear et al.'s work could involve manipulating the number of practice cycles to induce differential item strengthening instead of varying item difficulty.

Finally, thus far, studies on individual differences have solely relied on duration-based procedures (Pyc & Rawson, 2009). These procedures can be seen as a one-size-fits-all approach, assuming that that the experimenter's chosen dosage of retrieval practice is suitable for all types of learners. Again, variability in performance during the practice phase suggest that this

assumption is unrealistic. The scenario could be different if criterion-based procedures were employed. In these procedures, additional practice manipulation occurs after each item has been successfully recalled a predetermined number of times (Karpicke & Roediger, 2008; Vaughn & Rawson, 2011). For instance, participants repeatedly cycle through rereading–retrieval-practice blocks, but, after a successful retrieval, an item can be dropped from additional rereading blocks, dropped from additional retrieval practice blocks, dropped from both blocks, or not dropped at all (Friedman et al., 2017; Karpicke & Roediger, 2007, 2008; Soderstrom et al., 2016). It is important to note that the total exposure time is not equated across learners in this approach; rather, the learning criterion is what is equated across learners (Karpicke & Roediger, 2008). In fact, less efficient learners are likely to make more attempts to reach the criterion than more efficient learners (Zerr et al., 2018). Criterion-based procedures, by equating performance across participants, render this variable no longer a potential mediator variable. In this scenario, could gF have direct and indirect effects on the magnitude of the retrieval practice effect?

Concluding Comments

In this study, we observed an indirect effect—but not a direct—effect of gF on the retrieval practice effect. However, due to the heterogeneity of procedures used in the extant literature, it is not possible to speak of a single effect. Therefore, we advocate for new studies using different procedures. Here, we recommend three possibilities: measuring the production, shift, and retrieval of mediators; manipulating the number of retrieval practice opportunities and other retention intervals; and adopting criterion-based procedures. Pursuing this research agenda has the potential to sharpen our understanding of the conditions and cognitive mechanisms underlying the relationship between individual-difference variables and the benefits of retrieval practice.

Chapter 4 – Testing the Dual-Memory Framework: Individual Differences in the Magnitude of the Retrieval Practice Effect and Fluid Intelligence

Manuscript 3

Abstract

Retrieving information from memory enhances long-term retention. However, our understanding of the moderating role of individual differences in the retrieval practice effect is still in its early stages. A complicating factor is that contemporary accounts of the retrieval practice effect typically do not address potential individual differences moderating the effect, at least in their original formulations. In this study, we explore the dual-memory framework (Rickard & Pan, 2018) as a promising candidate to instantiate specific quantitative models in the study of individual differences of the retrieval practice effect. After outlining the framework, we describe our approach to simulate various scenarios and make empirical predictions. We derive two simple models from the dual-memory framework, namely, the fixed-threshold model and the random-threshold model. The random-threshold model yielded a point estimate closer to the empirical value we obtained than the fixed-threshold model, although the empirical confidence interval overlapped with estimates from both models. Our discussion focuses on differences between the conceptual description and the mathematical implementation of the dual-memory framework, potential connections between the dual-memory framework and the bifurcation-based framework, and the goodness of fit at both the distribution and participant levels.

Keywords: retrieval practice, testing effect, test-enhanced learning, memory, quantitative model

Testing the Dual-Memory Framework: Individual Differences in the Magnitude of the Retrieval Practice Effect and Fluid Intelligence

Retrieving information from memory enhances long-term retention (Karpicke & Roediger, 2008; Roediger & Butler, 2011). In an influential review assessing the utility of 10 learning strategies, Dunlosky et al. (2013, p. 32) concluded that retrieval practice "may benefit individuals with varying levels of knowledge or ability, but the extent to which the magnitude of the benefit depends on these factors remains an open question." A decade later, our understanding on the moderating role of individual-difference variables on the retrieval practice effect is still in its early stages. A complicating factor is that contemporary accounts of the retrieval practice effect, at least in their original formulations, typically do not address learners' characteristics as potential moderators of the effect (for a review, see Karpicke, 2017).[9] In this chapter, our main objective is to introduce and explore the dual-memory framework (Rickard & Pan, 2018) as a promising candidate to instantiate specific quantitative models capable of predicting different relationships between individual-difference variables and the magnitude of the retrieval practice effect.

The Dual-Memory Framework

So far, contemporary accounts of the retrieval practice effect have been described solely in verbal terms (Carpenter, 2009; Lehman et al., 2014; Pyc & Rawson, 2009, 2010, 2012; for a notable exception, see Mozer et al., 2004). This stands in contrast to the views of several scholars who advocate for the advantages of mathematical theorizing, such as reducing ambiguity,

[9] However, contemporary hypotheses might be capable of accommodating evidence of the moderating role of individual differences. For instance, considering the elaborative retrieval hypothesis (Carpenter, 2009), Minear et al. (2018, p. 1476) argue: "One might speculate that individuals higher in crystallized intelligence (e.g., vocabulary knowledge) would have more elaborate semantic networks and this would be most evident for the more difficult items, yielding a larger [retrieval practice] effect on difficult items for individuals high in this measure than those scoring low." In the same vein, Buchin and Mulligan (2023) claimed that the elaborative retrieval hypothesis predicts a greater retrieval practice effect for high-prior knowledge information than for low-prior knowledge information. One can argue that the same reasoning applies when prior knowledge is an individual-difference, instead of an experimentally manipulated, variable.

generating more specific predictions, and clarifying theoretical assumptions (Bjork, 1973; Farrell & Lewandowsky, 2018; Franco & Iglesias, 2023).

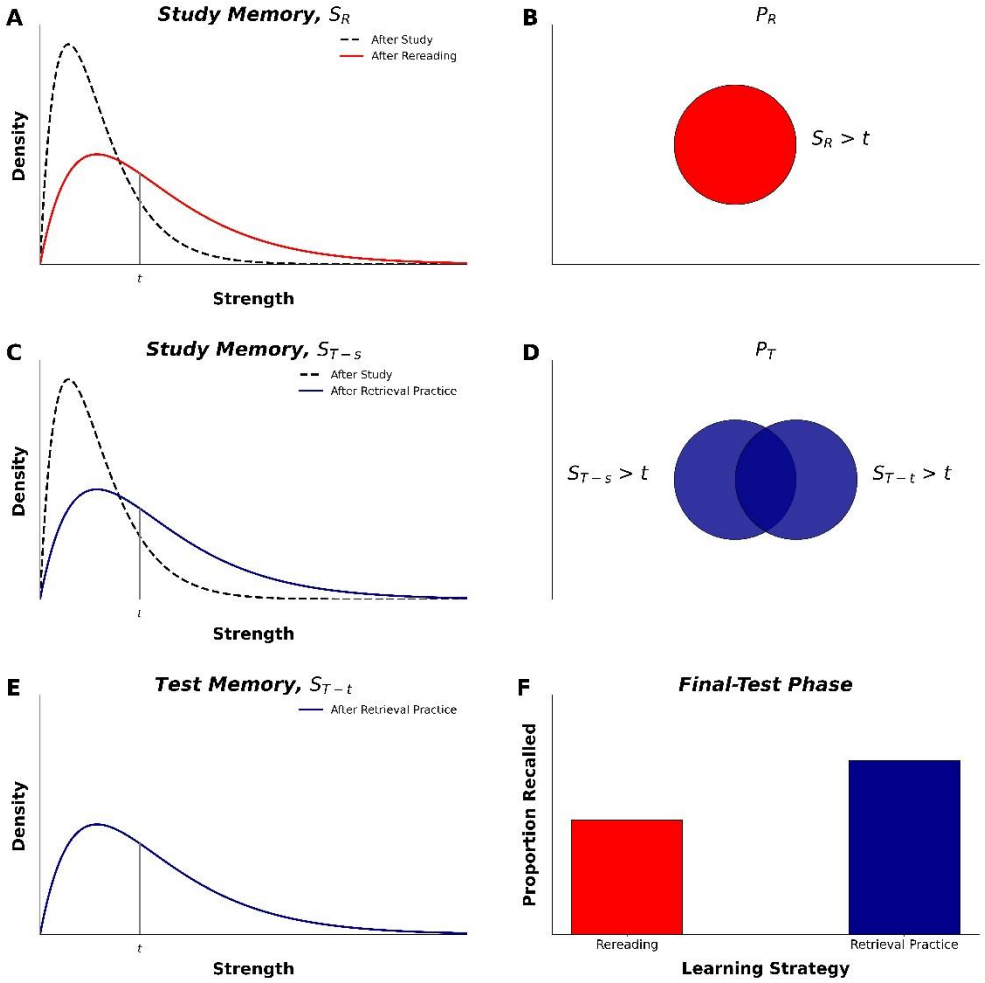
The dual-memory framework (Rickard & Pan, 2018) is a descriptive account that relies on the idea of strength of memory traces. Strength based-accounts have a long history in memory research (Wixted, 2007; Yonelinas, 1994), with at least one predecessor in the retrieval practice literature, the bifurcation-based framework (Halamish & Bjork, 2011; Kornell et al., 2011). The conceptual description of the dual-memory framework includes the following claims: (a) initial study of items leads to the encoding of a study memory; (b) rereading and retrieval practice (plus feedback) lead to increments in the strength of study memory; (c) retrieval practice (plus feedback) uniquely leads to the encoding of a new test memory; (d) the final-test performance after rereading is supported by the study memory trace, whereas final-test performance after retrieval practice is supported by two distinct and independent memory traces, study memory and test memory.

Figure 4.1 illustrates the dual-memory framework. We first consider the case for an ideal participant, with an infinite number of items randomly assigned either to rereading or to retrieval practice. Let S_R represent the memory strength of an item assigned to the rereading condition, and S_{T-s} represent the memory strength of an item assigned to the retrieval practice condition. After the study phase, these two random variables are independent and identically distributed (see the dotted black lines in Figure 4.1, panels A and C). During the practice phase, the ideal participant is exposed to rereading and retrieval practice trials, boosting these two memory strengths to a similar degree (Figure 4.1, see the solid red and blue lines in panels A and C, respectively). Importantly, the framework posits that a retrieval practice trial constitutes an event distinct enough to encode a new test memory trace (Rickard & Pan, 2018). Let S_{T-t} represent the strength of this new memory test. The framework assumes that S_R, S_{T-s}, and S_{T-t} are independent and identically

distributed after the practice phase (in Figure 4.1, see the solid red line in panel A, and the two solid blue lines in panels C and E).

Figure 4.1

Summary of the Dual-Memory Framework



Note. Panels A, C, and E depict the strength distributions under the dual-memory framework for study memory (rereading), study memory (retrieval practice), and test memory (retrieval practice), respectively. For didactic purposes, we follow Rickard and Pan (2018) and model these random variables using gamma distributions. Panels B and D depict the proportion of items with memory strength above the response threshold (t) for rereading and retrieval practice, respectively. Panel F depicts the predicted proportion recalled in a final-test phase after rereading and retrieval practice.

As illustrated in Figure 4.1, panel B, the dual-memory framework posits that successful recall in a final test for a randomly chosen item in the rereading condition depends on items with memory strength surpassing the threshold t, $P_R = (S_R > t)$. In the retrieval practice condition, successful recall could be supported either by study memory, $P_{T-s} = (S_{T-s} > t)$, or by test memory, $P_{T-t} = (S_{T-t} > t)$, as depicted in Figure 4.1, panel D. Assuming independence between P_{T-s} and P_{T-t}, the union rule for independent events results in the following probability of successful recall in the final test for a randomly chosen item (Ross, 2007):

$$P_T = P_{T-s} + P_{T-t} - P_{T-s} \times P_{T-t}. \tag{4.1}$$

Under the assumption of identical and independent distributions, Equation 4.1 can be reexpressed in terms of P_R:

$$P_T = 2P_R - P_R^2. \tag{4.2}$$

In summary, the dual-memory framework predicts recall in the retrieval practice condition as a quadratic function of the proportion recalled in rereading. In research on individual differences, the retrieval practice effect at the participant level is often quantified as a difference score (Agarwal et al., 2017; Pan et al., 2015). Hence, the framework predicts the magnitude of the retrieval practice effect solely based on rereading recall probability on the final test:

$$\begin{aligned} TE &= P_T - P_R, \\ TE &= P_R - P_R^2. \end{aligned} \tag{4.3}$$

This quadratic function suggests larger retrieval practice effects when $P_R = .50$; decreasing as P_R approaches 0 or 1. In real-world scenarios with a finite item count, Equations 4.2 and 4.3, respectively, are adjusted as follows:

$$\widehat{PC_T} = 2PC_R - PC_R^2, \tag{4.4}$$

$$\widehat{TE} = PC_R - PC_R^2. \tag{4.5}$$

Here, PC_R is the observed proportion correct in the rereading condition, $\widehat{PC_T}$ is the predicted proportion correct in the retrieval practice condition, and $\widehat{TE}$ is the predicted retrieval practice effect (Rickard & Pan, 2018).

Two Models Based on the Dual-Memory Framework

The dual-memory framework, while not directly addressing individual differences, can be applied for this purpose through simulations. Firstly, for each participant, a value for PC_R may be sampled. Secondly, $\widehat{PC_T}$ and $\widehat{TE}$ may be estimated with Equations 4.4 and 4.5. This process may be repeated *N* times—*N* being the intended sample size in the simulation. Thirdly, an individual-difference variable may be simulated with a desired correlation with PC_R. For example, an algorithm can be used to simulate 144 participants recalling an average of .42 in the rereading condition, and a correlation between rereading performance and Raven scores of $r = .39$, as found in the data from Chapter 3. Finally, the correlation between $\widehat{TE}$ and Raven scores can be tested against the empirical correlation we obtained, namely, $r = .12$. This was the approach we used in this study. Appendix E provides an overview of this approach, indicating how it can be used in future studies.

Our simulations considered $P_R = .42$, $r_{PC_R,Raven} = .39$, and $N = 144$. Each iteration represented one participant. In an iteration, one Raven score was sampled from a normal distribution, $Raven \sim Normal(0, 1)$. Additionally, 20 values for memory strength for rereading items in the final test were sampled from a normal distribution, $S_R \sim Normal(0, 1)$.[10] The same procedure was independently followed for S_{T-s} and S_{T-t}, but since they were not relevant for the

[10] Rickard and Pan (2018) used gamma distributions throughout their model description, which are psychologically more plausible than normal distributions. In gamma distributions, memory strengths are constrained to be always positive, and the distributions are allowed to exhibit positive skewness. However, the authors argue that predictions do not critically depend on the chosen distribution. For this reason, we deviate from the original article and the representations in Figure 4.1 here and use normal distributions. Negative values can be considered as standardized values below the distribution mean. We will revisit the distribution issue in the Final Comments section.

present purposes, it will not be mentioned further. The proportion of items in the rereading condition recalled in the final test was computed as $PC_R = p(S_R > t)$. After iterating for all participants, PC_R and Raven scores were independent. Therefore, a correlation adjustment was necessary. We achieved this by replacing the original values for Raven with adjusted values:

$$Raven_{adjusted} = .39 \times zscore(PC_R) + \sqrt{1 - .39^2} \times Raven_{original} \quad (4.6)$$

Equation 4.6 implements a *copula approach* for deriving joint distributions, given the marginal distributions (Trivedi & Zimmer, 2005). The current implementation forces Raven scores to have a correlation with PC_R approximately equal to .39. Then, $\widehat{PC_T}$ and $\widehat{TE}$ were estimated with Equations 4.4 and 4.5, respectively. Finally, $r_{\widehat{TE},Raven}$ was computed using the Pearson correlation coefficient.

We derived two models from the dual-memory framework. In the *fixed-threshold model*, the threshold t was fixed across participants, and it was defined as the critical *z*-value associated with $1 - P_R$. To ensure that the average recall in the rereading condition converged to .42, as observed in Chapter 3, z_{crit} was set to approximately 0.20. A fixed-threshold model assumes uncorrelated final-test performances for rereading and retrieval practice conditions, based on identical and independent distributions and fixed-threshold assumptions across participants. Yet, studies on individual differences report strong correlations between rereading and retrieval practice performance. For example, Robey (2019) found *r*s of .67 and .78 in her Experiments 1 and 2, respectively. For this reason, we also tested a *random-threshold model*, which relaxes the fixed-threshold assumption. The second set of simulations mirrored the first, with one exception: A participant-specific threshold, t_i, was sampled for each participant from a normal distribution centered at the critical *z*-value associated with $1 - P_R$, $t_i \sim Normal(critical\ z_{(1-P_R)}, 1)$. This allowed thresholds to vary across participants. The fixed- and the random-threshold models can

be thought of as akin to the fixed-effect and random-effects models of meta-analysis, respectively (Hedges & Vevea, 1998; Lima & Buratto, 2023a). That is, by keeping the threshold t fixed across participants in the fixed-threshold model, deviations of each participant's PC_R from P_R are purely due to random error. On the other hand, by allowing t to vary across participants, it is implicitly assumed that each participant's PC_R estimates a true, specific P_R for that participant.

To generate interval estimates for the prediction, each model simulation was repeated 100,000 times, similar to "replicating" the experiment described in Chapter 3. For each of the 100,000 simulations under each model, $r_{\widehat{TE},Raven}$ values were sorted, and the lowest 2.5% and highest 2.5% were excluded to create an empirical confidence interval.

Datasets

The dataset used for testing the fixed- and the random-threshold models was introduced in Chapter 3. However, in order to validate the appropriateness of our design, we also conducted a preliminary analysis using additional datasets collected in our lab. The main characteristics of these datasets are summarized in Table 4.1.

The test of the dual-memory models presented here is based on the final-cued recall test (as a function of the learning strategy) and participants' scores on the Raven. Before comparing the data from the experiment described in Chapter 3 with the two models, a series of analyses based on the equations of the dual-memory framework are presented. It is important to note that these analyses are based on previous reports (Gupta et al., 2022; Rickard, 2020; Rickard & Pan, 2018), and they do not offer conclusive support or refutation for either of the two models presented earlier.

Table 4.1

Laboratory Datasets

Experiment	Reference	Design Summary
1	Lima, Venâncio, et al. (2020), Experiment 1	• Sample size: 51; • Number of practice cycles: 4; • Retention interval: 48 hr.
2	Lima, Venâncio, et al. (2020), Experiment 2	• Sample size: 28; • Number of practice cycles: 4 or 6; • Retention interval: 48 hr.
3	Lima and Buratto (2023b), Session 1	• Sample size: 54; • Number of practice cycles: 6; • Retention interval: 5 min.
4	Lima and Buratto (2023b), Session 2	• Sample size: 54; • Number of practice cycles: 6; • Retention interval: 5 min.
5	Current dissertation, Chapter 3	• Sample size: 144; • Number of practice cycles: 4; • Retention interval: 1 day or 2 days.

Results

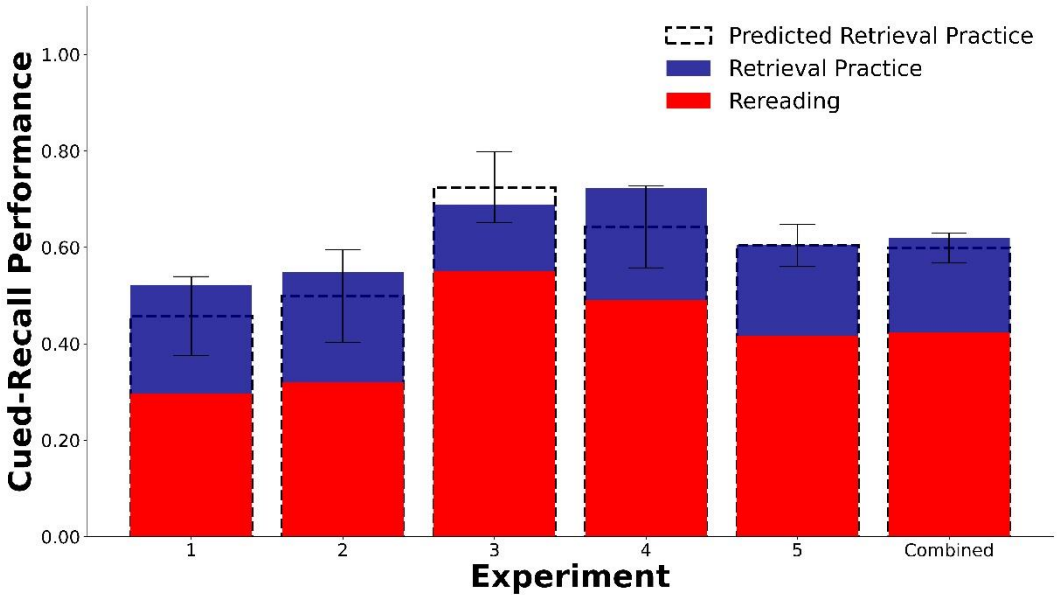
Laboratory Datasets

In line with Rickard and Pan (2018), we computed $\widehat{PC}_T$ for each participant using Equation 4.4, then we averaged these values for all participants within each experiment. As illustrated in Figure 4.2, the six 95% CIs for the equation predictions captured the observed cued-recall performance for the retrieval practice condition. This result replicates a prior analysis (Rickard &

Pan, 2018), and it suggests the adequacy of our experimental paradigms for testing predictions of models derived from the dual-memory framework.

Figure 4.2

Observed and Predicted Results for Five Datasets Collected in Our Laboratory and for All Five Datasets Combined



Note. Error bars represent 95% CIs for the Equation 4.4 predictions.

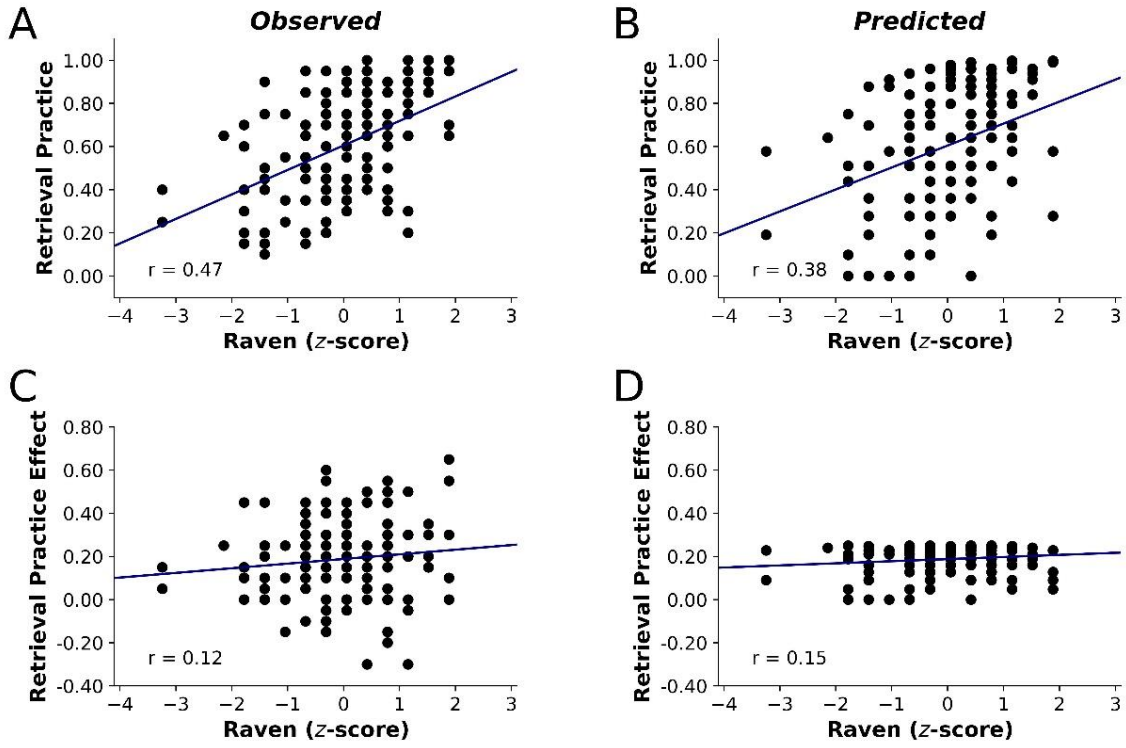
Pairwise Correlations

Figure 4.3 displays Raven's correlations with final cued-recall test performance for retrieval practice items and the retrieval practice effect. Raven scores exhibited positive correlations with both observed and predicted retrieval practice performance. The correlation between observed and predicted retrieval practice performance (not shown in Figure 4.3) was strong, r = .72 [.63, .79]. However, Raven scores exhibited almost no correlation with retrieval practice effect scores; both 95% CIs included zero, as illustrated in Figure 4.3, panels C and D.

The Equation 4.5 results in retrieval practice effect scores ranging from 0 to .25, due to the mathematical constraints of the dual-memory framework, which does not allow for negative or greater than .25 retrieval practice effects (but see Rickard & Pan, 2018, for a modeling approach of negative retrieval practice effects).

Figure 4.3

Raven Correlations with Retrieval Practice and Retrieval Practice Effect



Note. Predicted refers to values computed from Equations 4.4 (retrieval practice) and 4.5 (retrieval practice effect).

Cumulative Distribution Analysis

The cumulative distribution analysis was conducted following Rickard's (2020) procedure. In summary, the proportion recalled in rereading, the proportion recalled in retrieval practice, and the proportion recalled in retrieval practice as predicted by the Equation 4.4 were independently

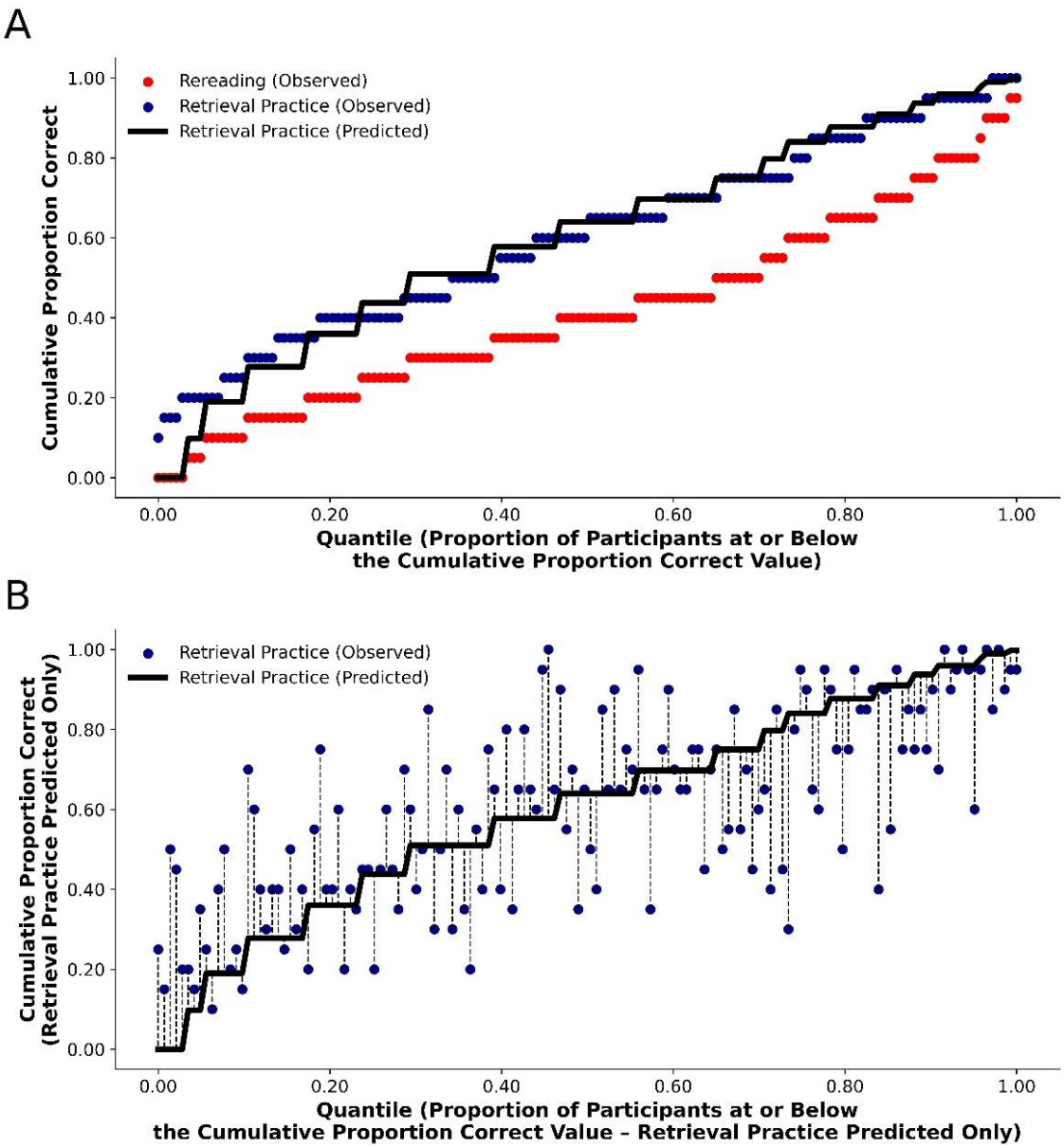
transformed into quantiles. For instance, the five participants with a rereading recall proportion of 0 received the lowest quantile values (between 0 and .028), while the two participants with a rereading recall proportion of 1 received the highest quantile values (.993 and 1).

Figure 4.4, panel A, depicts the final cued-recall test cumulative distribution results, adopting the presentation method employed by Gupta et al. (2022) and Rickard (2020). Figure 4.4, panel A, can be conceptualized as a grouped scatterplot comparing recall quantiles (*x*-axis) against cumulative proportion recalled (*y*-axis). The predictions from Equation 4.4 generally fit cumulative proportion recalled well, although there is a slight underestimation of recall in retrieval practice in the left tail of the distribution. A two-sample Kolmogorov–Smirnov test for goodness of fit indicated that the retrieval practice and the predicted retrieval practice distributions did not differ, $D(144) = 0.10$, $p = .42$.

It is essential to highlight that Figure 4.4, panel A, shows a distribution-level analysis. While Figure 4.4, panel A, aligns retrieval practice predictions based on different rereading performances at the participant level, the same alignment is not seen between observed and predicted retrieval practice values. In Figure 4.4, panel B, predicted retrieval practice is plotted against quantiles and cumulative proportion correct, while observed retrieval practice datapoints are plotted in alignment with their respective predictions for each participant. The vertical dashed lines represent the residuals. Equation 4.4 underestimates recall in the retrieval practice condition in the first tercile of the distribution ($M_{residual} = -.09, SD = .17,\ n = 48$) and overestimates it in the third tercile ($M_{residual} = .11, SD = .16,\ n = 48$); predictions tend to be, on average, more accurate in the second tercile ($M_{residual} = -.02, SD = .18, n = 48$). In summary, while the Equation 4.4 fits the data reasonably well at the distribution level, its accuracy diminishes at the participant level.

Figure 4.4

Final Cued-Recall Test Cumulative Distribution Results for the Full Dataset



Note. RP = retrieval practice. Dashed lines in panel B represent residuals.

Comparing Fixed- and Random-Threshold Models

The modeling approach we employed is the one described in the Two Models Based on the Dual-Memory Framework section. The fixed-threshold model predicted a correlation between the

retrieval practice effect and Raven scores of r = .28 [.12, .43], whereas the random-threshold model predicted a correlation of r = .08 [–.09, .24]. The observed value was r = .12 [–.04, .28]. The result suggests that the random-threshold model provides an overall better fit to the data.

Discussion

We introduced the dual-memory framework as a promising tool for instantiating models in studies on individual differences. After outlining the framework in the Introduction (see also Appendix E), we described our approach to simulate various scenarios and made empirical predictions. The random-threshold model produced a point estimate closer to the empirical value we obtained than the fixed-threshold model, although the empirical confidence interval overlapped with estimates from both models. While it is premature to strongly favor any model based solely on these simulation results, some considerations are important here.

It is essential to distinguish between the conceptual description and the mathematical implementation of the dual-memory framework. The conceptual description comprises statements about three memory strength distributions and how they interact to produce the retrieval practice effect. In line with the conceptual description, the random-threshold model holds the assumption of identical and independent distributions across S_R, S_{T-s}, and S_{T-t}. By allowing the response threshold, t, to vary across participants, this model allows for correlation between PC_R and PC_T. Considering the literature indicating strong correlations between final-test performance for rereading and retrieval practice conditions (Brewer & Unsworth, 2012; Minear et al., 2018; Pan et al., 2015; Robey, 2019), the random-threshold model appears more plausible than the fixed-threshold model.

On the other hand, the mathematical implementation concerns equations modeling interactions between different memory strength distributions hypothesized by the dual-memory

framework. Note that the assumption of fixed threshold, when it applies, occurs across participants, but the threshold is fixed across the three distributions for the same participant. In this sense, Equation 4.4 is used indiscriminately for both fixed- and random-threshold models. Although PC_R and PC_T are correlated in the random-, but not in the fixed-, threshold model, PC_R and $\widehat{PC}_T$ generally correlate because the latter is a quadratic function of the former.[11] In our simulations, we calculated both $\widehat{PC}_T$ and $\widehat{TE}$ as well as PC_T and TE, although we used the former to compare model predictions to the data.

Crucially, estimates generated using the conceptual description and the mathematical implementation may not always align. This divergence is a characteristic we discovered upon delving into the dual-memory framework. While this might be perceived as a weakness of the framework, we view it as an advantage of mathematical modeling—the potential to reveal previously overlooked points for clarifying theoretical assumptions (Farrell & Lewandowsky, 2018). An important avenue for future research lies in understanding when and why the conceptual description and the mathematical implementation of the dual-memory framework lead to distinct predictions.

The dual-memory framework has a precursor in the retrieval practice research. Building upon the new theory of disuse (Bjork & Bjork, 1992), the bifurcation-based framework (Halamish & Bjork, 2011; Kornell et al., 2011) was originally developed to account for the phenomenon of a negative retrieval practice effect commonly observed in studies with short retention intervals (e.g., 5 min) and when corrective feedback after retrieval practice is absent (e.g., Roediger & Karpicke, 2006b). In contrast to the dual-memory framework, which posits different memory strengths with

[11] Note that $PC_T = p(S_{T-s} > t_i) + p(S_{T-t} > t_i) - p(S_{T-s} > t_i) \times p(S_{T-t} > t_i)$, while $\widehat{PC}_T$ is given by Equation 4.4. Similarly, $TE = [p(S_{T-s} > t_i) + p(S_{T-t} > t_i) - p(S_{T-s} > t_i) \times p(S_{T-t} > t_i)] - p(S_R > t_i)$, while $\widehat{TE}$ is given by Equation 4.5; $t_i \in \mathbb{R} \mid t = \text{constant}\ \forall\ i$ in the fixed-threshold model, but it is allowed to vary in the random-threshold model.

identical distributions, the bifurcation-based framework posits that successful retrieval attempts lead to a greater increase in memory strengths compared to rereading attempts and assumes a single memory strength dimension. Unsuccessful retrieval attempts do not alter memory strengths, creating a bifurcation in the memory strength distribution between successful and unsuccessful retrieval practice items. Consequently, this framework predicts positive or negative retrieval practice effects depending on the initial rate of retrieval success and final-test difficulty (Halamish & Bjork, 2011).

Recently, a study compared four computational models based on the dual-memory framework, with a condition-dependent boost model showing the best fit to the data (Guðmundsdóttir & Ragnarsdóttir, 2023). This model allowed study-memory (S_R and S_{T-s}) and test-memory (S_{T-t}) traces to increase their strengths to different degrees. This model seems to incorporate elements from both the bifurcation-based framework (i.e., differential boost after rereading and after retrieval practice) and the dual-memory framework (i.e., two memory traces supporting retrieval practice). Future studies could benefit by contrasting distinct predictions arising from these two frameworks and exploring hybrid models based on both frameworks.

Additionally, we conducted a series of analyses based on previous reports (Gupta et al., 2022; Rickard, 2020; Rickard & Pan, 2018). These analyses indicate that the dual-memory framework is valuable for capturing general patterns or distribution-level effects. However, a closer examination reveals notable deviations in the predictions. For example, Figure 4.2 suggests that the correlation between Raven scores and the retrieval practice effect was closer to the correlation between Raven scores and the predicted retrieval practice effect. However, this comes at the cost of a narrow range in predicted values, restricted between 0 and .25. To address this limitation, Rickard and Pan (2018) introduced a one-parameter model, allowing c to vary across

participants within the range of 0 to 1. This parameter was then multiplied by P_{T-t}—an alternative approach to the condition-dependent boost model (Guðmundsdóttir & Ragnarsdóttir, 2023), except that the Rickard and Pan approach require cP_{T-t} to be equal to or lower than P_R and P_{T-s}. This one-parameter model permits negative retrieval practice effects. However, the authors did not provide a substantive interpretation for the parameter c.

A second example involves the cumulative distribution analysis. We were able to replicate the adequate fit at the distribution level reported in previous studies (Gupta et al., 2022; Rickard, 2020). However, when we considered the residuals, the model exhibited important discrepancies at different quantiles. Of course, the direction of the residuals—positive or negative—could purely result from scaling issues (e.g., for a participant with perfect recall after rereading, the dual-memory framework predicts perfect recall after retrieval practice, so the residuals should be nonpositive).

Final Comments

This study represents an initial endeavor to theoretically predict various relationships between individual-difference variables and the magnitude of the retrieval practice effect. While we view our approach as promising, we believe that our main contribution was methodological. We did not observe a difference between rereading in the 1-day and in the 2-day retention intervals. This precluded us to provide a stronger test for the fixed- and random-threshold models (with observed values below and above the theoretically-relevant proportion of .50). Regarding the models themselves, we assumed normality in the distribution of memory trace strengths. Although Rickard and Pan (2018) claimed that the original framework's predictions are not affected by the chosen distribution, their focus was primarily on estimating recall after rereading and retrieval practice. Hence, their claim might not hold true when predictions involve correlations. Future

simulation studies will need to investigate whether the models' predictions are sensitive to the choice of the distribution. Additionally, we encourage future studies to contrast different strength-based accounts, as the dual-memory and the bifurcation-based frameworks.

Chapter 5 – General Discussion

General Discussion

Assessment of Main Contributions

In experimental research, statements about effects hold true for aggregated statistics, but not for individual cases (Borsboom et al., 2009). Individual differences in treatment are commonly viewed as undesirable and, if possible, should be eliminated. However, what constitutes noise in experimental research is precisely the focus of individual-difference research (Cronbach, 1957). Studies on individual differences in the retrieval practice effect integrate experimental and individual-difference approaches, aiming to understand whether retrieval practice benefits some learners more than others and, if so, which learners' characteristics moderate the retrieval practice effect (McDermott, 2021; Roediger & Yamashiro, 2020).

The overall goal of this dissertation was to investigate direct and indirect effects of gF on the magnitude of the retrieval practice effect. In Chapter 2, a narrative literature review revealed that consistent links between individual differences and the retrieval practice effect remain elusive. While this state of affairs might seem discouraging, we prefer to view this provisional conclusion optimistically. In other words, the main contributions of the review were to pinpoint hypotheses for observed inconsistencies (e.g., heterogeneity of procedures) and to suggest avenues for future research. We will return to this point in the last section of the dissertation.

In Chapter 3, we presented an experiment examining direct and indirect effects of gF on the magnitude of the retrieval practice effect. Initially, our analyses closely followed Minear et al.'s (2018) analytical procedures. Subsequently, we extended our analyses by emphasizing retrieval practice during the practice phase. In hindsight, it is surprising that prior studies did not consider performance during the practice phase, given evidence that duration-based procedures do not ensure equal learning across individuals. We demonstrated that participants with higher gF

benefit more from initial encoding (study phase) and corrective feedback after retrieval attempts (practice phase). Most importantly, in a mediation model, we tested and found an indirect effect of gF on the retrieval practice effect mediated by performance during the practice phase.

The relationship of individual-difference variables with performance during the practice phase has implications for research and application. For research, future studies interested in investigating individual differences in the benefits of retrieval practice on retention should suppress—or at least minimize—differences between learners during the practice phase. One approach we suggested in earlier chapters was the use of criterion-based procedures. Instead of implementing a one-size-fits-all approach, those procedures allow different learners to achieve the performance criterion established by the experimenter in a self-paced manner (Friedman et al., 2017; Karpicke & Roediger, 2007, 2008).

On the applied side, it is important to consider that learners may differ in learning efficiency (Zerr et al., 2018; Zerr et al., 2021). Therefore, teachers and educators should implement a moderate dosage of retrieval practice but be flexible to deviate from this average treatment depending on each learner's and material's specificities. For example, more sophisticated learners or those learning an easier material may require fewer retrieval practice opportunities than less sophisticated learners or those learning a more difficult material.

In Chapter 4, we introduced the dual-memory framework applied to individual-difference research and presented a modeling approach to investigate individual differences in the retrieval practice effect. Given that we had only one data point for correlation testing, we consider our main contribution in the chapter to be methodological. The approach we introduced allowed us to propose two models, one assuming a fixed threshold across participants (fixed-threshold model) and another allowing the threshold to vary across participants (random-threshold model). This

approach is flexible enough to allow researchers to explore other possibilities, including testing other distributions (e.g., gamma), manipulating variance in threshold distributions, and relaxing the identical distributions assumption (Guðmundsdóttir & Ragnarsdóttir, 2023; Halamish & Bjork, 2011).

In the dual-memory framework, predictions of different correlations between individual differences and the retrieval practice effect under certain scenarios lack substantive interpretation. This primarily stems from the descriptive nature of the framework, positing the existence and interaction of different memory traces without specifying the underlying cognitive mechanisms of the retrieval practice effect. Therefore, it is possible that the predominantly null results presented in Chapter 2 reflect the fact that retrieval practice is genuinely "A Learning Method for All: The Testing Effect is Independent of Cognitive Ability," as stated in the title of Jonsson et al.'s (2020) article. Under this view, significant correlations would emerge artifactually as a consequence of specific methodological combinations, in line with the ideas of "hidden moderators" (Klein et al., 2018) and the contextualist approach presented by Roediger (2008).

It is possible that the assumptions of the dual-memory framework could be reconciled with contemporary accounts in the field. For instance, the notion of memory strength distribution in the dual-memory framework has a precedent in the literature (Halamish & Bjork, 2011; Kornell et al., 2011) and can also be linked to the concepts of activation and strengthening of information, as posited by the elaborative retrieval hypothesis (Carpenter, 2009). Furthermore, the idea of increased number of retrieval routes after retrieval practice aligns with elaboration- and mediator-based accounts (Carpenter, 2009; Pyc & Rawson, 2010, 2012).

As a final remark, in Chapter 3, the direct effect of gF on the retrieval practice effect was negative, $c' = -.006$, but positive in Chapter 4, $r = .12$. Although both confidence intervals included

zero, it is noteworthy that the same data yielded estimates with distinct signs. Both values, however, align with observations in much of the literature, where results often hover around zero (Brewer & Unsworth, 2012; Pan et al., 2015). In Chapter 3, the c' coefficient represents the effect of gF on the retrieval practice effect after statistically controlling for the effect of practice recall performance in the model. In Chapter 4, the r statistic represents the strength of the linear relationship between variables, disregarding other potential important variables.

The reviewed literature in Chapter 2 exhibited not only methodological but also analytical heterogeneity. These analytical differences hinder the direct comparison of results, such as those obtained from studies employing bivariate correlation versus multiple regression or quartile-based ANOVAs. In some sense, the seemingly disparate results we presented in Chapters 3 and 4 are representative of the broader literature. The impact of analytical decisions on results is well-documented in the literature (Carpenter et al., 2008; Silberzahn et al., 2018). The key point here is that authors do not always report analyses in a manner that directly addresses the reader's interest. We strongly advocate for adherence to open science practices, such as making data publicly accessible whenever possible. By embracing these practices, it becomes possible to reanalyze data in ways most conducive to the goals of a literature review.

Limitations

Some limitations of the dissertation need to be acknowledged. First, we conducted a narrative review, which might have resulted in the omission of some relevant literature on the topic of interest. In the future, an updated review will be needed, clearly defining research questions, literature search mechanisms, and criteria for assessing the quality of studies.

Second, we did not find a significant difference between rereading in the 1-day and the 2-day retention intervals. This limitation prevented us from conducting a more robust test for the

fixed- and random-threshold models. In hindsight, we recognized that 1-day and 2-day retention intervals tend to result in minor performance differences (Carpenter et al., 2008). Moreover, Swahili–Brazilian-Portuguese word pairs proved to be more challenging than word pairs in the participants' native language. A shorter retention interval or more practice cycles might have been necessary for one of the conditions to yield performance above .50.

Third, we only compared retrieval practice with rereading, possibly a weak control condition. Of course, when scholars discuss the benefits of retrieval practice, it should be asked, "compared to what?" (Kornell et al., 2012, p. 257). This issue is particularly relevant in applied research, where the goal is to assess whether retrieval practice outperforms teaching activities or rehabilitation techniques available for practitioners (Middleton et al., 2015; Moreira, Pinto, Starling, & Jaeger, 2019). We acknowledge that our retrieval practice effect is contingent on the specifics of our experimental paradigm, including the choice of rereading as the control condition. Different choices can be made in the future.

Fourth, we used only one measure of gF, based on the Raven test (Raven et al., 1998). Psychometricians claim that a single task imperfectly represents a construct and that the term *cognitive ability* assumes consistent performance across a set of tasks that supposedly recruit that ability domain (Conway et al., 2005; Lakin & Kell, 2020). Our choice of a single task was driven by practical considerations, namely, the concern that incorporating multiple tasks would make the design lengthy enough to discourage participants from volunteering for the experiment.

Fifth, the models we derived from the dual-memory framework assumed normality in the distribution of memory trace strengths. Future simulation studies should investigate whether model predictions critically depend on the chosen distribution.

Research Agenda

The current dissertation included a narrative literature review, an experimental investigation, and an initial attempt to derive models from the dual-memory framework. Drawing from these efforts, we propose the following research agenda:

1. Examine whether the impact of individual differences on the retrieval practice effect depends on other individual differences (e.g., participants' spontaneous encoding and retrieval strategy use) and contextual factors (e.g., lag, extended practice).
2. Investigate the reliability of the retrieval practice effect (e.g., test–retest reliability, reliability across tasks, materials, and retention intervals).
3. Adopt alternative procedures to suppress—or at least minimize—differences across learners during the practice (e.g., criterion-based procedures).
4. Replicate our results using mediation analysis.
5. Test whether individual differences impact the retrieval practice effect indirectly through the learners' ability to generate and retrieve mediators.
6. Test whether individual differences impact the retrieval practice effect indirectly through the learners' ability to monitor and shift mediators after unsuccessful retrieval attempts.
7. Test different models derived from the dual-memory framework with various data points (e.g., by crossing factorially different values of P_R and $r_{PC_R,ID}$).
8. Investigate why and under which conditions the conceptual description and the mathematical implementation of the dual-memory framework lead to distinct predictions.

9. Propose critical experiments contrasting distinct predictions arising from the dual-memory and bifurcation-based frameworks, and explore hybrid models as well.
10. Explore whether predictions based on the fixed-threshold and the random-threshold models are sensitive to the choice of distribution.

References

References

Abbot, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs*, *11*(1), 159–177. https://doi.org/10.1037/h0093018

Abel, M., & Roediger, H. L., III. (2017). Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affected by practice format. *Memory & Cognition*, *45*(1), 81–92. https://doi.org/10.3758/s13421-016-0641-8

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K. (2019). Retrieval practice and Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L., III. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, *25*(6), 764–771. https://doi.org/10.1080/09658211.2016.1220579

Alamri, A., & Higham, P. H. (2022). The dark side of corrective feedback: Controlled and automatic influences of retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(5), 752–768. https://doi.org/10.1037/xlm0001138

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*, 415–445. https://doi.org/10.1016/j.jml.2003.08.006

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. https://doi.org/10.1037//0278-7393.20.5.1063

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*(1), 43–56. https://doi.org/10.1037/xap0000133

Arnold, K. M., & McDermott, K. B. (2013a). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, *20*, 507–513. https://doi.org/10.3758/s13423-012-0370-3

Arnold, K. M., & McDermott, K. B. (2013b). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940–945. https://doi.org/10.1037/a0029199

Batsell, W. R., Jr., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology*, *44*(1), 18–23. https://doi.org/10.1177/0098628316677492

Bertilsson, F., Stenlund, T., Wiklund-Hörnqvist, C., & Jonsson, B. (2021). Retrieval practice: Beneficial for all students or moderated by individual differences? *Psychology Learning & Teaching*, *20*(1), 21–39. https://doi.org/10.1177/1475725720973494

Bertilsson, F., Wiklund-Hörnqvist, C., Stenlund, T., & Jonsson, B. (2017). The testing effect and its relation to working memory capacity and personality characteristics. *Journal of Cognitive Education and Psychology*, *16*(3), 241–259. https://doi.org/10.1891/1945-8959.16.3.241

Bjork, R. A. (1973). Why mathematical models? *American Psychologist*, *28*(5), 426–433. https://doi.org/10.1037/h0034623

Bjork, R. A. (1975). Retrieval as memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Erlbaum.

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

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudhary (Eds.), *Dynamics process methodology in the social and developmental sciences* (pp. 67–97). Springer.

Brewer, G., Robey, A., & Unsworth, N. (2021). Discrepant findings on the relation between episodic memory and retrieval practice: The impact of analysis decisions. *Journal of Memory and Language*, *116*, Article 104185. https://doi.org/10.1016/j.jml.2020.104185

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*(3), 407–415. https://doi.org/10.1016/j.jml.2011.12.009

Brown, R., & McNeill, D. (1966). The “tip of the tongue” phenomenon. *Journal of Verbal Learning & Verbal Behavior*, *5*(4), 325–337. https://doi.org/10.1016/S0022-5371(66)80040-3

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Buchin, Z. L., & Mulligan, N. W. (2017). The testing effect under divided attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1934–1947. https://doi.org/10.1037/xlm0000427

Buchin, Z. L., & Mulligan, N. W. (2019). Divided attention and the encoding effects of retrieval. *Quarterly Journal of Experimental Psychology*, *72*(10), 2474–2494. https://doi.org/10.1177/1747021819847141

Buchin, Z. L., & Mulligan, N. W. (2023). Retrieval-based learning and prior knowledge. *Journal of Educational Psychology*, *115*(1), 22–35. https://doi.org/10.1037/edu0000773

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448. https://doi.org/10.3758/MC.36.2.438

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language*, *92*, 128–141. https://doi.org/10.1016/j.jml.2016.06.008

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633–642. https://doi.org/10.3758/bf03202713

Cavendish, B. A., Lima, M. F. R., Perícoli, L., & Buratto, L. G. (2022). Effects of combining retrieval practice and tDCS over long-term memory: A randomized controlled trial. *Brain and Cognition, 156,* Article 105807. https://doi.org/10.1016/j.bandc.2021.105807

Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition*, *2*, 95–100. https://doi.org/10.1016/j.jarmac.2013.04.001

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition*, *40*, 528–539. https://doi.org/10.3758/s13421-011-0168-y

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/bf03196772

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684. https://doi.org/10.1037/h0043943

Crowder, R. G. (2015). *Principles of learning and memory (Classic edition)*. Psychology Press. (Original work published 1976)

Crutcher, R. J., & Ericsson, K. A. (2000). The role of mediators in memory retrieval as a function of practice: Controlled mediation to direct access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1297–1317. https://doi.org/10.1037//0278-7393.26.5.1297

Dienes, Z. (2014). Using Bayes to get the most out of nonsignificant results. *Frontiers in Psychology*, *5*, Article e00781. https://doi.org/10.3389/fpsyg.2014.00781

Dikmans, M. E., van den Broek, G. S. E., & Klatter-Folmer, J. (2020). Effects of repeated retrieval on keyword mediator use: Shifting to direct retrieval predicts better learning outcomes. *Memory*, *28*(7), 908–917. https://doi.org/10.1080/09658211.2020.1797094

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Ekuni, R., & Pompeia, S. (2020). Improving retention by placing retrieval practice at the end of class: A naturalistic study. *Revista Latinoamericana de Psicología*, *52*, 22–32. https://doi.org/10.14349/rlp.2020.v52.3

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331. https://doi.org/10.1037/0096-3445.128.3.309

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press. https://doi.org/10.1017/CBO9781316272503

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language, 64*, 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Franco, V. R., & Iglesias, F. (2023). Invitation to mathematical psychology: Models and benefits of formal theorizing. *Psicologia: Teoria e Pesquisa*, *39,* Article e39515. https://doi.org/10.1590/0102.3772e39515.en

Friedman, R. B., Sullivan, K. L., Snider, S. F., Luta, G., & Jones, K. T. (2017). Leveraging the test effect to improve maintenance of the gains achieved through cognitive rehabilitation. *Neuropsychology*, *31*(2), 220–228. https://doi.org/10.1037/neu0000318

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

Gates, A. I. (1918). Correlations of immediate and delayed recall. *Journal of Educational Psychology*, *9*, 489–496. https://doi.org/10.1037/h0074445

Gazzaniga, M. S. (1991). Interview with Endel Tulving. *Journal of Cognitive Neuroscience, 3*(1), 89–94. https://doi.org/10.1162/jocn.1991.3.1.89

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. https://doi.org/10.1037/0022-0663.81.3.392

Goldstein, E. B., & Cacciamani, S. (2022). *Sensation and perception* (11th ed.). Cengage Learning.

Guðmundsdóttir, K. Á., & Ragnarsdóttir, S. R. (2023). *Repeated testing enhances memory through differential strength for test and study traces: Dual memory vs. instance theory model analysis* [Bachelor's thesis, University of Akureyri]. Skemman. https://skemman.is/handle/1946/45118

Gupta, M. W., Pan, S. C., & Rickard, T. C. (2022). Prior episodic learning and the efficacy of retrieval practice. *Memory & Cognition*, *50,* 722–735. https://doi.org/10.3758/s13421-021-01236-4

Gupta, M. W., Pan, S. C., & Rickard, T. C. (2024). Interaction between the testing and forward testing effects in the case of cued-recall: Implications for theory, individual difference studies, and application. *Journal of Memory and Language, 134*, Article 104476. https://doi.org/10.1016/j.jml.2023.104476

Guran, C.-N. A., Lehmann-Grube, J., & Bunzeck, N. (2020). Retrieval practice improves recollection-based memory over a seven-day period in younger and older adults. *Frontiers in Psychology, 10*, Article 2997. https://doi.org/10.3389/fpsyg.2019.02997

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 801–811. https://doi.org/10.1037/a0023219

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, *76*(4), 408–420. https://doi.org/10.1080/03637750903310360

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). The Guilford Press.

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*(2), 575–596. https://doi.org/10.1037/a0033715

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989x.3.4.486

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. https://doi.org/10.1016/S0022-5371(71)80029-4

Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language, 102,* 1–15. https://doi.org/10.1016/j.jml.2018.04.005

Horton, K. D. (1987). The incongruity effect in memory for generated targets: Fact or artifact? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 172–174. https://doi.org/10.1037/0278-7393.13.1.172

Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educational Psychology, 35*(4), 513–521. https://doi.org/10.1080/01443410.20.2014.963030

JASP Team. (2018). *JASP (Version 0.17.1) [Computer software]*. https://jasp-stats.org/

Jonsson, B., Wiklund-Hörnqvist, C., Nyroos, M., & Börjesson, A. (2014). Self-reported memory strategies and their relationship to immediate and delayed text recall and working memory capacity. *Education Inquiry, 5,* 385–404. https://doi.org/10.3402/edui.v5.22850

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M., & Nyberg, L. (2020). A learning method for all: The testing effect is independent of cognitive ability. *Journal of Educational Psychology*, *113*(5), 972–985. https://doi.org/10.1037/edu0000627

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference (J. H. Byrne, Series Ed.)* (pp. 487–514). Academic Press.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology, 7,* Article 350. https://doi.org/10.3389/fpsyg.2016.00350

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, *67*(1), 17–29. https://doi.org/10.1016/j.jml.2012.02.004

Kenney, K. L., & Bailey, H. (2021). Low-stakes quizzes improve learning and reduce overconfidence in college students. *Journal of the Scholarship of Teaching and Learning*, *21*(2), 79–92. https://doi.org/10.14434/josotl.v21i2.28650

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/doi.org/10.1177/25152459188102

Klier, C., & Buratto, L. G. (2023). The benefit of retrieval practice on cued recall under stress depends on item difficulty. *Neuroscience Letters, 797,* Article 137066. https://doi.org/10.1016/j.neulet.2023.137066

Knouse, L. E., Rawson, K. A., Vaughn, K. E., & Dunlosky, J. (2016). Does testing improve learning for college students with attention-deficit/hyperactivity disorder? *Clinical Psychological Science*, *4*(1), 136–143. https://doi.org/10.1177/2167702614565175

Kole, J. A., & Healy, A. F. (2013). Is retrieval mediated after repeated testing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 462–472. https://doi.org/10.1037/a0028880

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Rabelo, V. C., & Klein, P. J. (2012). Test enhance learning—Compared to what? *Journal of Applied Research in Memory and Cognition*, *1*, 257–259. https://doi.org/10.1016/j.jarmac.2012.10.002

Lakin, J. M., & Kell, H. J. (2020). Intelligence and reasoning. In R. J. Sternberg (Ed.), *The Cambridge handbook of intelligence* (2nd ed., pp. 528–552). Cambridge University Press. https://doi.org/10.1017/9781108770422

LeBel, E., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*, Article MP.2018.2843. https://doi.org/10.15626/MP.2018.843

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Leeming, F. C. (2005). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*(3), 210–212. https://doi.org/10.1207/S15328023TOP2903_06

Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1573–1591. https://doi.org/10.1037/xlm0000267

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794. https://doi.org/10.1037/xlm0000012

Lima, M. F. R., & Buratto, L. G. (2021). Norms for familiarity, concreteness, valence, arousal, wordlikeness, and recall accuracy for Swahili–Portuguese word pairs. *SAGE Open*, *11*(1), 1–13. https://doi.org/10.1177/2158244020988524

Lima, M. F. R., & Buratto, L. G. (2023a). Metanálises em psicologia: Uma introdução conceitual e prática [Meta-analyses in psychology: A conceptual and hands-on introduction]. *Psico-USF*, *28*(2), 267–279. https://doi.org/10.1590/1413-82712023280205

Lima, M. F. R., & Buratto, L. G. (2023b). The test–retest reliability of the retrieval practice effect. *Quarterly Journal of Experimental Psychology*, *76*(9), 2028–2036. https://doi.org/10.1177/17470218221141586

Lima, M. F. R., Cavendish, B. A., Deus, J. S., & Buratto, L. G. (2020). Retrieval practice in memory- and language-impaired populations: A systematic review. *Archives of Clinical Neuropsychology*, *35*(7), 1078–1093. https://doi.org/10.1093/arclin/acaa035

Lima, M. F. R., Venâncio, S., Feminella, J., & Buratto, L. G. (2020). Does item difficulty affect the magnitude of the retrieval practice effect? An evaluation of the retrieval effort hypothesis. *The Spanish Journal of Psychology*, *23*, Article e31. https://doi.org/10.1017/SJP.2020.33

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43,* 14–26. https://doi.org/10.3758/s13421-014-0452-8

Liu, X. L., Tan, D. H., & Reder, L. M. (2018). The two processes underlying the testing effect—Evidence from event related potentials (ERPs). *Neuropsychologia*, *112*, 77–85. https://doi.org/10.1016/j.neuropsychologia.2018.02.022

Ludowicy, P., Paz-Alonso, P. M., Lachmann, T., & Czernochowski, D. (2023). Performance feedback enhances test-potentiated encoding. *Frontiers in Behavioral Neuroscience, 17,* Article 1100497. https://doi.org/10.3389/fnbeh.2023.1100497

Mair, P. (2018). *Modern psychometrics with R*. Springer. https://doi.org/10.1007/978-3-319-93177-7

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*, 335–384. https://doi.org/10.1016/j.cogpsych.2008.02.004

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. https://doi.org/10.1037/a0021782

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371–385. https://doi.org/10.1037/0278-7393.11.2.371

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology*, *72*(23), 609-633. https://doi.org/10.1146/annurev-psych-010419-051019

Melton, A. W. (1967). Repetition and retrieval from memory. *Science*, *158*(3800), 532. https://doi.org/10.1126/science.158.3800.532-b

Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging*, *28*(1), 142–147. https://doi.org/10.1037/a0030890

Middleton, E. L., Schwartz, M. F., Rawson, K. A., & Garvey, K. (2015). Test-enhanced learning versus errorless learning in aphasia rehabilitation: Testing competing psychological principles. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1253–1261. https://doi.org/10.1037/xlm0000091

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1474–1486. https://doi.org/10.1037/xlm0000486

Minear, M. E., Coane, J. H., Cooney, L. H., Boland, S. C., & Serrano, J. W. (2023). Is practice good enough? Retrieval benefits students with ADHD but does not compensate for poor encoding in unmedicated students. *Frontiers in Psychology*, *14*, Article 1186566. https://doi.org/10.3389/fpsyg.2023.1186566

Moreira, B. F. T., Pinto, T. S. S., Justi, F. R. R., & Jaeger, A. (2019). Retrieval practice improves learning in children with diverse visual word recognition skills. *Memory*, *27*(10), 1423–1437. https://doi.org/10.1080/09658211.2019.1668017

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Psychology*, *4*, Article 5. https://doi.org/10.3389/feduc.2019.00005

Moscovitch, M. (2007). Memory: Why the engram is elusive. In H. L. Roediger, III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 17–22). Oxford University Press.

Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 975–980). Erlbaum.

Mulligan, N. W., Buchin, Z. L., & Zhang, A. L. (2022). The testing effect with free recall: Organization, attention, and order effects. *Journal of Memory and Language. , 125*, Article 104333. https://doi.org/10.1016/j.jml.2022.104333

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition 41*(3), 859–871. https://doi.org/10.1037/xlm0000056

Mulligan, N. W., Rawson, K. A., Peterson, D. J., & Wissman, K. T. (2018). The replicability of the negative testing effect: Differences across participant populations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(5), 752–763. https://doi.org/10.1037/xlm0000490

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. https://doi.org/10.1037//1082-989X.S.2.241

Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill, Inc.

Nunes, C. S. S., & Nunes, M. F. O. (2015). *Matrizes progressivas avançadas de Raven: APM-Raven - Manual técnico*. Casa do Psicólogo.

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. https://doi.org/10.1016/j.jml.2015.04.001

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pastötter, B., & Bäuml, K.-H. T. (2019). Testing enhances subsequent learning in older adults. *Psychology and Aging*, *34*(2), 242–250. https://doi.org/10.1037/pag0000307

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 287–297. https://doi.org/10.1037/a0021801

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1287–1293. https://doi.org/10.1037/a0031337

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*, 178–192. https://doi.org/10.1037/1082-989X.10.2.178

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335–335. https://doi.org/10.1126/science.1191465

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. https://doi.org/10.1037/a0026166

R Core Team. (2023). *R: A language and environment for statistical computing.* Vienna, Austria. https://www.R-project.org/

Racsmány, M., Szőllősi, Á., & Bencze, D. (2018). Retrieval practice makes procedure from remembering: An automatization account of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 157–166. https://doi.org/10.1037/xlm0000423

Rajaram, S., & Barber, S. J. (2008). Retrieval processes in memory. In H. L. Roediger, III (Ed.), *Learning and memory: A comprehensive reference* (Vol. 2: Cognitive psychology of memory, pp. 261–284). Academic Press.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Advanced progressive matrices*. Oxford Psychologists Press.

Rawson, K. A., Wissman, K. T., & Vaughn, K. E. (2015). Does testing impair relational processing? Failed attempts to replicate the negative testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1326–1336. https://doi.org/10.1037/xlm0000127

Rawson, K. A., & Zamary, A. (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. *Journal of Memory and Language, 105,* 141–152. https://doi.org/10.1016/j.jml.2019.01.002

Rickard, T. C. (2020). Extension of the dual-memory model of test-enhanced learning to distributions and individual differences. *Psychonomic Bulletin & Review, 27,* 783–790. https://doi.org/10.3758/s13423-020-01734-7

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, *25*(3), 847–869. https://doi.org/10.3758/s13423-017-1298-4

Robey, A. (2019). The benefits of testing: Individual differences based on student factors. *Journal of Memory and Language, 108,* Article 104029. https://doi.org/10.1016/j.jml.2019.104029

Roediger, H. L., III. (2000). Why retrieval is the key process in understanding human memory. In E. Tulving (Ed.), *Memory, consciousness, and the brain: The Tallinn conference* (pp. 52–75). Psychology Press.

Roediger, H. L., III. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59,* 225–254. https://doi.org/10.1146/annurev.psych.57.102904.190139

Roediger, H. L., III, Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 13–49). Psychology Press.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., III, 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* (Vol. 55, pp. 1–36). Elsevier. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Roediger, H. L., III, & Yamashiro, J. K. (2020). Evaluating experimental research. In R. J. Sternberg & D. F. Halpern (Eds.), *Critical thinking in psychology* (pp. 249–279). Cambridge University Press. https://doi.org/10.1017/9781108684354.012

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, *76*, 45–48. https://doi.org/10.1037/h0031355

Ross, S. (2007). *Introduction to probability models* (9th ed.). Academic Press.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*, 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Schacter, D. L. (2021). *The seven sins of memory: How the mind forgets and remembers* (Updated 2nd ed.). Houghton Mifflin Harcourt.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Shaffer, R. A., & McDermott, K. B. (2020). A role for familiarity in supporting the testing effect over time. *Neuropsychologia*, *138*, Article 107298. https://doi.org/10.1016/j.neuropsychologia.2019.107298

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Silva, F. V., Ekuni, R., & Jaeger, A. (2023). Retrieval practice benefits for spelling performance in fifth-grade children. *Memory*, *31*(9), 1197–1204. https://doi.org/10.1080/09658211.2023.2248420

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. https://doi.org/10.1037/0278-7393.4.6.592

Smith, A. M., Floerke, V. A., & Thomas, A. K. (2016). Retrieval practice protects memory against acute stress. *Science*, *354*(6315), 1046–1048. https://doi.org/10.1126/science.aah5067

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. https://doi.org/10.1177/0956797615617778

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641–656. https://doi.org/10.1037/h0063404

Starling, D. S. V., Moreira, B. F. T., & Jaeger, A. (2019). Retrieval practice as a learning strategy for individuals with Down syndrome: A preliminary study. *Dementia & Neuropsychologia*, *13*(1), 104–110. https://doi.org/10.1590/1980-57642018dn13-010012

Stenlund, T., Jönsson, F. U., & Jonsson, B. (2017). Group discussions and test-enhanced learning: Individual learning outcomes and personality characteristics. *Educational Psychology*, *37*(2), 145–156. https://doi.org/10.1080/01443410.2016.1143087

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 115–124. https://doi.org/10.1037/a0034252

Sumowski, J. F., Leavitt, V. M., Cohen, A., Paxton, J., Chiaravalloti, N. D., & DeLuca, J. (2013). Retrieval practice is a robust memory aid for memory-impaired patients with MS. *Multiple Sclerosis Journal*, *19*(14), 1943–1946. https://doi.org/10.1177/1352458513485980

Sumowski, J. F., Wood, H. G., Chiaravalloti, N., Wylie, G. R., Lengenfelder, J., & DeLuca, J. (2010). Retrieval practice: A simple strategy for improving memory after traumatic brain injury. *Journal of the International Neuropsychological Society*, *16*, 1147–1150. https://doi.org/10.1017/S1355617710001128

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392–1399. https://doi.org/10.1037/a0013082

Trivedi, P. K., & Zimmer, D. M. (2005). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, *1*(1), 1–115. https://doi.org/10.1561/0800000005

Trumbo, M. C. S., McDaniel, M. A., Hodge, G. K., Jones, A. P., Matzen, L. E., Kittinger, L. I., Kittinger, R. S., & Clark, V. P. (2021). Is the testing effect ready to be put to work? Evidence from the laboratory to the classroom. *Translational Issues in Psychological Science, 7*(3), 332–355. https://doi.org/10.1037/tps0000292

Tse, C.-S., Balota, D. A., & Roediger, H. L., III. (2010). The benefits and costs of repeated testing on the learning of face–name pairs in healthy older adults. *Psychology of Aging*, *25*(4), 833–845. https://doi.org/10.1037/a0019933

Tse, C.-S., Chan, M. H.-M., Tse, W.-S., & Wong, S. W.-H. (2019). Can the testing effect for general knowledge facts be influenced by distraction due to divided attention or experimentally induced anxious mood? *Frontiers in Psychology*, *10*, Article 969. https://doi.org/10.3389/fpsyg.2019.00969

Tse, C.-S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, *18*(3), 253–264. https://doi.org/10.1037/a0029190

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. https://doi.org/10.1016/S0022-5371(67)80092-6

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior*, *5*(4), 381–391. https://doi.org/10.1016/S0022-5371(66)80048-8

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. https://doi.org/10.1037/h0020071

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, *145*(1), 79–139. https://doi.org/10.1037/bul0000176

van den Bergh, D., Wagenmakers, E.-J., & Aust, F. (2022). Bayesian repeated-measures ANOVA: An updated methodology implemented in JASP. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/fb8zn

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science*, *22*, 1127–1131. https://doi.org/10.1177/0956797611417724

Walrath, R., Willis, J. O., Dumont, R., & Kaufman, A. S. (2020). Factor-analytic models of intelligence. In R. J. Sternberg (Ed.), *The Cambridge handbook of intelligence* (2nd ed., pp. 75–98). Cambridge University Press. https://doi.org/10.1017/9781108770422

Wenzel, K., & Reinhard, M.-A. (2019). Relatively unintelligent individuals do not benefit from intentionally hindered learning: The role of desirable difficulties. *Intelligence*, *77*, Article 101405. https://doi.org/10.1016/j.intell.2019.101405

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *34*(4), 240–245. https://doi.org/10.1111/j.1467-9280.1992.tb00036.x

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1036–1046. https://doi.org/10.1037/xlm0000379

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*, 10–16. https://doi.org/10.1111/sjop.12093

Wiklund-Hörnqvist, C., Stillesjö, S., Andersson, M., Jonsson, B., & Nyberg, L. (2022). Retrieval practice is effective regardless of self-reported need for cognition - Behavioral and brain imaging evidence. *Frontiers in Psychology*, *12*, Article 797395. https://doi.org/10.3389/fpsyg.2021.797395

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–172. https://doi.org/10.1037/0033-295X.114.1.152

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341–1354. https://doi.org/10.1037//0278-7393.20.6.1341

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. https://doi.org/10.3758/MC.38.8.995

Zerr, C. L., Berg, J. J., Nelson, S. M., Fishell, A. K., Savalia, N. K., & McDermott, K. B. (2018). Learning efficiency: Identifying individual differences in learning rate and retention in

healthy adults. *Psychological Science*, *29*(9), 1436–1450. https://doi.org/10.1177/0956797618772540

Zerr, C. L., Spaventa, T., & McDermott, K. B. (2021). Are efficient learners of verbal stimuli also efficient and precise learners of visuospatial stimuli? *Memory*, *29*(5), 675–692. https://doi.org/10.1080/09658211.2021.1933039

Appendices

Appendix A – Approval by the Research Ethics Committee

INSTITUTO DE CIÊNCIAS HUMANAS E SOCIAIS DA UNIVERSIDADE DE BRASÍLIA - UNB

PARECER CONSUBSTANCIADO DO CEP

DADOS DA EMENDA

Título da Pesquisa: Uma Investigação Sobre a Estabilidade Temporal do Efeito de Prática de Recuperação

Pesquisador: MARCOS FELIPE RODRIGUES DE LIMA
Área Temática:
Versão: 2
CAAE: 50528121.1.0000.5540
Instituição Proponente: Instituto de Psicologia -UNB
Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 5.233.793

Apresentação do Projeto:
Nessa proposta de Extensão de Pesquisa aprovada pelo CEP/CHS, o Pesquisador alterou o título da Pesquisa "Uma Investigação sobre a Estabilidade Temporal do Efeito de Prática de Recuperação" (aprovada pelo CEP) para "Uma Investigação Sobre o Efeito de Prática na Memória". Além da troca de título, também foram feitas alterações no número de estímulos, de sessões e no tempo de exposição do participante em cada tentativa da tarefa, e o TCLE foi atualizado.

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

BRASILIA, 09 de Fevereiro de 2022

Assinado por:
MARCIO CAMARGO CUNHA FILHO
(Coordenador(a))

Appendix B – Written Informed Consent (in Brazilian Portuguese)

(Em acordo às Normas da resolução 466/12 do Conselho Nacional de Saúde-MS)

Você está sendo convidado(a) a participar como voluntário(a), da pesquisa "Uma Investigação Sobre o Efeito de Prática na Memória", cujo pesquisador responsável é Marcos Felipe Rodrigues de Lima, estudante de doutorado do Programa de Pós-Graduação em Ciência do Comportamento, do Departamento de Processos Psicológicos Básicos, Instituto de Psicologia, Universidade de Brasília, sob a orientação do Prof. Dr. Luciano Grüdtner Buratto.

O estudo tem como objetivo investigar como diferentes formas de praticar um material afetam a memória. Os procedimentos da pesquisa envolvem a realização de tarefas de memória e de um teste de inteligência. A pesquisa terá três sessões presenciais. A primeira sessão terá duração estimada de 40 minutos. A segunda sessão terá duração estimada de 15 minutos. A terceira sessão terá duração estimada de 20 minutos. O intervalo entre a primeira e a segunda sessões será de 1 ou 2 dias. O intervalo entre a segunda e a terceira sessões é mais flexível, de modo que a terceira sessão poderá ocorrer na semana seguinte à segunda sessão, conforme sua disponibilidade. Sua participação na pesquisa não implica nenhum risco. No final da terceira sessão, você terá a oportunidade de ser instruído acerca das tarefas que participou na pesquisa. Além disso, você poderá manifestar interesse em receber notificações acerca das publicações decorrentes da pesquisa na qual está participando.

O estudo será realizado no Laboratório Integrado de Pós-Graduação e Pesquisa Experimental em Psicologia com Humanos (LIPSI), no Instituto de Psicologia (UnB, campus Darcy Ribeiro). Sua participação é voluntária e livre de qualquer remuneração. Você é livre para recusar-se a participar, retirar seu consentimento ou interromper sua participação a qualquer momento. A recusa em participar não irá acarretar qualquer penalidade ou perda de benefícios. Além disso, na publicação dos resultados do estudo, será mantido o sigilo sobre a sua identidade.

Seus dados ficarão sob a guarda do pesquisador responsável, sendo que somente os integrantes da equipe de pesquisa terão acesso a seus dados pessoais.

Os resultados dessa pesquisa serão divulgados sob a forma de tese de doutorado do pesquisador responsável, o qual ficará disponível no Repositório Institucional da UnB (http://repositorio.unb.br). Além disso, os resultados poderão culminar em artigos científicos e em apresentação de trabalhos em eventos científicos. Esclarecimentos poderão ser feitos a qualquer momento da pesquisa, mediante contato com o pesquisador responsável [telefone e WhatsApp: [redacted]; e-mail: [redacted]].

Este projeto foi revisado e aprovado pelo Comitê de Ética em Pesquisa em Ciências Humanas e Sociais (CEP/CHS) da Universidade de Brasília. As informações com relação à assinatura do TCLE ou aos direitos do participante da pesquisa podem ser obtidas por meio do e-mail do CEP/CHS: cep_chs@unb.br ou pelo ou pelo telefone: (61) 3107 1592.

Este documento foi elaborado em duas vias, uma ficará com o pesquisador responsável pela pesquisa e a outra com você.

Assinatura do/da participante	Assinatura do pesquisador

Brasília, _____ de _____ de _____.

Appendix C – Swahili–Brazilian-Portuguese Word Pairs

For the purpose of counterbalancing, the word pairs used in the experiment described in Chapter 3 were divided into two sets (A and B). For a given participant, items from one set were assigned to the rereading condition, and items from the other set were assigned to the retrieval practice condition. The assignment of sets to experimental conditions was counterbalanced across participants.

Table C1 presents the attribute estimates of the stimuli used in the experiment described in Chapter 3. For Brazilian Portuguese words, Table C1 contains estimates for familiarity, concreteness, valence, and arousal. Familiarity ranges from 1 (*I never saw/heard that word*) to 7 (*I see/hear that word almost daily*). Concreteness ranges from 1 (*Highly abstract*) to 7 (*Highly concrete*). Valence ranges from 1 (*Negative emotional valence*) to 9 (*Positive emotional valence*). Arousal ranges from 1 (*Relaxing*) to 9 (*Exciting*). The wordlikeness column refers to the perceived similarity of the Swahili word to Brazilian Portuguese words, ranging from 1 (*Not like a word at all*) to 5 (*Very like a word*). The average recall column refers to the average recall accuracy of each Brazilian Portuguese word across three test blocks (i.e., where the Swahili word was presented on the screen, and the participant was asked to recall its Brazilian Portuguese translation). Average accuracy values can range from 0 to 1, with lower and higher values indicating, respectively, more difficult and easier word pairs.

Table C1

Attributes of Swahili–Brazilian-Portuguese Word Pairs Used in the Experiment

Set	Swahili	Brazilian Portuguese	Familiarity	Concreteness	Valence	Arousal	Wordlikeness	Average Recall
A	adhama	honra	5.55	3.07	7.03	5.03	2.13	.42
A	ambo	cola	5.83	6.36	5.10	4.70	3.60	.25
A	ankra	fatura	6.15	5.98	2.54	6.83	2.61	.22
A	bustani	jardim	6.10	6.35	7.99	2.59	2.20	.26
A	duara	roda	6.03	6.08	5.79	4.73	2.68	.33
A	fagio	vassoura	6.51	6.80	4.74	5.29	3.29	.31
A	handaki	trincheira	4.18	5.68	2.77	6.54	1.70	.14
A	kaa	caranguejo	5.13	6.59	5.37	5.08	1.38	.46
A	malkia	rainha	5.53	5.65	6.08	4.78	2.07	.73
A	mashua	barco	5.41	6.52	6.17	4.16	1.87	.32
A	nabii	profeta	5.14	4.39	5.52	5.16	2.03	.49
A	nyanya	tomate	6.57	6.92	7.26	4.07	1.42	.75
A	pazia	cortina	6.08	6.62	6.65	3.36	3.41	.42
A	roho	alma	6.16	2.52	6.72	4.45	2.49	.67
A	ruba	sanguessuga	4.57	5.63	2.12	6.96	3.15	.28
A	sahani	prato	6.72	6.85	6.97	4.60	2.01	.35
A	tumbili	macaco	5.75	6.60	5.93	4.86	1.85	.32

Set	Swahili	Brazilian Portuguese	Familiarity	Concreteness	Valence	Arousal	Wordlikeness	Average Recall
A	vumbi	poeira	6.50	6.29	2.27	6.50	3.58	.31
A	yay	ovo	6.68	6.80	7.20	4.29	2.65	.81
A	zulia	tapete	6.01	6.70	6.55	3.42	3.45	.44
B	adha	problema	6.58	3.51	1.88	7.72	1.74	.41
B	bahasha	envelope	5.39	6.53	5.75	4.18	2.01	.21
B	buu	larva	5.08	6.41	2.41	6.45	2.52	.53
B	desturi	costume	5.98	3.50	5.75	4.84	3.93	.30
B	embe	manga	6.08	6.58	7.23	3.86	1.91	.35
B	farasi	cavalo	5.80	6.68	6.72	4.89	2.18	.39
B	gharika	enchente	5.05	5.85	1.93	7.45	1.63	.31
B	hariri	seda	5.25	5.91	6.90	3.32	1.70	.28
B	jani	folha	6.43	6.53	6.83	3.80	2.68	.39
B	jibini	queijo	6.50	6.84	7.59	4.08	2.12	.40
B	joko	forno	6.34	6.59	6.39	5.03	3.00	.32
B	kamba	corda	5.65	6.49	5.27	5.30	3.91	.33
B	kasuku	papagaio	5.43	6.74	6.74	4.57	1.67	.35
B	leso	cachecol	5.20	6.60	7.08	3.10	4.20	.43
B	mbwa	cachorro	6.74	6.72	7.62	4.12	1.30	.65
B	pombe	cerveja	6.37	6.68	5.77	5.03	4.25	.65
B	punda	burro	5.90	4.64	3.28	5.95	4.34	.70

Set	Swahili	Brazilian Portuguese	Familiarity	Concreteness	Valence	Arousal	Wordlikeness	Average Recall
B	samadi	estrume	4.62	5.86	2.40	6.13	2.21	.25
B	wakili	agente	5.26	5.22	5.50	5.47	1.61	.32
B	wingu	nuvem	6.30	5.91	7.72	3.01	2.01	.57
Mean (Set A)	—	—	5.83	5.92	5.54	4.87	2.48	.41
Mean (Set B)	—	—	5.80	5.99	5.54	4.91	2.55	.41
Mean (Overall)	—	—	5.81	5.95	5.54	4.89	2.51	.41

Note. Estimates based on Lima and Buratto's (2021) norms. Sets A and B were created for counterbalancing purposes.

Appendix D – Performance on the Final Tests Divided by Retention Intervals

Figure D1 shows the proportion of targets correctly recalled during the final cued-recall test. Figure D2 shows the proportion of word pairs correctly recognized during the final associative-recognition test, that is $(hits + correct\ rejections)/total\ items$. Figures D1 and D2 consist of new versions of Figures 3.2 and 3.3, respectively (see Chapter 3). except that here they are also broken down by retention interval.

Figure D1

Final Cued-Recall Test Performance as a Function of Learning Strategy and Retention Interval

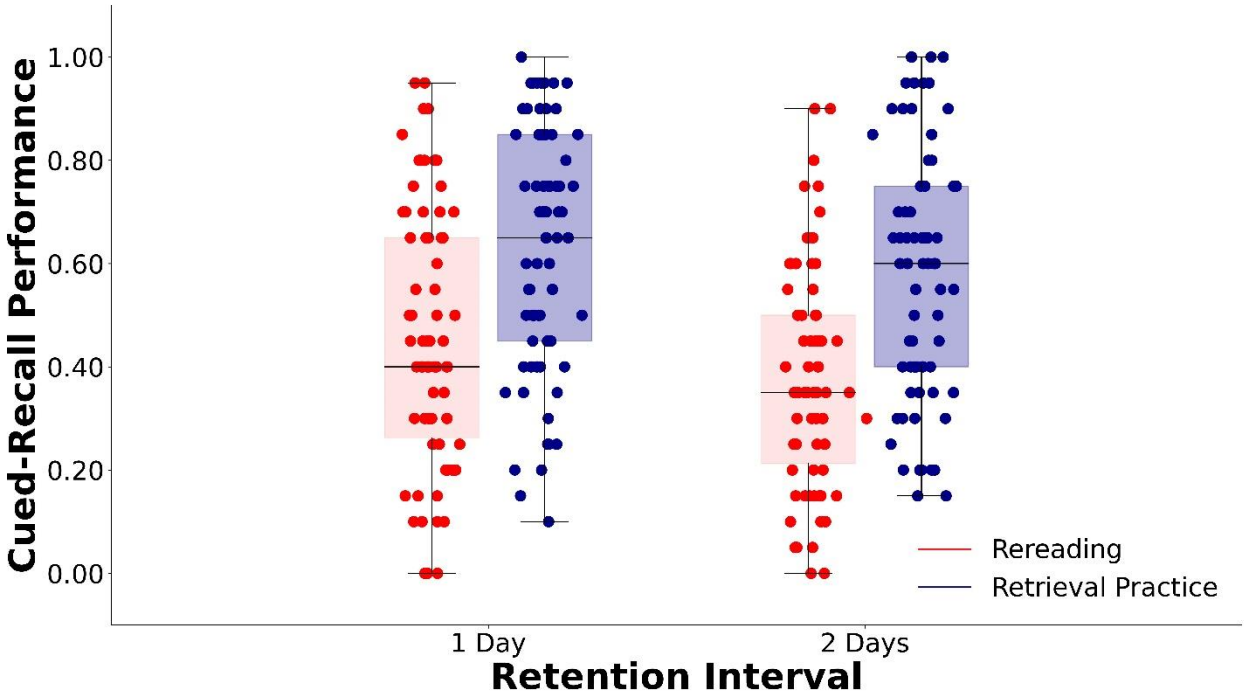
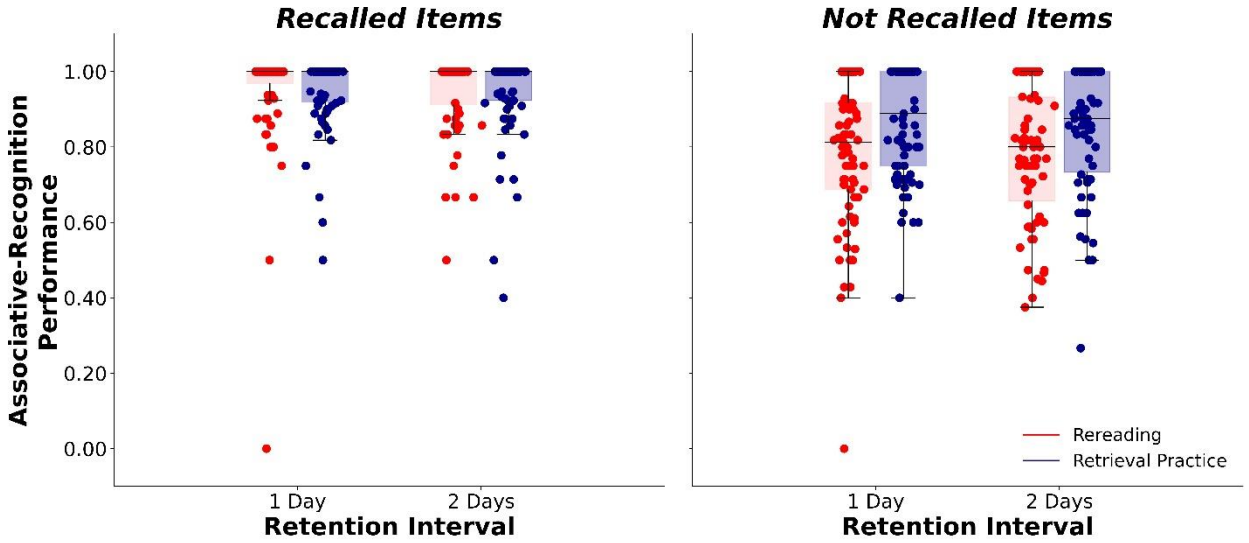


Figure D2

Final Associative-Recognition Test Performance as a Function of Learning Strategy and Retention Interval



Appendix E – Modeling Approach Advanced in Chapter 4

Assume that we want to model different empirical scenarios, considering two distinct variables, namely, P_R and the correlation between PC_R and Raven scores across participants. Regarding P_R, modeling a more difficult final test can be achieved by using a higher response threshold, whereas modeling an easier test can be achieved by using a lower response threshold. For instance, assume we manipulate the difficulty of the items such that, in the final test, we obtain $P_R = .40$ for the more difficult items, and $P_R = .60$ for the easier items. This difference, around .50, is theroretically important, as shown below. In practice, we tried to achieve these P_R values by manipulating the retention interval (1 day vs. 2 days). Because the results with these retention intervals yielded nonsignificant differences in P_R, we collapsed the data in the retention interval condition for most of the subsequent analyses. As for the correlation between PC_R and Raven observed in the literature, different positive correlations have been found between rereading performance in the final-test phase and Raven across studies, ranging from.14 (Robey, 2019) to .40 (Minear et al., 2018). For this reason, in our simulations we only considered scenarios in which Raven positively correlated with memory.

We next advanced two models derived from the dual-memory framework. The first model is referred to as the *fixed-threshold model*, which directly follows from the conceptual description presented by Rickard and Pan (2018). The simulation for the fixed-threshold model crossed the two levels of P_R (.40 and .60) with the 11 levels of $r_{PC_R,Raven}$ (0 to 1, in steps of .10). Thus, 22 scenarios were considered.

Each scenario was repeated 100,000 times, representing a very large sample size. In each iteration, representing one participant, one Raven score was sampled from a normal distribution, $Raven \sim Normal(0, 1)$. Additionally, 20 values for memory strength for rereading items in the

final test were sampled from a normal distribution, $S_R \sim Normal(0, 1)$. The same procedure was independently followed for S_{T-s} and S_{T-t}, but since they were not relevant for the present purposes, they will not be mentioned further. The proportion of items in the rereading condition recalled in the final test was computed as $PC_R = p(S_R > t)$. The value for t varied in tandem with P_R. More specifically, in each scenario, t was fixed across participants, and it was defined as the critical *z*-value associated with $1 - P_R$, ensuring that recall across participants converged for P_R—although PC_R values were allowed to vary across participants.

After iterating for all participants, PC_R and Raven scores were independent. Therefore, a correlation adjustment was necessary. We achieved this by replacing the original values for Raven with adjusted values:

$$Raven_{adjusted} = r_{PC_R,Raven} \times zscore(PC_R) + \sqrt{1 - r^2_{PC_R,Raven}} \times Raven_{original} \qquad \text{(E1)}$$

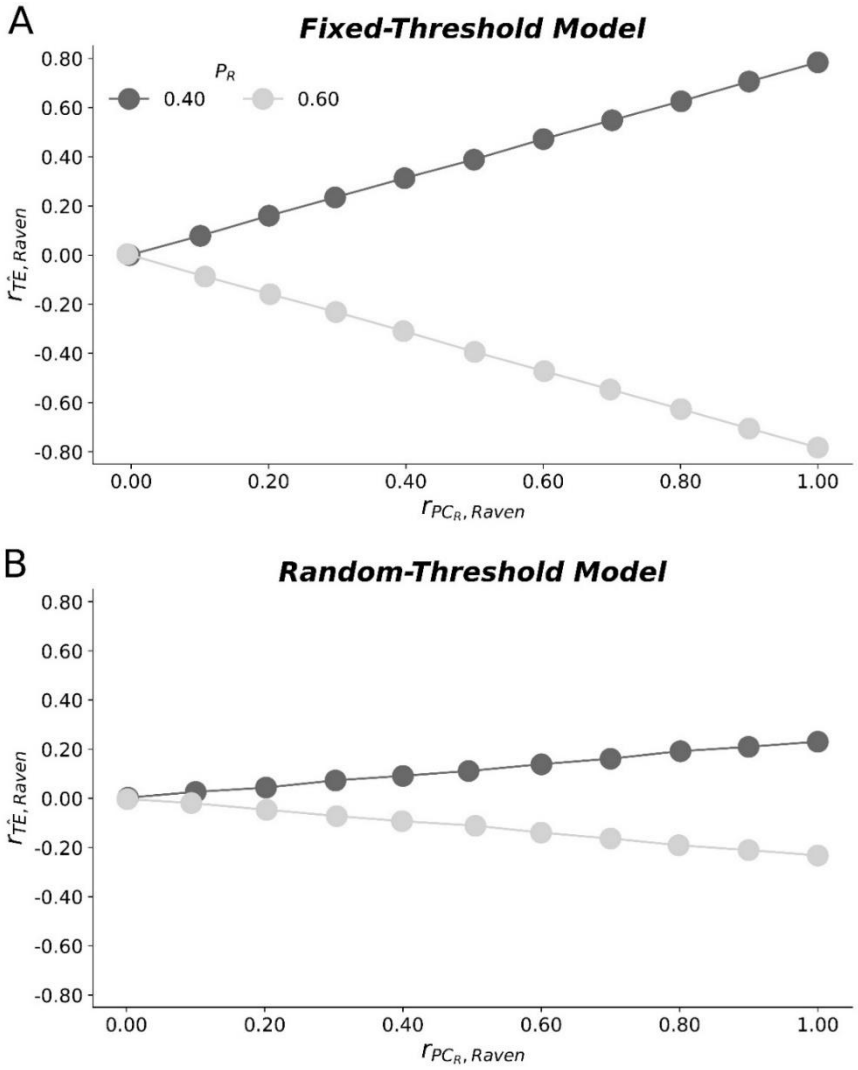
Essentially, Equation E1 consists of a *copula approach* for deriving joint distributions, given the marginal distributions (Trivedi & Zimmer, 2005). It forces Raven to have a correlation with PC_R approximately equal to the desired value of $r_{PC_R,Raven}$ (which changed in each scenario). Then, $\widehat{PC_T}$ and $\widehat{TE}$ were estimated with Equations 4.4 and 4.5, respectively. Finally, $r_{\widehat{TE},Raven}$ was computed using the Pearson correlation coefficient.

Figure E1, panel A, depicts the predicted correlations between the retrieval practice effect and the Raven scores from these simulations. As depicted in the Figure E1, panel A, the fixed-threshold model makes two key predictions about the relationship between Raven scores and the retrieval practice effect. Firstly, this correlation increases as the correlation between PC_R and Raven increases. Secondly, the fixed-threshold model predicts positive correlations when $P_R = .40$, but it predicts negative correlations when $P_R = .60$. These results agree with Rickard (2020)

statement, indicating that the correlation between an individual-difference variable and the retrieval practice effect is a joint function of P_R and $r_{PC_R,Raven}$.

Figure E1

Predictions Derived from the Fixed-Threshold and the Random-Threshold Models



Note. PC_R = observed proportion correct in the rereading condition. P_R = probability of successful recall in the final test for a randomly chosen item in the rereading condition (ideal participant). *r* = Pearson correlation coefficient. $\widehat{TE}$ = predicted retrieval practice effect. The fixed-threshold model assumes a fixed response threshold across participants, whereas the random-threshold model allows response threshold to vary across participants.

We term the second model derived from the dual-memory framework as the *random-threshold model*, which relaxes the fixed-threshold assumption. The second set of simulations mirrored the first, with one exception: A participant-specific threshold, t_i, was sampled for each participant from a normal distribution centered at the critical z-value associated with $1 - P_R$, $t_i \sim Normal(critical\ z_{1-P_R}, 1)$. This allowed thresholds to vary across participants. The fixed- and the random-threshold models can be thought as akin to the fixed-effect and random-effects models of meta-analysis, respectively (Hedges & Vevea, 1998; Lima & Buratto, 2023a). That is, by keeping the threshold t fixed across participants in the fixed-threshold model, deviations of each participant's PC_R from P_R are purely due to random error. On the other hand, by allowing t to vary across participants, it is implicitly assumed that each participant's PC_R estimates a true, specific P_R for that participant.

Figure E1, panel B, depicts the predicted correlations between the retrieval practice effect and the individual-difference variable based on the second set of simulations. As can be seen in the Figure E1, panel B, the random-threshold model makes qualitatively similar predictions than the fixed-threshold model. The main difference lies in the magnitude of the values associated with predictions. In particular, the random-threshold model predicts more modest correlations between an individual-difference variable and the magnitude of the retrieval practice effect.

What if a study yields values that differ from the values for average performance in rereading and correlation presented here? Our approach in Chapter 4 suggests that it is simply a case of simulating this scenario and comparing it with the empirical data. Our take-home message here is that the dual-memory framework has the advantage of being, at least in principle, able to describe how different empirical patterns would emerge from the complex interaction between characteristics of learners, materials, and tasks.