University of Brasília

Institute of Psychology

Department of Basic Psychological Processes

Postgraduate Program in Behavioral Sciences

Master's Thesis

Retrieval Practice and Item Difficulty: A Normative and an Experimental Study

by

Marcos Felipe Rodrigues de Lima

Brasília, July 22nd, 2019

Universidade de Brasília

Instituto de Psicologia

Departamento de Processos Psicológicos Básicos

Programa de Pós-Graduação em Ciências do Comportamento

Dissertação de Mestrado

Prática de Recuperação e Dificuldade do Item: Um Estudo Normativo e um Estudo Experimental

Marcos Felipe Rodrigues de Lima

Brasília, 22 de julho de 2019

Retrieval Practice and Item Difficulty: A Normative and an Experimental Study

by

Marcos Felipe Rodrigues de Lima

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Behavioral Sciences in the Postgraduate Program in Behavioral Sciences at the University of Brasília (Area of research concentration: Cognition and Behavioral Neurosciences).

Supervisor: Luciano Grüdtner Buratto, Ph.D.

Brasília, July 22nd, 2019

Thesis Committee:

Luciano Grüdtner Buratto, Ph.D. (President)

Postgraduate Program in Behavioral Sciences

University of Brasília - UnB

Antônio Jaeger, Ph.D. (External member)

Postgraduate Program in Psychology: Cognition and Behavior

Federal University of Minas Gerais - UFMG

Goiara Mendonça de Castilho, Ph.D. (Internal member)

Postgraduate Program in Behavioral Sciences

University of Brasília - UnB

Ricardo Basso Garcia, Ph.D. (Substitute member)

Postgraduate Program in Behavioral Sciences

University of Brasília - UnB

Brasília, July 22nd, 2019

Acknowledgments

Agradeço ao professor Luciano Grüdtner Buratto, pelas orientações, preocupações, incentivo, e por incutir em mim o desejo de realizar um trabalho de excelência. Fico feliz por trabalharmos juntos desde a graduação, por seus muitos ensinamentos e por estender nossa parceria até o mestrado. Planejar e implementar um projeto de pesquisa não é uma missão fácil. Tê-lo como meu orientador tornou essa missão mais agradável, instrutiva e encorajadora. Muito obrigado!

Também deixo meus sinceros agradecimentos aos professores Antônio Jaeger, Goiara Mendonça de Castilho e Ricardo Basso Garcia, por gentilmente aceitarem participar de minha banca examinadora. Suas contribuições certamente serão fundamentais para mim e para a melhoria dos estudos aqui reportados.

Agradeço aos professores e às professoras do Programa de Pós-Graduação em Ciências do Comportamento; aos servidores, aos terceirizados e aos estagiários, que atuam dentro e fora do Instituto de Psicologia, cujo trabalho é fundamental para o funcionamento da Universidade. Minha gratidão aos servidores Daniel Milke, Daniel Oliveira e Rodolfo Santos, pela prestatividade e bom atendimento de sempre. Em especial, registro meus agradecimentos ao Breno dos Santos e à Ester dos Santos, estagiários do LIPSI, que foram sempre gentis e solícitos, dando-me todo o suporte necessário nos bastidores das coletas de dados.

Dezenas de outros professores tiveram importância em minha formação. Uma série deles merece um registro especial. Agradeço à professora Virgínia Turra, por ter influenciado fortemente minha decisão de vir para a UnB; à professora Carla Antloga, pelos gentis e bem humorados atendimentos acadêmicos de sempre; e aos professores Marina Kohlsdorf e Gerson Janczura, por terem influenciado meu interesse por psicologia cognitiva com suas excelentes aulas. Registro aqui também a minha gratidão àqueles professores com quem tive o privilégio de trabalhar diretamente, em monitorias, estágios ou pesquisas: Denise Fleith,

Fabio Iglesias, Gardênia Abbad, Josemberg Andrade, Laércia Vasconcelos e Sérgio Oliveira. Por fim, agradeço ao Jairo E. Borges-Andrade, por quem possuo profundo carinho e admiração. Obrigado a todos vocês por terem me introduzido ao mundo da ciência psicológica!

Aos membros e ex-membros de meu grupo de pesquisa, Cadu Klier, Cláudia Pietrobon, Giulia Melo, Igor Santos, Ivan Grebot, Jonathan Jones, Júlia Valle, Nadia Alcoragi, Nathani Ramos, Ricardo Rocha, e Vitória Dias. Obrigado pelas contribuições dadas durante as reuniões e pelas trocas que me proporcionaram ao longo de todo o nosso período de convívio. Em especial, Tatiana Litvin, obrigado pela contribuição nas reuniões e na fase inicial de coletas de dados. Também agradeço à Júlia Feminella e ao Sebastião Venâncio, que realmente "compraram" a ideia desse projeto, participaram das coletas e foram verdadeiros companheiros de pesquisa! Vocês são lindos e eu tive muito orgulho de trabalhar com vocês, que também são responsáveis pelo produto final desta dissertação. Agradeço à Gabriela Iwama, por seu carinho, preocupação e pelas boas conversas que sempre tivemos. Admiro muito sua competência! Agradeço à Juliana de Deus, pelos vários anos de amizade e por ter me incentivado a fazer pesquisa em psicologia cognitiva. Obrigado por ter insistido tanto! E, é claro, deixo o meu enorme agradecimento à Beatriz Cavendish, a minha cara-metade acadêmica, pela parceria, por sentar ao meu lado nas reuniões e por nossas conversas sobre psicologia da memória, sobre ideias de projetos e sobre coisas da vida, em geral. Obrigado pelos puxões de orelha, obrigado pelas importantes contribuições em meus textos e obrigado por fazer parte dessa história!

A UnB é um lugar incrível. Aqui conheci muitas (muitas!) pessoas que se tornaram importantes para mim, desde a graduação até o mestrado. Fiz amigos e colegas de disciplinas e de pesquisa muito competentes. Agradeço a cada um de vocês por tornarem a minha caminhada mais leve: Adalberto Costa, Amanda Calmon, Amanda Cordeiro, Ana Karolina

Costa, Bruna França, Clara Dias, Daniele Paiva, Darlene Cruz, Desirée Américo, Elis Martins, Emilly Lima, Érika Vieira, Estéfane Andriny, Ezequiel Ruiz, Fernanda Drummond, Guilherme Novaes, Helena Miguez, Isabella Levino, João Moreira, Leandro Moreira, Leonardo Martins, Letícia Versiani, Lisa Miranda, Luana Veiga, Lucas Heiki, Lucas Marengo, Marcelo Lima, Maria Julia Bueno, Maria Luiza Barbosa, Maria Luíza Rodrigues, Maria Paula Fernandes, Mariana Rodrigues, Marina Barros, Marina Bittar, Mateus Fabrício, Matheus Damascena, Matheus Montalvão, Mikaelly Araújo, Miriã Carvalho, Nicolly Magrin, Raphaella Christine, Ravena Bufolo, Renata Musa, Romulo Lima, Sidney Rodrigues, Simone Cassiano, Sueli Martins, Suzane Garcia, Teresa Clara, Tiago Cunha, Victória Palmerston, Víthor Rosa Franco e Vitória Lima. Em especial, agradeço à Aline Freitas, por ser uma amiga tão linda, querida e por me tirar de minha rotina de vez em quando; à Any Esther, por me proporcionar as melhores conversas filosóficas da vida, por seu acolhimento e por ter segurado a minha barra em momentos difíceis; ao Augusto de Carolina, por sua preocupação, cuidado e pelas parcerias em busca de alguns trocados; à Caroline Feital, por sempre manifestar sua torcida pelas minhas conquistas e por estar comigo em momentos fundamentais; à Charlise Albrecht, por ter alegrado meus dias com seu sorriso, senso de humor e empolgação pelas minhas nerdices; à Débora Gramkow, por sua presença diária, gentileza e paciência; e à Lara Perícoli, por trazer doçura e leveza à minha rotina com sua maravilhosa companhia.

Agradeço também aos amigos e colegas cujas relações foram construídas ao longo da vida. Aos meus amigos e companheiros de república, Carlos Biagolini Jr., Eduardo Fernando, Lucas Damásio, Michel Garcia, Ranier Cardoso, Vitor Alves e Wellington Araújo. Obrigado pelas conversas e por compartilharem um pouco de minha rotina! Às pessoas queridas espalhadas pelo DF, Brasil e mundo: Ana Luisa Amaral, Ana Gabriela Sandoval, Beatriz Neves, Bruna Andreia, Carolina Maria, Cinthia Rocha, Denyse Furuhashi, Dominique

Schellnock, Franciele Leal, Gabriel Sales, Geovana Sales, Heloísa Adlung, Jhany Araújo, Joneilton Araújo, Kaio Bussmann, Khamilla Batista, Leonardo Belquiman, Lilian Santos (e família), Lucas Polo, Marcelo Koba, Natália Apolônio, Paula Bolzan, Pedro Leme, Pedro Peton, Rafael Felix, Ricardo Cury, Richard do Valle, Rodrigo Leme, Sarah de Sousa, Simone Lima, Thallita Oliveira e Wagner Santos. Em especial, ao Bruno Boretti, amigo de longa data, cuja distância jamais foi um impedimento para fazermos parte da vida um do outro.

Agradeço à família Sales, por ter me dado suporte desde que cheguei a Brasília. Em especial, faço esses agradecimentos para Débora, Pia e Raquel, por sempre me darem suporte, sobretudo em meu primeiro ano de capital.

Meus pais, David Rodrigues de Lima e Miryam de Freitas, deram-me suporte instrumental, emocional, carinho e torcida. Agradeço-os por isso. Também agradeço à minha irmã, Mirella Lima, pelo companheirismo e pela amizade, que torna nossa ligação maior do que apenas por laços sanguíneos. Vocês três são fundamentais em minha vida. Amo muito vocês e espero ser capaz de retribuir um pouco daquilo que fizeram e continuam fazendo por mim. Também agradeço a todos os meus parentes, tia Mônica e tio Dito, tia Maria, tia Márcia, tio Maurício, Vando e Fátima, Luzia e Chiquinho, bem como suas respectivas famílias. Obrigado por torcerem por mim!

Por fim, concluo agradecendo a dois atores fundamentais nesses dois anos de caminhada. O primeiro é o CNPq, a quem agradeço pelo apoio financeiro, que foi muito importante para que eu conseguisse implementar meu projeto. O segundo se refere, de fato, a um grupo: Os participantes de pesquisa, a quem agradeço sinceramente por me cederem, de forma generosa, um pouco de seu tempo.

A curious peculiarity of our memory is that things are impressed better by active than by passive repetition. I mean that in learning (by heart, for example), when we almost know the piece, it pays better to wait and recollect by an effort within, than to look at the book again. If we recover the words the former way, we shall probably know them the next time; if in the latter way, we shall likely need the book once more.

William James

Table of Contents

List of Figures

Manuscript 1

Manuscript 2

List of Tables

Manuscript 1

Manuscript 2

Appendix C

List of Symbols and Abbreviations

α	Cronbach's alpha (internal consistency index) and significance level.
ANCOVA	Analysis of covariance.
ANOVA	Analysis of variance.
β	Standardized coefficient.
b	Unstandardized coefficient.
BCa CI	Bias-corrected and accelerated bootstrapping confidence interval.
BDI	Beck Depression Inventory.
$\boldsymbol{BF_{10}}$	Bayes Factor.
$\mathbf{C_1}$	Cycle 1.
$\mathbf{C_2}$	Cycle 2.
$\mathbf{C_3}$	Cycle 3.
CI	Confidence interval.
d	Cohen's *d* (measure of effect size on *t*-tests).
f	Cohen's *f* (one of several measures of effect size on ANOVAs).
F	*F*-ratio (ANOVA's statistic).
g	Hedges's *g* (an adjusted measure of effect size, commonly reported in meta-analyses).
$\boldsymbol{H_0}$	Null hypothesis.
$\boldsymbol{H_1}$	Alternative hypothesis.
hr	Hour.
ITI	Intertrial interval.
JOLs	Judgments of learning.
κ	Cohen's kappa (measure of inter-rater agreement).
LMM	Linear mixed model.

M	Sample mean.
Mdn	Median.
min	Minutes.
$\boldsymbol{\eta_p^2}$	Partial eta-squared (one of several measures of effect size on ANOVAs).
NILC	Interinstitutional Center for Computational Linguistics (Núcleo Interinstitucional de Linguística Computacional).
OR	Odds ratio.
p	*p*-value.
r	Pearson's *r* (Pearson product-moment correlation coefficient).
REH	Retrieval effort hypothesis.
RT	Reaction time.
s	Seconds.
SAM	Self-Assessment Manikin.
SD	Sample standard deviation.
SE	Standard error.
STAI	State–Trace Anxiety Inventory.
t	*t*-statistic.
$\boldsymbol{\omega_p^2}$	Partial omega-squared (one of several measures of effect size on ANOVAs).
WIC	Written Informed Consent.
$\boldsymbol{\chi^2}$	Chi-square statistic (measure of goodness-of-fit in the likelihood-ratio test).
z	*z*-score (standardized score) and Wilcoxon signed-rank test's statistic.
Z	Wald statistic.

Abstract

Retrieval practice promotes long-term retention. In fact, retrieving information by testing improves retention more than repeated study of that same information, a phenomenon known as the *retrieval practice effect*. One account suggests that tests involve greater cognitive effort than restudy, and that such additional effort explains the memory benefits afforded by testing. If this *retrieval effort hypothesis* (REH) is correct, then difficult study items (which require more retrieval effort) should benefit more from retrieval practice than easy study items (which require less retrieval effort). Here we tested this prediction by using item memorability as an estimate of item difficulty. First, we conducted a normative study (Manuscript 1) to obtain item difficulty estimates. In Study 1, participants judged 80 Portuguese words for familiarity, concreteness, valence, arousal and 80 corresponding Swahili words for wordlikeness (similarity to Portuguese); in Study 2, participants underwent three study–test cycles on Swahili–Portuguese associations. Multiple regressions showed that familiarity, wordlikeness, and previous memorability predicted current item memorability. The word pairs normed in these two studies were then used in two retrieval practice experiments (Manuscript 2). After the initial study of easy and difficult items, participants repeatedly restudied half of the pairs and retrieval practiced the other half. In both experiments, we replicated the retrieval practice effect and the item difficulty effect. More importantly, we also found (a) a smaller retrieval practice effect for difficult items (Experiment 1) and, after controlling for practice-phase recall levels, (b) a (non-significant) trend toward a greater retrieval practice effect for difficult items, particularly for positive testers (Experiment 2). The mixed results provide only weak evidence for the REH and are discussed in relation to alternative accounts of the retrieval practice effect.

Keywords: retrieval practice, testing effect, retrieval effort, desirable difficulties, cued recall

Prática de Recuperação e Dificuldade do Item: Um Estudo Normativo e um Estudo Experimental

Marcos Felipe Rodrigues de Lima

Orientador: Luciano Grüdtner Buratto

Resumo

A prática de recuperação de informações da memória promove a retenção em longo prazo. De fato, recuperar informações por meio de testagem melhora a retenção mais que o estudo repetido dessa mesma informação, um fenômeno conhecido como *efeito de prática de recuperação*. Uma teoria sugere que testes envolvem maior esforço cognitivo que o reestudo, e que tal esforço adicional explica os benefícios de memória proporcionados pela testagem. Se essa *hipótese de esforço de recuperação* (HER) estiver correta, então itens de estudo difíceis deveriam se beneficiar mais da prática de recuperação que itens de estudo fáceis. Aqui, testou-se essa predição usando a memorabilidade do item como uma estimativa de sua dificuldade. Primeiro, conduzimos um estudo normativo (Manuscrito 1) para obter estimativas de dificuldade do item. No Estudo 1, os participantes julgaram o grau de familiaridade, concretude, valência e alerta de 80 palavras em português, bem como a *wordlikeness* (grau de similaridade com palavras em português) de suas 80 palavras suaíli correspondentes; no Estudo 2, os participantes realizaram três ciclos de estudo–teste de associações suaíli–português. Regressões múltiplas mostraram que familiaridade, *wordlikeness* e a memorabilidade prévia predizem a memorabilidade do item. Os pares de palavras normatizados nesses dois estudos foram então usados em dois experimentos de prática de recuperação (Manuscrito 2). Depois do estudo inicial de items fáceis e difíceis, os participantes repetidamente reestudaram metade dos pares e recuperaram a outra metade. Em ambos os experimentos, nós replicamos o efeito de prática de recuperação e o efeito de

dificuldade do item. Mais importante, observou-se (a) um menor efeito de prática de recuperação para itens difíceis (Experimento 1) e, depois de controlar os níveis de recordação na fase de prática, (b) uma tendência (não significativa) em direção a um maior efeito de prática de recuperação para itens difíceis, especialmente para participantes que se beneficiaram da testagem (Experimento 2). Os resultados mistos fornecem somente evidências fracas para a HER, sendo discutidos à luz de teorias alternativas do efeito de prática de recuperação.

Palavras-chave: prática de recuperação, efeito de testagem, esforço de recuperação, dificuldades desejáveis, recordação com pistas

Resumo Expandido

Recuperar a informação por meio de testagem melhora a retenção mais que o estudo repetido da mesma informação, fenômeno conhecido como *efeito de prática de recuperação* (Whiffen & Karpicke, 2017). A *hipótese de esforço de recuperação* (HER) sugere que testes envolvem maior esforço cognitivo que o reestudo, e que tal esforço adicional explica os benefícios mnemônicos proporcionados pela testagem (Pyc & Rawson, 2009). Se a HER está correta, então itens de estudo difíceis (que exigem maior esforço de recuperação) deveriam se beneficiar mais da prática de recuperação que itens de estudo fáceis (que exigem menor esforço de recuperação). Esta dissertação descreve dois manuscritos. No *Manuscrito 1*, um estudo normativo de pares de palavras suaíli–português permitiu estimar diferentes características dos estímulos. No *Manuscrito 2*, usando um subconjunto desses pares, dois experimentos testaram uma predição da HER.

Manuscrito 1: Normas de Familiaridade, Concretude, Valência, Alerta, *Wordlikeness* e Memorabilidade para Pares de Palavras Suaíli–Português

Introdução

Estudos normativos permitem a estimativa de características dos estímulos e, em consequência disso, um melhor controle em pesquisas experimentais futuras. Os estudos aqui relatados tiveram quatro objetivos: (1) obter estimativas de familiaridade, concretude, valência e alerta para um conjunto único de palavras do português brasileiro; (2) estimar a *wordlikeness* (grau de similaridade com palavras em português) de um conjunto de palavras estrangeiras (suaíli); (3) estimar a memorabilidade de pares de palavras suaíli–português, a partir do desempenho em uma tarefa de recordação com pistas; e (4) avaliar se alguma medida estimada nos objetivos (1) e (2) prediz a memorabilidade dos itens, estimada no objetivo (3).

Método

Estimativas das características de pares de palavras suaíli–português foram obtidas por meio de dois estudos. No Estudo 1, os participantes julgaram sequencialmente o grau de familiaridade, concretude, valência e alerta de 80 palavras em português. Em seguida, julgaram a *wordlikeness* de suas 80 palavras suaíli correspondentes. No Estudo 2, os participantes realizaram três ciclos de estudo–teste de associações suaíli–português, a partir do qual foram obtidas estimativas da memorabilidade desses pares. Regressões múltiplas usando as estimativas obtidas nos dois estudos permitiram investigar se alguma característica das palavras é preditora da memorabilidade dos pares suaíli–português.

Resultados e Discussão

O conjunto de estímulos apresentou alta familiaridade e concretude (palavras em português) e alta *wordlikeness* (palavras em suaíli). Padrões de correlações e análises de fidedignidade indicaram a consistência dos julgamentos realizados. Os pares de palavras retiveram sua memorabilidade relativa ao longo dos ciclos de estudo–teste, de maneira similar ao observado em estudos prévios (e.g., Nelson & Dunlosky, 1994). Regressões múltiplas indicaram que as variáveis preditoras da memorabilidade foram a *wordlikeness* (no ciclo 1), a familiaridade e a memorabilidade no ciclo 1 (no ciclo 2) e a memorabilidade no ciclo 2 (no ciclo 3).

Manuscrito 2: A Dificuldade do Item Afeta a Magnitude do Efeito de Prática de Recuperação? Uma Avaliação da Hipótese de Esforço de Recuperação

Introdução

No contexto da prática de recuperação, a HER tem recebido suporte empírico por meio de manipulações da dificuldade da tarefa de recuperação (e.g., Carpenter & DeLosh, 2006). No entanto, resultados mistos emergem quando a variável manipulada é a dificuldade do item

(e.g., Carpenter, 2009; Vaughn, Rawson, & Pyc, 2013). Os experimentos aqui reportados investigaram as seguintes questões: A dificuldade do item afeta a magnitude do efeito de prática de recuperação? Em caso positivo, quais itens se beneficiam mais da prática de recuperação: fáceis ou difíceis? Buscou-se (a) replicar o efeito de prática de recuperação e o efeito de dificuldade do item; e (b) investigar se a dificuldade do item afeta a magnitude do efeito de prática de recuperação.

Método

Em dois experimentos, após o estudo inicial de pares de palavras suaíli–português (fáceis e difíceis), os participantes repetidamente reestudaram metade dos pares e repetidamente fizeram prática de recuperação da outra metade. No Experimento 1, quatro ciclos de prática foram realizados para itens fáceis e difíceis; no Experimento 2, itens difíceis foram praticados por dois ciclos adicionais. Dois dias depois, os participantes retornaram ao laboratório e realizaram um teste de recordação com pistas de todo material previamente estudado.

Resultados e Discussão

Em ambos os experimentos, replicaram-se os efeitos de prática de recuperação (Whiffen & Karpicke, 2017) e de dificuldade do item (Cull & Zechmeister, 1994). No Experimento 1, observou-se um maior efeito de prática de recuperação para itens fáceis, ao contrário do que prediz a HER (Pyc & Rawson, 2009). No entanto, identificou-se que, na fase de prática, a aprendizagem foi superior para itens fáceis. No Experimento 2, após igualar os níveis iniciais de recordação, observou-se uma tendência (não significativa) de maior efeito de prática de recuperação para itens difíceis. Em uma análise subsequente, restrita a participantes que se beneficiaram da testagem no Experimento 2, essa tendência atingiu significância estatística. Em ambos os experimentos, análises condicionais indicaram que itens fáceis são mais prováveis de serem recordados no teste final. Esses resultados mistos fornecem somente

evidências fracas em favor da HER e são discutidos à luz de teorias explicativas do efeito de prática de recuperação.

Considerações Finais da Dissertação

A ausência de evidência forte apoiando a HER indica que é possível implementar a prática de recuperação em contextos educacionais e de reabilitação cognitiva com materiais mais difíceis. De uma perspectiva teórica, é importante que se relacione o construto esforço a outros processos engajados durante a prática de recuperação e que são úteis para a memória. Estudos futuros se beneficiarão de investigações sobre os mecanismos cognitivos subjacentes aos efeitos positivos de certos tipos de dificuldades, como o espaçamento (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), relacionando-os tanto à dificuldade do item como a variáveis de diferenças individuais (Unsworth, 2019). Esta dissertação possui relevância (a) metodológica, pois introduziu uma série de medidas normativas de pares de palavras suaíli–português, até então inexistentes no Brasil; (b) teórica, pois testou predições da HER por meio de novos procedimentos, enfatizando a dificuldade do item; e (c) aplicada, na medida em que investigou se o efeito de prática de recuperação também está presente em materiais de maior dificuldade, uma informação de fundamental importância para professores e estudantes, dada a ampla gama de dificuldade conceitual dos materiais ensinados em salas de aula.

Referências

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., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect.

Memory & Cognition, *34*(2), 268–276. Retrieved from https://link.springer.com/journal/13421

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory & Cognition*, *22*, 249–257. Retrieved from https://link.springer.com/journal/13421

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, *2*(3), 325–335. https://doi.org/10.108/09658219408258951

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

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

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review*, *20*, 1239–1245. https://doi.org/10.3758/s13423-013-0434-z

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

General Introduction

Teachers deliberately administer closed-book tests, in order to assess how much students have learned in a given knowledge domain. In addition to assessing learning however, tests also enhance later retention of tested and related information (Roediger & Karpicke, 2006). Since the benefits of testing are driven "by the retrieval processes that learners engage in when they take tests" (Karpicke, 2017, p. 487), this phenomenon has been referred to as the *retrieval practice effect* (Whiffen & Karpicke, 2017).

Retrieval practice is an effective learning strategy for a wide range of learner characteristics, learning conditions, study materials, and criterion tests (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). With a growing body of literature demonstrating that this is a robust effect, theories have been proposed in order to describe (e.g., Halamish & Bjork, 2011) and to account for (e.g., Carpenter, 2009, 2011; Karpicke, Lehman, & Aue, 2014) the retrieval practice effect (for a comprehensive review, see Karpicke, 2017).

It seems established in the literature that difficult tests tend to produce greater memory benefits than easy ones (e.g., Pyc & Rawson, 2009). An influential idea related to this finding is the *desirable difficulties framework*, which posits that greater memory enhancement are derived from learning conditions that require more learner effort (Bjork, 1994). A testable instance from this idea is the *retrieval effort hypothesis* (REH), which states that difficult but successful retrievals will produce more benefits for memory than easy, successful retrievals (Pyc & Rawson, 2009). The assumption is that manipulation of retrieval difficulty will be reflected in the level of *cognitive effort* undertaken by the learner. In attentional capacity models, cognitive effort is understood as the proportion used of a limited capacity processing central, which can be allocated in a highly flexible manner (see Kahneman, 1973; Tyler, Hertel, McCallum, & Ellis, 1979). Thus, cognitive effort is a construct that mediates the

relationship between task difficulty (i.e., the type of practice) and capacity, on the one hand, and performance, on the other hand (Shenhav et al., 2017).

In order to test predictions of REH, it is possible to design experiments that manipulate either retrieval task difficulty (e.g., Pyc & Rawson, 2009) or item difficulty (e.g., Vaughn, Rawson, & Pyc, 2013). Item difficulty estimates can be obtained from norms, which present measures of memorability of the to-be-learned material obtained from representative samples (e.g., Nelson & Dunlosky, 1994). While normed items have been used in several retrieval practice studies (e.g., Pyc & Rawson, 2009, 2010), few investigations have used these normative measures as an independent variable (e.g., Minear, Coane, Boland, Cooney, & Albat, 2018; Vaughn et al., 2013).

Although the REH has gained empirical support in investigations that have manipulated retrieval task difficulty (e.g., Pyc & Rawson, 2009), mixed results emerged when item difficulty was the manipulated variable (Carpenter, 2009; Vaughn et al., 2013). The aim of this Master's thesis is to further explore the impact of item difficulty on the retrieval practice effect. Specifically, we investigate the following research question: Does item difficulty affect the magnitude of the retrieval practice effect? If so, which items benefit the most from retrieval practice: easy or difficult ones?

Following the guidelines of the Postgraduate Program in Behavioral Sciences and the guidelines of the American Psychological Association (2010), this Master's thesis was written in a manuscript format to be submitted to scientific journals and it was organized in four main sections. After this *General Introduction*, the *Manuscript 1* describes a normative study of Swahili–Portuguese word pairs – stimuli that can be used in future studies by memory researchers. Then, the *Manuscript 2* is presented, describing two experiments that investigated predictions of REH. Finally, the *Final Considerations* integrate the contribution of both manuscripts. The studies reported here were previously approved by the Research Ethics

Committee (Appendix A) and participants gave written informed consent before starting the tasks (Appendix B).

We believe this thesis has methodological, theoretical, and applied relevance. It has methodological relevance, as it introduces a series of normative measures of Swahili–Portuguese word pairs, previously non-existent in Brazil. In addition, it has theoretical and empirical relevance, as it tests predictions of the REH through novel variations from procedures adopted in previous studies. Finally, this thesis has applied relevance, as it investigates whether item difficulty affects the magnitude of the retrieval practice effect, a fundamental piece of information to teachers and students alike given the wide range of conceptual difficulty in the materials taught in the classroom.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: APA.

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

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. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547–1552. https://doi.org/10.1037/a0024140

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

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

Kahneman, D. (1973). *Attention and effort.* Upper Saddle River, NJ: Prentice Hall.

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). Oxford: Academic Press.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 237–284). San Diego, CA: Elsevier Academic Press.

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

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, *2*(3), 325–335. https://doi.org/10.108/09658219408258951

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

Roediger, H. L., III, & Karpicke, J. D. (2006). 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

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*, 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 607–617. https://doi.org/10.1037/0278-7393.5.6.607

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review, 20*, 1239–1245. https://doi.org/10.3758/s13423-013-0434-z

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

Manuscript 1: Norms of Familiarity, Concreteness, Valence, Arousal, Wordlikeness, and Memorability for Swahili–Portuguese Word Pairs

Norms of Familiarity, Concreteness, Valence, Arousal, Wordlikeness, and Memorability for Swahili–Portuguese Word Pairs

Marcos Felipe Rodrigues de Lima and Luciano Grüdtner Buratto

University of Brasília

Author Note

Marcos Felipe Rodrigues de Lima and Luciano Grüdtner Buratto, Department of Basic Psychological Processes, Institute of Psychology, University of Brasília.

This manuscript is based on a Master's thesis to-be-submitted to the University of Brasília by the first author under the supervision of the second author. The study was supported by the National Council for Scientific and Technological Development (CNPq). We thank Carlos Eduardo D. C. Lage (stimulus translation), Sebastião V. Pereira-Jr. and Tatiana Litvin (data collection), Carlos H. Biagolini-Jr. (data analysis), and Beatriz A. Cavendish (data collection, response scoring, and comments on a previous version of this manuscript).

Correspondence concerning this article should be addressed to Marcos Felipe Rodrigues de Lima, Departamento de Processos Psicológicos Básicos, Instituto de Psicologia, Universidade de Brasília, Campus Darcy Ribeiro, ICC Sul, Sala ASS-012/5, CEP 70.910-900, Brasília, DF, Brasil. E-mail: lima.piraju@gmail.com

Abstract

Normative studies allow the estimation of stimulus characteristics and better control in experimental research. The studies reported here had four aims: (1) to obtain estimates of familiarity, concreteness, valence, and arousal for a single set of words in Brazilian Portuguese; (2) to estimate the wordlikeness (similarity to Portuguese) of a set of foreign words (Swahili); (3) to estimate the memorability of Swahili–Portuguese word pairs from cued-recall performance; and (4) to assess if any measure predicts the memorability of the items. Sixty-one participants took part in one of two studies. In Study 1, participants judged 80 Portuguese words for familiarity, concreteness, valence, arousal and 80 corresponding Swahili words for wordlikeness; in Study 2, participants underwent three study–test cycles on Swahili–Portuguese associations. Results showed that the stimulus set was high in familiarity and concreteness (Portuguese) and high in wordlikeness (Swahili). Word pairs also retained their relative degree of memorability over the course of the study–test cycles. Multiple regressions showed that familiarity, wordlikeness, and previous memorability predicted current item memorability. These norms can benefit memory research conducted in Portuguese, particularly research on retrieval-based learning with foreign–native language cued-recall paradigms.

Keywords: word pairs, concreteness, emotionality, familiarity, cued-recall

Norms of Familiarity, Concreteness, Valence, Arousal, Wordlikeness, and Memorability for Swahili–Portuguese Word Pairs

Normative studies are common in cognitive psychology because they allow us to estimate with more precision the characteristics of the stimuli used in empirical studies (e.g., Janczura, Castilho, Rocha, Van Erven, & Huang, 2007; Nelson & Dunlosky, 1994). Normed stimuli can, for example, be distributed in different experimental conditions in a balanced manner, increasing internal validity of the experiments (e.g., Pyc & Rawson, 2009). Norms also allow us to evaluate how specific aspects of the stimuli affect performance (e.g., Jia et al., 2016). In fact, certain stimuli properties have a great impact on their *memorability*, here understood as how likely an item is to be retrieved in a given memory task. In free recall tests, participants recall more concrete words than abstract ones (Witherby & Tauber, 2017), more familiar words than unfamiliar ones (Jia et al., 2016), and more emotional words than neutral words (Johnson & MacKay, 2019).

In recent years, norms for several word characteristics, such as emotionality (Kristensen, Gomes, Justo, & Vieira, 2011; Oliveira, Janczura, & Castilho, 2013), concreteness (Janczura et al., 2007), and free association (Janczura, Castilho, Keller, & Oliveira, 2017) have been produced for Brazilian Portuguese. Norms for frequency-of-occurrence of words in prose texts – an indirect index of objective familiarity – are also available for Brazilian Portuguese (Núcleo Interinstitucional de Linguística Computacional [NILC], 2005). Norms of familiarity ratings – an indirect index of subjective familiarity – are available for European Portuguese (Leitão, Figueira, & Almeida, 2010). These two indexes are strongly correlated (Balota, Pilotti, & Cortese, 2001), but are based on distinct sources of information (written vs. spoken). Although some studies have evaluated different word characteristics for the same stimulus set (e.g., Janczura et al., 2017; Janczura et al., 2007; Oliveira et al., 2013), none of them have evaluated familiarity ratings for a set of stimuli in

Brazilian Portuguese. Since our interest was on several word characteristics for a single stimulus database, we sought to fill this gap. Thus, the first aim of this study was to obtain estimates for familiarity, concreteness, valence, and arousal for a single set of words in Brazilian Portuguese.

The available norms in Brazilian Portuguese allow researchers both to control and to manipulate specific aspects of words, in different memory tasks, such as free recall and recognition. Another task commonly used in memory studies is *cued recall*, in which the experimenter provides cues to the participant at the time of testing (Baddeley, Eysenck, & Anderson, 2015). Tasks that use paired-associates (e.g., foreign–native word pairs) fall into this category. In a cued-recall task, initially both elements of the pair are presented to the participant (study phase). Next, in the test phase, only the first element of the pair is provided to the participant (i.e., the *cue*), who must respond with the second (i.e., the *target*; Wilson & Criss, 2017). When the to-be-learned material is foreign–native word pairs, one factor that can influence their memorability is *wordlikeness*, the extent to which a sound sequence is typical in words in the learner's native language (Gomes, Mendes, Silva, Esteves, & Gomes, 2015). Given this possible influence, the second aim of this study was to estimate the wordlikeness of a set of foreign words.

Normative measures of memorability have been obtained for Swahili–English (Nelson & Dunlosky, 1994), English–Swahili (Bangert & Heydarian, 2017), and Lithuanian–English word pairs (Grimaldi, Pyc, & Rawson, 2010). However, to date, no study has produced normative measures of memorability for word pairs in which either cue or target, or both are in Brazilian Portuguese. In order to fill this gap, the third aim of this study was to estimate the memorability of Swahili–Portuguese word pairs in a cued-recall task. Swahili is a language spoken in East African countries, and its use in memory studies is based on a series of arguments posed by Nelson and Dunlosky (1994) that support Swahili's suitability as a

potential source of stimuli. Following Nelson and Dunlosky's reasoning, we chose Swahili because: (a) native speakers of Brazilian Portuguese are unlikely to have been exposed to words in Swahili. This ensures that learners know little about the to-be-learned material (Bjork & Kroll, 2015); (b) like Brazilian Portuguese, Swahili's writing is based on the Latin alphabet. Thus, in an experimental task, the learner is not burdened with the additional demand of having to learn new symbols from the foreign language; (c) Swahili words are unlikely to produce floor effects on memory tasks, allowing additional learning in the multitrial learning paradigm (for similar arguments, see Nelson & Dunlosky, 1994).

In previous studies, researchers have attempted to relate how word variables predict their memorability (e.g., Bangert & Heydarian, 2017; Nelson & Dunlosky, 1994) and here we also pursue this goal. The fourth aim of this study was to investigate if any of the assessed measures predicts item memorability. The use of foreign–native word pairs allows the investigation of both second-language acquisition and vocabulary learning, which are relevant topics for applied areas such as School and Educational Psychology.

In sum, the studies reported here had four aims: (1) to obtain estimates of several characteristics for a single set of words in Brazilian Portuguese; (2) to estimate wordlikeness of a set of foreign words; (3) to estimate the memorability of Swahili–Portuguese word pairs from a cued-recall task; and (4) to investigate if any measure predicts the memorability of the items. To achieve our proposed aims, two studies were conducted. In Study 1, estimates of familiarity, concreteness, valence, arousal, and wordlikeness were collected. In Study 2, estimates of memorability were collected. Finally, we ran cross-study regression analyses to investigate if any measure predicts item memorability.

Method

Participants

Sixty-one participants were recruited for these studies. The sample had the following characteristics: (a) Study 1: 21 participants; 52.4% female; age range = 18–23 years (M = 19.38, SD = 1.28); (b) Study 2: 40 participants; 55% female; age range = 17–28 years (M = 20.60, SD = 2.85). Data collection of both studies was conducted simultaneously. Each individual took part in only one of the two studies. Participants gave written informed consent before starting the tasks. Research was approved by the Research Ethics Committee before data collection.

Instruments

Word pairs. Eighty Swahili–Portuguese word pairs were selected and adapted from Nelson and Dunlosky's (1994) norms. Additionally, six pairs were used both as examples during instructions (Study 1) or as filler items (Study 2). Stimuli were presented on a computer screen and stimulus presentation was controlled with PsychoPy (Peirce, 2007). Word pairs were presented in white color centered in a black screen (lowercase, boldface Arial 18 font). When both Swahili and Portuguese words were simultaneously presented (only in Study 2), the Swahili word was always presented on top.

Beck Depression Inventory (BDI). Self-report questionnaire of depression intensity with 21 items ranging from 0 to 3 points and Cronbach's α ranging from .70 to .92 (Cunha, 2001).

State–Trait Anxiety Inventory (STAI). Self-report questionnaire of transitional and dispositional aspects of anxiety (state and trait, respectively) with 40 items ranging from 1 (*absolutely not* or *almost never*) to 4 (*very much* or *often*; Biaggio & Natalicio, 1979).

Procedure

In both studies, participants were tested individually in a single session. They

answered the BDI and STAI questionnaires and then proceeded to the main task.[1] Order was counterbalanced across participants. As we did not find order effects, we will ignore this factor later. Below we describe separately the procedure for each study.

Study 1. Figure 1 shows a schematic representation of the judgment task. The label of the judgment and the word to be judged were presented at the top and center of the screen, respectively. A scale was presented at the bottom along with labels identifying the extreme values. Participants first practiced the task with a non-normed word.

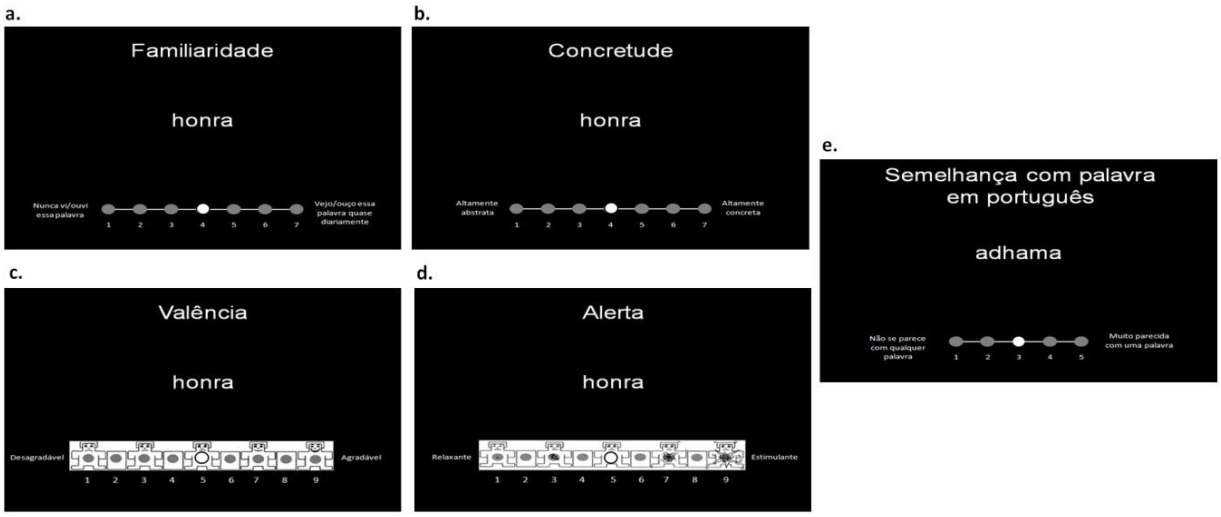

Figure 1. Schematic representation of judgments made in Study 1. In the first phase, participants judged 80 Portuguese words for (a) familiarity, (b) concreteness, (c) valence, and (d) arousal. One word was judged per time. After the last word was judged, in the second phase, participants judged 80 Swahili words for (e) wordlikeness. At the top of each screen appears judgment label. At the center of each screen appears to-be-judged word. At the bottom of each screen appears judgment scale, with labels identifying the extreme values of the scale (see text to details).

Study 1 was divided into two phases. In the first phase, participants judged 80

[1] Both instruments were applied in order to exclude participants with higher depression and anxiety scores, since judgments from such participants could bias our normative measures.

Portuguese words, presented in a random order. Each word was always judged in the same fixed order: familiarity, concreteness, valence, and arousal. We chose to begin trials with a familiarity judgment to avoid familiarity overestimation due to prolonged exposure to the word. A new word was presented only after the participant judged the previous word for all aforementioned characteristics. After concluding judgment of all 80 Portuguese words, in the second phase, participants made judgments about the wordlikeness of 80 Swahili words. In both phases, there was no time limite on participant's responses, although they were instructed to work fast. At the midpoint of each scale there was a red circle that could be moved either to the left or to the right. Participants should select the point in the scale that best represented their response and press the "Enter" key to confirm it. The estimated time for completion of all tasks was 40 minutes.

Familiarity ratings ranged from 1 (*I never saw/heard that word*) to 7 (*I see/hear that word almost daily*) corresponding to how unfamiliar or familiar they considered each word. Concreteness ratings ranged from 1 (*highly abstract*) to 7 (*highly concrete*) corresponding to how abstract or concrete they considered each word. Valence was assessed with the Self-Assessment Manikin scale (SAM; Kristensen et al., 2011) ranging from 1 (*negative emotional valence*) to 9 (*positive emotional valence*) corresponding to how unpleasant or pleasant they considered each word. Arousal was also assessed with the SAM scale, ranging from 1 (*relaxing*) to 9 (*exciting*) corresponding to how relaxed or aroused they considered each word. Wordlikeness ranged from 1 (*Not like a word at all*) to 5 (*Very like a word*) corresponding to how similar a Swahili word is from any Brazilian Portuguese word.

Study 2. The 80 Swahili–Portuguese word pairs were divided into two lists, each with 40 pairs. The six filler pairs were included in both lists. Thus, each list comprised 46 word pairs, and each participant was exposed to only one list. In each list, the word pairs were divided into three sets, the first with the six filler pairs and the other two with twenty pairs

each. Filler items were added to control for possible primacy effects (i.e., better recall for items at the beginning of the study list). Word pairs were split in sets of twenty items to control for lag effects (i.e., better recall for items presented closer in time). We made sure that there were at least 26 word pairs between study and test of a given word pair. Within each set, presentation occurred in random order. From the participants' point of view, only one set was studied, since there was no indication of the end of one set and the beginning of the other set.

In the multitrial learning paradigm, participants performed three study–test cycles without feedback. Each cycle was composed of a study block and a test block, as depicted in Figure 2. Each block started with a brief instruction (e.g., *Study Block: Try to learn the association between Swahili and Portuguese words.*) In a study block trial, participants saw a word pair (e.g., *wingu–cloud*) for 10 s and were asked to learn the association between that pair (Figure 2a). In a test block trial, participants saw only the Swahili word (e.g., *wingu*) and were asked to recall its meaning in Portuguese by typing the corresponding word on a keyboard. A trial ended either after participants pressed the "Enter" key or after 10 s, regardless of whether a response was given or not (Figure 2b). On both study and test blocks, intertrial interval (ITI) was 1 s. The estimated time for completion of all tasks was 60 minutes.

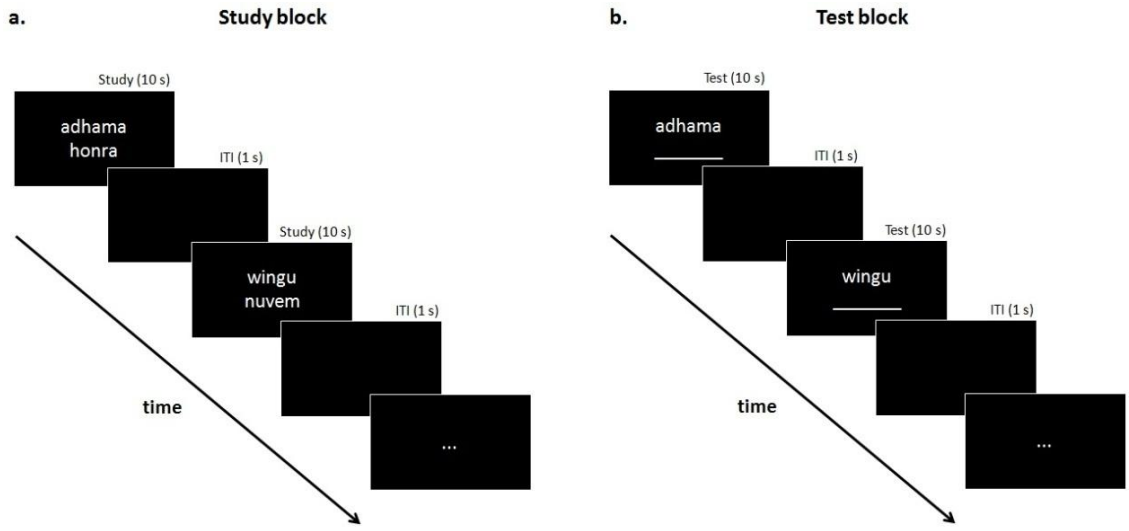

Figure 2. Schematic representation of (a) a study block and (b) a test block of the multitrial learning paradigm. ITI corresponds to intertrial interval (see text to details).

Statistical Analyses

Initial screening. Participants with extreme scores on BDI and STAI questionnaires were excluded. Extreme scores were defined as those participants with *z*-scores > 2.24 (1.25% of the area under the normal curve) and were assessed separately for each study.

Study 1. Means and standard deviations for each judgment and for each word were computed. To check the reliability of these judgments, we carried out a participant-based, instead of an item-based, split-half procedure (for similar analyses, see Janczura et al., 2007; Oliveira et al., 2013). We randomly split the sample into two subsamples and correlated the average word judgments for these subsamples (separately for each judgment). As the observed coefficient may vary depending on the way the sample is split (see Hutz, Bandeira, & Trentini, 2015), we used a bootstrapping approach (number of samples = 1,000) in order to overcome this limitation. This allowed us to compute average Pearson's *r*s and their corresponding bootstrapping confidence intervals (CIs), which increased the precision of our reliability estimates.

Study 2. Two judges independently rated participants' answers. Since the focus of these norms was on memorability, typing and spelling errors were not counted as errors. Answers of eight participants (20% of sample) were rated by both judges. Cohen's kappa (κ) was computed to assess the level of agreement between judges. Next, we computed *z*-scores for average performance across the three test blocks. We excluded from subsequent analyses *z*-scores > ± 2.50 (see Hair, Black, Babin, & Anderson, 2014, p. 65). The main analysis was carried out on the proportion of participants who correctly recalled the Portuguese word across the three test cycles.

Studies 1 and 2. We ran three multiple linear regression models to investigate if any measure predicts item memorability. Familiarity, concreteness, valence, arousal, wordlikeness, and word length (Swahili and Portuguese) were entered into each model as

predictors with the stepwise method. For each one of three models, the dependent variables were the proportion of participants that correctly recalled a given word in the test blocks of cycles 1, 2, and 3 (hereafter, C_1, C_2, and C_3, respectively). Additionally, performance on C_1 was entered into Model 2 as a predictor, and both performance on C_1 and C_2 were entered into Model 3 as predictors.

Results

Excluded Cases in the Initial Analyses

Seven participants were excluded. Two participants were excluded because they were positive outliers on BDI, three were positive outliers on STAI, one was a positive outlier in the multitrial learning paradigm, and one participant chose the same response on all judgments, indicating lack of compliance. Thus, subsequent analyses were based on 18 cases in Study 1 and on 36 cases in Study 2.

Study 1

Table 1 (left) shows the means and standard deviations of participants' judgments across all dimensions for the 80 normed words. In general, the words in our sample had high familiarity (M = 5.61, SD = 0.78). Familiarity estimates were correlated with the base 10 logarithm of frequency-of-occurrence per million words (NILC, 2005), r = .63, p < .001. Words also were high on concreteness (M = 5.33, SD = 1.52), which is related to the sensory experience with the meaning of a given word. Words like *dog* and *cheese* have referents in the world that can be experienced from the senses. On the other hand, *honor* and *flavor* are abstract ideas, without a specific sensory match in the world. Scores given by participants reflect this property of the words.

Table 1

Means and Standard Deviations of Words Judgments and Proportion of Participants that Correctly Recalled Them in Each Test Cycle

Swahili–Portuguese word pair		Familiarity		Concreteness		Valence		Arousal		Wordlikeness		Memorability		
(English translation in brackets)		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	C_1	C_2	C_3
wakili	agente [agent]	5.06	1.21	4.83	2.01	5.33	1.14	5.67	1.41	1.61	0.85	.05	.37	.53
mshoni	alfaiate [tailor]	4.11	1.81	5.28	1.84	5.56	0.98	4.17	1.50	1.83	1.29	.12	.47	.59
pamba	algodão [cotton]	5.89	0.96	6.39	0.78	7.00	1.53	3.06	1.47	4.39	0.50	.12	.35	.59
roho	alma [soul]	6.22	1.44	1.89	1.88	7.28	1.45	4.22	2.21	2.78	1.48	.35	.76	.88
lozi	amêndoa [almond]	5.11	1.41	5.94	1.73	6.44	1.34	4.17	1.50	2.56	1.62	.00	.18	.47
nanga	âncora [anchor]	4.72	1.27	6.22	1.35	5.89	1.37	4.78	1.35	4.72	0.46	.12	.24	.47
zeituni	azeitona [olives]	5.78	1.17	6.78	0.55	6.00	2.83	5.11	2.40	3.78	1.48	.71	.88	.88
ndoo	balde [bucket]	6.00	1.03	6.78	0.65	4.89	0.96	4.72	1.32	1.39	0.50	.16	.42	.68
mashua	barco [boat]	5.50	1.34	6.50	1.04	6.56	1.69	3.67	2.11	2.06	1.30	.11	.32	.53
pipa	barril [barrel]	4.33	1.28	6.22	0.94	5.06	1.30	5.11	0.96	5.00	0.00	.53	.71	1.00
punda	burro [donkey]	5.94	0.94	3.44	2.04	3.00	1.46	6.72	1.71	4.61	0.50	.42	.79	.89
leso	cachecol [scarf]	5.44	1.25	6.39	0.92	7.39	1.33	3.06	1.70	4.11	1.23	.24	.47	.59
mbwa	cachorro [dog]	6.61	0.70	6.61	0.78	7.44	2.06	4.22	2.96	1.06	0.24	.26	.68	1.00
maiti	cadáver [corpse]	4.83	1.62	5.22	1.96	1.83	1.25	7.44	1.29	2.50	1.20	.11	.37	.68
kaa	caranguejo [crab]	4.78	1.35	6.61	0.78	5.39	1.58	5.44	1.34	1.56	1.15	.16	.47	.74
farasi	cavalo [horse]	5.50	1.25	6.94	0.24	6.89	1.45	5.17	2.18	2.56	1.42	.12	.35	.71

(Table 1 continues)

Table 1. continuation

Swahili–Portuguese word pair		Familiarity		Concreteness		Valence		Arousal		Wordlikeness		Memorability		
(English translation in brackets)		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	C_1	C_2	C_3
pombe	cerveja [beer]	6.33	0.77	6.83	0.51	4.56	3.35	5.44	2.43	4.17	1.10	.35	.71	.88
elimu	ciência [science]	6.61	0.61	2.67	1.97	6.78	1.96	6.72	2.08	2.56	1.29	.12	.53	.76
ambo	cola [glue]	5.39	1.09	5.78	1.40	4.89	0.68	5.22	1.22	3.83	1.15	.00	.24	.53
godoro	colchão [mattress]	6.56	0.70	6.78	0.55	8.06	1.35	2.28	2.05	3.00	1.33	.00	.53	.88
chakula	comida [food]	6.89	0.47	6.33	1.37	8.61	0.98	5.44	3.24	2.28	1.23	.11	.42	.74
kamba	corda [rope]	5.39	1.33	6.28	1.23	5.28	1.36	5.78	1.73	3.83	1.20	.26	.32	.42
pazia	cortina [curtain]	5.78	1.00	6.56	0.70	6.33	1.50	3.83	1.72	3.83	1.04	.16	.47	.63
desturi	costume [custom]	5.94	1.11	2.50	1.95	5.78	1.52	4.94	1.76	4.39	0.78	.11	.37	.42
talaka	divórcio [divorce]	5.00	1.78	3.89	2.22	2.83	1.86	6.78	1.59	1.67	0.69	.11	.47	.47
iktisadi	economia [economy]	5.94	1.00	2.89	1.94	4.61	2.12	6.06	2.10	1.28	0.46	.05	.26	.47
gharika	enchente [flood]	5.00	1.53	5.39	1.82	1.56	0.86	7.50	1.15	1.83	0.99	.06	.24	.65
bahasha	envelope [envelope]	5.44	1.46	6.56	0.98	5.50	1.04	4.67	1.50	2.56	1.46	.11	.21	.32
samadi	estrume [manure]	3.94	1.86	5.72	1.67	2.28	1.49	6.67	1.81	2.44	1.34	.11	.16	.47
ankra	fatura [invoice]	5.83	1.20	5.50	2.04	2.50	1.47	7.28	1.41	2.78	1.52	.06	.18	.41
jeraha	ferida [wound]	5.39	1.54	5.17	2.12	2.17	1.25	7.39	1.38	2.22	1.17	.06	.29	.47
jani	folha [leaf]	6.61	0.61	6.61	0.85	6.89	1.60	4.00	2.14	2.67	1.37	.05	.47	.63
joko	forno [kiln]	6.17	0.99	6.33	1.19	6.39	1.42	5.11	1.75	4.00	1.33	.05	.37	.53

(Table 1 continues)

Table 1. continuation

Swahili–Portuguese word pair		Familiarity		Concreteness		Valence		Arousal		Wordlikeness		Memorability		
(English translation in brackets)		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	C_1	C_2	C_3
hadithi	história [story]	6.44	0.78	2.50	1.58	6.67	1.68	5.44	2.23	1.50	0.79	.05	.26	.53
adhama	honra [honor]	5.67	1.08	1.78	1.48	6.61	2.55	5.22	2.78	2.33	1.46	.26	.42	.58
adui	inimigo [enemy]	5.56	1.50	3.78	2.39	1.89	0.90	7.50	1.15	1.89	0.90	.00	.59	.71
bustani	jardim [garden]	6.22	0.81	6.22	1.11	7.89	1.13	2.94	2.34	2.50	1.34	.00	.26	.53
goti	joelho [knee]	5.72	1.45	6.61	1.04	5.78	1.11	4.50	1.38	3.44	1.34	.35	.47	.71
yamini	juramento [oath]	5.06	1.26	3.00	2.00	5.89	1.28	5.72	1.56	2.56	1.46	.16	.26	.32
ziwa	lago [lake]	6.17	1.04	6.11	1.60	7.06	1.66	3.94	2.60	1.89	1.23	.11	.32	.42
buu	larva [maggot]	4.72	1.64	6.33	1.08	2.56	1.69	7.00	1.64	2.44	1.29	.26	.58	.74
wasaa	lazer [leisure]	6.39	0.98	4.11	2.19	8.72	0.57	3.67	2.54	1.56	0.70	.24	.65	.88
hamira	levedura [yeast]	3.61	1.61	4.72	2.02	5.00	1.19	5.06	1.55	2.44	1.38	.00	.21	.32
tumbili	macaco [monkey]	5.56	1.04	6.56	0.78	6.06	1.51	4.89	1.45	2.39	1.38	.00	.37	.58
inda	malícia [spite]	5.22	1.17	2.28	1.45	4.06	2.24	6.67	1.33	4.17	1.04	.06	.29	.53
embe	manga [mango]	6.00	1.50	6.61	1.04	6.61	2.25	4.11	2.25	2.39	1.33	.05	.42	.58
tabibu	médico [doctor]	6.11	0.96	5.89	1.28	5.50	2.28	5.61	2.30	2.06	1.11	.11	.37	.58
nafaka	milho [corn]	6.22	1.00	6.72	0.67	7.61	1.61	4.50	2.41	2.94	1.39	.06	.65	.82
fumbo	mistério [mystery]	5.78	1.22	1.78	1.06	5.44	1.54	7.00	1.53	4.17	0.92	.18	.29	.41
theluji	neve [snow]	4.78	1.73	6.11	1.45	7.22	1.59	4.39	2.52	1.56	0.70	.12	.47	.76

(Table 1 continues)

Table 1. continuation

Swahili–Portuguese word pair		Familiarity		Concreteness		Valence		Arousal		Wordlikeness		Memorability		
(English translation in brackets)		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	C_1	C_2	C_3
wingu	nuvem [cloud]	6.33	1.14	5.56	1.65	7.50	1.10	2.94	1.76	2.11	1.13	.29	.65	.76
yatima	órfão [orphan]	4.89	1.57	4.83	2.07	2.00	1.24	7.00	1.24	3.83	1.42	.16	.47	.68
mfupa	osso [bone]	5.94	1.00	6.39	1.50	5.06	1.06	5.39	1.04	1.39	0.50	.12	.47	.76
yai	ovo [egg]	6.56	0.98	6.78	0.55	7.33	1.33	4.22	1.90	3.06	1.63	.53	.95	.95
kasuku	papagaio [parrot]	5.39	1.24	6.89	0.47	6.78	1.11	4.89	1.88	1.94	1.26	.11	.37	.58
lulu	pérola [pearl]	4.50	1.47	5.61	1.33	6.67	1.37	4.50	2.15	4.11	1.23	.47	.82	.88
vumbi	poeira [dust]	6.28	0.89	5.67	1.68	1.94	1.11	6.78	1.48	4.22	1.06	.12	.24	.59
utenzi	poema [poem]	5.89	1.13	4.17	1.92	7.67	1.24	3.50	2.50	3.56	1.46	.11	.21	.37
lango	portão [gate]	6.22	0.88	6.83	0.51	5.39	0.85	5.22	0.81	4.06	0.94	.05	.26	.53
sahani	prato [plate]	6.72	0.57	6.83	0.51	6.89	1.91	4.56	2.09	2.00	1.46	.21	.37	.47
adha	problema [trouble]	6.78	0.55	3.00	2.11	1.83	1.29	8.06	1.11	2.00	1.19	.12	.41	.71
nabii	profeta [prophet]	5.06	1.55	3.11	1.41	5.61	2.00	5.72	2.14	2.11	1.18	.18	.53	.76
pafu	pulmão [lung]	5.56	1.29	6.33	1.50	5.61	1.75	5.11	2.08	3.22	1.26	.06	.41	.65
jibini	queijo [cheese]	6.56	1.04	7.00	0.00	8.44	1.04	4.50	3.11	2.28	1.49	.26	.42	.53
malkia	rainha [queen]	5.39	1.38	5.00	1.50	6.22	1.66	4.94	1.92	2.17	1.15	.47	.76	.94
duara	roda [wheel]	6.00	0.77	5.22	2.02	6.06	1.43	4.89	1.45	3.28	1.36	.06	.41	.53
ladha	sabor [flavor]	6.06	0.87	2.33	1.33	7.11	1.64	4.61	2.30	2.39	1.14	.00	.05	.37

(Table 1 continues)

Table 1. continuation

Swahili–Portuguese word pair		Familiarity		Concreteness		Valence		Arousal		Wordlikeness		Memorability		
(English translation in brackets)		*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	C_1	C_2	C_3
ruba	sanguessuga [leech]	4.06	1.66	5.61	1.82	2.22	1.52	7.17	1.95	3.44	1.04	.11	.21	.53
chura	sapo [frog]	5.28	1.49	6.56	0.92	3.72	1.45	6.44	1.38	4.00	1.03	.12	.59	.71
hariri	seda [silk]	5.22	1.56	5.67	1.64	7.11	1.45	3.83	2.12	1.78	0.94	.11	.21	.53
chama	sociedade [society]	6.56	0.78	3.28	2.08	5.11	1.60	6.72	1.99	5.00	0.00	.29	.71	.94
usingizi	sono [sleep]	6.94	0.24	3.83	2.50	5.50	3.17	3.17	2.79	1.72	0.96	.18	.53	.71
rushwa	suborno [bribe]	4.89	1.28	3.89	1.84	2.33	1.61	7.11	1.23	1.22	0.43	.06	.35	.47
zulia	tapete [carpet]	6.17	1.38	6.61	0.85	6.72	1.71	3.67	1.97	4.06	1.11	.26	.42	.63
dafina	tesouro [treasure]	5.22	1.56	4.78	1.77	7.83	1.04	5.17	2.66	3.94	1.21	.37	.68	.84
gutu	toco [stump]	4.67	1.08	4.44	1.85	4.50	1.29	5.28	1.36	3.44	1.46	.06	.41	.71
nyanya	tomate [tomato]	6.28	1.07	6.94	0.24	6.94	2.10	4.39	1.88	1.61	0.85	.41	.88	.94
handaki	trincheira [trench]	3.50	1.42	5.17	1.92	3.11	1.57	6.67	1.33	1.89	1.02	.00	.16	.26
kaburi	túmulo [grave]	4.72	1.53	5.89	1.41	2.22	1.40	6.50	1.69	1.78	1.06	.06	.35	.47
fagio	vassoura [broom]	6.22	0.81	6.89	0.47	5.11	1.64	5.17	1.76	3.33	1.28	.06	.35	.53

Average indexes of valence and arousal were slightly above the midpoint of the scale (valence: M = 5.48, SD = 1.91; arousal: M = 5.23; SD = 1.30). We found a strong negative correlation between these two measures, r = –.83, p < .001. This result indicates that words that have a higher valence (i.e., positive) tend to be more relaxing (e.g., *mattress*), whereas words with a lower valence (i.e., negative) tend to be more arousing (e.g., *corpse*). Nonetheless, there were exceptions, such as *science*, that despite having a positive valence is also considered an arousing word (participants were sampled from a university setting, so this effect can be population-specific.) There were no cases showing the opposite pattern (i.e., negatively-valenced, but relaxing, words).

Average wordlikeness scores were slightly below midpoint of the scale (M = 2.79, SD = 1.04). Scores appear to have apparent validity, since the words *chama* and *pipa*, two Swahili words that are homonymous to words in Portuguese, were the only ones which approached ceiling (Ms = 5.00, SDs = 0.00). In contrast, *mbwa*, *rushwa*, and *iktisadi* had the lowest wordlikeness scores. These words have consonant clusters that do not occur in Brazilian Portuguese, thereby presenting an objective index of low wordlikeness.

We found other significant correlations between measures: concreteness and arousal, r = –.36, familiarity and arousal, r = –.32, and familiarity and valence, r = .45 (all ps ≤ .004). Finally, average Pearson's r suggests reliability of judgments for familiarity, r = .86, CI 95% [.82, .90], concreteness, r = .93, CI 95% [.90, .95], valence, r = .94, CI 95% [.90, .95], arousal, r = .84, CI 95% [.75, .89], and wordlikeness, r = .89, CI 95% [.85, .92].

Study 2

Agreement between judges. Judges showed high level of agreement, κ = .98, p < .001, which can be considered almost perfect (Landis & Koch, 1977). As to rating answers as hits or errors, judges disagreed on only one out of 1,104 answers.

Memorability. In this study, memorability was operationally defined as the proportion of participants who correctly recalled a given word in the test blocks of C_1, C_2, and C_3. As an example, consider the following word pairs: 76% of participants recalled the target *cloud*, given the cue *wingu*, on the test block of C_3; on the other hand, only 58% of them recalled the target *honor*, given the cue *adhama*. Thus, based on the definition of memorability adopted here, it can be said that *wingu–cloud* has greater memorability than *adhama–honor*, after three study–test cycles. Table 1 (right) shows the memorability indexes of Swahili–Portuguese word pairs across the three study–test cycles. In only two cases (*pipa–barrel* and *mbwa–dog*), memorability index approached ceiling. We found a moderate-to-strong positive correlation between memorability in C_1 and C_2, $r = .78$, C_1 and C_3, $r = .65$, and C_2 and C_3, $r = .88$, all ps $< .001$.

Studies 1 and 2: Multiple Regressions

In the Model 1 (memorability in C_1), the wordlikeness was the only significant predictor, $\beta = .28$, $b = 0.039$, CI 95% [0.009, 0.070], $SE = .02$, $p = .011$, accounting for 8% of variance in memorability in C_1. In the Model 2 (memorability in C_2), the memorability in C_1 was a significant predictor, $\beta = .76$, $b = 1.01$, CI 95% [0.83, 1.20], $SE = .09$, $p < .001$. Familiarity was also a significant predictor, $\beta = .17$, $b = 0.04$, CI 95% [0.008, 0.076], $SE = .02$, $p = .02$. Taken together, these two variables explained 63.5% of variance in memorability in C_2. Finally, the memorability in C_2 was the only significant predictor in the Model 3 (memorability in C_3), $\beta = .88$, $b = 0.82$, CI 95% [0.72, 0.92], $SE = .05$, $p < .001$, accounting for 77.5% of variance in memorability in C_3.

Discussion

The use of word pairs in studies has been common in cognitive psychology literature (e.g., Pyc & Rawson, 2009). Knowing the characteristics of these stimuli and how they affect memorability is important for experimental planning. The two studies reported here present

estimates of familiarity, concreteness, valence, and arousal for words in Portuguese, and wordlikeness for words in Swahili. In addition, we obtained memorability estimates for Swahili–Portuguese word pairs. Finally, we assessed whether the estimated measures predict item memorability.

We found convergent validity for the familiarity measure through its moderate-to-strong correlation with a related construct (i.e., frequency-of-occurrence of words in prose texts). This result replicates a previous finding of strong correlation between familiarity and frequency measures (Balota et al., 2001). We identified similar patterns for the concreteness measure in the present and in a previous study (Janczura et al., 2007), suggesting that participants were able to discriminate concrete from abstract words. The negative relationship between valence and arousal found in the present study replicates the pattern reported in a previous normative study (see Oliveira et al., 2013). Lastly, the moderate-to-strong positive relationship of memorability index in C_1, C_2, and C_3 suggests that pairs tend to retain their relative degree of memorability over the course of the study–test cycles (for similar patterns, see Bangert & Heydarian, 2017; Grimaldi et al., 2010; Nelson & Dunlosky, 1994).

The multiple regression results indicate that none of the predictor variables can alone explain memorability across the three cycles. Although wordlikeness and familiarity were both significant predictors in memorability in C_1 and C_2, respectively, the coefficient values were low. These low values imply that a unit change in wordlikeness and in familiarity cause only a small change in memorability. For C_2 and C_3, memorability in the previous cycle is the most important predictor. Wordlikeness explained only 8% of variance in Model 1, possibly through its facilitating effect on word processing and recognition (Gomes et al., 2015). This suggests that native words that look like foreign words can serve as an association cue to improve performance on memory tasks.

Unlike Nelson and Dunlosky (1994) and Grimaldi et al. (2010), which found weak but significant correlations (r = .25 in both studies) between frequency-of-occurrence and recall in C_1, we did not observe the same pattern in the present study (r = –.008, p = .94). Nelson and Dunlosky argued that extra-experimental objective familiarity could contribute to their association with foreign words. In the present study, only familiarity had predictive power (in Model 2). Nelson and Dunlosky (1994) and Grimaldi et al. (2010) did not collect subjective familiarity ratings. Thus, our results raise the hypothesis that familiarity is more important than objective frequency to explain memorability.

Some limitations of the present study should be noted. First, the sample size used here was lower than usual in normative studies (Bangert & Heydarian, 2017; Grimaldi et al., 2010; Janczura et al., 2007; Kristensen et al., 2011; Leitão et al., 2010; Nelson & Dunlosky, 1994; Oliveira et al., 2013). Despite this, our reliability analyses using bootstrapping CIs showed satisfactory results. The bootstrapping approach allowed us to overcome the limitation of the traditional split-half procedure (i.e., potential biases in correlation coefficient related to the way that sample was split; see Hutz et al., 2015). Second, due to sample homogeneity (i.e., mostly undergraduate students), some estimates may be idiosyncratically related to the population sampled. Third, due to Study 2's design, it was not possible to investigate the role of the list context in which a given item was inserted on the memorability of the item. One possibility for future studies is to create random lists of items to assess whether the list context in which an item appears contributes significantly to the increase or decrease of the memorability of the item.

This study makes an important methodological contribution, as it introduces a series of normative measures of Swahili–Portuguese word pairs, previously non-existent in Brazil. These norms will enable the design and implementation of language and memory studies (e.g., cued-recall studies) in Brazilian Portuguese with sets of controlled stimuli.

Memorability, for example, can be used as an index of item difficulty for investigations of various phenomena, like the *retrieval practice effects*, the finding that retrieving previously studied information is better for long-term retention than its restudy (Karpicke, 2017). Retrieval effort is one of the moderators of these effects (Pyc & Rawson, 2009). The present norms can be used to assess whether low memorability items are benefited more strongly from active retrieval than high memorability items, an important test for the retrieval effort hypothesis of the retrieval practice effects. An interconnected line of research involves pupil size, a physiological index of mental effort (Mathôt, 2018), and its relationship with item memorability and retrieval effort. The stimuli normed here can be used in pupillometric studies. Based on retrieval effort hypothesis (Pyc & Rawson, 2009), it is predicted that low memorability items will elicit larger pupil sizes than high memorability items during retrieval practice, but not during restudy.

These norms may also be useful in applied settings relevant for School and Educational Psychology. Vocabulary learning is an important pre-requisite for school achievement. The stimulus set normed in this study can be used, for example, to design and implement tightly controlled studies in second-language vocabulary learning (Bjork & Kroll, 2015).

References

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2015). *Memory* (2nd ed.). London: Psychology Press.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition, 29*(4), 639–647. Retrieved from https://link.springer.com/journal/13421

Bangert, A. S., & Heydarian, N. M. (2017). Recall and response time norms for English–Swahili word pairs and facts about Kenya. *Behavior Research Methods*, *49*, 124–171. https://doi.org/10.3758/s13428-015-0701-1

Biaggio, A. M. B., & Natalicio, L. (1979). *Manual para o Inventário de Ansiedade Traço–Estado (IDATE) [Manual for the State–Trait Anxiety Inventory (STAI)]*. Rio de Janeiro: Centro Editor de Psicologia Aplicada-CEPA.

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, *128*, 241–252. Retrieved from https://www.jstor.org/journal/amerjpsyc

Cunha, J. A. (2001). *Manual da versão em português das escalas Beck [Portuguese version of the Beck's scale manual]*. São Paulo: Casa do Psicólogo.

Gomes, C. A., Mendes, S. C., Silva, M. B., Esteves, C. O., & Gomes, G. C. (2015). Efeito de wordlikeness no processamento de não-palavras por falantes do português brasileiro. [Effect of wordlikeness in processing of nonwords by Brazilian Portuguese speakers]. *Revista de Estudos da Linguagem, 23*(1), 195–210. https://doi.org/10.17851/2237-2083.23.1.195-210

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian–English

paired associates. *Behavior Research Methods*, *42*(3), 634–642. https://doi.org/10.3758/BRM.42.3.634

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Harlow, Essex: Pearson Education Limited.

Hutz, C. S., Bandeira, D. R., & Trentini, C. M. (Orgs.). (2015). *Psicometria [Psychometrics]*. Porto Alegre: Artmed.

Janczura, G. A., Castilho, G. M., Keller, V. N., & Oliveira, N. R. (2017). Free association norms for 1,004 Brazilian Portuguese words. *Psicologia: Teoria e Pesquisa*, *32*(n.esp.), 1–7. https://doi.org/10.1590/0102-3772e32ne23

Janczura, G. A., Castilho, G. M., Rocha, N. O., Van Erven, T. J. C., & Huang, T. P. (2007). Normas de concretude para 909 palavras da língua portuguesa [Concreteness norms for 909 Portuguese words]. *Psicologia: Teoria e Pesquisa*, *23*(2), 195–204. https://doi.org/10.1590/S0102-37722007000200010

Jia, X., Li, P., Li, X., Zhang, Y., Cao, W., Cao, L., & Li, W. (2016). The effect of word frequency on judgments of learning: Contributions of beliefs and processing fluency. *Frontiers in Psychology*, *6*, 1995. https://doi.org/10.3389/fpsyg.2015.01995

Johnson, L. W., & MacKay, D. G. (2019). Relations between emotion, memory encoding, and time perception. *Cognition and Emotion*, *33*(2), 185–196. https://doi.org/10.1080/02699931.2018.1435506

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). Oxford: Academic Press.

Kristensen, C. H., Gomes, C. F. A., Justo, A. R., & Vieira, K. (2011). Brazilian norms for the Affective Norms for English words. *Trends in Psychiatry and Psychotherapy*, *33*(3), 135–146. https://doi.org/10.1590/S2237-60892011000300003

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Leitão, J. A. G., Figueira, A. P. C., & Almeida, A. C. F. (2010). Normas de imaginabilidade, familiaridade e idade de aquisição para 252 nomes comuns [Imageability, familiarity, and age-of-acquisition norms for 252 common words]. *Laboratório de Psicologia*, *8*(1), 101–119. https://doi.org/10.14417/lp.651

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, *1*(1), 1–23. https://doi.org/10.5334/joc.18

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, *2*(3), 325–335. https://doi.org/10.1080/09658219408258951

Núcleo Interinstitucional de Linguística Computacional. (2005). *Corpus NILC*. Retrieved from http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm

Oliveira, N. R., Janczura, G. A.,& Castilho, G. M. (2013). Normas de alerta e valência para 908 palavras da língua portuguesa [Norms of arousal and valence for 908 Portuguese words]. *Psicologia: Teoria e Pesquisa*, *29*(2), 185–200. https://doi.org/10.1590/S0102-37722013000200008

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

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

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78–88. https://doi.org/10.1016/j.jml.2017.01.006

Witherby, A. E., & Tauber, S. K. (2017). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & Cognition*, *45*, 639–650. https://doi.org/10.3758/s13421-016-0681-0

Manuscript 2: Does Item Difficulty Affect the Magnitude of the Retrieval Practice Effect? An Evaluation of the Retrieval Effort Hypothesis

Does Item Difficulty Affect the Magnitude of the Retrieval Practice Effect? An Evaluation of the Retrieval Effort Hypothesis

Marcos Felipe Rodrigues de Lima, Sebastião Venâncio Pereira Júnior, Júlia Feminella Duarte da Costa, and Luciano Grüdtner Buratto

University of Brasília

Author Note

Marcos Felipe Rodrigues de Lima, Sebastião Venâncio Pereira Júnior, Júlia Feminella Duarte da Costa, and Luciano Grüdtner Buratto, Department of Basic Psychological Processes, Institute of Psychology, University of Brasília.

This article is based on a Master's thesis submitted to the University of Brasília. The study was supported by the National Council for Scientific and Technological Development (CNPq). We thank Beatriz A. Cavendish for their helpful comments on a previous version of the article.

Correspondence concerning this article should be addressed to Marcos Felipe Rodrigues de Lima, Departamento de Processos Psicológicos Básicos, Instituto de Psicologia, Universidade de Brasília, Campus Darcy Ribeiro, ICC Sul, Sala ASS-012/5, CEP 70.910-900, Brasília, DF, Brasil. E-mail: lima.piraju@gmail.com

Abstract

Retrieving information by testing improves its subsequent retention more than restudy, a phenomenon called retrieval practice effect. According to the retrieval effort hypothesis (REH), difficult items require more retrieval effort than easier items and, consequently, should benefit more from retrieval practice. In two experiments, we tested this prediction. After the initial study of easy and difficult Swahili–Portuguese word pairs, participants repeatedly restudied half of the pairs and retrieval practiced another half. In both experiments, we replicated both the retrieval practice effect and the item difficulty effect. In Experiment 1, we found greater retrieval practice effect for easy items. However, different recall performance for easy and difficult items in the practice phase clouds the interpretation of this finding. In Experiment 2, after ensuring similar recall levels at practice, we found a (non-significant) trend toward a greater retrieval practice effect for difficult items, particularly for positive testers (i.e., participants who benefit from retrieval practice). The results provide only weak evidence for the REH and they are discussed in relation to the episodic context account and the automatization account of the retrieval practice effect.

Keywords: retrieval practice, testing effect, retrieval effort, desirable difficulties, cued recall

Does Item Difficulty Affect the Magnitude of the Retrieval Practice Effect? An Evaluation of the Retrieval Effort Hypothesis

A student has just read an introductory chapter in a cognitive psychology textbook. She plans to have another study session the next day and wonders what would be the best way to go through said chapter again in order to boost her long-term memory. Most students, when faced with a similar situation, tend to choose to reread the chapter (Karpicke, Butler, & Roediger, 2009). However, a growing body of research has shown that retrieving information by testing improves its subsequent retention more than restudy, a phenomenon called *testing effect* (Roediger & Karpicke, 2006b), also known as *retrieval practice effect* (Whiffen & Karpicke, 2017). It is assumed that tests are useful because they allow learners to engage in retrieval processes (Karpicke, 2017), which alter memory representations of the practiced items, making them more recallable in the future (Bjork, 1975, 1994).

The traditional procedure used to investigate retrieval practice effect involves three phases. After the initial study of the items (*study phase*), a *practice phase* takes place, in which learners either restudy them or perform an initial test that aims to induce the retrieval practice (Halamish & Bjork, 2011). In the *final test phase*, the learners perform a final memory test (*criterion test*) that refers to all previously studied items. Mnemonic benefits of the retrieval practice are indicated by better performance in the criterion test for previously retrieval practiced items than for restudied items. A meta-analysis indicated that, in 81% of the studies analyzed, retrieval practice led to a better performance in the criterion test than restudy (Hedges' g = 0.50; Rowland, 2014). The benefits of retrieval practice have been observed across a wide range of materials, contexts, criterion tests, and learners' characteristics (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Furthermore, retrieval practice seems promising in clinical contexts, such as in the cognitive rehabilitation of patients with aphasia (see, e.g., Middleton, Schwartz, Rawson, Traut, & Verkuilen, 2016).

Several hypotheses have been proposed to account for the retrieval practice effect, ranging from descriptive, such as the *bifurcation model* (Halamish & Bjork, 2011), to explanatory accounts, such as the *elaborative retrieval hypothesis* (Carpenter, 2009), and the *episodic context account* (Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014). These accounts differ on the purported cognitive mechanism underlying the benefits of the retrieval practice. However, they agree on the idea that an initial test involves greater *cognitive effort* than restudy. Some authors suggest that it is this effort that is responsible for the beneficial effects of retrieval practice (e.g., Bjork, 1994; Glover, 1989; Pyc & Rawson, 2009). *Effort* is usually an abstract and vaguely defined construct. Roediger and Butler (2011, p. 24), for example, argued that "retrieval effort can be thought of as an index of the amount of reprocessing of the memory trace that occurs during retrieval". In attentional capacity models, cognitive effort is understood as the proportion of processing dedicated to perform a task given a limited capacity central, which can allocate processing capacity in a highly flexible manner (Kahneman, 1973; Tyler, Hertel, McCallum, & Ellis, 1979). Although these definitions are abstract, it is understood that the allocation of effort varies between tasks, depending on the manipulation of task difficulty (Beatty, 1982; Tyler et al., 1979).

The desirable difficulties framework, an influential idea in the learning and memory literature, posits that greater memory gains are expected in conditions that require greater retrieval effort from the learner (Bjork, 1994). Such conditions include spaced practice (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), interleaved practice (Kang, 2017), and retrieval practice (Roediger & Karpicke, 2006b). In the latter case, long-term benefits of retrieval practice tend to be greater due to the fact that retrieval is, allegedly, a more difficult process (Bjork, 1975). This is exactly the core of the *retrieval effort hypothesis* (REH), a descriptive account derived from the desirable difficulties framework. The REH predicts that successful, more difficult retrievals will yield greater memory benefits than successful, easier

ones (Pyc & Rawson, 2009). Rowland's (2014) meta-analysis supported this prediction, showing that more difficult initial tests (i.e., free and cued-recall) produce greater retrieval practice effects (g = 0.81 and 0.72, respectively) than less difficult ones (i.e., recognition; g = 0.36).

Task and Item Difficulty

The difficulty of a given retrieval task has been variously operationalized by manipulating (a) the degree of informativeness of a cue at the practice phase (Carpenter & DeLosh, 2006; Fiechter & Benjamin, 2017; Finley, Benjamin, Hays, Bjork, & Kornell, 2011), (b) the time interval between successive retrieval attempts (Agarwal, Finley, Rose, & Roediger, 2017; Karpicke & Bauernschmidt, 2011; Middleton et al., 2016; Pyc & Rawson, 2009), (c) the attentional demands imposed during the practice phase (Buchin & Mulligan, 2017, 2019; Gaspelin, Ruthruff, & Pashler, 2013; Mulligan & Picklesimer, 2016), and (d) the number of times an item was required to be correctly recalled (i.e., *criterion level*; Pyc & Rawson, 2009; Vaughn & Rawson, 2011; Vaughn, Rawson, & Pyc, 2013). A series of results that indicate better performance on the criterion test for items initially tested under more difficult conditions support the REH (e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2009).

A distinct way of operationalizing difficulty is selecting items from normative studies that provide information about item difficulty. Currently, there are normed data sets for English–English (Underwood, 1982), Swahili–English (Nelson & Dunlosky, 1994), English–Swahili (Bangert & Heydarian, 2017), Lithuanian–English (Grimaldi, Pyc, & Rawson, 2010), and more recently, Swahili–Portuguese word pairs (Lima & Buratto, 2019). When the data set comprises two languages, the choice of the foreign language is based on its desirable features, namely, it should (a) be unknown or unusual to the learners, (b) share few cognates with the learners' native tongue, (c) be written in the same alphabet as the learner's native language, and (d) not produce floor and ceiling effects (Grimaldi et al., 2010; Nelson & Dunlosky,

1994). The aforementioned normative studies used a multitrial learning paradigm, in which the learner engages in a predetermined number of successive study–test cycles. In a study block, the learner must study a set of intact pairs (e.g., *wingu–cloud*), whereas in a test block, she must recall the target word (*cloud*), given the cue (e.g., *wingu–?*). The main measure in these studies is the proportion of participants who correctly recalled the target word across test blocks (Nelson & Dunlosky, 1994). This proportion provides an estimate of the relative difficulty of the pair. Given that the remaining variables are kept constant, these tests tend to reflect estimated item difficulty, with easy items leading to greater retention than difficult ones (e.g., Cull & Zechmeister, 1994).

Nelson and Dunlosky's (1994) norms have been widely used in the retrieval practice research (e.g., Karpicke & Roediger, 2008; Pyc & Rawson, 2009). Most studies have used normative measures to establish experimental control, balancing item difficulty between experimental conditions. Only a few studies have used normative measures as an independent variable (Carpenter, 2009; Karpicke, 2009; Minear, Coane, Boland, Cooney, & Albat, 2018; Vaughn et al., 2013).[2] Vaughn et al argued that at high criterion levels, retrieval practice would benefit difficult items more than easy items, since across practice cycles there would be a decrease in retrieval effort for easy items but not for difficult ones. In two experiments, they found that across several criterion levels, performance in the criterion test was better for easy items than for difficult ones, contrary to REH predictions.

Carpenter (2009) manipulated the associative strength between cue and target word pairs and, consequently, the difficulty of item retrieval. These pairs were either restudied or retrieval practiced. In two experiments, it was observed that the advantage that pairs with a

[2] Carpenter (2009) used norms of associative strength, which provide an estimate of the probability of producing the target given the cue. Although this estimate was not based on performance in a multitrial learning paradigm, it is also a measure of target difficulty, since paired associate learning is affected by the degree of relatedness between cues and targets.

strong cue had over pairs with a weak cue in the initial test was either reversed (Experiment 1) or eliminated (Experiment 2) on the criterion test. Moreover, in Carpenter's Experiment 1, the proportion of items correctly recalled on final test was greater for pairs with weak cues than for pairs with strong cues, but only in the retrieval practice condition. This suggests that different levels of retrieval effort induced by the initial test were responsible for this effect (see Carpenter, 2009, Table 3). In an investigation about metamemory, Cull and Zechmeister (1994, Experiment 2) found an analogous interaction, although this should be interpreted with caution, since a self-paced procedure was used, leading to different exposure times under different conditions, which may partly explain the results (see Glover, 1989).

In sum, difficulty can be operationalized either as the amount of demands placed on the learner in the practice phase (e.g., Pyc & Rawson, 2009) or as the relative item difficulty, based upon normative studies (e.g., Vaughn et al., 2013). In both cases, it is assumed that more difficult tasks and items tend to require more effort than their easier counterparts (Kahneman, 1973). Although the REH has gained empirical support in studies that operationalized difficulty under the former (Carpenter & DeLosh, 2006; Pyc & Rawson, 2009; Vaughn & Rawson, 2011), mixed results were obtained under the latter (Carpenter, 2009; Vaughn et al., 2013). A possible explanation for these divergent results is the way evidence was produced: on the one hand, Carpenter's experiments factorially crossed type of practice (restudy, retrieval practice) with item difficulty (weak and strong cues); on the other hand, Vaughn et al. compared easy and difficult items across several criterion levels, under the assumption that only difficult items would benefit further at higher criterion levels, since retrieval would still involve effort on later retrieval attempts.

Present Study

Here, we adopted a similar approach to Carpenter's (2009) experiments, crossing factorially type of practice and item difficulty. Unlike Carpenter, we used a longer retention

interval (48 hr instead of 5 min) and repeated practice for each item (instead of only one presentation). Like Vaughn et al. (2013), we tested REH predictions, albeit addressing a slightly different question: Does item difficulty affect the magnitude of the retrieval practice effect? If so, which items benefit most from retrieval practice: easy or difficult items? These questions are theoretically relevant because they provide a new test of the REH with effort manipulated via item difficulty rather than task difficulty, which has been more usually investigated. These questions are also of particular applied relevance for teachers and educators, as some materials are more difficult to learn than other and require more effort. If retrieval practice can benefit learning of these materials to a similar or higher extent than learning easier materials, teachers and educators may decide to invest time and resources differently to such difficult materials.

In two experiments we sought (a) to replicate both the retrieval practice and item difficulty effects, and (b) to investigate whether item difficulty affects retrieval practice effect sizes. Participants learned a set of word pairs (study phase) and repeatedly restudied half of this set and repeatedly retrieval practiced another half (practice phase). Forty-eight hours later, they took a cued-recall test (final test phase). Three predictions were made about the experimental results. First, it was hypothesized that participants would recall more retrieval practiced items than restudied ones (Dunlosky et al., 2013; Rowland, 2014). Second, it was hypothesized that participants would recall more easy items than difficult ones (Cull & Zechmeister, 1994; Underwood, 1982). Third, based on the REH (Pyc & Rawson, 2009) and on previous experimental results (e.g., Carpenter, 2009), it was hypothesized that the retrieval practice effect would be greater for difficult items when compared to easy ones.

Experiment 1

Method

Participants and design. Fifty-two undergraduates were recruited from the University

of Brasília (Brazil) volunteered to take part in the experiment. Sample size was based on Pyc and Rawson's (2012) Experiment 1b, whose stimulus type, retention interval, and both initial and final tests are the same as the ones used there. One participant was excluded prior to data analysis because she failed to return for the second session. Thus 51 participants comprised the final sample (females = 46; age range = 18–32 years, M = 20.29, SD = 3.05). All participants were native Brazilian Portuguese speakers and gave written informed consent. The experiment followed a 2 × 2 factorial design, with the factors of type of practice (restudy, retrieval practice) and difficulty (easy, difficult) both manipulated within-participants.

Materials. Forty Swahili–Portuguese word pairs were selected. Based on the memorability normative measures provided by Lima and Buratto (2019), twenty pairs were labeled as *easy* (M = .60, SD = .10), and 20 pairs as *difficult* (M = .24, SD = .05; see Appendix C). Word pairs were divided into two sets, each one with 10 easy and 10 difficult items. Both sets were equated in terms of familiarity, concreteness, arousal, valence for Portuguese words; wordlikeness (similarity to Portuguese) for the Swahili words, and difficulty (ts ≤ 0.78, ps ≥ .44). The assignment of both sets to experimental conditions was counterbalanced across participants. Additionally, twenty math problems, 10 easy (e.g., 7 × 8) and 10 difficult (e.g., 17 × 18), were created.[3] Instructions and materials were presented on a computer screen controlled with PsychoPy (Peirce, 2007).

Procedure. Figure 1a depicts a general schematic representation of Experiments 1 and 2. In Experiment 1, at the beginning of the first session, participants completed an initial training task, whose stimuli were unrelated to Swahili–Portuguese word pairs. This task aimed to train participants on how to use the keyboard and to help them understand the

[3] We chose easy and difficult math problems because we originally intended to measure participants' pupil size as a function of task difficulty. There is evidence that the eye's pupil dilates more while participants perform more difficult mathematical tasks than easier ones (Hess & Polt, 1964). The math problems would thus serve both as a retention interval filler and as a control task to assess pupil size sensitivity for our eye-tracker.

feedback that would be provided throughout this session. Next, in the study phase, participants were presented with 40 Swahili–Portuguese word pairs in random order. Each trial began with a "+" symbol on the center of the screen for 4 s, which was followed by the presentation of a word pair, also on the center of the screen (Swahili word on top; Portuguese word below). Participants were instructed to study the pairs. After the study phase and after each practice cycle, participants engaged on a distracter task, which consisted of four math problems. Each distracter task cycle lasted 1 min.

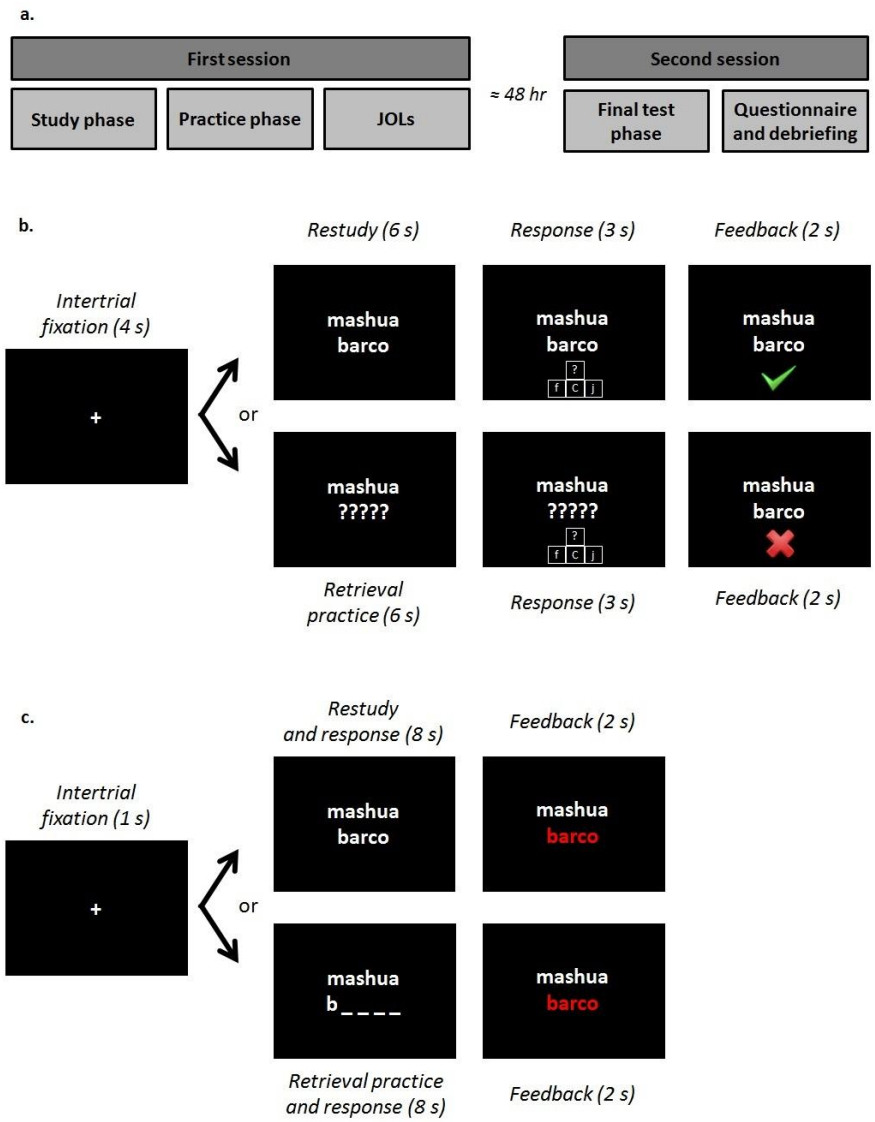

Figure 1. (a) General schematic representation for Experiments 1 and 2. Examples of trials on practice phase for restudy and retrieval practice conditions are depicted for both (b) Experiment 1 and (c) Experiment 2.

After the first distracter task cycle, participants were informed that all word pairs would be practiced again by one of two different methods ("method A" or "method B"). Assignment of method name (A, B) to type of practice (restudy, retrieval practice) was counterbalanced across participants. Examples of trials on both types of practice are depicted in Figure 2b. In both types of practice, each trial started with a "+" symbol, with the same location and presentation time of the study phase. On retrieval practice trials, the Swahili word, alongside question marks that replaced the Portuguese word, were presented for 9 s. Participants were then asked to covertly recall the Portuguese translation of the Swahili word. After 6 s of the sole presentation of the Swahili word, four alternatives (three letters and one question mark) were presented on the bottom of screen. Participants were then asked to press the keyboard arrow that represented the penultimate letter of the Portuguese word they recalled.[4] They had 3 s to indicate their response, which was followed by a 2 s feedback (symbols indicating *correct*, *incorrect*, or *no response*). Additionally, during the feedback, the correct Portuguese word replaced the question marks on the screen, giving participants a new opportunity to encode the correct translation of the given Swahili word. We chose the penultimate letter for two reasons: (1) the first letter could potentially encourage a strategy in which alternatives could be used as retrieval cues (see Wing et al., 2013); and (2) given the fact that Portuguese words tend to end with a strict set of letters, when asked to indicate the last letter of a word, the participant's range of potential responses would be rather limited. On restudy trials, word pairs were presented intact for 11 s each. Participants were instructed to use these trials as an additional opportunity to study the pairs. After 6 s onset of the word pair, four alternatives (similar to the retrieval practice trials) were presented on the bottom of the

[4] This procedure was adapted from Wing, Marsh, and Cabeza (2013), as our original aim was to measure pupil size with an eye-tracker. Consequently, we collected discrete responses instead of a full typed response, in both conditions. This minor design feature should not affect the results, as retrieval practice effects also occur when participants make covert recall (Putnam & Roediger, 2013; Smith, Roediger, & Karpicke, 2013) or when they emit a discrete response on the keyboard (Racsmány, Szőllősi, & Bencze, 2018, Experiment 1).

screen, amongst which participants should indicate the penultimate letter of the Portuguese word. They had 3 s to indicate their response, which was once more followed by a 2 s feedback. In both types of practice, participants were encouraged to select "?" if they were not sure of the answer. Word pairs were presented in a random order, the position of the three alternatives (only the letters) was also randomized, and type of practice was mixed across trials. Four cycles of practice were performed on practice phase.

After the last cycle of distracter task (after the fourth practice cycle), we assessed participants' metacognitive knowledge of the effectiveness of both types of practice by having them make two judgments of learning (JOLs). Participants estimated what percentage of Swahili words they believed they would remember two days later. These two judgments were made on a 0–100 scale (0 = *I think I'll remember nothing*; 100 = *I think I'll remember all*). Participants saw images representing each method ("A" and "B") to ensure that they would make the judgment based on the appropriate method. Upon finishing the JOLs, participants were dismissed and reminded to return to the lab two days later. Forty-eight hours after the first session (range = 42–53 hr), participants returned to the lab. The second session started with a criterion cued-recall test. On each trial, participants were prompted with a Swahili word and were asked to recall its Portuguese translation. Participants typed each word onto the computer, and after they pressed the "Enter" key, the next trial began. The maximum response time allowed on each trial was 15 s. All 40 items previously studied were tested and no feedback was provided. The order of items was randomized. After this task, participants answered a brief questionnaire, were debriefed, thanked, and dismissed.

Statistical analyses. An alpha level of .05 was used for all statistical tests, unless otherwise stated. When the assumption of sphericity was violated, as indicated by Mauchly's test, the Greenhouse–Geisser correction was applied to adjust for degrees of freedom (Greenhouse & Geisser, 1959). Measures of effect size were reported as Cohen's *d* (*t*-tests),

as partial eta-squared (η_p^2; ANOVAs), or as log odds (β; mixed logit models), when appropriate. Exploratory analyses unrelated to the main hypotheses of this study are presented in Appendix D.

Results

Practice phase.

Performance on practice cycles. Figure 2a depicts the proportion of correct answers on cycles of the practice phase. It should be noted that, for retrieval practiced items, correct answers represent learning across cycles, whereas for restudied items, correct answers only indicate that participants paid attention on the task across cycles. Since correct answers indexed different cognitive processes for each type of practice, we conducted two 2 (difficulty) × 4 (cycle) repeated measures ANOVA, separately for each type of practice. For retrieval practiced items, there were main effects of difficulty, $F(1, 50) = 127.14$, $p < .001$, $\eta_p^2 = .72$, and cycle, $F(1.91, 98.53) = 204.83$, $p < .001$. The effect of difficulty indicates that easy items ($M = .58$, $SD = .19$) were better learned than difficult items ($M = .35$, $SD = .20$), whereas the effect of cycle indicates an increasing linear trend across the four cycles. Importantly, there was a Difficulty × Cycle interaction, $F(2.34, 117.13) = 4.29$, $p = .01$, $\eta_p^2 = .08$. Paired-sample *t*-tests indicated that the advantage for easy items over difficult ones was lower in the first cycle, $t(50) = 5.27$, $p < .001$, $d = 0.74$, than in the other cycles, $ts(50) \geq 8.26$, $ps < .001$, $ds \geq 1.16$. We will return to this interaction later in the discussion. For restudied items, there were no significant effects, $Fs \leq 2.32$, $ps \geq .10$.

Reaction time (RT) on practice cycles. Following previous investigations (e.g., Karpicke & Bauernschmidt, 2011; Vaughn et al., 2013), we used RT as a measure of task difficulty during practice cycles. RT here represented the time between the onset of the screen showing response alternatives and the participant's response. We computed median RT for

each participant (by condition), considering all trials.[5] Figure 2b depicts average RT on cycles of the practice phase. We entered these data on two 2 (difficulty) × 4 (cycle) repeated measures ANOVA. For restudied items, there was only a main effect of cycle, $F(2.52, 126.03) = 4.07$, $p = .01$, $\eta_p^2 = .08$. These effect showed that, on average, median RT was shorter for the fourth cycle (998 ms) than for the first, second, and third cycles (1,084 ms, 1,072 ms, and 1,058 ms, respectively), all *p*s < .05. For retrieval practiced items, there were main effects of difficulty, $F(1, 50) = 26.68$, $p < .001$, $\eta_p^2 = .35$, and cycle, $F(3, 150) = 5.622$, $p = .001$, $\eta_p^2 = .10$. These effects showed that, on average, median RT was shorter (a) for easy items than for difficult ones (1,083 ms vs. 1,244 ms), and (b) for the fourth cycle (1,077 ms) than for the first and third cycles (1,224 ms and 1,187 ms, respectively), all *p*s < .02. Other comparisons were not significant, *F*s ≤ 2.14, *p*s ≥ .10. Taken together, these results suggest that retrieval was more effortful for difficult items, as indexed by RTs.

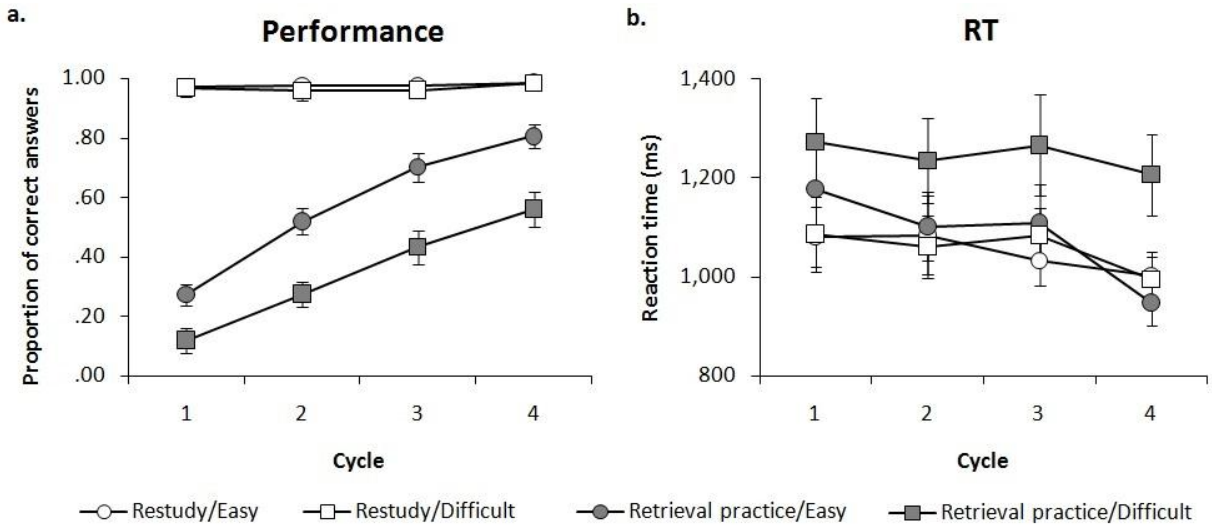

Figure 2. Data from practice cycles of Experiment 1. (a) Proportion of correct answers across the four cycles of the practice phase. (b) Reaction time across the four cycles of the practice phase. Error bars represent the 95% within-participant confidence interval (Cousineau, 2005).

[5] Unless otherwise stated, analyses using only correct answers yielded similar pattern of results. Analyses with all the answers were conducted to prevent the loss of statistical power due to listwise exclusion of missing cases in repeated-measures ANOVA.

JOLs. The upper side of Table 1 shows the number of participants that judged they would remember more restudied items, more retrieval practiced items, or an equal number of items on both conditions (i.e., a tie). Participants' average JOLs (converted into proportions) for restudied and retrieval practiced items were virtually identical (.41 vs. .42). Paired-sample *t*-test showed that there was no difference between them, $t(50) = .37$, $p = .71$, $d = 0.05$.

Table 1

Number of Participants Showing Different Patterns of JOLs and Performance in Experiments 1 and 2

Measure	Advantage		
	Retrieval practice	Restudy	Tie
JOLs			
Experiment 1	21 (.41)	25 (.49)	5 (.10)
Experiment 2	21 (.75)	4 (.14)	3 (.11)
Performance			
Experiment 1—Easy	43 (.84)	0 (.00)	8 (.16)
Experiment 1—Difficult[a]	38 (.75)	4 (.08)	9 (.18)
Experiment 1—All items	50 (.98)	1 (.02)	0 (.00)
Experiment 2—Easy[a]	20 (.71)	6 (.21)	2 (.07)
Experiment 2—Difficult	23 (.82)	1 (.04)	4 (.14)
Experiment 2—All items	25 (.89)	1 (.04)	2 (.07)

Note. Sample proportions are reported in parentheses.

[a]Sum of proportions does not total 1.00 due to rounding.

Final test phase.

Scoring. Two independent judges were trained to assess the participants' responses on the final test phase. They were blinded to what condition each item pertained. To assess inter-

rater agreement, Cohen's kappa (κ) was computed. The two judges showed high level of agreement on scorings, $\kappa = .97$, $p < .001$, which could be considered almost perfect (Landis & Koch, 1977). Thus, the scores of one of the judges were randomly selected and used on subsequent analyses.

Performance on final test. Figure 3a depicts recall performance on the final test. A 2 (type of practice) × 2 (difficulty) repeated-measures ANOVA revealed significant main effects of type of practice, $F(1, 50) = 159.70$, $p < .001$, $\eta_p^2 = .76$, and difficulty, $F(1, 50) = 171.62$, $p < .001$, $\eta_p^2 = .77$. The main effect of type of practice reflects overall higher recall in the retrieval practice condition ($M = .52$, $SD = .26$) than in the restudy condition ($M = .30$, $SD = .22$), whereas the main effect of difficulty shows that recall was higher for easy items ($M = .56$, $SD = .27$) than for difficult ones ($M = .26$, $SD = .23$). The Type of Practice × Difficulty interaction was significant, $F(1, 50) = 4.05$, $p = .05$, $\eta_p^2 = .08$, which revealed that retrieval practice effect was greater for easy items (.69 vs. .43), $t(50) = 10.84$, $p < .001$, $d = 1.52$, than for difficult ones (.35 vs. .16), $t(50) = 6.95$, $p < .001$, $d = 0.97$ (see Figure 3a). Following Minear et al.'s (2018) recommendations, we also report the proportion of participants showing different patterns of performance (see the bottom of Table 1). Retrieval practice effects were more frequent for easy items than for difficult ones (.84 vs .75). Moreover, when we consider all items, virtually all participants showed retrieval practice effects.

RT on final test. We also used RT as an alternative index of performance on the final test (Racsmány et al., 2018). RT represented the time between stimulus onset (cue word) and participant's first key press. Figure 3b depicts RT on the final test. A 2 (type of practice) × 2 (difficulty) repeated-measures ANOVA revealed significant main effects of type of practice, $F(1, 46) = 16.82$, $p < .001$, $\eta_p^2 = .27$, and difficulty, $F(1, 46) = 37.98$, $p < .001$, $\eta_p^2 = .45$. The effect of type of practice reflects overall shorter RT in the retrieval practice condition than in the restudy condition (3,338 ms vs. 4,171 ms), whereas the effect of difficulty shows that RT

was shorter for easy items than difficult ones (3,110 ms vs. 4,399 ms). The Type of Practice × Difficulty interaction was not significant, $F(1, 46) = 1.42$, $p = .24$, $\eta_p^2 = .03$.

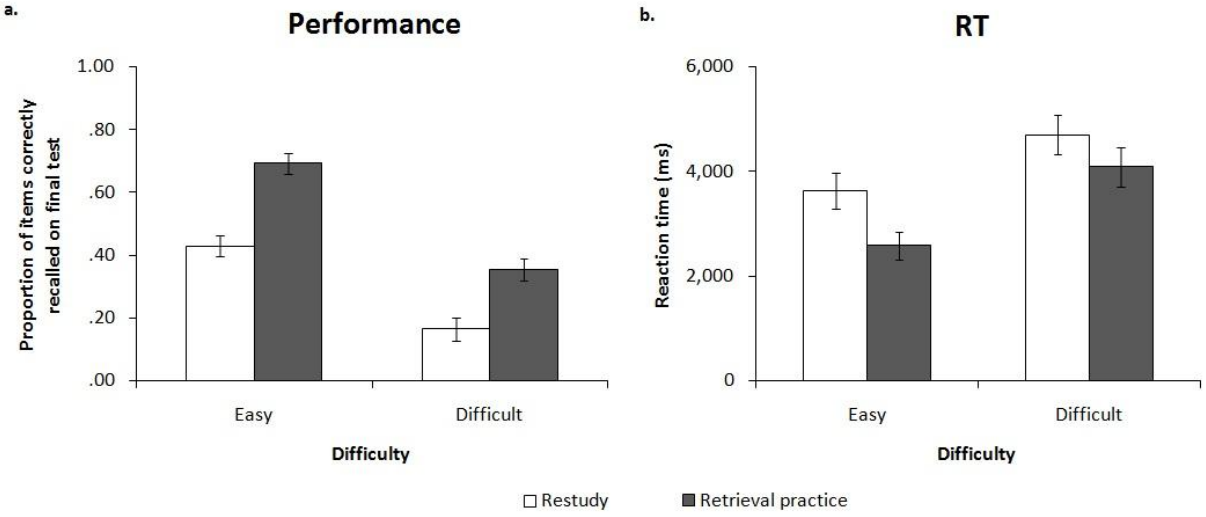

Figure 3. Proportion of correct recall (a) and RT (b) on the final test of Experiment 1. Error bars represent the 95% within-participant confidence interval (Cousineau, 2005).

Conditional analyses. Figure 4 depicts the probability of correct recall on the final test, given the difficulty and the number of correct answers on practice cycles (see Finley et al., 2011, for a similar procedure). The center of each bubble corresponds to the probability of recall for a given item, whereas bubble diameter represents the proportion of cases falling into each category. We entered these data into two mixed logit models, for restudied and retrieval practiced items (for rationale, see Jaeger, 2008; Sommet & Morselli, 2017). Fixed effects for difficulty and number of correct answers were entered into the model (mean centered), with random-participant level intercepts. These models were compared with empty models (i.e., which estimate whether the odds of recall and non-recall varies between participants). The likelihood-ratio tests indicated that the addition of fixed and random terms improved the prediction for both restudy, $\chi^2(3) = 113.80$, $p < .001$, and retrieval practice models, $\chi^2(3) = 357.22$, $p < .001$.

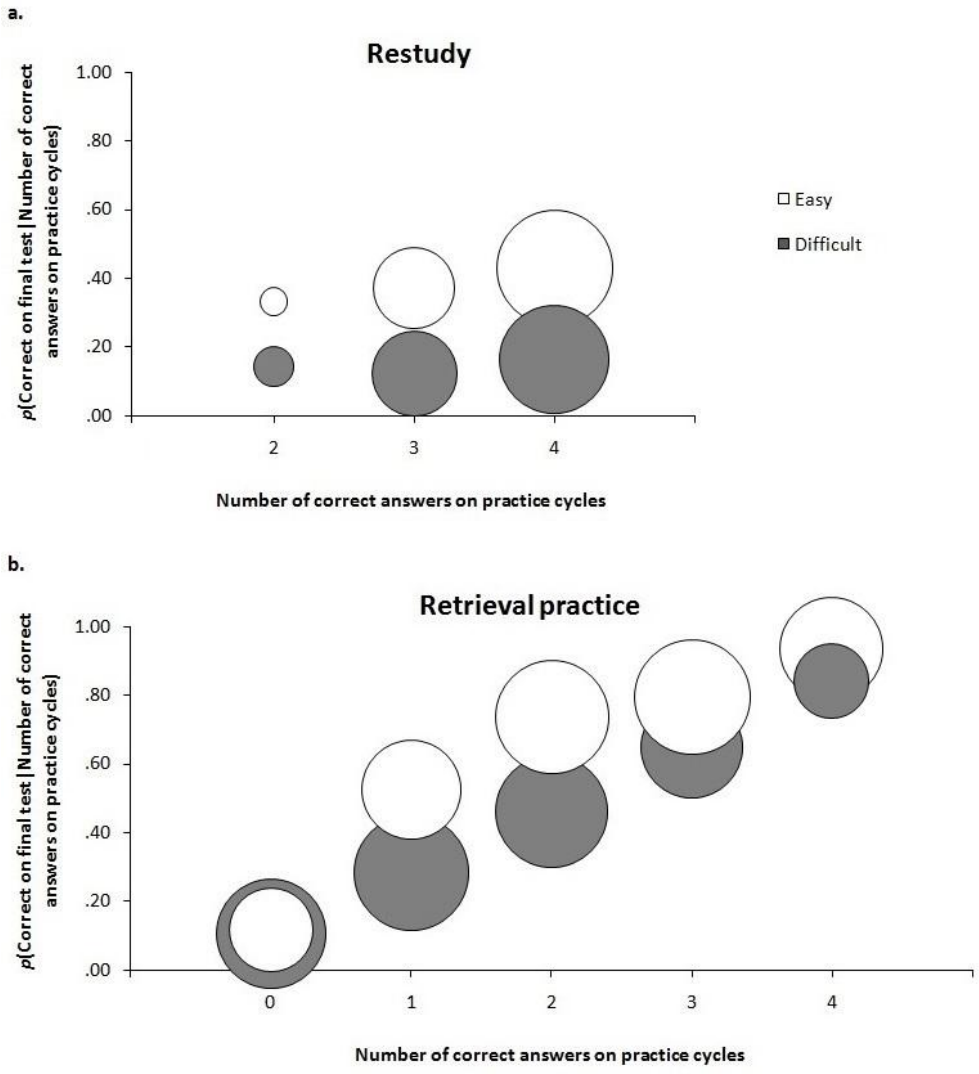

Figure 4. Conditional probabilities depicting the probability of correct recall on final test, given number of correct answers on practice cycles (range = 0–4 times). (a) Restudy condition and (b) retrieval practice condition.

Table 2 shows model summaries. For the restudy model, difficulty was a significant predictor of successful final recall. The odds favored recall of easy items over difficult ones on the final test (*OR* = 5.61). Other predictors were nonsignificant. For the retrieval practice model, difficulty was also a significant predictor, with the odds of recalling an easy item on the final test greater than the odds of recalling a difficult one (*OR* = 3.71). Furthermore, the odds of correct recall increased with the number of correct answers (*OR* = 2.92). In sum,

difficulty predicts successful recall on the final test for both models, but only for the retrieval practice model the number of correct answers had a predictive value. This result suggests that the benefits of retrieval practice are partially conditional on successful retrieval during the practice phase, although this result should be interpreted with caution, because this is a post-hoc analysis.

Table 2

Fixed Effects for the Mixed Logit Models Predicting Final Recall in Experiment 1

Fixed effects	β	*SE*	Wald *Z*	*p*
Restudy model				
Intercept	–1.28	0.22	–5.96	< .001
Number of correct answers	0.19	0.27	0.70	.48
Difficulty	1.72	0.18	9.79	< .001
Interaction	0.71	0.52	1.37	.17
Retrieval practice model				
Intercept	0.21	0.19	1.14	.25
Number of correct answers	1.07	0.09	12.41	< .001
Difficulty	1.31	0.19	6.96	< .001
Interaction	–0.04	0.15	–0.26	.80

Note. Beta represents log odds.

Discussion

Experiment 1 sought (a) to replicate both the retrieval practice and item difficulty effects, and (b) to investigate whether item difficulty affects the retrieval practice effect sizes. The first aim was achieved, replicating previous studies (Cull & Zechmeister, 1994; Dunlosky et al., 2013; Rowland, 2014). Contrary to our hypothesis, however, the magnitude of the retrieval practice effect favored easy over difficult items, not the other way around. One limitation of Experiment 1 is that easy and difficult items had very different proportions of correct answers in the four practice cycles. This indicates that difficult items probably were

not as well-learned as easy ones (see Figure 2a). The fixed number of practice cycles used in the present study tends to favor easier items due to the greater proportion of initial recall in the retrieval practice condition (Vaughn et al., 2013). Although the presence of feedback could partially overcome this problem (Kang, McDermott, & Roediger, 2007), the design of Experiment 1 does not allow disentangling the relative contributions of retrieval effort (indexed by item difficulty) and retrieval success (indexed by proportion of correct answers on practice cycles) to final recall performance. Experiment 2 sought to balance the initial performance in the last practice cycle in order to eliminate this confounding factor present in the Experiment 1.

Experiment 2

In this experiment, easy items were practiced four times, whereas difficult items were practiced six times. If the direction of the interaction effect observed on Experiment 1 was due to different learning rates for easy and difficult items, a reversion of the interaction direction would be expected on Experiment 2, supporting the REH (Pyc & Rawson, 2009). If retrieval effort is not related to retrieval practice effect, then either no interaction should be found or an interaction favoring easy items over difficult ones should be found.

Method

Participants, design, and materials. We ran an a priori power analysis using G*Power.1.9.4 (Faul, Erdfelder, Lang, & Buchner, 2007), with power set at. .95, $\alpha = .05$, and Cohen's $f = .29$ (equivalent to our Type of Practice × Difficulty effect size, $\eta_p^2 = .08$), which suggested a sample of 28 participants. Thirty-three undergraduates were recruited from the University of Brasília. In total, five participants had to be excluded, three of them because they failed to return for the second session, one participant due to power outage during the session, and one due to failure to follow the instructions. Thus, the final sample size consisted of 28 participants, as suggested by the a priori power analysis (females = 17; age range = 17–

34 years, *M* = 20.29, *SD* = 3.87). All participants were native Brazilian Portuguese speakers and gave written informed consent. Design and materials were the same as those used in the Experiment 1.

Procedure. The study phase was the same as in Experiment 1, except that the "+" symbol lasted only 1 s. Participants then engaged on a distracter task for 1 min (also presented after each practice cycle). At the beginning of the practice phase, participants were informed that all word pairs would be practiced again by one of two different methods. In both types of practice, each trial started with a "+" symbol, with the same location and presentation time of the study phase. On retrieval practice trials, the Swahili word, alongside the cue containing the first letter of the Portuguese word (for an example, see Figure 1c), were presented for 8 s. Participants were then asked to recall and type the Portuguese translation of the Swahili word, which was followed by feedback (2 s), consisting of the replacement of the given cue for the correct Portuguese word in red color. On restudy trials, word pairs were presented for 10 s each. Participants were instructed to use these trials as an additional opportunity to study the pairs. In the last 2 s, the Portuguese word's color changed from white to red, to balance this feature between conditions. Participants practiced all word pairs for four cycles, but difficult pairs were practiced for two additional cycles. After the last cycle of distracter task, participants made JOLs on a 0–100 scale and were dismissed and asked to return to the lab two days later (range = 46–53 hr). The second session was identical to that in Experiment 1.

Statistical analyses. The same analyses were conducted in Experiment 2. Exploratory analyses unrelated to the main hypothesis of this study are presented in Appendix D.

Results

Practice phase.

Performance on practice cycles. Figure 5a depicts the proportion of correct answers on cycles of the practice phase. Increasing linear trends, observed on Figure 5a, were

significant in all conditions, *F*s ≥ 5.00, *p*s ≤ .03, all $\eta_p^2 \geq .16$. Furthermore, differences in performance in the final cycle for each type of practice (i.e., easy items, cycle 4 vs. difficult items, cycle 6) were nonsignificant, *t*s(27) ≤ |1.99|, *p*s ≥ .06. The observed differences in performance between restudied and retrieval practiced conditions were .10 and .08 for easy and difficult items, respectively, whereas these same differences were .27 and .42 in Experiment 1. Consequently, comparisons between the retrieval practice effects for easy and difficult items in Experiment 2 are more accurate than in Experiment 1.

RT on practice cycles. Figure 5b depicts average RT on cycles of the practice phase. We again computed median RT for each participant, considering all trials.[6] We conducted two 2 (difficulty) × 4 (cycle) repeated measures ANOVA, separately for each type of practice. Since our aim was to check the effectiveness of retrieval effort manipulation and the RT was available for both difficulty levels up to the fourth cycle, we entered only the first four cycles in these analyses. For restudied items, there was only a main effect of cycle, $F(1.65, 44.66) = 34.53$, $p < .001$, $\eta_p^2 = .56$. These effect showed that, on average, median RT decreased across cycle 1 (2,053 ms), cycle 2 (1,811 ms), cycle 3 (1,706 ms), and cycle 4 (1,622 ms), all *p*s < .001. For retrieval practiced items, there were main effects of difficulty, $F(1, 27) = 11.01$, $p < .001$, $\eta_p^2 = .29$, and cycle, $F(2.14, 57.79) = 69.09$, $p < .001$, $\eta_p^2 = .72$. These effects showed that, on average, median RT (a) was shorter for easy items than for difficult ones (2,804 ms vs. 3,239 ms), and (b) decreased across cycle 1 (4,069 ms), cycle 2 (3,074 ms), cycle 3 (2,609 ms), and cycle 4 (2,335 ms), all *p*s < .001. There was a Difficulty × Cycle interaction, $F(2.40, 64.79) = 3.78$, $p = .02$, $\eta_p^2 = .12$, which indicates that decreasing linear trend was steeper for easy items ($\eta_p^2 = .80$) than for difficult ones ($\eta_p^2 = .62$). These results suggest again that retrieval was more effortful for difficult items, as indexed by RTs.

[6] In four cases, participants had missing RT data because they did not respond in any trials of a given condition. In these cases, the missing values were replaced by the mean of all participants in that condition.

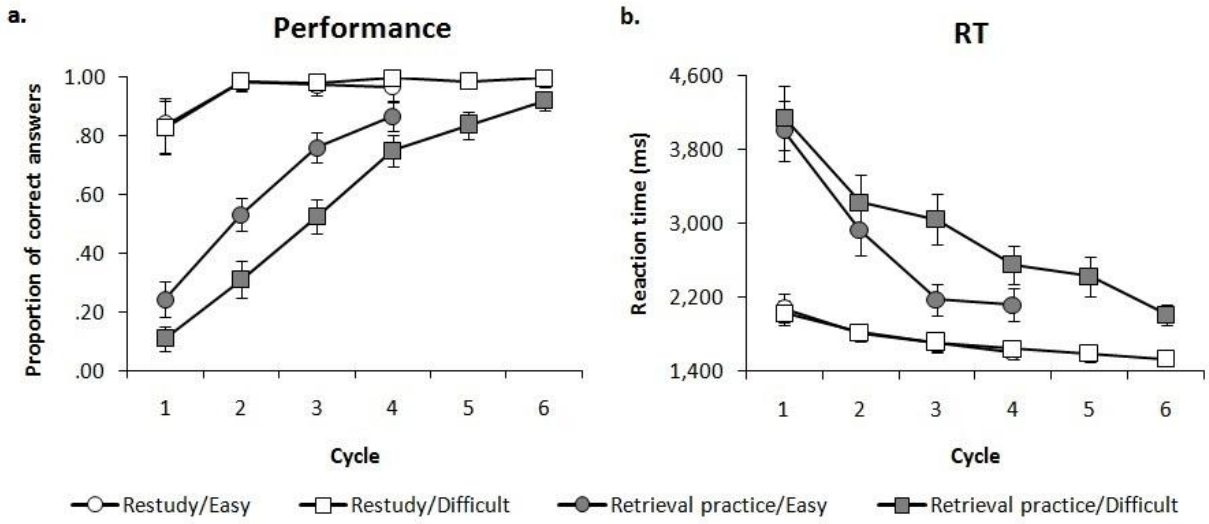

Figure 5. Data from practice cycles of Experiment 2. (a) Proportion of correct answers across the six cycles of the practice phase. (b) Reaction time across the six cycles of the practice phase for correct responses. Error bars represent the 95% within-participant confidence interval (Cousineau, 2005).

JOLs. The upper side of Table 1 shows the number of participants that judged they would remember more restudied items, more retrieval practiced items, or an equal number of items on both conditions (i.e., a tie). Contrary to Experiment 1, participants' average JOLs (converted into proportions) was higher for retrieval practiced items than for restudied ones (.55 vs. .38), $t(27) = 3.62$, $p = .001$, $d = 0.69$.

Final test phase.

Scoring. Two judges independently scored participants' responses. Judges had an almost perfect level of agreement on scoring, both in the practice ($\kappa = .97$) and in the final test phases ($\kappa = .98$), ps $< .001$ (Landis & Koch, 1977). The scores of one of the judges were randomly selected and used on subsequent analyses.

Performance on final test. Figure 6a depicts recall performance on the final test. A 2 (type of practice) × 2 (difficulty) repeated-measures ANOVA revealed main effects of type of practice, $F(1, 27) = 72.19$, $p < .001$, $\eta_p^2 = .73$, and difficulty, $F(1, 27) = 61.75$, $p < .001$, $\eta_p^2 =$

.70. The main effect of type of practice reflects overall higher recall in the retrieval practice condition (*M* = .55, *SD* = .18) than in the restudy condition (*M* = .32, *SD* = .20), whereas the main effect of difficulty shows that recall was higher for easy items (*M* = .54, *SD* = .20) than for difficult ones (*M* = .33, *SD* = .18). More important, the Type of Practice × Difficulty interaction was not significant, $F(1, 27) = 2.86$, $p = .10$, $\eta_p^2 = .10$. Nonetheless, a trend towards interaction was observed in the opposite direction to that in Experiment 1. We ran *t*-tests to further explore this trend, which suggested that the retrieval practice effect was higher for difficult items (retrieval practice: *M* = .46, *SD* = .22; restudy: *M* = .19, *SD* = .19), $t(27) = 6.80$, $p < .001$, $d = 1.29$, than for easy ones (retrieval practice: *M* = .63, *SD* = .22; restudy: *M* = .45, *SD* = .24), $t(27) = 4.58$, $p < .001$, $d = .87$. The bottom side of Table 1 shows different patterns of performance. Unlike Experiment 1, retrieval practice effects in Experiment 2 were more frequent for difficult items than for easy ones (.82 vs. 71). Moreover, when all items are considered, most of participants (.89) showed retrieval practice effects.

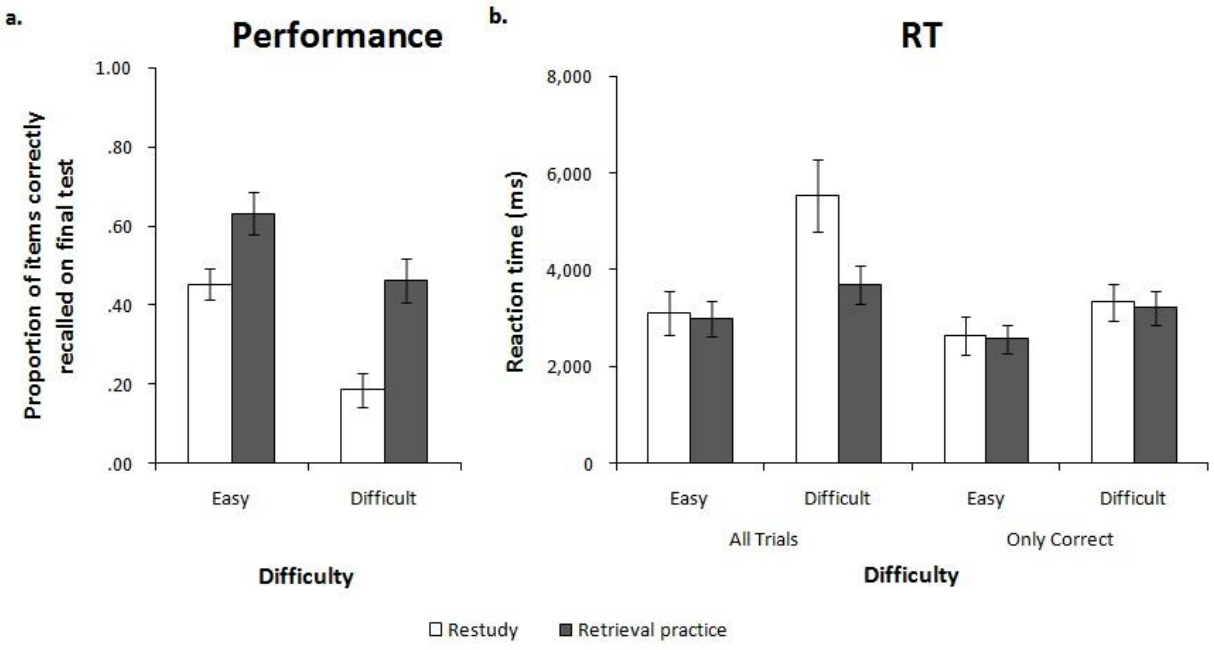

Figure 6. Proportion of correct recall (a) and RT (b) on the final test of Experiment 2. Error bars represent the 95% within-participant confidence interval (Cousineau, 2005).

Response times on final test. Figure 6b depicts RT for all trials on the final test. A 2 (type of practice) × 2 (difficulty) repeated-measures ANOVA revealed significant main effects of type of practice, $F(1, 25) = 11.38$, $p = .002$, $\eta_p^2 = .31$, and difficulty, $F(1, 25) = 26.97$, $p < .001$, $\eta_p^2 = 0{,}52$. The main effect of type of practice reflects overall shorter RT for retrieval practiced items than for restudied ones (3,349 ms vs. 4,322 ms), whereas the main effect of difficulty reflects overall shorter RT for easy items than for difficult ones (3,053 ms vs. 4,618 ms). The Type of Practice × Difficulty interaction was also significant, $F(1, 25) = 19.12$, $p = .008$, $\eta_p^2 = .25$. Paired comparisons indicates that, for difficult items, RT was significantly shorter for retrieval practiced items than for restudied ones (3,702 ms vs. 5,533 ms), $t(25) = –3.56$, $p = .002$, $d = 0.70$. For easy items, there were no significant differences in RT between retrieval practiced and restudied items (2,995 ms vs. 3,110 ms), $t(25) = –0{,}41$, $p = .69$, $d = 0.08$. When analysis was restricted to correct trials,[7] only the difficulty effect remained significant, $F(1, 25) = 11.69$, $p = .002$, $\eta_p^2 = 0.32$, reflecting overall shorter RT for easy items than for difficult ones (2,608 ms vs. 3,279 ms). The other effects were nonsignificant, Fs$(1, 25) \leq 0.15$, ps $\geq .70$, all $\eta_p^2 \leq .006$ (see Figure 6b).

Conditional analyses. Figure 7 depicts the probability of correct recall on the final test, given the difficulty and the number of correct answers on practice cycles. We entered difficulty and number of correct answers into two mixed logit models (mean centered), with random-participant level intercepts (Jaeger, 2008; Sommet & Morselli, 2017). Models were once more compared with empty models. The likelihood-ratio tests indicated that the addition of fixed and random terms improved the prediction for both restudy, $\chi^2(3) = 57.69$, $p < .001$,

[7] Six participants did not recall any difficult words in the restudy condition. To reduce the loss of statistical power due to missing cases, they were replaced by the mean of all participants in that condition. We also entered our data into a Linear Mixed Model (LMM). Such models not require listwise deletion for missing cases and thus allow the inclusion of data from all participants (Hoffman & Rovine, 2007). The results from the LMM analyses and the ANOVAs led to the same conclusions.

and retrieval practice models, $\chi^2(3) = 94.81$, $p < .001$.

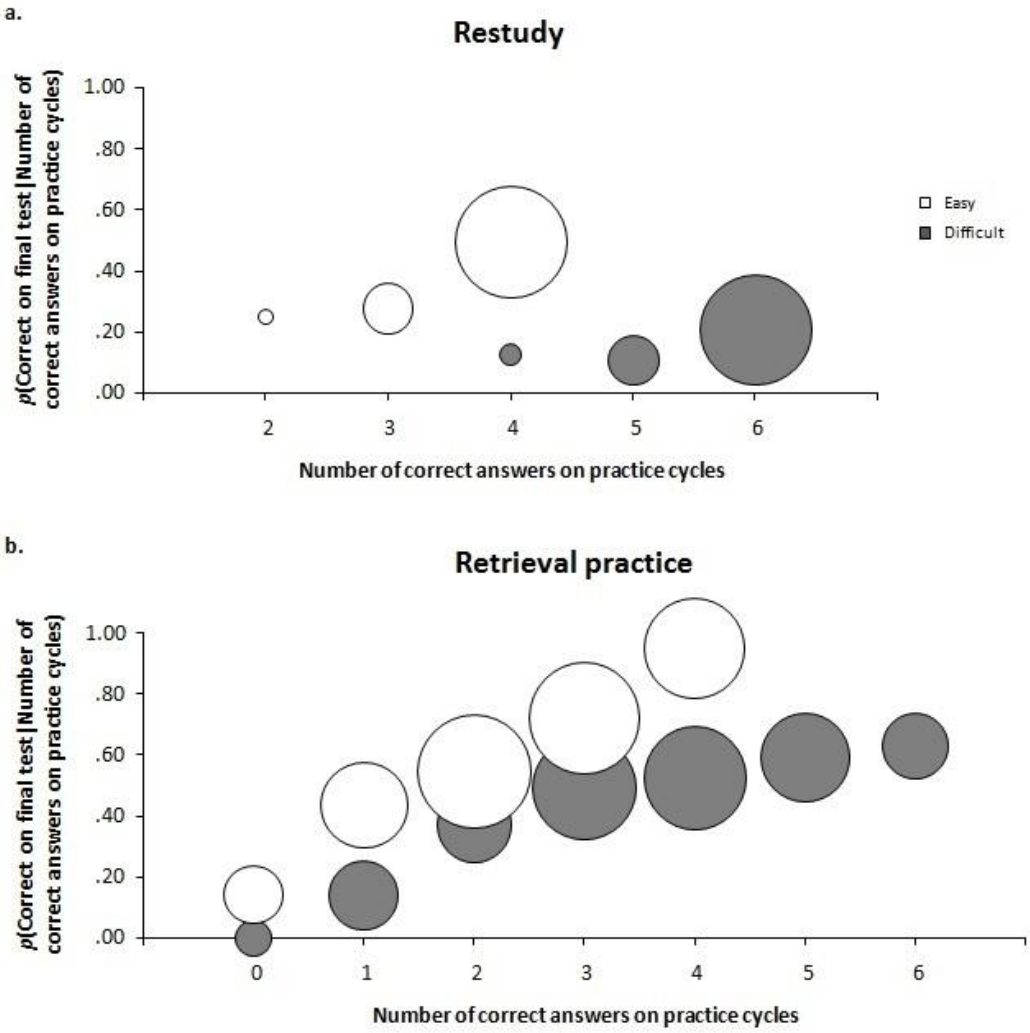

Figure 7. Conditional probabilities depicting the probability of correct recall on final test, given number of correct answers on practice cycles (range = 0–6 times). (a) Restudy condition and (b) retrieval practice condition.

Table 3 shows summaries for these models. Like in Experiment 1, for the restudy model, difficulty was a significant predictor of successful final recall. The odds of recalling an easy item on the final test was more than ten times greater than the odds of recalling a difficult item. For the retrieval practice model, difficulty was also a significant model. The odds of recalling an easy item was five times greater than the odds of recalling a difficult item. Again, the odds of correct recall increased with the number of correct answers (*OR* =

1.97). The Difficulty × Number of Correct Answers interaction was also significant, which indicates that slopes for easy and difficult items are different in this model (see Figure 7b). In sum, the logit mixed models showed convergent results in both experiments.

Table 3

Fixed Effects for the Mixed Logit Models Predicting Final Recall in Experiment 2

Fixed effects	β	*SE*	Wald *Z*	*p*
Restudy model				
Intercept	–0.81	0.34	–2.36	.02
Number of correct answers	0.41	0.31	1.30	.19
Difficulty	2.32	0.66	3.53	< .001
Interaction	0.36	0.53	0.68	.50
Retrieval practice model				
Intercept	0.40	0.15	2.64	.008
Number of correct answers	0.68	0.08	8.00	< .001
Difficulty	1.61	0.23	6.98	< .001
Interaction	0.54	0.17	3.28	< .001

Note. Beta represents log odds.

Discussion

Experiment 2 balanced initial performance in the last of the practice cycles to eliminate the confounding factor present in the Experiment 1. In the last cycle, the observed differences in performance were .06 and .03 for easy and difficult items, making more interpretable the comparisons of retrieval practice effects between conditions. We replicated both the retrieval practice and the item difficulty effects found in Experiment 1. Unlike Experiment 1, we found a Type of Practice × Difficulty interaction trend (p = .10) in the hypothesized direction, with a slightly greater retrieval practice effect for difficult items than for easy items.

General Discussion

In two experiments, we replicated both the retrieval practice effect and the item difficulty effect. The retrieval practice effect was large and reliable, having been observed in most participants in both Experiments 1 and 2 (.98 and .89, respectively). We also investigated whether item difficulty affect the magnitude of the retrieval practice effect. According to the REH, difficult items require more retrieval effort than easier items and, consequently, should benefit more from retrieval practice. In Experiment 1, we found the opposite pattern (greater retrieval practice effect for easy items). However, different recall performance in the practice phase clouds the interpretation of this finding. When in Experiment 2 performance for easy and difficult items was matched in the practice phase, a trend emerged in the predicted direction. In the following, we elaborate on these findings, discussing in more detail the validity of the item difficulty manipulation, the relationship of our findings with findings in the literature, and implications for current memory theory.

Item Difficulty Manipulation

Two conditions must be satisfied to test the REH: (a) difficulty should vary between retrieval attempts and (b) retrieval must bem successful (Pyc & Rawson, 2009). The selection of Swahili–Portuguese word pairs previously normed for item difficulty (Lima & Buratto, 2019) sought to satisfy the first condition. Two findings suggest that this was an effective strategy. First, Lima and Buratto's normative measure strongly correlated with performance (on an item-by-item basis) in the final test, both in Experiment 1, r = .88, 95% BCa CI [.82, .93], and 2, r = .72, 95% BCa CI [.56, .83], ps < .001. Second, like in previous studies (Karpicke & Bauernschmidt, 2011; Vaughn et al., 2013), RT was used as measure of task difficulty during practice cycles. In both experiments, we found greater RT for difficult items than for easy ones during retrieval practice (but not during restudy), indicating that retrieval effort induction was indeed effective.

Recall Performance During Practice

Experiment 1 showed greater retrieval practice effect for easy items on recall performance, but not on RTs. However, recall analyses on practice cycles suggested that difficult items were not as well learned as easy ones. Despite the possibility that feedback could mitigate the negative effects of the retrieval failures (Kang et al., 2007), the initial low performance for difficult items indicates a failure to satisfy the first condition for the appropriate test of the REH. Conditional analyses have suggested that easy items benefit more from retrieval practice than difficult items. However, contrary to previous studies (Pyc & Rawson, 2009; Vaughn & Rawson, 2011; Vaughn et al., 2013), in which criterion level was directly manipulated, our analyses were post-hoc and therefore should be interpreted with caution. One reason for this concern is that these analyses are susceptible to item selection effects, that is, more retrievable items are also recalled more often in the practice phase (see Buchin & Mulligan, 2017, 2019).

Experiment 2 sought to balance initial performance in the last practice cycle in order to eliminate low initial performance as a confounding factor. The goal was to make sure that easy and difficult items were recalled at approximately the same rate in the final cycle of the practice phase. Easy items were practiced four times, whereas difficult items were practiced six times. After ensuring similar recall levels at practice, we found a trend pointing to a greater retrieval practice effect for difficult items. The trend is consistent with the REH. One possible explanation for the nonsignificant result is lack of statistical power. Our power analysis was based on η_p^2, a biased estimator, instead of the partial omega-squared, (ω_p^2; see Albers & Lakens, 2018). Had we used ω_p^2 as our estimator, the estimated sample size for Experiment 2 would increase from 28 to 38, and we might have observed a significant interaction in the predicted direction.

Alternatively, since *p*-values do not provide support for H_0 (Masson, 2011), we entered

our data into a Bayesian repeated-measures ANOVA on JASP (JASP Team, 2018). Here, we report *Bayes factor* (BF_{10}), which informs the strength of evidence in favor of H_1 relative to H_0 – in this case, evidence for a model including an interaction term against evidence for a model including only the two main effects. Values smaller than 0.33 provide support for H_0, values higher than 3 provide support for H_1 (Dienes, 2014). We obtained a $BF_{10} = 0.95$, which suggests data insensitivity, that is, we did not find positive evidence in favor of the H_0 (see Dienes, 2014). When we analyzed the RT measure for all trials, a significant interaction emerged on the expected direction. However, when the analysis was restricted to correct trials, this interaction disappeared (see Figure 6b). Lastly, conditional analyses once again indicated that easy items always benefit more from retrieval practice.

Relation to Previous Studies

Carpenter (2009) manipulated both the cue–target associative strength (weak, strong) and the type of practice (restudy, retrieval practice). Carpenter's Table 3 shows a greater retrieval practice effect for pairs with weak cues than for pairs with strong ones – the difference was .24 and .16 for pairs with weak cues against .14 and .13 for pairs with strong cues in Experiments 1 and 2, respectively. The results were discussed in light of the *elaborative retrieval hypothesis*, which posits that learners are more likely to activate related information during retrieval when pairs are weakly associated. Subsequently, this activated information will facilitate later retrieval (Carpenter, 2009). Our experiments used foreign–native word pairs (instead of weakly and strongly associated cue–targets), a longer retention interval (48 hr instead of 5 min), and repeated practice for each item (instead of only one presentation). Our results neither reject the null hypothesis (frequentist approach) nor provide strong evidence for an effect of item difficulty on the retrieval practice effects (Bayesian approach).

Vaughn et al. (2013) investigated whether retrieval practice benefits easy and difficult

items differently. Two assumptions underpin their studies. First, it is assumed that successful retrieval involves greater effort for difficult items than for easy ones. Second, it is assumed that easy items will reach an asymptote at lower criterion levels, while retrieving difficult items will remain effortful even at higher criterion levels, leading to later mnemonic gains (see Pyc & Rawson, 2009). In two experiments, Vaughn et al. found that across several criterion levels, performance in the criterion test was always better for easy items, contrary to REH prediction. Although we did not directly manipulate criterion level, our conditional analyses provided convergent evidence for their results, since more effortful successful retrievals did not reverse the advantage of easy items over difficult items.

Theoretical Implications

A recent study published by Minear et al. (2018) found a (non-significant) trend toward a greater retrieval practice effect for easy items. However, when they constrained the subsequent analyses to positive testers (i.e., individuals who benefit from retrieval practice), a different result emerged: Participants with high fluid intelligence (gF) showed a greater retrieval practice effect for difficult items than for easy items, whereas low gF showed the opposite pattern. Following Minear et al.'s procedure, we also reanalyzed Experiment 2's recall performance on the final test only for positive testers (considering all items; see Table 1). In this constrained analysis, the Type of Practice × Difficulty interaction was significant, $F(1, 24) = 4.34$, $p = .05$, $\eta_p^2 = .15$, indicating that retrieval practice effect was greater for difficult items (.47 vs. .15), $t(24) = 8.88$, $p < .001$, $d = 1.78$, than for easy items (.64 vs. .44), $t(24) = 4.88$, $p < .001$, $d = 0.98$. This finding should be taken cautiously, however, as the corresponding Bayesian repeated-measures ANOVA ($BF_{10} = 2.10$) suggests that there is no strong evidence to choose a model including the interaction term against a model including only the two main effects (see Dienes, 2014).

In light of the foregoing considerations, we conclude that the REH was partially supported by our results. In Experiment 2, after ensuring similar recall levels at practice, we found a (non-significant) trend toward a greater retrieval practice effect for difficult items. When the analysis was constrained to positive testers, this trend was significant, although Bayes Factor did not provide strong evidence for H_1. At a first look, these mixed results seem at odds with previous studies showing that task difficulty affects the magnitude of the retrieval practice effect (Agarwal et al., 2017; Carpenter & DeLosh, 2006; Karpicke & Bauernschmidt, 2011; Pyc & Rawson, 2009). However, Minear et al.'s (2018) findings suggest that we should take into account both item difficulty and learners' skills when introducing desirable difficulties.[8] To date, little is known about the relationship between individual differences and the retrieval practice effect. A recent review states that relations observed "are inconsistent and are potentially moderated by factors such as the ability range of the sample, difficulty of the items, presence or absence of feedback, delay length, and potentially other factors" (Unsworth, 2019, p. 118). Future research is needed to better examine relations between learners' skills and item difficulty.

The only individual differences variable measure in our study was the participants' number of fluent languages. One of our exploratory analyses showed that the benefits of retrieval practice are greater for multilinguals than for monolinguals (see Appendix D). Bjork and Kroll's (2015) review proposes that the learner's known languages are active and competing, which may be a desirable difficulty for learning a new vocabulary. However, this finding was not replicated in Experiment 2. Our question about number of fluent languages did not provide a clear operational definition of what participants should consider as "fluent".

[8] In their study, negative testers (i.e., individuals who benefit from restudy) outperformed positive testers in the overall final recall performance. Positive and negative testers did not differ in working memory, gF and crystallized intelligence, but they did differ in self-reported encoding strategies: Positive testers more frequently reported using shallow processing strategies, whereas self-testing was used more frequently by negative testers (Minear et al., 2018).

Future studies on individual differences may benefit from a greater control over several language variables (e.g., number of fluent languages and age-of-acquisition).

One criticism to the REH is that this account does not provide a cognitive mechanism to explain why effortful retrieval benefits memory (Karpicke, 2017). Karpicke et al. (2014) argue that not all difficult (or effortful) retrievals are beneficial to memory. As an example, they mentioned that dividing attention during retrieval practice did not improve memory (Gaspelin et al., 2013). In fact, some difficulties are *undesirable* difficulties, "if the learner, by virtue of prior knowledge and current cues, is not equipped to respond to them successfully" (Bjork & Kroll, 2015, p. 242; see also Bjork, 1994). To put it another way, it is imperative to balance retrieval success on the one hand and retrieval effort on the other (Karpicke, 2017).

One possible way to reconcile our results with explanatory accounts is to consider the episodic context account, which proposes that effort is important as it leads to context reinstatement (Karpicke et al., 2014). In this sense, we can argue that a longer interval between successive retrieval attempts leads to a greater degree of context reinstatement than a smaller interval, because contextual cues change more during longer intervals than during shorter intervals. It is not clear, however, how item difficulty could engage different degrees of context reinstatement during retrieval practice. Thus, according to the episodic context account, there is no clear reason to expect a greater retrieval practice effect for difficult items.

The episodic context account also suggests that retrieval practice (but not restudy) helps the learner to restrict the search set of potential targets in subsequent retrieval attempts. This should be reflected in a shorter RT for previously retrieval practiced items than for previously restudied ones (see Lehman et al., 2014). Our Experiment 2 indicates longer RT for restudied items in the final test phase, as predicted by the episodic context account. When the analysis was restricted to correct trials, differences in RT disappeared. This suggests that the additional time to recall targets previously restudied did not imply retrieval success, that

is, the search set was unrestricted in the absence of retrieval practice.

In a similar vein, both the decreasing linear trends in RT across practice cycles (see Figures 2b and 5b) and the shorter RT in the criterion test for retrieval practiced items (see Figures 3b and 6b) are consistent with an automatization account of the retrieval practice effect (Racsmány et al., 2018). As automatization takes place, effortful processing decreases. This should entail in shorter RT across cycles as well as shorter RT for retrieval practiced items in the final test, as observed in our experiments. However, it is not yet clear how these two concepts (effort and automatization) are interrelated in the accounts of retrieval practice.

Judgments of Learning

It is noteworthy that participants' JOLs did not differ in Experiment 1 and favored retrieval practice in Experiment 2. Although several between-participant designs showed that JOLs favored restudied over retrieval practiced condition (e.g., Roediger & Karpicke, 2006b), Tullis, Finley, and Benjamin (2013) have suggested that one factor (out of four) that is important for accurate metacognitive judgments is the opportunity to compare different conditions of processing. Hence, JOLs should be more accurate in both within-participant and mixed-list designs, which is the design employed here. Moreover, in self-regulated scenarios, participants usually believe they have "learned" an item as long as they can recall it (Karpicke, 2009). Possibly, higher JOLs for retrieval practice condition in Experiment 2 can be an artifact from higher retrieval success for difficult items achieved in Experiment 2 compared to Experiment 1. This hypothesis should be further explored in future studies.

Our design did not allow us to assess learner's metacognitive judgments as a function of item difficulty. To do so, it would be necessary to separate easy and difficult items (i.e., a "blocked-by-difficulty design"), so that participants knew which set of items was being judged. Another way to conduct this analysis would be by using an item-by-item-JOL procedure instead of an aggregate-JOL procedure (Karpicke, 2009; Tullis et al., 2013).

Although the REH is silent about the metacognitive results after practicing easy and difficult items, it is an interesting question whether these different judgments are accurate, since they can guide subsequent time allocation by learners in self-regulated scenarios.

Limitations

Our study has some limitations. First, the presence of feedback introduces indirect effects of retrieval practice (Roediger & Karpicke, 2006a). Retrieval practice can benefit subsequent memory both through retrieval processes themselves and through feedback, exposing learners to correct answers (Karpicke, 2017). Despite this drawback, we chose to provide feedback like a previous study (Minear et al., 2018), both to mitigate the negative effects of retrieval failures (Kang et al., 2007) and to ensure that participants would have new opportunities to encode the word pairs. Due to the features of our experimental design, it is not possible to disentangle the relative contributions of the direct benefits of retrieval practice and indirect benefits introduced by feedback. Future studies may attempt to replicate our results addressing the REH by not providing feedback to participants. Second, we used a cued-recall task in both practice and final test phases, as the REH has been primarily tested through this type of task. It is important, however, to test this hypothesis using other tasks, such as free recall and recognition. Free recall, in particular, may prove an interesting test for the REH because it yields a larger retrieval practice effect than cued-recall and recognition, and it is associated with less frequent use of feedback (Rowland, 2014).

The results presented in this study are informative for practitioners. An equal retrieval practice effect size for easy and difficult materials suggests that this technique can be extended to a wide range of materials, not only easy materials, which is important in both educational and clinical contexts. The evidence that retrieval practice boosts retention for easy and difficult items alike has important implications for its use both as a learning tool and as a rehabilitation technique.

References

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

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *24*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Bangert, A. S., & Heydarian, N. M. (2017). Recall and response time norms for English–Swahili word pairs and facts about Kenya. *Behavior Research Methods*, *49*, 124–171. https://doi.org/10.3758/s13428-015-0701-1

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. Retrieved from https://psycnet.apa.org/PsycARTICLES/journal/bul

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). Hillsdale, NJ: Erlbaum.

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

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, *128*(2), 241–252. Retrieved from https://www.jstor.org/journal/amerjpsyc

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.* Advance online publication. https://doi.org/10.1177/1747021819847141

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., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. Retrieved from https://link.springer.com/journal/13421

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory & Cognition*, *22*, 249–257. Retrieved from https://link.springer.com/journal/13421

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

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

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

Fiechter, J. S., & Benjamin, A. S. (2017). Diminishing-cues retrieval practice: A memory-enhancing technique that works when regular testing doesn't. *Psychonomic Bulletin & Review*, *25*(5), 1868–1876. https://doi.org/10.3758/s13423-017-1366-9

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

Gaspelin, N., Ruthruff, E., & Pashler, H. (2013). Divided attention: An undesirable difficulty in memory retention. *Memory & Cognition*, *41*(7), 978–988. https://doi.org/10.3758/s13421-013-0326-5

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

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112. https://doi.org/10.1007/BF02289823

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian–English paired associates. *Behavior Research Methods*, *42*(3), 634–342. https://doi.org/10.3758/BRM.42.3.634

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

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117. Retrieved from https://link.springer.com/journal/13428

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007

JASP Team. (2018). JASP (Version 0.7.5.5) [Computer software].

Kahneman, D. (1973). *Attention and effort.* Upper Saddle River, NJ: Prentice Hall.

Kang, S. H. K. (2017). The benefits of interleaved practice for learning. In J. C. Horvath, J. Lodge, & J. A. C. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 79–93). New York, NY: Routledge.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469–486. https://doi.org/10.1037/a0017341

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). Oxford: Academic Press.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250–1257. https://doi.org/10.1037/a0023436

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*, 417–479. https://doi.org/10.1080/09658210802647009

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 237–284). San Diego, CA: Elsevier Academic Press.

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

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. (2019). *Norms of familiarity, concreteness, valence, arousal, wordlikeness, and memorability for Swahili–Portuguese word pairs.* Manuscript under preparation.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690. https://doi.org/10.3758/s13428-010-0049-5

Middleton, E. L., Schwartz, M. F., Rawson, K. A., Traut, H., & Verkuilen, J. (2016). Towards a theory of learning for naming rehabilitation: Retrieval practice and spacing effects. *Journal of Speech, Language, and Hearing Research*, *59*(5), 1111–1122. https://doi.org/10.1044/2016_JSLHR-L-15-0303

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

Mulligan, N. W., & Picklesimer, M. (2016). Attention and the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 938–950. https://doi.org/10.1037/xlm0000227

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, *2*(3), 325–335. https://doi.org/10.108/09658219408258951

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Putnam, A. L., & Roediger, H., L., III. (2013). Does response mode affect amount recalled or the magnitude of testing effect? *Memory & Cognition*, *41*, 36–48. https://doi.org/10.3758/s13421-012-0245-x

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. (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

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

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

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

Smith, M. A., Roediger, H. L., III, & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1712–1725. https://doi.org/10.1037/a0033569

Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, *30*(1), 203–218. https://doi.org/10.5334/irsp.90

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*, 429–442. https://doi.org/s13421-012-0274-5

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 607–617. https://doi.org/10.1037/0278-7393.5.6.607

Underwood, B. J. (1982). Paired associate learning: Data on pair difficulty and variables that influence difficulty. *Memory & Cognition*, *10*, 610–617. Retrieved from https://link.springer.com/journal/13421

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

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

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review*, *20*, 1239–1245. https://doi.org/10.3758/s13423-013-0434-z

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

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia*, *51*(12), 2360–2370. https://doi.org/10.1016/j.neuropsychologia.2013.04.004

Final Considerations

This thesis tested the predictions of the retrieval effort hypothesis (REH; Pyc & Rawson, 2009), addressing the following question: Does item difficulty affect the magnitude of the retrieval practice effect? If so, which items are most benefited from retrieval practice: easy or difficult ones?

Before testing the REH itself, we conducted a normative study (Manuscript 1), in order to select stimuli whose difficulty was known. This study makes an important methodological contribution by allowing the design of carefully controlled memory studies with foreign–native word pairs in Brazilian Portuguese. In two subsequent retrieval practice experiments (Manuscript 2), we replicated both the retrieval practice (Roediger & Karpicke, 2006; Rowland, 2014) and item difficulty effects (Cull & Zechmeister, 1994; Underwood, 1982). We found greater retrieval practice effect for easy items (Experiment 1) and a (non-significant) trend toward a greater retrieval practice effect for difficult items, particularly for positive testers (Experiment 2). The results provide only weak evidence for the REH.

Although the absence of a strong support for the REH seems discouraging at a first glance, we prefer to view these results in a positive light. First, from an applied perspective, the evidence that retrieval practice boosts retention for both easy and difficult items has important implications for its use as a learning tool and as a rehabilitation technique, as long as the practitioner guarantees the successful retrieval by the learner. Second, from a theoretical standpoint, it is possible that difficult (or effortful) retrievals are beneficial to memory only to the extent that it engages the processes that are useful for subsequent retention (Karpicke, Lehman, & Aue, 2014). Future studies will benefit from investigating cognitive mechanisms underlying the positive effects of certain types of difficulties (e.g., spacing; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), relating them both to item difficulty and to individual difference variables (Unsworth, 2019).

References

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory & Cognition*, *22*, 249–257. Retrieved from https://www.psychonomic.org/page/MC

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 237–284). San Diego, CA: Elsevier Academic Press.

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

Roediger, H. L., III, & Karpicke, J. D. (2006). 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

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

Underwood, B. J. (1982). Paired associate learning: Data on pair difficulty and variables that influence difficulty. *Memory & Cognition*, *10*, 610–617. Retrieved from https://www.psychonomic.org/page/MC

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

Appendix A: Approval by the Research Ethics Committee (Manuscripts 1 and 2)

UNB - INSTITUTO DE CIÊNCIAS HUMANAS E SOCIAIS DA UNIVERSIDADE

Plataforma Brasil

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Normas de características de palavras para uso em experimentos de memória
Pesquisador: Luciano Grüdtner Buratto
Área Temática:
Versão: 2
CAAE: 80056117.1.0000.5540
Instituição Proponente: Instituto de Psicologia -UNB
Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 2.616.770

Apresentação do Projeto:
Trata-se de projeto de pesquisa já avaliado anteriormente, conforme se depreende de parecer pregresso. Uma pesquisa do campo da psicologia cognitiva com 60 pessoas mais ou menos sobre a memória de palavras em português e sualí. Uma pesquisa a ser realizada em software já existente e de manuseio do Instituto de Psicologia.

Situação do Parecer:
Aprovado
Necessita Apreciação da CONEP:
Não

Continuação do Parecer: 2.616.770

BRASILIA, 24 de Abril de 2018

Assinado por:
Érica Quinaglia Silva
(Coordenador)

Figure A1. Approval by the Research Ethics Committee (Manuscript 1).

UNB - INSTITUTO DE CIÊNCIAS HUMANAS E SOCIAIS DA UNIVERSIDADE

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: O papel do esforço cognitivo no efeito de testagem: Evidências comportamentais e fisiológicas

Pesquisador: MARCOS FELIPE RODRIGUES DE LIMA

Área Temática:

Versão: 1

CAAE: 87030218.5.0000.5540

Instituição Proponente: Instituto de Psicologia -UNB

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 2.681.537

Apresentação do Projeto:

Trata-se de projeto de pesquisa de mestrado acadêmico no Instituto de Psicologia da Universidade de Brasília.

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Continuação do Parecer: 2.681.537

BRASILIA, 29 de Maio de 2018

Assinado por:
Érica Quinaglia Silva
(Coordenador)

Figure A2. Approval by the Research Ethics Committee (Manuscript 2).

Appendix B: Written Informed Consents (WICs; Manuscripts 1 and 2)

WIC (Manuscript 1, Study 1)

Você está sendo convidado a participar da pesquisa "Normas de características de palavras para uso em experimentos de memória", de responsabilidade de Luciano Grüdtner Buratto, professor do Instituto de Psicologia da Universidade de Brasília. O objetivo desta pesquisa é investigar características de palavras em português (ex., concretude, familiaridade) e em uma língua pouco familiar. Assim, gostaria de consultá-lo/a sobre seu interesse e disponibilidade de cooperar com a pesquisa.

Você receberá todos os esclarecimentos necessários antes, durante e após a finalização da pesquisa. Seu nome não será divulgado, sendo mantido o mais rigoroso sigilo mediante a omissão total de informações que permitam identificá-lo/a. Os dados provenientes de sua participação na pesquisa ficarão sob a guarda do pesquisador responsável pela pesquisa.

A coleta de dados será realizada por meio de tarefas de julgamento de palavras. É para estes procedimentos que você está sendo convidado a participar. Sua participação na pesquisa não implica nenhum risco. O tempo total dessa pesquisa será de aproximadamente 30 minutos.

Espera-se com esta pesquisa contribuir com a identificação de características de palavras que são importantes para a memorização. Ao final da sessão, você terá a oportunidade de esclarecer suas dúvidas sobre os procedimentos que realizou, tendo a oportunidade de aprender sobre o processo de pesquisa.

Sua participação é voluntária e livre de qualquer remuneração ou benefício. 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. Se você tiver qualquer dúvida em relação à pesquisa, você pode me contatar através do telefone [redacted] ou pelo e-mail lburatto@unb.br.

Objetiva-se divulgar os resultados encontrados no projeto por meio de congressos e de publicações científicas em periódicos nacionais e/ou internacionais, respeitando-se o sigilo de suas informações individuais. A equipe de pesquisa garante que os resultados do estudo divulgados por estes canais serão devolvidos a você, mediante sua solicitação no e-mail acima.

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 desse Termo de Consentimento Livre e Esclarecido ou aos direitos do participante da pesquisa podem ser obtidas por meio do e-mail do CEP/CHS: cep_chs@unb.br.

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 _____

WIC Consent (Manuscript 1, Study 2)

Você está sendo convidado a participar da pesquisa "Normas de características de palavras para uso em experimentos de memória", de responsabilidade de Luciano Grüdtner Buratto, professor do Instituto de Psicologia da Universidade de Brasília. O objetivo desta pesquisa é investigar características do processo de memorização de palavras em uma língua pouco familiar. Assim, gostaria de consultá-lo/a sobre seu interesse e disponibilidade de cooperar com a pesquisa.

Você receberá todos os esclarecimentos necessários antes, durante e após a finalização da pesquisa, e lhe asseguro que o seu nome não será divulgado, sendo mantido o mais rigoroso sigilo mediante a omissão total de informações que permitam identificá-lo/a. Os dados provenientes de sua participação na pesquisa ficarão sob a guarda do pesquisador responsável pela pesquisa.

A coleta de dados será realizada por meio de tarefa de memorização de pares de palavras. É para estes procedimentos que você está sendo convidado a participar. Sua participação na pesquisa não implica nenhum risco. O tempo total dessa pesquisa será de aproximadamente 60 minutos.

Espera-se com esta pesquisa contribuir com a construção de normas de memorização pares de palavras. Ao final da sessão, você terá a oportunidade de esclarecer suas dúvidas sobre os procedimentos que realizou, tendo a oportunidade de aprender sobre o processo de pesquisa.

Sua participação é voluntária e livre de qualquer remuneração ou benefício. 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. Se você tiver qualquer dúvida em relação à pesquisa, você pode me contatar através do telefone [redacted] ou pelo e-mail lburatto@unb.br.

Objetiva-se divulgar os resultados encontrados no projeto por meio de congressos e de publicações científicas em periódicos nacionais e/ou internacionais, respeitando-se o sigilo de suas informações individuais. A equipe de pesquisa garante que os resultados do estudo divulgados por estes canais serão devolvidos a você, mediante sua solicitação no e-mail acima.

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 desse Termo de Consentimento Livre e Esclarecido ou aos direitos do participante da pesquisa podem ser obtidas por meio do e-mail do CEP/CHS: cep_chs@unb.br.

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 _____

WIC (Manuscript 2, Experiment 1)

(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 "O papel do esforço cognitivo no efeito de testagem: Evidências comportamentais e fisiológicas", cujo pesquisador responsável é Marcos Felipe Rodrigues de Lima, estudante de mestrado 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 de estudo afetam a memória. Investigações sobre memória são importantes, pois permitem compreender quais variáveis influenciam como as pessoas aprendem e retêm informações a longo prazo. Os procedimentos da pesquisa envolvem a realização de tarefas de memória, que serão realizados em duas sessões: a primeira será realizada hoje, com duração estimada de 80 minutos; e a segunda será realizada em 48 horas, com duração estimada de 30 minutos. Sua participação na pesquisa não implica em nenhum risco. Ao final da segunda sessão, você terá a oportunidade de esclarecer suas dúvidas sobre os procedimentos que realizou na pesquisa.

O estudo será realizado no LIPSI (Laboratório Integrado de Pós-Graduação e Pesquisa Experimental em Psicologia com Humanos), 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 osintegrantes da equipe de pesquisa terão acesso a seus dados pessoais.

Os resultados dessa pesquisa serão divulgados sob a forma de dissertação de mestrado

do pesquisador responsável, o qual ficará disponível no Repositório Institucional da UnB (http://repositorio.unb.br), provavelmente a partir do início do segundo semestre de 2019. Esclarecimentos poderão ser feitos a qualquer momento da pesquisa, mediante contato com o pesquisador responsável [telefone: [redacted]; e-mail: lima.piraju@gmail.com].

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.

Este documento encontra-se redigido em duas vias, sendo uma para o participante e outra para o pesquisador.

Brasília, _____ de_____ de _____.

_____	_____
Assinatura do Participante	Assinatura do Pesquisador Responsável

WIC (Manuscript 2, Experiment 2)

(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 "O papel do esforço cognitivo no efeito de testagem: Evidências comportamentais e fisiológicas", cujo pesquisador responsável é Marcos Felipe Rodrigues de Lima, estudante de mestrado 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 de estudo afetam a memória. Investigações sobre memória são importantes, pois permitem compreender quais variáveis influenciam como as pessoas aprendem e retêm informações a longo prazo. Os procedimentos da pesquisa envolvem a realização de tarefas de memória, que serão realizados em duas sessões: a primeira será realizada hoje, com duração estimada de 60 minutos; e a segunda será realizada em 48 horas, com duração estimada de 30 minutos. Sua participação na pesquisa não implica em nenhum risco. Ao final da segunda sessão, você terá a oportunidade de esclarecer suas dúvidas sobre os procedimentos que realizou na pesquisa.

O estudo será realizado no LIPSI (Laboratório Integrado de Pós-Graduação e Pesquisa Experimental em Psicologia com Humanos), 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 dissertação de mestrado

do pesquisador responsável, o qual ficará disponível no Repositório Institucional da UnB (http://repositorio.unb.br), provavelmente a partir do início do segundo semestre de 2019. Esclarecimentos poderão ser feitos a qualquer momento da pesquisa, mediante contato com o pesquisador responsável [telefone: [redacted]; e-mail: lima.piraju@gmail.com].

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.

Este documento encontra-se redigido em duas vias, sendo uma para o participante e outra para o pesquisador.

Brasília, _____ de_____ de _____.

_____	_____
Assinatura do Participante	Assinatura do Pesquisador Responsável

Appendix C: Experimental Stimuli Used in Experiments 1 and 2 (Manuscript 2)

Table C1

Parameters of Swahili–Portuguese Word Pairs Used in Experiments 1 and 2

Difficulty	Swahili	Portuguese	English[a]	Memorability[b]
Easy	roho	alma	soul	.67
	pipa	barril	barrel	.75
	punda	burro	donkey	.70
	mbwa	cachorro	dog	.65
	pombe	cerveja	beer	.65
	elimu	ciência	science	.47
	godoro	colchão	mattress	.47
	goti	joelho	knee	.51
	buu	larva	maggot	.53
	wasaa	lazer	leisure	.59
	nafaka	milho	corn	.51
	wingu	nuvem	cloud	.57
	lulu	pérola	pearl	.73
	nabii	profeta	prophet	.49
	malkia	rainha	queen	.73
	chura	sapo	frog	.47
	chama	sociedade	society	.65
	dafina	tesouro	treasure	.63
	nyanya	tomate	tomato	.75
	usingizi	sono	sleep	.47

(Table C1 continues)

Table C1. continuation

Difficulty	Swahili	Portuguese	English[a]	Memorability[b]
Difficult	lozi	amêndoa	almond	.22
	nanga	âncora	anchor	.27
	ambo	cola	glue	.25
	iktisadi	economia	economy	.26
	bahasha	envelope	envelope	.21
	samadi	estrume	manure	.25
	ankra	fatura	invoice	.22
	jeraha	ferida	wound	.27
	hadithi	história	story	.28
	bustani	jardim	garden	.26
	yamini	juramento	oath	.25
	ziwa	lago	lake	.28
	hamira	levedura	yeast	.18
	inda	malícia	spite	.29
	utenzi	poema	poem	.23
	lango	portão	gate	.28
	ladha	sabor	flavor	.14
	ruba	sanguessuga	leech	.28
	hariri	seda	silk	.28
	handaki	trincheira	trench	.14

[a] Original English word normed for Nelson and Dunlosky (1994).

[b] Memorability was computed as average proportion of participants that correctly recalled items over three cycles of tests (see Lima & Buratto, 2019).

Appendix D: Exploratory Analyses (Manuscript 2)

For completeness, exploratory analyses conducted are reported next, for both Experiments 1 and 2.

Experiment 1

Distracter task. We found that participants had a higher proportion of correct answers on easy problems (*Mdn* = 1.00) than on difficult ones (*Mdn* = .10). Wilcoxon signed-rank test showed that there was a highly reliable problem difficulty effect, $z = 6.25$, $p < .001$, $r = .88$.

Relationship between performance on main and distracter tasks. It is possible that some participants used subvocal rehearsal or another kind of retrieval practice strategy during distracter task periods. The use of such strategies could possibly lead to lower performance in the distracter task and to produce an enhancement in performance on main task. To check this possibility, we correlated performance on distracter task with performance on the immediately subsequent practice cycle. We hypothesize that if some participants engaged in retrieval practice strategy during distracter task, a negative correlation could be expected between performance in this task and in main task. No significant correlations were observed across cycles, $rs \leq .06$, $ps \geq .69$. This suggests that, in fact, participants engaged in the distracter task, instead of some kind of retrieval practice strategy.

Analysis using covariates. Retention interval differed between participants (range = 42–53 hr). Moreover, the number of fluent languages differed according to each participant's report (range = 1–5). Therefore, we reanalyzed our main dependent variable (i.e., proportion of items correctly recalled on final test) considering this variation. We entered our data in a 2 (type of practice) × 2 (difficulty) repeated-measures analysis of covariance (ANCOVA), including both retention interval and number of fluent languages as covariates. Although the two covariates were nonsignificant, ANCOVA revealed a significant Type of Practice × Number of Fluent Languages interaction, $F(3, 32) = 5.74$, $p = .003$, $\eta_p^2 = .35$. Pearson's

correlation indicated that there was a positive relationship between retrieval practice effect size and number of fluent languages, r = .41, p = .003. After recoding participants as monolinguals (i.e., speakers of the native language only) and multilinguals (i.e., speakers of two or more languages), a *t*-test showed that the retrieval practice effects are greater for multilinguals than for monolinguals, $t(49) = 2.92$, p = .005, d = 0.34.

Experiment 2

Distracter task. Participants had a significant higher proportion of correct answers on easy problems (*Mdn* = 1.00) than on difficult ones (*Mdn* = .36), $z = 4.63$, $p < .001$, r = .87.

Relationship between performance on main and distracter tasks. We again correlated performance on distracter task with performance on the immediately subsequent practice cycle. No significant correlations were observed across cycles, rs ≤ .32, ps ≥ .10, suggesting that participants engaged in distracter task, instead of some kind of retrieval practice strategy.

Analysis using covariates. Retention interval differed between participants (range = 46–53 hr). Similarly, the number of fluent languages differed according to each participant's report (range = 1–3). Like in Experiment 1, we reanalyzed our main dependent variable entering our data in a 2 (type of practice) × 2 (difficulty) repeated-measures ANCOVA, with both retention interval and number of fluent languages as covariates. Unlike Experiment 1, the Type of Practice × Number of Fluent Languages interaction was nonsignificant, $F(2, 18) = 0.25$, p = .79, η_p^2 = .03. All other effects were nonsignificant, Fs ≥ 1.60, ps ≤ .21.