Optimizing Memory in Non-Traditional Settings with Evidence-based Learning Strategies

Mihaylova, Mariela

UNIVERSITÉ DE GENÈVE
FACULTÉ DE PSYCHOLOGIE ET DES SCIENCES DE L'ÉDUCATION

Section de Psychologie

Sous la direction de Prof. Dr. Matthias KLIEGEL (UNIGE) et Prof. Dr. Nicolas ROTHEN (UniDistance Suisse)

Optimizing Memory in Non-Traditional Settings with Evidence-based Learning Strategies

THESE

Présentée à la
Faculté de psychologie et des sciences de l'éducation
de l'Université de Genève
pour obtenir le grade de Docteur en Psychologie

par
MIHAYLOVA, Mariela

de
Sofia (Bulgarie) / New York (États-Unis)

Thèse No: 909

14-306-104
(numéro d'étudiant)

GENÈVE

9 JUILLET, 2025

,

Committee Members:

Prof. Dr. Matthias Kliegel

University of Geneva, Geneva, Switzerland

Prof. Dr. Nicolas Rothen

UniDistance Suisse, Brig, Switzerland

Prof. Dr. Ulrike Rimmele

University of Geneva, Geneva, Switzerland

Jury Members:

Prof. Dr. Matthias Kliegel

University of Geneva, Geneva, Switzerland

Prof. Dr. Nicolas Rothen

UniDistance Suisse, Brig, Switzerland

Prof. Dr. Andreas Ihle

University of Geneva, Geneva, Switzerland

Dr. Alexandra Hering

Tilburg University, Tilburg, Netherlands

Acknowledgments

This work was conceived while staying home for 2 years during COVID lockdown between 2020-2022, conducted while commuting 18 hrs/week from Geneva to Brig for 1.5 years between 2022-2023, and completed while working full time in a fast-paced corporate world for the last 2 years. All these phases brought unique perspectives, challenges, and lessons to my PhD experience and showed me the true meaning of *"Every challenge can be turned into an opportunity,"* and *"When you love what you do, you don't work a day in your life."* Yet, none of this would have been possible or worthwhile without the faith and support of those around me.

First and foremost, a millions thanks and a world of appreciation for my co-supervisors, Prof. Matthias Kliegel and Prof. Nicolas Rothen. Thank you both for your guidance, patience, and the incredible trust you've placed in me over the years. Matthias, thank you for welcoming me into the CAL in 2021, for your insights and advice, and for your generous support at every step. Nicolas, thank you for your mentorship, for pushing me to grow, and for helping me discover my passion for scientific writing. I wouldn't be the researcher I am today without you both.

I want to sincerely thank Prof. Ulrike Rimmele for being part of my committee and my graduate academic journey since I first came to Switzerland over ten years ago—from supervising my Master's work and my first peer-reviewed publication to now, coming full circle. I also sincerely thank Prof. Andreas Ihle and Dr. Alexandra Hering for agreeing to be part of my jury, thank you both for your support and interest in my research.

To all my colleagues at UniDistance Suisse - thank you for your collaboration, discussions, conferences, encouragement, hikes in Valais, journal clubs, and all the warm memories made in Brig - they will always hold a dear place in my heart. Special thanks to Prof. Thomas Reber, Dr. Simon Gorin, and Prof. Alodie Rey-Mermet for their contributions to the papers in this thesis and valuable guidance during my days at UniDistance.

To everyone at the CAL, thank you for the kindness you've shown me since joining your lab in 2021, the motivation, writing weeks in Gstaad, lunches & drinks, advice, and all the support these last two months as I finish this thesis. Special shout-out to Dr. Max Haas for being my go-to person anytime I needed lab access or had a random question :)

I want to thank the Master students I had the opportunity to work with during my time at the CIGEV between 2022-2023, Matthias Leitner and Marie Zocca. Working with you both was so inspiring and I wish you success and happiness in everything you do. I also want to thank all the participants at the UniDistance Suisse participant pool and the University of Geneva's Cognitive Aging Course for their contributions to my research between 2022-2023.

A huge thanks goes out to Dr. Erika Yashiro for working round the clock to help me proof this thesis right before submission - you're a star!

I could not have done the last leg of my PhD journey these last two years without my amazing colleagues at Weber Shandwick Switzerland/IPG DXTRA. Thank you all for your flexibility, support, and cheering me on as I complete this major milestone and for seeing the value in my work brought forth to the clients we service every day.

Lastly, though words will never be enough, I want to thank my favorite people in the world, my family–Nicolas, my parents, Lubomir and Nina, Laura, Alain, and my sister, Mila. Your love, pride, and faith in me means everything and I'm so lucky to have you. Endless thanks to my parents for their countless sacrifices that made this moment possible, their unwavering belief in me, and for always reminding me of my purpose when I felt like giving up. Nicolas, thanks for keeping me from going in guns blazing at every minor inconvenience, for always making sure I had a steady supply of homemade bone broth to keep my brain working, and for standing by me through it all <3

Statement

The present work is a cumulative dissertation based on three articles (Thèse sur dossier d'articles) representing three studies. It has been prepared as a self-contained work and all chapters were composed specifically for this thesis. I confirm that I wrote the three papers as first author according to the APA author guidelines with feedback and edits from co-authors. This thesis includes the published paper of Study 1, in-principle acceptance of Study 2 (registered report), and a pre-print version of Study 3. Terminological and formatting inconsistencies may occur due to different journal publishing guidelines, as well as due to the use of Latex/Overleaf for this thesis.

This thesis includes the following three articles:

Study 1/Article 1: Mihaylova, M., Gorin, S., Reber, T. P., & Rothen, N. (2022). A meta-analysis on mobile-assisted language learning applications: Benefits and risks. *Psychologica Belgica, 62*(1), 252.

Study 2/Article 2: Mihaylova, M., Zocca, M., Kliegel, M., & Rothen, N. (2024). Does retrieval practice protect memory against stress? A meta-analysis [Stage 2 Registered Report].

Study 3/Article 3: Mihaylova, M., Rey-Mermet, A., & Rothen, N. (2025). Does retrieval practice protect memory against test anxiety? [Pre-print, see OSF: https://osf.io/aefy4/]

Abstract

Retrieval practice, spacing, feedback, and multisensory learning are the most effective learning strategies we know of today that can optimize learning. Their benefits on memory and learning have previously been shown using traditional pen-and-paper approaches, but little is known about their ability to enhance memory in non-traditional, underexplored, and real-world learning contexts such as mobile applications. Study 1 filled this knowledge gap by conducting a meta-analysis on the effectiveness of mobile learning applications and found a strong effect size ($g = 0.88$, 95% CI [0.62, 1.14]) for mobile applications on learning, as well as a strong effect size ($g = 0.95$, 95% CI [0.56, 1.34]) for learning outcomes using mobile applications that featured retrieval practice. Retrieval practice specifically has also been shown to aid memory in other underexplored contexts, such as during stress, where it is shown to have protective effects. To date, few studies have adequately investigated its protective benefits in general and for other real-life stressors, like test anxiety. Study 2 conducted a meta-analysis to determine whether retrieval practice can make memory less sensitive to the detrimental effects of stressors. Study 2 found moderate evidence for retrieval practice compared to restudy in stress situations ($g = 0.45$, 95% CI [0.19, 0.71]), though the effect of the stressor was not confirmed overall. Study 3 experimentally tested whether learning with retrieval practice can protect memory against the detrimental effects of test anxiety by subjecting participants to evaluative or control testing conditions. Results showed that retrieval practice was effective regardless of evaluative condition, though the effect of the test anxiety induction was again unsupported. Together, these results highlight a complex interplay among learning strategies and stress and advance our understanding of how these different factors and contexts shape memory.

Résumé en français

Le "retrieval practice", le feedback, l'espacement et l'apprentissage multisensoriel sont aujourd'hui reconnus comme les stratégies d'apprentissage les plus efficaces pour optimiser la mémorisation. Leurs bienfaits sur la mémoire et l'apprentissage ont été démontrés dans des contextes traditionnels utilisant papier et crayon, mais on sait peu de choses sur leur efficacité dans des contextes non traditionnels, peu explorés ou en conditions réelles, comme les applications mobiles. L'étude 1 a comblé cette lacune en menant une méta-analyse sur l'efficacité des applications mobiles d'apprentissage. Elle a révélé un effet important (g = 0,88, IC 95% [0,62, 1,14]) des applications mobiles sur l'apprentissage, ainsi qu'un effet fort (g = 0,95, IC 95% [0,56, 1,34]) pour les applications intégrant la pratique de récupération. La "retrieval practice" a également montré des bénéfices sur la mémoire dans d'autres contextes sous-explorés, comme en situation de stress, où elle semble avoir un effet protecteur. À ce jour, peu d'études ont examiné de manière rigoureuse ses bénéfices protecteurs en général et face à d'autres facteurs de stress réels, comme l'anxiété liée aux examens. L'étude 2 a mené une méta-analyse pour déterminer si le "retrieval practice" rend la mémoire moins sensible aux effets néfastes du stress. Elle a trouvé une évidence modérée en faveur de le "retrieval practice" comparée à la relecture en situation de stress (g = 0,45, IC 95% [0,19, 0,71]), bien que l'effet du facteur de stress n'ait pas été confirmé de manière générale. L'étude 3 a testé expérimentalement si l'apprentissage avec "retrieval practice" peut protéger la mémoire contre les effets négatifs de l'anxiété liée aux tests, en exposant les participants à des conditions d'évaluation ou de contrôle. Les résultats ont montré que le "retrieval practice" était efficace, quelle que soit la condition évaluative, mais là encore, l'effet de l'induction de l'anxiété liée au test n'a pas été confirmé. Pris ensemble, ces résultats mettent en évidence une interaction complexe entre les stratégies d'apprentissage et le stress, et approfondissent notre compréhension de la manière dont ces facteurs et contextes influencent la mémoire.

Table of Contents

List of Figures

List of Tables

"A memory is what is left when something happens and does not completely unhappen."

- Edward de Bono, 1933-2021

1. Introduction

Learning is at the core of everything we do. Cognitive learning strategies like retrieval practice, feedback, spacing, and multisensory learning can enhance our ability to learn and remember. At the same time, information technologies like mobile applications are changing the way we learn information. Identifying how these experimentally backed learning strategies can improve mobile learning is critical to successfully adapt and integrate new learning technologies (Reber & Rothen, 2018). On the other hand, stressful events, such as exams, lead to decreased memory retrieval (Shields, Doty, et al., 2017; Shields, Sazma, et al., 2017; Vogel & Schwabe, 2016). Whether learning strategies can protect memory during these situations is a key concern for the memory and learning field, and exploring these strategies further may help design effective cognitive interventions that protect memory in the face of stressors. The aim of this thesis is to explore whether these learning strategies can optimize learning in two areas that are critical for everyday life but have not yet been examined sufficiently, namely mobile applications and stress. This aim will be addressed through three main research questions: 1) Are learning strategies implemented in mobile learning applications, and can they enhance learning in this context? 2) Can learning strategies protect memory in the context of stress? 3) Can learning strategies protect memory in the context of test anxiety?

1.1 Learning Strategies

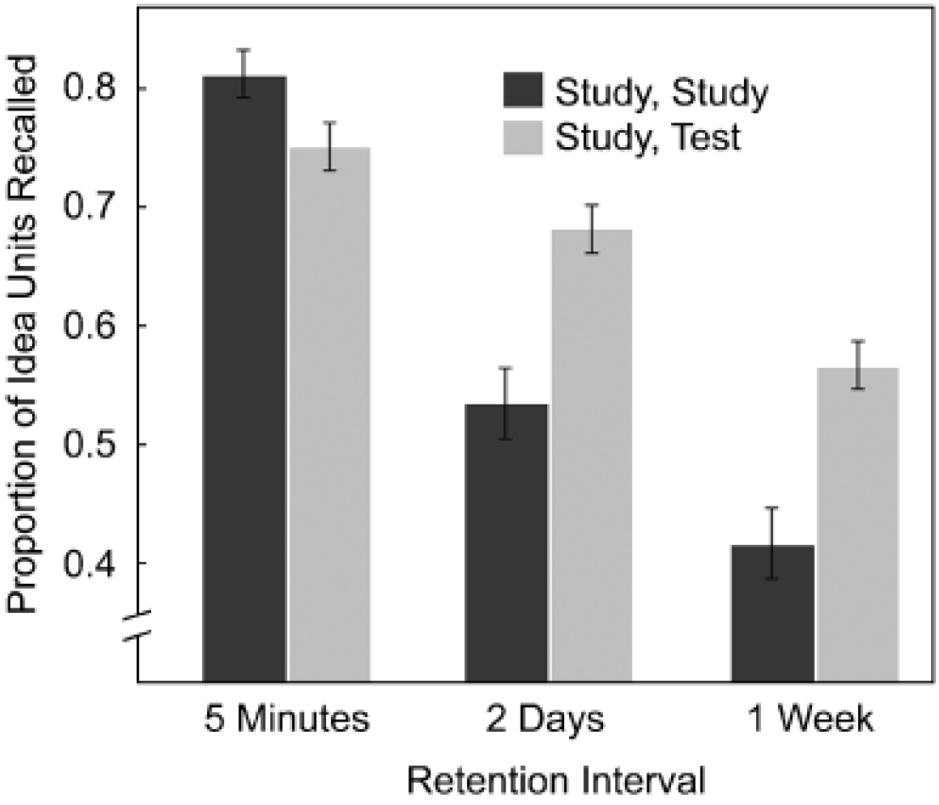
A learning strategy is a process of organizing and selecting the appropriate way to learn that best suits the current task demands (Gettinger & Seibert, 2002). Learning strategies improve learning because they recruit a wide range of cognitive competencies, such as acquiring, analyzing, organizing, synthesizing, and using information in a goal-oriented and skillful way that enable understanding and learning and allow students to retain information and think critically (Akinleke & Adeaga, 2014; Gettinger & Seibert, 2002; Kleijn et al., 1994; Mendezabal, 2013). The field of memory and learning has identified four highly effective learning strategies that lead to better learning outcomes and

memory performance, namely retrieval practice, spacing, feedback, and multisensory learning (Carpenter et al., 2022; Dunlosky et al., 2013; Reber & Rothen, 2018; Weinstein et al., 2018). The following sections elaborate on these strategies and the theories that explain their effectiveness.

1.1.1 Retrieval Practice

Retrieval practice involves actively retrieving information from memory (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006). It is consistently shown to improve memory more so than simple restudying, a phenomenon known as the testing effect (Karpicke & Roediger, 2008; Roediger & Pyc, 2012; Rowland, 2014; Yang et al., 2018; Zaromb & Roediger, 2010). In a seminal study, Roediger and Karpicke (2006) presented students (N = 120) with two study passages to read for seven minutes. Then, participants were asked to either restudy (reread) the passage or take a short test, in which they wrote down as much of the passage as they could remember (retrieval practice). After a delay of five minutes, two days or one week, participants were asked to retrieve as much as they could from both passages. Results revealed that for the longer retention intervals of two days and one week, participants who used retrieval practice performed significantly better than those in the restudy condition, highlighting the long-term effectiveness of this strategy and illustrating the testing effect (Figure 1).

Figure 1
Sustained benefits of retrieval practice



Note. Results of the seminal study from Roediger & Karpicke (2006) showing the mean proportion of idea units recalled on the final test after retention intervals of five minutes, two days, or one week, as a function of learning condition. Figure adapted from Roediger & Karpicke, 2006.

Since Roediger and Karpicke's (2006) seminal study, the benefits of retrieval practice on memory have been demonstrated for a variety of learning materials, content types and types of learning, including visuo-spatial learning (Kang, 2010), vocabulary learning (Goossens et al., 2014; Kang et al., 2013), in classroom learning and across all levels of schooling (Agarwal, D'Antonio, Roediger III, et al., 2014; Moreira et al., 2019; Ritchie et al., 2013; Schwieren et al., 2017). Retrieval practice has also benefited patients suffering memory impairments after brain injury (Coyne et al., 2015) and multiple sclerosis (Sumowski et al., 2010b). Retrieval practice is effective across different test formats, like

free recall, cued recall, and short-answer (Carpenter et al., 2006; Moreira et al., 2019). Retrieval practice can also benefit the learning and retention of new information encountered after the initial practice, a phenomenon known as the forward testing effect (Pastötter & Bäuml, 2014; Pastötter & Frings, 2019; Yang et al., 2018) and enhance retention of previously studied material, known as the backward testing effect (Roediger & Karpicke, 2006). Additionally, retrieval practice can be covert, which involves silently recalling information internally, and it can be overt, or requiring external expression of the recalled content (M. A. Smith et al., 2013).

Retrieval practice consistently outperforms other frequently used learning strategies. For example, many studies have shown its effectiveness over highlighting or restudying (Moreira et al., 2019; Roediger & Karpicke, 2006; Rowland, 2014; Schwieren et al., 2017). Carpenter et al. (2016) found that retrieval practice with feedback was more effective than simply copying notes in a biology course. When compared to mind-mapping, another commonly used technique that supposedly requires more effort than mere note-taking, Karpicke and Blunt (2011) found that retrieval practice produced significantly better learning outcomes, a result that has been replicated in other studies (O'Day & Karpicke, 2020; Ritchie et al., 2013).

Several theories have been proposed to explain why retrieval practice is so effective. The episodic context account, the leading theory on the topic, states that contextual memory cues become bound to the learned items during learning (Lehman et al., 2014). This contextual information becomes updated with new information each time the item gets retrieved, thereby making it easier to access the memory traces at each retrieval. These changes involve strengthening the links between the memory trace and its contextual cues, as well as integrating additional retrieval-relevant features, which together enhance the accessibility of the memory in future recall attempts (Karpicke, 2017; Lehman et al., 2014).

The transfer-appropriate processing (TAP) theory posits that memory performance is optimal when the encoding processes during initial learning align with those used during

recall. Because retrieval practice involves actively recalling information, this mirrors the cognitive demands of future assessments, thereby strengthening memory traces and improving retention (Karpicke, 2017). Other theories suggest that retrieval practice creates "desirable difficulties" by requiring effortful retrieval, which strengthens memory consolidation and improves retention (Adesope et al., 2017; E. L. Bjork & Bjork, 2011; R. A. Bjork & Bjork, 2020; Wenzel & Reinhard, 2021).

Recently, the bifurcation model entered the scene to explain the benefits of retrieval practice (Kornell et al., 2011; Racsmány et al., 2020). This model posits that retrieval attempts selectively strengthen the memories of successfully retrieved items, while unretrieved items remain unaffected. This process creates a divergence or "bifurcation" in memory strength, with retrieved items becoming more durable and accessible over time, while unretrieved items fail to benefit. The model also emphasizes that retrieval attempts are extremely important and suggests that strategies or conditions that maximize retrieval success can improve long-term retention and benefit retrieval practice.

1.1.2 Spacing

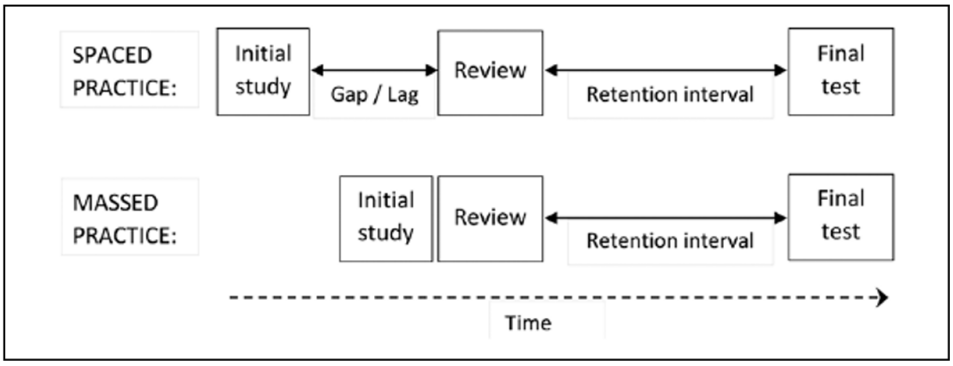
Spacing, synonymous with distributed learning, is when learning is conducted over several intervals with time in-between each interval. This approach dates back to the foundational research of Ebbinghaus, who concluded it to be more effective than learning with short intervals between sessions, known as massed training or massed learning (or more commonly, "cramming"). Ebbinghaus himself stated that, "With any significant number of repetitions, spreading them out over time is significantly more beneficial than clustering them together at one time" (Ebbinghaus, 2013). Today, this theory is known as the spacing effect (Figure 2).

Spaced learning has been shown to significantly improve long-term memory formation. In Sobel et al. (2011), students who reviewed vocabulary words one week after initial learning performed 177% better on a test than those who practiced massed. Despite believing massed learning was more effective, Kornell and Bjork (2008) found that students

who used spaced practice had enhanced long-term conceptual understanding, supporting better retention and category formation.

Figure 2

Experimental paradigm for the spacing effect



Note. Figure adapted from Kang (2016) illustrating a spaced practice schedule and a massed practice schedule.

One theory explaining why spacing is so effective is the retrieval theory. This theory suggests that when an item is repeated after a delay, it triggers active recall of the prior occurrence, which, in turn, strengthens memory. In contrast, massed repetition eliminates this retrieval process because the information is only presented and does not require retrieval (Wahlheim et al., 2014). Bahrick and Hall (2005) proposed that learners experience more retrieval failures when sessions are spaced over time, which push them to change and improve their memory strategies, like using better mental images or connections. This strategy shift leads to stronger, longer-lasting memory. In contrast, when practice is massed, fewer failures occur, so learners stick to less effective strategies. As a side note, this theory was later evolved into the mediator-shift hypothesis to also explain the benefits of testing with restudying (Pyc & Rawson, 2012). Another perspective is the contextual variability theory, which proposes that spaced repetitions occur in different

learning contexts, both internally and externally, providing a greater diversity of retrieval cues that enhance memory retention (Glenberg, 1979). These theories are not mutually exclusive, and it is likely that multiple mechanisms contribute to the memory advantage observed with spaced practice.

1.1.3 Corrective Feedback

Corrective feedback is a type of feedback that provides information to a learner about how well a task is being accomplished or performed (Hattie & Timperley, 2007; Metcalfe, 2017). In the Power of Feedback, Hattie and Timperley (2007) state that the "purpose of feedback is to reduce discrepancies between current understandings and performance and a goal." It typically involves distinguishing between correct and incorrect answers, which is needed for feedback to be effective (Pashler et al., 2005), acquiring additional or different information, or building more surface knowledge (Figure 3).

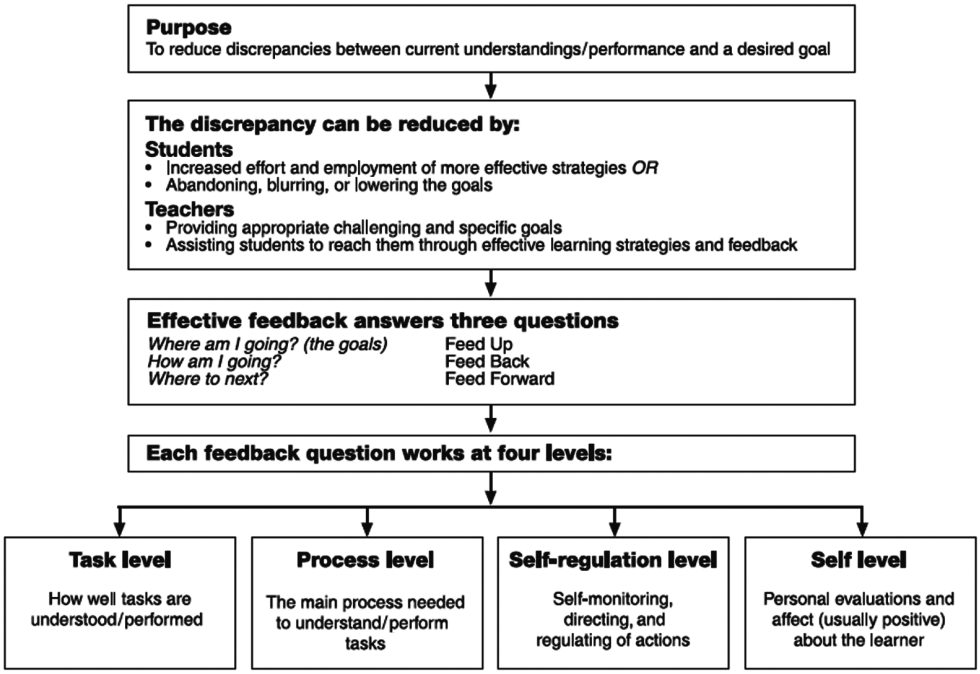
In Butler and Roediger (2008), participants studied reading passages during a study phase and were then given a multiple-choice test with either immediate feedback, delayed feedback (at the end of the test), or no feedback. A final cued recall memory test was administered after a one-week delay. Results showed that participants in the immediate and delayed feedback conditions recalled significantly better at the delayed memory test than participants who did not receive feedback (Butler & Roediger, 2008), suggesting that feedback improved learning. Meta-analyses also suggest that peer feedback has a highly positive effect on various learning outcomes, including second language learning (Vuogan & Li, 2023), medical education (Kim & Kim, 2023), and various school assignments (Fryer & Leenknecht, 2023).

The underlying mechanisms that explain why feedback is effective relate to the "prediction-error" signal in the brain (Friston, 2005; Wilkinson, 2014). When individuals receive feedback that calls out discrepancies, this feedback triggers a prediction-error signal that prompts learners to adjust their internal expectations to better align with reality. By comparing predictions with actual outcomes on an iterative basis and by updating internal

models accordingly, corrective feedback reduces prediction-error and facilitates learning. Another theory called recursive reminding suggests that errors made while learning can enhance memory retrieval by prompting individuals to recall the contextual details surrounding their mistakes and subsequent corrections (Jacoby & Wahlheim, 2013). The idea here is that individuals are likely to remember the context in which they made the error, which would then prompt them to remember the correct answer.

Figure 3

A model of giving feedback to improve learning



Note. Figure adapted from Hattie and Timperley (2007).

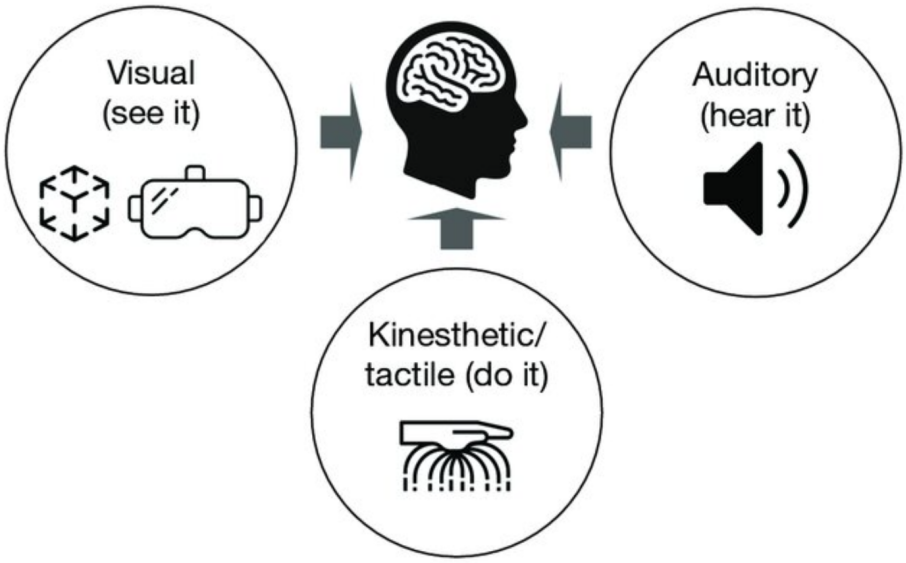
1.1.4 Multisensory Learning

Multisensory, or multi-modal, learning is the notion that learning from two or more streams of information is better than unimodal learning, or using one stream (Shams & Seitz, 2008). For instance, learning using both written diagrams and listening to a lecture is better than using only one of these sources of information (Figure 4). This strategy

aligns closest with how we behave in the real world because our brain is constantly perceiving and integrating information from multiple streams. Experimental evidence suggests that when courses are redesigned to incorporate multimedia components such as discussion forums and projects as opposed to a textual delivery format, significant improvements in student engagement and outcomes are observed (Schilling, 2009). Favorable learning outcomes are also been achieved when language learning material is presented in different modes (Ajayi, 2012).

Multisensory strategies are thought to enhance learning by providing multiple neural processing routes for the storage and retrieval of that information (Shams & Seitz, 2008). In other words, because information is stored in various forms (i.e., visual, auditory, ect), it can be accessed and retrieved more easily using any number of neural access routes. Additionally, audio-visual presentations of learning material activate larger areas of the brain, including those responsible for processing auditory and visual information, compared to unimodal presentations. This utilization of multiple processing pathways is believed to aid retrieval by leveraging redundant information stored across distributed brain regions (Shams & Seitz, 2008), thus facilitating learning.

Figure 4
Multisensory inputs for learning



Note. Figure adapted from Sanfilippo et al. (2022).

1.1.5 Summary

Retrieval practice, spacing, corrective feedback, and multisensory learning are four of the most effective learning strategies we know of today (Dunlosky et al., 2013; Reber & Rothen, 2018; Weinstein et al., 2018). Moreover, retrieval practice and spacing receive the highest utility ratings because they benefit learners of all ages and abilities (Dunlosky et al., 2013). Studies exploring the effectiveness of these learning strategies have been conducted in classrooms or laboratory-based settings (Moreira et al., 2019; Rowland, 2014; Weinstein et al., 2018). Despite their effectiveness in these contexts, these learning strategies have not been adequately explored in non-traditional learning contexts, such as when using mobile applications or during stressful situations like acute stress or exams. Therefore, investigating whether these strategies can enhance learning in non-traditional contexts is a critical next step in memory and learning research.

1.2 Mobile Learning

Non-traditional learning environments encompass a wide range of learning contexts that depart from the traditional classroom pen-and-paper setting. One such environment is mobile-assisted learning, also known as mobile learning, or m-learning, which refers to the use of mobile devices for learning (Viberg & Grönlund, 2012). Mobile-assisted learning has emerged as a popular form of non-traditional learning since the turn of the millennium with the proliferation of mobile devices like smartphones, tablets, or other handheld mobile devices. One type of mobile learning that has become popular over the last 20 years is mobile-assisted language learning (MALL), which is the use of mobile learning technologies to learn a language (Bano et al., 2018; Bozdoğan, 2015; Sung et al., 2015, 2016). This chapter describes the advances, benefits, and challenges of mobile learning. Note that while MALL is a well-researched area of mobile learning, this thesis does not focus on language learning itself but instead uses MALL as a lens for examining learning technologies in a different setting, allowing us to explore their broader applications and effectiveness.

1.2.1 Benefits of Mobile Learning

Mobile learning such as MALL offers many advantages over traditional learning approaches, and these advantages make it a highly attractive mode of learning. For example, it offers high levels of flexibility and convenience because learners can access educational materials anytime and anywhere (Kukulska-Hulme & Bull, 2009). This flexibility allows learners to tailor their learning experience to their own needs and preferences, enabling them to learn at their own pace and on their own schedule. Additionally, mobile devices allow learners to take advantage of "micro-learning" opportunities, or short bursts of learning that can be completed during small pockets of free time, such as during a commute or while waiting in line (Bozdoğan, 2015; Viberg & Grönlund, 2012).

Another advantage of mobile learning is its ability to engage learners through multimedia content. Mobile devices can display text, images, videos, and audio, providing

a range of media for learners to interact with and learn from. This variety of content formats allows learners to engage with the material in ways that are most effective for them, increasing motivation and enhancing retention (Shadiev et al., 2018).

Mobile learning also facilitates real-time communication, and allows learners to connect with their peers and teachers for discussion, feedback, and support. This collaborative aspect of mobile learning can enhance the learning experience and foster a sense of community (Bano et al., 2018; Bozdoğan, 2015; Sung et al., 2016; Tseng et al., 2007). Finally, mobile learning can offer cost savings for both learners and educators, as these devices are typically less expensive than computers and the access to educational materials they offer can reduce the need for costly resources like textbooks.

1.2.2 Advances in Mobile Learning

Over the last 20 years, thousands of applications have been developed by individuals, professionals and companies to aid learning. For instance, there has been a proliferation of mobile applications to support foreign language vocabulary learning, such as mobile flashcards and dictionaries (Bozdoğan, 2015; Rahimi & Miri, 2014; Q. Wu, 2014). These mobile applications typically feature a learning system in which a target word is presented on one side and the learner has several seconds to guess the correct answer before it appears.

Additionally, a broad range of MALL applications have been developed and utilized to facilitate other language skills, including grammar, reading abilities, writing skills and pronunciation. For example, Li and Hegelheimer (2013) developed an application called Grammar Clinic to assist users with their writing by helping them identify and correct sentence-level errors. The application targeted 15 common grammatical error types, which were identified based on their prevalence in a locally developed learner writing corpus. Users were tasked with identifying and correcting errors, and feedback was provided for each item. Similarly, Cavus (2016) developed an interactive mobile application to help learners improve their skills in a multitude of language components, including vocabulary,

pronunciation, listening, and comprehension, without the need for teacher assistance. The authors integrated a speech recognition engine on mobile phones to identify words spoken by learners, thereby facilitating the correction of pronunciation errors.

Mobile learning applications have also been designed to offer personalized content tailored to the user's learning levels, interests, and learning cycles. For example, Chen and Chung (2008) developed a personalized mobile learning system that adjusts learning modes based on learners' prior knowledge and memory cycles, enhancing learning performance and interests. Hsieh et al. (2012) proposed a personalized English article recommendation system that suggests articles for learners to read based on their profiles, effectively improving their English proficiency levels. Similarly, Hsu et al. (2013) designed a mobile learning system featuring a reading material recommendation workflow that suggests articles based on learners' preferences and knowledge levels, together with a reading annotation module that allows them to take notes from reading content.

1.2.3 Effectiveness and Challenges

Studies have reported conflicting results on the effectiveness of mobile learning. Several meta-analyses have indicated that mobile learning is more effective than traditional classroom-based approaches. Taj et al. (2016) and H. Cho et al. (2016) reported medium effect sizes ($d = 0.43$ and $d = 0.51$, respectively) for MALL on foreign language learning, while Mahdi (2018) demonstrated a medium effect size ($d = 0.67$) for vocabulary learning. Garzón et al. (2023) found a medium-level ($g = 0.89$) positive effect on students' learning outcomes compared to traditional lectures, pedagogical tools, or other multimedia resources, particularly for Bachelor's level education. Sung et al. (2016) also observed a medium effects favoring MALL for language learning methods, albeit with uncertain long-term effectiveness. These findings, reinforced by studies in math and science areas (Bano et al., 2018; Drigas & Pappas, 2015; Güler et al., 2022; Tlili et al., 2023), highlight the potential of MALL to enhance learning outcomes.

However, other studies have not found a meaningful impact of mobile learning on

learning outcomes. Martin and Ertzberger (2013) explored the effectiveness of mobile learning compared to computer-based instruction in a group of 109 undergraduate students. They found that, although the mobile learning group had positive attitudes about the experience, they performed less well on a post-test than a computer-based learning group. When Rachels and Rockinson-Szapkiw (2018) examined the mean difference in performance between students who learned via a popular language learning application compared to traditional learning methods, they observed a negative small effect size ($g = 0.05$), suggesting mobile learning may not always be effective.

Research also points to several challenges in the mobile learning field that limit any conclusions that can be drawn on its overall effectiveness. For example, Baran (2014) noted the lack of clear best practices for how to best implement mobile learning in learning outcomes. Scholars also agree that the field is fraught with a high degree of heterogeneity and moderating factors among learning outcomes (Tamim et al., 2011), making it difficult to distinguish whether the learning benefits originate from the applications themselves or from other sources. Learners are exposed to a wide array of mobile applications with varying levels of effectiveness, making it challenging to discern which options are most beneficial for their educational needs (Tamim et al., 2011).

Moreover, experts highlight the need for mobile learning to be based on scientifically-backed learning strategies, including retrieval practice, spacing, feedback, and multisensory learning (Bano et al., 2018; Reber & Rothen, 2018; Roediger & Pyc, 2012). Currently, a plethora of mobile applications are developed every year for different learning purposes, but there is little effort to standardize and agree on how to optimize these applications based on proven learning principles (Bano et al., 2018; Sung et al., 2016). In their review studies, Sung et al. (2015, 2016) point out how important learning strategies are missing from mobile applications and that their inclusion may improve learning outcomes and effectiveness, but the authors do not mention which strategies are found to be missing. This reveals a gap in our understanding of the effectiveness of mobile learning

applications and the integration of effective learning strategies in these applications.

1.2.4 Summary

The inconsistent findings in the literature on the effectiveness of mobile learning and its challenges indicates that a different approach to mobile learning needs to be assessed. Experts emphasize that mobile learning must be grounded in well-established learning strategies, including retrieval practice, spacing, feedback, and multisensory learning, to maximize its effectiveness and promote meaningful learning outcomes (Bano et al., 2018; Reber & Rothen, 2018; Roediger & Pyc, 2012). These four learning strategies are proposed because they have been extensively researched and proven effective in traditional educational settings (Dunlosky et al., 2013; Rowland, 2014; Schwieren et al., 2017).

1.3 Stressful Learning

Another type of non-traditional learning environment that affects our learning is stressful learning. Stress is how organisms respond to any real or anticipated threats (McEwen & Gianaros, 2011; Wolf, 2017). Stress encompasses psychological stressors like uncontrollable situations or socio-evaluative threat like performance being judged or evaluated (Cassady & Johnson, 2002; Dickerson & Kemeny, 2004; Wenzel & Reinhard, 2021). The detrimental effects of stressors on memory, particularly memory retrieval, have been well-documented (Kuhlmann, 2005; Oei et al., 2006; Shields, Sazma, et al., 2017). Retrieval practice, one of the four most effective learning strategies discussed in the previous section, has emerged as a potential means to counteract these negative effects (A. M. Smith et al., 2016). However, the extent to which it can enhance and protect memory in stressful settings remains unclear. This chapter aims to explore the role of psychological stressors in learning, their adverse effects on memory, and how retrieval practice may offer a path to mitigate these effects.

1.3.1 Stress and Test Anxiety

Test anxiety, also called "academic stress," is a type of psychological stressor that occurs during evaluative situations (Cassady & Johnson, 2002). Test anxiety is associated

with a wide range of unfavorable symptoms, such as memory retrieval issues, decreased motivation, increased potential for making mistakes, lack of concentration, disruptions in attention, higher cognitive load, and reduced effort and persistence (Cassady, 2004; Eysenck et al., 2007b; Hembree, 1988). Theories on test anxiety posit that it is triggered by perceptions of threat (Lazarus & Folkman, 1987), and this concept has been consistently supported by studies showing that students commonly perceive exams as stressful and unpleasant (Beilock & Carr, 2016; Cardozo et al., 2020; Sarason, 1961, 1984; Wenzel & Reinhard, 2021). This stress is not limited to graded assessments, and even tests used solely for learning purposes can induce anxiety due to their challenging nature (Wenzel & Reinhard, 2021). Factors like struggling to grasp certain subjects, fear of failure, and disappointment from not meeting expectations can also lead to anxiety and worry, which, in turn, contribute to academic stress and performance deficits (Cardozo et al., 2020; Monteiro et al., 2007).

Laboratory studies demonstrate that both low-stakes and high-stakes tests evoke feelings of pressure and anxiety, with high-stakes tests leading to even higher levels of stress (Hinze & Rapp, 2014). Schoofs et al. (2008) examined the relationship between stress biomarkers, such as cortisol and saliva, and oral exam performance. They exposed participants to two oral examination sessions several weeks apart. Significant increases in cortisol levels were observed on and during exam day (Schoofs et al., 2008), suggesting increased levels of stress. Stojanović et al. (2021) investigated the impact of psychological stress on salivary parameters by having participants perform cognitive tasks over three days and measuring stress biomarkers. They found a significant increase in salivary cortisol levels post-task, indicating heightened stress. Similarly, M. Cohen and Khalaila (2014) explored saliva as a stress biomarker and its ability to predict exam performance. They observed higher salivary pH levels post-exam, associated with reduced threat appraisal, stress, and test anxiety levels. Additionally, they found that pH levels could be used to predict perceived threat, stress, and test anxiety dimensions, meaning that these markers

are viable correlates for increased stress levels.

1.3.2 The Stress Response

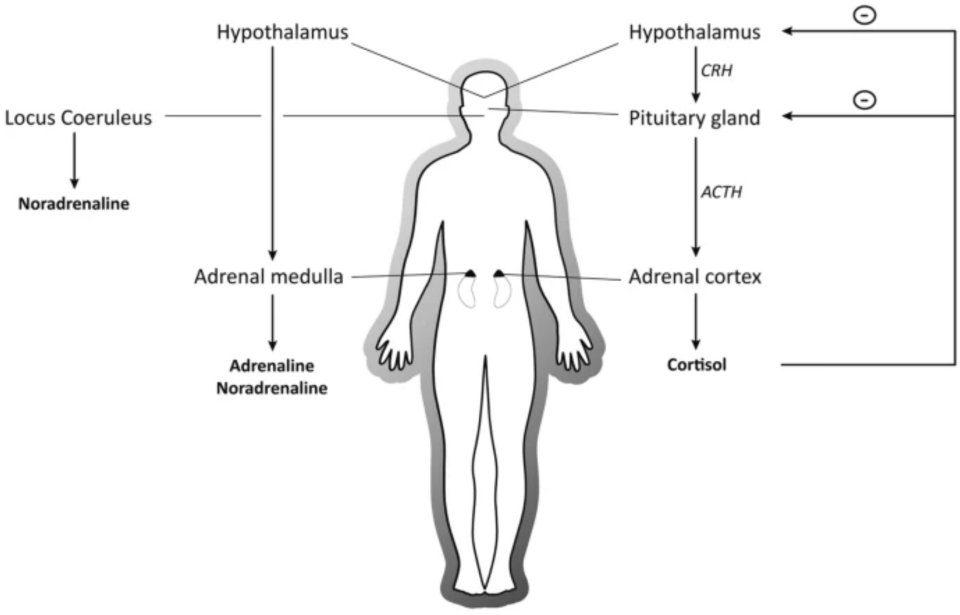
Psychological stressors like test anxiety can engage the body's stress response (Figure 5). The stress response is made up of two nervous systems, the fast-acting autonomic nervous system (ANS) and sympathetic nervous system (SNS), and the slower-acting hypothalamus-pituitary-adrenal axis (HPA axis) (Schwabe & Wolf, 2010b; Shields, Doty, et al., 2017; Wolf, 2017). Within seconds of stress onset, the ANS and SNS activate, triggering the release of neurotransmitters like noradrenaline from the locus coreulus. These neurotransmitters prepare the body for a fight or flight response and decrease activity in prefrontal regions of the brain responsible for decision-making and planning (Roozendaal, 2002). Then, on a slower timescale of around 20-60 minutes, the HPA axis activates, leading to increased production of glucocorticoids such as cortisol from the adrenal glands. These hormones freely cross the blood-brain barrier and directly influence neural activity in key memory centers in the brain like the hippocampus, amygdala, and prefrontal cortex by binding to receptors found on neurons in those regions (Cabeza & Nyberg, 2000; Joëls & Baram, 2009; Schwabe & Wolf, 2013; Shields, Sazma, et al., 2017; Smeets, 2011). On an even longer time-scale of several days and weeks following stress onset, neuroendocrine-immune interactions are observed (Pruett, 2001). These interactions lead to the production of proinflammatory proteins that alter activity in key memory centers of the brain, immunosuppression, autoimmune disoders, and other functions that result in negative health consequences.

When individuals experience stressors like test anxiety, the body's stress response activates, leading to the release of stress hormones such as cortisol. These hormones can cause various physical symptoms such as sweating, increased heart rate, and rapid breathing, which are frequently reported in individuals with test anxiety (Cassady & Johnson, 2002; Hembree, 1988; Schoofs et al., 2008). Additionally, cognitive symptoms associated with test anxiety, such as difficulty concentrating and racing thoughts, can

exacerbate the stress response even further (Cassady & Johnson, 2002). Moreover, when stress becomes excessively severe or persists over an extended period, the level of these stress hormones can remain elevated, hindering the body's ability to adapt effectively. Such situations can increase the likelihood of developing various health problems (McEwen & Gianaros, 2011).

Figure 5

Diagram of body systems involved in the stress response



Note. In response to stress, the autonomic nervous system (ANS) rapidly activates, triggering the release of catecholamines like noradrenaline from the adrenal medulla and brainstem, initiating the 'fight-or-flight' response and affecting attention and memory. Shortly after, the hypothalamic–pituitary–adrenal (HPA) axis is engaged: the hypothalamus releases CRH, leading to ACTH secretion from the pituitary, which stimulates cortisol release from the adrenal cortex. Cortisol peaks 20–30 minutes post-stressor and crosses the blood-brain barrier to influence cognition and emotion. It also exerts negative feedback on the HPA axis and brain areas like the hippocampus to help end the stress response. Figure and note adapted from Vogel and Schwabe (2016).

1.3.3 Stress Induction Methods

Stress can be induced in laboratory settings using various paradigms. The most well-established procedure is called the Trier Social Stress Test (TSST). The TSST is a timed protocol involving tasks where performance is judged, such as preparing for an interview and solving math problems (Kirschbaum et al., 1993). In addition to performance judgments, individuals acting as the "judges" also treat subjects coldly without showing emotion or enthusiasm, leading to psychological stress (Kirschbaum et al., 1993).

Stress can also be induced through instructions that hint at performance being judged or scrutinized, such as intelligence tests (Almazrouei et al., 2022; Wenzel & Reinhard, 2021). Another way stress can be induced is through physical procedures like the Cold Pressor Test (CPT), where individuals immerse their hands in cold water. The CPT task can be accompanied by socio-evaluative elements like solving math problems or being judged (Schwabe & Schächinger, 2018; Schwabe et al., 2008). These methods of stress induction are associated with increases in subjective anxiety measures and cortisol levels.

1.3.4 Effects of Stress on Memory

The type of memory that is explored in this thesis is episodic memory, which is the ability to remember past events (Tulving, 1993). The effects of stress on episodic memory depend on several factors, including timing. For example, stress experienced after encoding typically enhances memory (Roozendaal, 2002; Smeets et al., 2008), especially for emotional information (Buchanan & Lovallo, 2001), whereas stress induced before retrieval impairs memory performance (de Quervain et al., 2000; Guenzel et al., 2013; Kuhlmann, 2005; Smeets, 2011). This impairment is typically seen when memory is measured 15-20 minutes after stress induction when the cortisol level is at its highest (Schwabe & Wolf, 2014; Shields, Sazma, et al., 2017). In this thesis, we are most interested in retrieval stress.

Specifically, test anxiety is associated with memory retrieval deficits (Cassady, 2010; Huberty, 2009; Salend, 2011). When individuals experience test anxiety, their cognitive resources may become overwhelmed by intrusive thoughts, worries, and physiological

arousal, leading to difficulties in accessing and retrieving information stored in memory (Cassady & Johnson, 2002). Moreover, the stress response associated with test anxiety may impair the functioning of memory-related brain regions, further exacerbating memory retrieval deficits (Vogel & Schwabe, 2016). These findings echo results reported in large-scale meta-analyses examining the effects of test anxiety on memory, suggesting that stressors like test anxiety are negatively correlated with a wide range of educational performance outcomes, including memory performance on standardized tests, university entrance exams, and grade point average (Hembree, 1988; von der Embse et al., 2018).

1.3.5 Mechanisms of Action

One way through which retrieval stress impairs memory is through context-dependent mechanisms. The context-dependent memory theory suggests that memory is improved when information is recalled in the same context in which it was originally learned (S. M. Smith & Vela, 2001). However, a shift in context, such as experiencing stress, can disrupt this process (S. M. Smith & Vela, 2001). In their meta-analysis, Shields, Sazma, et al. (2017) proposed this context-shift theory as an underlying mechanism to explain how stress acts on memory. They suggested that stress could create a mental context shift that disrupts context-dependent memory. For example, individuals might learn information in a calm state but struggle to recall it when stressed.

Support for the context-shift hypothesis comes from studies demonstrating that stress-induced changes in mental state during memory retrieval can impair memory when the retrieval context differs from the learning context (Schwabe et al., 2009). Trammell and Clore (2014) discovered that post-encoding stress had a detrimental effect on memory, unlike the enhancing effect that is usually observed when stress is induced right after learning. The authors attributed this discrepancy to a primary methodological difference: their participants experienced a change in context between learning and stress, whereas most other studies exposed participants to stress in the same context as learning. This memory impairment effect was also *not* seen in cases where the memory acquisition phase

and stressor have the same context (e.g., Smeets et al., 2007). Together, these studies suggest that stress disrupts the context-dependent nature of memory by creating a mismatch between the encoding and retrieval contexts, leading to impaired memory.

1.3.6 Stress and Retrieval Practice

Stress can disrupt context-dependent memory by altering the internal state or mental context in which information was initially encoded (Shields, Sazma, et al., 2017; A. M. Smith & Thomas, 2018a). This means that when individuals experience stress during memory retrieval, the context in which they learned the information may not match their state at retrieval. As a result, it becomes more challenging to access and recall memories effectively. Retrieval practice may offer a potential solution to this deficit by improving memory accessibility. When individuals recall information in various contexts, they become better at reinstating the specific contextual cues associated with the initial learning environment, thereby making contextual details more readily available for future retrieval attempts (Akan et al., 2018; Karpicke, 2012; Lehman et al., 2014). As a result, even when stressors are present during retrieval, the strengthened memory traces and enhanced availability of contextual details facilitated by retrieval practice can mitigate the negative impact of stress on memory performance.

To date, research on the protective role of retrieval practice against retrieval stress is scarce. A.M. Smith et al. (2016) conducted the first such investigation. They demonstrated that participants who learned via retrieval practice and were exposed to stress via a TSST protocol outperformed those who restudied and were also exposed to stress. These results indicated that retrieval practice could create strong memories that are resilient to the negative effects of acute stress on memory. However, they were not able to replicate this finding in a next study (A. M. Smith et al., 2018), nor did other researchers (Szőllősi et al., 2017). More recently, Klier and Buratto (2023) investigated the interaction among retrieval practice, stress, and the difficulty level of study materials on memory retention. Participants studied Swahili-Portuguese word pairs through restudy or retrieval

practice. A week later, they took a cued-recall test before undergoing a stress or control procedure. Notably, stress reduced recall of easy items but enhanced recall of difficult items that had been successfully retrieved during encoding. Taken together, these studies suggest that retrieval practice may mitigate the memory retrieval deficit typically experienced during stress. However, the findings are inconsistent and sometimes inconclusive.

The literature for stressors like test anxiety and retrieval practice is mixed. Agarwal et al. (2014) found that students who underwent a retrieval practice intervention in school reported feeling less nervous about upcoming exams. This outcome suggested that retrieval practice may decrease some of the symptoms associated with test anxiety that lead to memory issues. Similarly, Szpunzar et al. (2013) found that participants who engaged in interim testing reported lower anxiety about cumulative tests than those who did not. These findings are consistent with other research by Brown and Tallon (2015), who observed that pre-lecture quizzes reduced test anxiety before exams, and Piroozmanesh and Imanipour (2018), who found that nursing students who took regular quizzes during a course experienced significantly lower test anxiety before a final exam compared to those who did not. More recently, Yang et al. (2023) looked at the overall impact of practice tests (i.e., quizzes) on test anxiety across 24 studies with 3,374 participants. The results indicated strong evidence that practice tests significantly reduce test anxiety to a moderate extent, with easy quizzes being more effective than difficult ones. Together, these studies indicate that retrieval practice may alleviate test anxiety, partly by familiarizing students with testing and partly by reducing fear of the unknown (Yang et al., 2023)

In other cases, results do not show protective effects of retrieval practice on test anxiety, particularly when assessed with other variables such as working memory capacity (Tse & Pu, 2012; Yang et al., 2020). Bertilsson et al. (2021) investigated whether individual differences in personality traits and working memory capacity moderated the benefits of retrieval practice on long-term memory retention. Participants learned word pairs through retrieval practice and restudying, with assessments at three time points: five

minutes, one week, and four weeks after practice. They found a significant testing effect at all time points, with no observable association between the testing effect and the examined personality traits or working memory capacity.

Hinze and Rapp (2014) investigated the impact of low- versus high-stakes tests and retrieval practice on memory retention. Undergraduate students read biology texts and were then placed in either a low-stakes condition or a high-stakes condition where a (monetary) reward was contingent on performance (i.e., the reward is given based on performance in the high-stakes condition but not contingent on performance in the low-stakes). Despite similar initial quiz performance, participants in the high-stakes condition performed worse on the final memory tests a week later compared to those in the low-stakes condition. This observation led the authors to propose the "retrieval disruption hypothesis," which suggested that stressful, or high-stakes, tests may disrupt memory formation following initial retrieval practice. Similarly, Wenzel and Reinhard (2021) found that learning tests, particularly in comparison to reading tasks, are perceived as more negative and stressful, with dispositional stress and anxiety amplifying these effects. Their findings suggested that testing (i.e., retrieval practice) in itself could trigger stress in participants because it is more challenging than re-reading.

1.3.7 Summary

Stressful experiences, including test anxiety, are known to decrease memory retrieval (Gagnon et al., 2019a; Vogel & Schwabe, 2016; Wolf, 2017) by shifting the mental context associated with the learned material (Shields, Sazma, et al., 2017). In contrast, retrieval practice increases contextual cues during each retrieval attempt (Lehman et al., 2014). As such, retrieval practice can help equalize the learning and recall environments by creating strong memories at encoding (A. M. Smith, 2018), thereby overcoming the negative impact of stress and test anxiety on memory (Mihaylova & Rothen, 2025).

The limited amount of research available in this area suggests that the role of retrieval practice in mitigating the negative impacts of stress on memory is promising, yet

mixed. Some studies provide evidence that retrieval practice can strengthen memory traces, making them more resistant to the detrimental effects of acute stress (Agarwal, D'Antonio, Roediger, et al., 2014; A. M. Smith et al., 2016). Yet the relationship between retrieval practice and test anxiety is not always straightforward (Clark et al., 2018; Hinze & Rapp, 2014). While certain studies found weak relationships between test anxiety and retrieval practice (Tse & Pu, 2012; Yang et al., 2020), others show that individual differences do not moderate the benefits of retrieval practice on memory (Bertilsson et al., 2021). This necessitates further investigation to fully understand the protective relationship between retrieval practice and stress.

2. Research Questions

The literature on mobile-assisted learning applications is growing, but systematic assessments of the specific advantages of mobile applications designed solely for learning purposes compared to traditional learning methods are lacking (Bano et al., 2018; Sung et al., 2016; Taj et al., 2016). Additionally, the integration and effectiveness of established learning strategies, such as retrieval practice, feedback, spacing, and multisensory learning, within these mobile applications remains underexplored (Reber & Rothen, 2018). Retrieval practice has demonstrated robust effects on memory (e.g., Rowland, 2014), but research examining its impact in stressful contexts like test anxiety, where memory may be impaired, is still limited. The existing evidence for the protective effect of retrieval practice against stressors is inconclusive (Hinze & Rapp, 2014; A. M. Smith et al., 2016; Szőllősi et al., 2017; Szpunar et al., 2013; Yang, Shanks, et al., 2023) and further investigation is needed. The overall aim of this thesis is thus to further explore the benefits of these learning strategies in mobile learning and then focus on the potential benefits of retrieval practice in stressful contexts.

2.1 **Question 1 (Study 1)**: What is the effectiveness of mobile applications built solely for learning purposes, and how do established learning strategies (e.g., retrieval practice, feedback, spacing, and multisensory learning) contribute to their effectiveness?

2.2 **Question 2 (Study 2)**: Is there evidence to support a cumulative benefit of retrieval practice on memory in stressful contexts, and what is the effect size of the protective benefits of retrieval practice on memory in these settings?

2.3 **Question 3 (Study 3)**: Can retrieval practice protect memory against stressors like test anxiety?

3. Study 1: A Meta-analysis on mobile-assisted language learning applications: Benefits and risks

3.1 Abstract

Mobile language learning applications are a pervasive facet of modern life, however evidence on their effectiveness on L2 learning outcomes is lacking. In the current work, we sought to determine the effect of mobile language learning applications on L2 proficiency between groups who used mobile language learning applications and control groups who learned with traditional methods on L2 achievement. We systematically searched journal articles and grey literature between 2007–2019 and performed a quantitative meta-analysis based on 23 synthesized effect sizes. We also performed risk of bias and quality of evidence assessments on our included papers. We found a moderate-to-strong overall effect ($g = 0.88$) of learning achievement using mobile language applications compared to control groups who learned with traditional approaches. At the same time, we found high risk of bias and low quality of evidence across all included studies. Our results provide evidence for mobile applications as a beneficial tool for second language learning. However, findings should be treated with caution due to risks of high bias and low quality of evidence. Improvements for future studies are discussed.

3.2 Introduction

The last few decades witnessed an explosion of mobile application use for personal, professional and educational purposes. Mobile-learning refers to the use of mobile or portable devices such as smartphones or handhelds for learning (Viberg & Grönlund, 2012). Mobile-assisted language learning (MALL) refers to the use of mobile or portable devices for second language (L2) learning and encompasses a wide variety software such as using SMS to send/receive L2 vocabulary words (Kukulska-Hulme & Bull, 2009; Thornton & Houser, 2005; Viberg & Grönlund, 2012). MALL-application is a subset of MALL that refers to software on mobile devices specifically developed for the purpose of L2 learning

(e.g., a mobile application developed specifically for vocabulary learning). Researchers and educators alike have recognized the potential benefits of MALL on L2 learning (Bano et al., 2018; Darmi & Albion, 2014; Kukulska-Hulme & Bull, 2009; Viberg & Grönlund, 2012). However, despite the exponential growth and popularity of MALL, research on the efficacy of MALL-application for L2 learning is lacking. The primary goal of this work is to conduct a meta-analysis on the efficacy of exclusively MALL-applications on L2 achievement in comparison to traditional L2 learning approaches used in classroom settings (e.g., pen-and-paper approaches, textbook learning, taking notes, doing worksheets etc.).

MALL has been rapidly and readily applied in the educational context and is favored over other types of learning approaches. The various advantages of MALL, such as immediate access to learning material, portability, and personalization make them attractive tools for learning and may increase time spent learning (Kukulska-Hulme & Bull, 2009; Viberg & Grönlund, 2012). For example, research suggests that students learning a second language actively engage with MALL tools (Yaman et al., 2015) and 70% of students own a mobile phone and prefer to learn with mobile-learning approaches (Oz, 2014). Qualitative interviews and reports with students and teachers generally reflect positive experiences with MALL and its perceived effectiveness for language learning, increased learner satisfaction, increased motivation, increased motivation to learn on one's own, and increased confidence (W.-Y. Hwang et al., 2014; Ibrahim et al., 2017; Kondo et al., 2012; Shadiev et al., 2018).

As the use of MALL in educational contexts increased, so did a market for MALL-applications, spurring the development of a multitude of MALL-applications created specifically for L2 learning. For example, mobile applications that support foreign vocabulary learning, such as mobile flashcards or mobile dictionaries, are now widespread (Bozdoğan, 2015; Fageeh, 2013; Mahdi, 2018; Rahimi & Miri, 2014; Q. Wu, 2014, 2015). A variety of MALL-applications have been designed and utilized to facilitate grammar learning and reading abilities (Ibrahim et al., 2017; Li & Hegelheimer, 2013; Ozer & Kılıç,

2018; Shadiev et al., 2018; W.-H. Wu et al., 2012), writing skills (G.-J. Hwang & Chang, 2011), as well as pronunciation and listening skills (Cavus & Ibrahim, 2017; Kondo et al., 2012). MALL-application systems that offer personalized content based on user learning levels, interest and learning cycles have also been developed (Chen & Chung, 2008; Chen & Li, 2010; Hsu et al., 2013; Zou & Xie, 2018).

Crucially, most MALL-applications developed by industries are based on established learning principles from fundamental memory research. For instance, retrieval-based learning benefits learning over re-studying (Agarwal, D'Antonio, Roediger, et al., 2014; Roediger III & Butler, 2011; Yang, Shanks, et al., 2023), corrective feedback is more beneficial for learning than non-corrective feedback (Metcalfe, 2017), spaced learning is more effective than massed learning (Dempster, 1987; Kapler et al., 2015; McDaniel et al., 2013; Sobel et al., 2011), and multisensory encoding leads to more robust memory traces than unisensory encoding (Kast et al., 2011; Shams & Seitz, 2008). While it is possible to apply these learning principles in traditional learning contexts (e.g., retrieval-based learning with flashcards), these and other learning principles can be enforced by means of MALL-application (for a discussion of learning principles in information and communication technologies, see Reber & Rothen, 2018). Hence, it can be reasonably assumed that MALL-application learning is more efficient for L2 learning than traditional learning approaches.

Despite the diverse range of MALL-application utilized for educational purposes, their specific advantages over traditional approaches on L2 learning has not been systematically assessed in a meta-analytic approach. No prior meta-analytic work has focused exclusively on MALL-applications, rather assessing other general-purpose aspects of MALL (e.g., exercises sent via text messages or social-networking sites rather than only applications which are built for L2 learning). For instance, Sung et al. (2016) reviewed 44 journal papers and dissertations on MALL over 20 years (1993–2013) and found a medium overall effect of $d = 0.55$ in favor of L2 learning approaches with MALL tools in

comparison to control groups who did not use the applications or used desktop computers. The authors also investigated the effects of different types of hardware (handheld devices, laptops, computers) and software (i.e., general purpose software and learning-oriented software) in their analysis, thus not concentrating solely on MALL-applications. Medium effects have also been reported in both Taj et al. (2016) and K. Cho et al. (2018) of $d = 0.43$ and $d = 0.51$ respectively for MALL on L2 language learning. These findings echo the results of other meta-analyses where medium effect sizes ($d = 0.67$) for vocabulary-learning with MALL compared to traditional-learning control groups were revealed (Mahdi, 2018). More recently, Chen and colleagues (2020) synthesized 80 experimental studies on MALL and found a medium-to-strong effect in favor of MALL over traditional-learning control groups. These studies all included different types of software, hardware and MALL tools (e.g. texting, gaming, social networking sites) as opposed to only mobile applications built specifically for L2 learning. Taken together, these data indicate compelling results in favor of adopting MALL for L2 learning. However, these data also demonstrate that the exclusive efficacy of MALL-application on L2 learning remains to be systematically investigated. Additionally, no previous work has considered both risk of bias and quality assessment of individual studies. These considerations are important to comply with current reporting standards and transparency for meta-analytic research (Maassen et al., 2020).

Therefore, a systematic literature search and quantitative meta-analysis performed in accordance with standard reporting guidelines is needed to better elucidate the effect of specifically MALL-application on L2 learning. The current work is, to our knowledge, the first meta-analysis to examine the effects of MALL-applications developed specifically for L2 learning and to assess the risk of bias and overall quality of the individual studies. Such an analysis is important to understand whether MALL-application can improve L2 language acquisition in comparison to traditional approaches. If this is the case, the analysis has the potential to further elucidate the most beneficial factors for L2 acquisition. We conducted a systematic meta-analysis using the Preferred Reporting Items for

Systematic Reviews and Meta-Analyses (PRISMA: (Page et al., 2021) to assess whether MALL-application compared to traditional learning approaches are more efficient when it comes to L2 acquisition. Furthermore, we explored which MALL-application factors are most beneficial for L2 acquisition.

3.3 Method

This meta-analysis followed the PRISMA (Page et al., 2021) guidelines for reporting. We ensured our meta-analyses abided by the transparency of data analysis and rigor of reporting as recommended by the field (for a review see Maassen et al., 2020). There is no protocol available for the current manuscript. All data and analysis have been made openly accessible on OSF (https://osf.io/htybd/).

3.3.1 Study Selection and Identification

Systematic Literature Search

A systematic literature search was conducted on the MALL literature. The search strategy was similar to the method employed in Bano et al. (2018). The systematic literature search was completed in early-middle 2019 and the last date the reported databases were checked was December 2020. To retrieve sufficient and comprehensive literature, this study probed scientific articles published in peer reviewed journals from 2007–2019. We chose 2007 as the start date for our literature search as that was the year in which the first Apple iPhone was released, markedly changing the mobile and communication landscape thereon after. The databases used in our search were Springerlink, Ovid, ISI, Scopus and Learntechlib. The terms used to search the databases were: [language OR vocabulary OR lingu*] NEAR [learn* OR train* OR acquisition OR teach* OR lecture OR edu*] AND [mobile OR wireless OR seamless OR ubiquitous OR electronic OR digital OR smart] NEAR [learn* OR pedagog* OR device OR app* OR phone] AND [“non native” OR “non-native” OR second* OR foreign] WITH [tongue OR speech OR language]. The complete list of search strings can be found on OSF (https://osf.io/htybd/). An additional manual literature search was conducted on reference

sections of prior meta-analyses and review papers published on MALL.

Grey literature was also probed to ensure we did not miss any relevant papers due to unpublished results, which could contribute to publication bias. We defined grey literature to extend to conference papers, dissertations, or other unpublished manuscripts on the field. The same search terms were used as our initial literature search. We also followed the same selection criteria as the general literature search, with the exception of the "published in a peer reviewed journal" criterion. We cross-referenced multiple pre-print archives on OSF (archives searched: OSF pre-prints, EdArXiv, MetaArXiv, Preprints.org, PsyArXiv) and unpublished dissertation repositories (Thesis Commons) in the education, psychology, and social sciences domains. One rater (MM) filtered through the searches for relevant titles and abstracts. In addition, we reached out to corresponding authors of included articles from our general literature search which met our inclusion criteria to inquire if they had any unpublished papers, null findings or any work in prep related to the topic. Authors were contacted by email and given 10 business days to respond to our request. They were informed that no answer by the end of the 10 business days meant a negative response from their part.

3.2.2 Eligibility Criteria

In this meta-analysis, we were solely interested in MALL-application studies where a control group learned with traditional pen and paper methods and an experimental group utilized a mobile language learning application or mobile learning system only to learn a foreign language. This is critical to ensure we observe the difference in learning outcome between MALL-application use versus traditional classroom learning. Our second primary inclusion criterion pertains to the use of a mobile application or a mobile language learning system which served the exclusive purpose of L2 learning. By contrast, we did not consider studies on other mobile tools, which can be used for exercises around second language learning, but whose original purpose is entirely different (e.g., SMS, gaming, video recoding, electronic notepads). Other inclusion criteria included:

Article is published in a peer-reviewed journal

Article language is English

Article is empirical, experimental or quasi-experimental

Contains enough statistical information (means, standard deviations, and sample sizes of pre/post tests for experimental and control groups) to obtain an effect size

L2 achievement is assessed as the main dependent variable in a post-test

Study Screening

A research assistant (EB) screened the database created from our initial literature search for relevant titles and abstracts. Relevant titles were coded with a "1" or a "2" and irrelevant papers with a "0." These titles were coded by three independent raters (EB, TR, NR) during the initial literature review stages. The abstracts and methods sections of papers coded with a "1" or "2" during screening were further read over and examined by MM. As a secondary step, key words ("mobile assisted language learning," "effects of mobile language learning," "language learning" and "vocabulary learning," "personalized," "learning system") were applied on the database to ensure no relevant titles were missed.

Data Extraction and Bias Risk Assessment

Predetermined information (school level, learning focus, application name and type, duration of intervention, learning principle used, country of study origin) was extracted from each study and coded by three independent raters (MM, TR, NR). Risk of bias of individual studies included in our meta-analysis was assessed with Cochrane Risk of Bias 2 tool (Higgins et al., 2020; Sterne et al., 2019). Risk of bias is assessed across the following domains: the randomization process, deviations from intended interventions, missing outcome data, measurement of outcome, selection of the reported results. Judgments regarding the risk of bias for each domain are based on answers to signaling questions, which are rated on the basis of "yes," "probably yes," "no," "probably no," or "no

information." The resulting judgments of "low," "some concerns," or "high" risk of bias are outputted by the risk of bias algorithm in the tool. Two raters (MM and SG) served as independent raters for the risk of bias assessments.

3.3.2 Analyses

Statistical analyses were performed in R version 4.1.3 (R Core Team, 2020) using the packages designed for meta-analysis: 'meta' (Balduzzi et al., 2019, v5.2-0), 'metafor' (Viechtbauer, 2010, v3.0-2), 'esc' (Lüdecke, 2019, v0.5.1) and 'dmetar' (Harrer et al., 2019, v0.0.9000). To perform the meta-analysis, we followed the handbook guide by Harrer, Cuijpers, Furukawa, and Ebert (2021) titled "Doing Meta-Analysis with R: A Hands-on Guide." All corresponding data and analysis, including raw data used to calculate effect sizes and analysis scripts, are available on OSF.

Effect Size Calculation

Effect sizes were computed to represent the impact of MALL-application interventions on language learning for experimental groups who used the L2 mobile application or learning system versus control groups who used traditional pen-paper, classroom approaches. Effect sizes were all calculated in Hedges' g (Hedges, 1981) using the R package 'esc' (Lüdecke, 2019). Studies containing more than one experimental group representing the same intervention category for the purposes of the review were pooled to create an overall intervention group and compared against the control to prevent unit-of-analysis error (J. P. Higgins et al., 2019). Papers containing several outcomes for experimental and control groups, an effect size was calculated separately for each outcome, then averaged to obtain one overall effect size for that article. Effect sizes were interpreted based on the Cohen standard specifications (small = 0.2 and above, medium = 0.5 and above, large = 0.8 and above, Borenstein et al., 2009).

In case of missing or unclear information in the articles, we reached out to authors to obtain the required information. Authors were contacted by the e-mail address indicated as the corresponding author in the original paper. All authors contacted for additional

information replied with the requested details.

Small Study Effects and Publication Bias

Publication bias was assessed with different approaches. The first two encompass what is known as small study effects (SSE), which is the notion that small studies with large standard errors are most likely to generate non-significant findings because only very large effects in small studies would become significant and hence lead to publication bias (Harrer et al., 2021). As such, these approaches are referred to as "Small Study Effects" rather than publication bias (Schwarzer et al., 2015).

To assess SSEs, we first conducted Egger's Test of the Intercept (Egger et al., 1997) which assesses the relationship between effect sizes and their corresponding standard errors. This relationship is illustrated with a funnel plot, and Egger's Regression calculates whether asymmetry exists in funnel plot that could be due to publication bias. Because Egger's Test is conducted on the effect sizes (i.e., standardized mean differences, SMDs), and the SMDs and standard error of included studies are independent, this process has been suggested to result in the inflation of false-positive results (Pustejovsky & Rodgers, 2019). To correct for this possibility, we conducted the Pustejovsky-Rodgers (2019). This approach uses a modified equation of the standard error when testing for funnel plot asymmetry which does not include the SMD itself, avoiding artificial correlation between the SMD and its standard error (Harrer et al., 2021).

To detect publication bias, we used three different quantitative methods recommended from the literature (Harrer et al., 2021). First, Duval and Tweedy's trim-and-fill procedure "trims" effect sizes with large standard errors from the funnel plot and "fills in" missing studies to maintain funnel plot symmetry (Duval & Tweedie, 2000). Second, we applied the PET-PEESE method (Stanley & Doucouliagos, 2014). In the PET method, the effect of small studies is controlled for by including the standard error as a predictor in a weighted regression model where the study's effect size is regressed on its standard error (Harrer et al., 2021). Similarly, the PEESE method uses the squared

standard error as a predictor. If the regression intercept calculated by PET is significantly larger than zero, the PEESE is used as the true effect estimate. If the PET intercept is not significantly larger than zero, the PET is used as the true effect estimate (Harrer et al., 2021). Lastly, we applied a selection model, which predicts how likely it is that a study is published (i.e., "selected") based on its results (i.e., its p-value) (McShane et al., 2016). We applied a three-parameter selection model, which is recommended if the number of studies is around 20 (Harrer et al., 2021). This model uses three parameters to assess publication bias: the effect size parameter, the heterogeneity parameter (τ^2) and the likelihood of selection. The model then "removes" the assumed bias due to selected publication and derives a corrected estimate of the true effect (Harrer et al, 2021).

Model Specification and Heterogeneity

We adopted a random-effects model approach to obtain an overall effect size measure based on the pooled weighted estimates from all the individual papers (Borenstein et al., 2009). The random-effects model assumes that the effects of individual studies deviate from the true intervention effect due to sampling variability and study variation because studies do not stem from the same population (Harrer et al., 2021). We used Knapp-Hartung adjustments (Knapp & Hartung, 2003) to calculate the confidence interval around the effect size, which is recommended to reduce false positives in case of a small number of studies (Harrer et al., 2021).

The random-effects model gives a measure of between-study heterogeneity, which is the extent to which effect sizes vary in a meta-analysis. Assessing heterogeneity is critical in a meta-analysis because there could be subgroups present in the data with a different true effect, or that there is no "real" effect behind the data meaning the studies included have nothing in common (Harrer et al., 2021). Under the random-effects model, heterogeneity is indicated by the Q, I^2, τ^2 statistics. The Q statistic is the weighted sum of squared differences between individual effect size and the overall pooled effect across all studies and represents heterogeneity; I^2 is the percentage of variation in the effects that is

not due to sampling error, or the degree of inconsistency in the meta-analysis; and τ^2 is the between-study variance in the meta-analysis and was calculated using the restricted maximum likelihood estimator (Viechtbauer, 2010) recommended by the field (Harrer et al., 2021). Typically, an I^2 of 25% signals low heterogeneity, 50% signals medium, and 75% indicates substantial heterogeneity (Harrer et al., 2021), which is the rule of thumb we will adopt in the current analysis. The Q statistic is sensitive to both precision (the sample size of the study) and number of studies (k), I^2 is not sensitive to k but to precision, and τ^2 is not sensitive to either. Due to the limitations of each of these measures, it is not generally recommended to rely on just one but rather consider all. The prediction interval (PI) is defined as the range for which we can expect future studies to fall (Harrer et al., 2021) such that if the PI is positive in favor of the intervention, we can expect future studies would reflect this benefit in their effects. The PI is the recommended way to overcome the limitations with the heterogeneity statistics described above since it considers the between-study variance (Harrer et al., 2021). As such, we will use the PI as a proxy for the results we can expect for future MALL interventions.

Outlier Analysis

To assess the robustness of our results, we conducted sensitivity analysis by investigating whether certain studies could be over-contributing to heterogeneity and therefore distorting our overall effect size. We tested for outliers using influential analysis in R (Harrer et al., 2021). Significant outliers are detected based on their respective weights on the pooled overall results and their contribution to the overall heterogeneity. The pooled overall effect was then recalculated with these outliers and influential cases removed and a corrected effect size is reported. These analyses were implemented using the 'dmetar' package in R (Harrer et al., 2019).

Subgroup Analysis

We planned to look at subgroup effects by comparing the effects for 1) vocabulary and other types of language learning skills; 2) school level of participants (elementary,

middle school, secondary school, university); 3) duration of intervention; 4) whether a pre-existing MALL-application or a language learning mobile application created by authors or researchers was used in the intervention; 4) learning principles. Assessments for subgroup inclusion were based on inclusion criteria (see Supplementary Materials, Table 3) which each rater was equipped with during the rating sessions. If the necessary information for subgroup classification could not be located in the respective article during rater deliberation sessions, the paper was excluded from classification and removed from further subgroup analysis. In case of disagreement, the three raters went through several rounds of deliberations to discuss differences in classifications until consensus was reached. Inter-rater reliability was assessed with Fleiss' Kappa scores before rater deliberation (see Supplementary Materials, Table 4). This deliberation process was identical for the learning principles exploratory analysis. Because some articles contained a mix of language learning skills such as listening and vocabulary ($N = 1$) or grammar, writing and reading ($N = 1$), these were combined into one category to denote mixed language learning skills. Subgroup analysis was performed by calculating the random-effects model to test for between subgroup differences using the 'dmetar' package in R (Harrer et al., 2019).

Learning Principles: Exploratory Analysis

We were interested to examine whether learning principles could be used to better understand, and potentially predict, the effectiveness of learning using mobile applications. We classified the intervention of each article based on whether learning principles were employed in the MALL-application. The learning principles included feedback, retrieval, distributed learning, and multisensory learning, all of which have been identified by memory research to be beneficial for learning (Reber & Rothen, 2018; Weinstein et al., 2018).

The same three raters (MM, TR, NR) rated each paper across the four principles based on inclusion criteria (see Supplementary Materials, Table 3). Effect sizes were computed for each learning principle by pooling all papers which were coded as having

included a particular learning principle and compared with the studies that did not include the learning principle in question.

3.3.3 Quality of Evidence

Quality of evidence was assessed in this meta-analysis with the Grading Recommendations Assessment, Development and Evaluation (GRADE; Guyatt et al., 2008) using the GRADEpro Guideline Development Tool available online (gdt.gradepro.org). Quality of evidence is assessed across five domains: risk of bias, inconsistency of trial results, indirectness of measure, imprecision of effect size estimate, and possible publication bias. Judgments across the domains are rated as "not serious," "serious," and "very serious" based on the likelihood of studies to be upgraded or downgraded for quality on each criterion.

3.4 Results

3.4.1 Systematic Literature Search

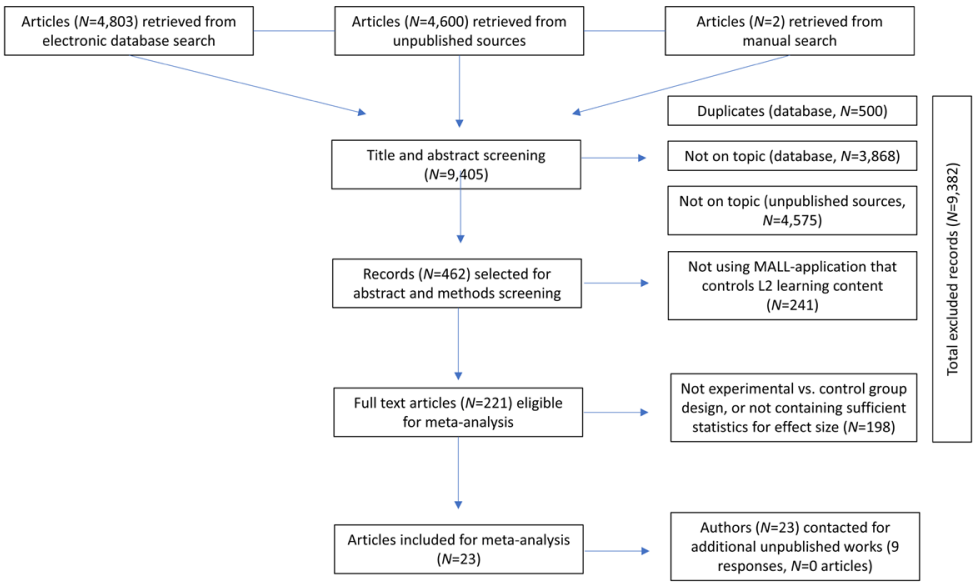
Figure 6 depicts the PRISMA flowchart of the literature review and screening process. We located a total of 4,803 research articles. After deletion of doubles, 4,303 articles remained, establishing our initial literature database for screening. Following article screening and abstract and method check, we were left with 18 papers that fit all inclusion parameters. Two additional papers had already been identified through references check were integrated for a total of 20 articles that fit all inclusion criteria. Articles which initially fit our inclusion criteria were excluded upon further inspection due to either utilizing other features of MALL (i.e., movie-maker or notepads) instead of MALL-application specific for L2 learning (Khodabandeh, Soleimani, et al., 2017; Lin & Lin, 2019) or due to lack of control group (Chen & Chung, 2008; Chen & Hsu, 2008; Li & Hegelheimer, 2013; Zou & Zie, 2018).

Probing the grey literature search yielded 4,600 unpublished studies from pre-print archive servers. Twenty-five titles were identified as potentially relevant, however after more thorough examination of the methods section, no papers met our inclusion criteria for

grey literature. We additionally scanned our initial literature search database (4,303 titles) and references checks for any relevant unpublished articles or conference papers. This yielded three results. The unpublished conference papers were integrated in our overall batch for a total of 23 articles for quantitative analysis. A total of 9 authors out of the 25 contacted responded to our inquiry regarding unpublished work or null findings, however all the respondents confirmed they did not have any such data to share. Thus, our total articles for inclusion were 23.

Figure 6

Flow chart of the literature search process



Note. Flow chart depicts the literature search process in this meta-analysis.

3.4.2 Characteristics of Included Studies

Descriptive information for all studies included in this meta-analysis for statistical analysis is presented in Table 1. All included studies are preceded by a code beginning with letter 'A' and will be referred to with this code for the rest of the work. A total of 23 articles total were identified for quantitative synthesis, with a sample size of 963 participants in the experimental group and 910 participants in the control group (total N = 1,873). Of these 23 articles, 20 were published in scientific journals and three were

retrieved from conference papers (i.e., grey literature).

Over half (65%, $N = 15$) of included papers were conducted in Asia, with Taiwan ($N = 8$) and China ($N = 4$) being the most common areas, followed by Malaysia ($N = 1$) and Japan ($N = 1$). The second most common geographic area was the Middle East (26%) comprising of Turkey ($N = 4$), Iran ($N = 1$), and Saudi Arabia ($N = 1$). Finally, Western countries contributed to only 8% of included studies with one from the USA (A2) and only one from Europe (Netherlands, A3). The target language to be learned was English in nearly all papers, with one paper learning Spanish (A4) and two Chinese Mandarin (A1, A7). All studies were published after 2010.

The majority of papers (70%, $N = 16$) were conducted on university aged students. The remaining 30% of included studies were conducted on younger children at the elementary school level (until 6th grade; $N = 3$) and middle school aged children (6–9th; $N = 2$), and high school (13–14+ years, $N = 2$). In one instance (A14), there was unclear information for participant ages. The MALL-application intervention durations varied greatly – with the shortest intervention comprising of 1 day (A5) and the longest duration was 4 months (A21). The most frequent interventions were between 2–6 weeks ($N = 7$) and an entire semester ($N = 7$).

Mobile language learning applications were used in 18 studies, while the rest ($N = 5$) employed a mobile language learning system approach which also ran as a mobile application. A considerable amount (43%, $N = 10$) of all applications used were created or designed by the authors and the other 13 papers featured already pre-existing mobile applications (57%). Roughly half of papers (47%, $N = 11$) targeted vocabulary learning, while the second major learning focus was reading (13%, $N = 3$). Grammar, listening comprehension, and writing followed, with each focus being about 4% each. The remaining studies featured a combination of language learning aspects such as communication ($N = 1$); vocabulary, pronunciation, listening, comprehension ($N = 1$); vocabulary and grammar ($N = 1$); listening and reading ($N = 1$); reading and vocabulary ($N = 1$); and grammar,

writing and reading ($N = 1$).

All studies featured a between-subject experimental or quasi-experimental design. Only 5 studies (22%) explicitly stipulated that participants were randomized in experimental and control groups. The remaining 78% of studies ($N = 18$) reported convenience or purposeful sampling (N = 4), or unclear sampling and randomization methods ($N = 14$). In these studies, participants were frequently allocated to the experimental or control group based on what classroom they were in, whether their mobile phone was compatible with the MALL-application to be utilized in the study, and whether they wanted to use a mobile device or work with traditional approaches.

All studies featured a pre-test and post-test design. Vast differences in L2 measurements were seen across all studies, with each paper utilizing a different language test to measure performance (i.e., no article utilized the same language learning measurement tool). The measurement scales were either instructor-created or an academic test.

Out of the 23 included papers, only three (13%) reported follow up results to assess language learning after the primary intervention had taken place. In Kiliçkaya and Krajka (2010), a follow-up post-test measure was conducted three months after the main intervention to assess performance between the experimental and control group once more. Lee (2014) reported delayed post-test results for experimental and control groups taken one week after the intervention. Lastly, Kondo and colleagues (2012) conducted a second experiment where 15 out of the 42 original participants from the experimental group in their first study assessing the MALL-application intervention were recruited to determine whether students continued using the mobile application. This second experiment did not contain a control group so we did not compute an effect size for it.

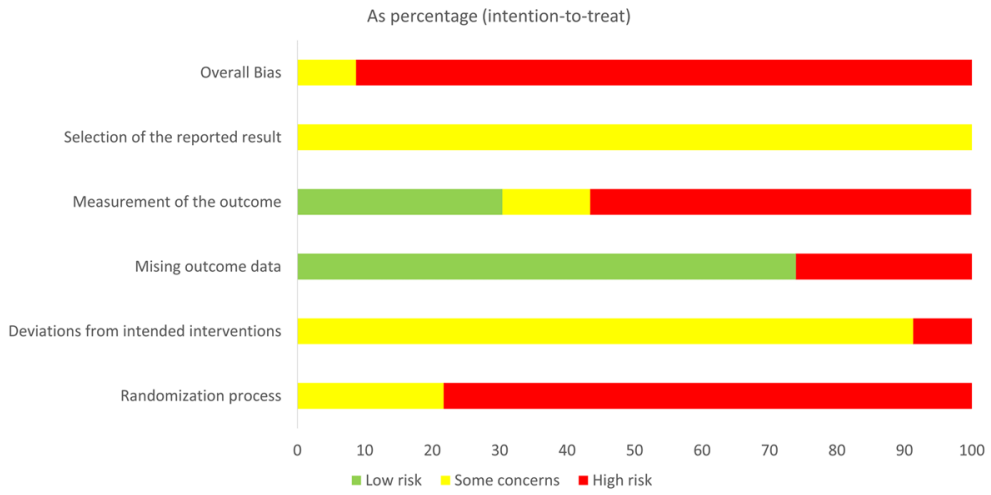
3.4.3 Risk of Bias

Risk of Bias was assessed with the Cochrane RoB 2 tool (Higgins et al., 2020; Sterne et al., 2019; see Figure 7). The total risk of bias was “high” across all studies

overall. No study featured true random allocation (i.e., random number generator or table) of participants into experimental or control groups, opting for convenience sampling, no information, or merely stating that randomization occurred. The majority of papers (91.3%) posed as "some concerns" for deviations from intended intervention, due to all participants who were recruited for the intervention or control groups be analysed as such (i.e., no participant switched groups during the intervention) and due to researchers being aware of what intervention was administered to which group of students (i.e., not blinded). Most papers (73.9%) did not include enough information on missing outcome data. If missing outcome data was reported, there was usually poor justification for why it was removed, and no sensitivity analyses were done on the data to test how the removed data points impacted overall results. In terms of measurement of the outcome, over half (56.5%) of articles did not describe what the language measurements entailed during the post test, mentioning only that a post-test was carried out. Overall, there was little rationale explaining how or why the chosen pre- or post-test questions were chosen to measure the particular language learning facet that they were measuring, nor was there any reliability measures done on these tests in most cases. However, the measurement of the outcome was comparable across both experimental and control groups at all time points measured (i.e., all participants received a pre- and post-test at the same time). No study referred to a study protocol established prior to the intervention, although several ($N = 3$) stipulated an analysis plan prior to presentation of experimental results. Two (MM and SG) completed RoB ratings independently and then discussed discrepancies until a consensus was reached.

Figure 7

Risk of bias across studies



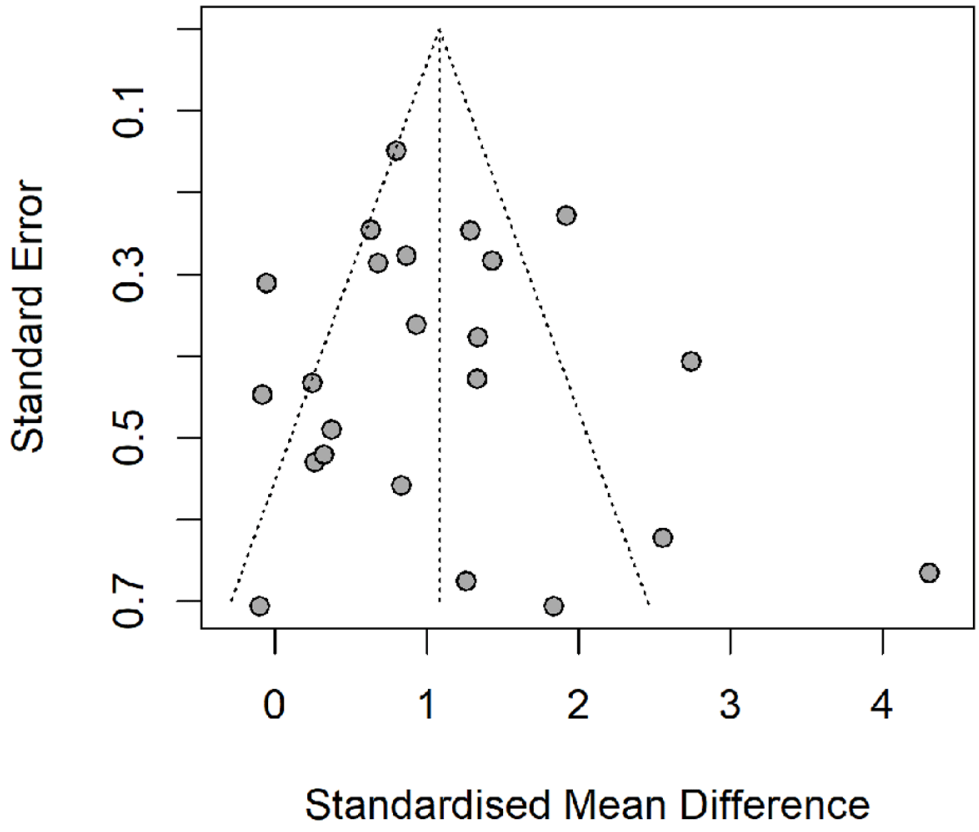
Note. Summary table of the risk of bias in all included studies overall and across each of the five domains: overall bias (high risk), selection of the reported result (some concerns), measurement of the outcome (a mix of low bias and some concerns, but mostly high risk), missing outcome data (predominantly low risk), deviations from intended interventions (largely some concerns), randomization process (mostly high risk). The bias domain is seen on the y-axis, and the score out of 100 is illustrated on the x-axis.

3.4.4 Meta-Analytic Results

Publication Bias

Publication bias was assessed in this meta-analysis with several methods. We first report the results of the SSE tests. Egger's test indicates no significant asymmetry present ($p = 0.58$), suggesting no significant publication bias is present in the current sample (Figure 8). Additionally, the Pustejovsky-Rodgers (2019) correction revealed no funnel plot asymmetry ($p = 0.23$), mirroring the results of Egger's Test. Taken together, these measures suggest no funnel plot asymmetry present in our analysis and thus no publication bias due to small study effects.

Figure 8
Funnel plot of all studies



Note. Figure shows funnel plot of studies included in meta-analysis. The x-axis is the effect size and the y-axis is the standard error. The dots represent individual studies. If publication bias exists, dots would typically cluster asymmetrically, with missing small studies showing negative or non-significant results, which is not seen here.

Additional measures of publication bias we conducted were the trim-fill procedure (Duval & Tweedy, 2000), the PET-PEESE approach (Stanley & Doucouliagos, 2014) and the three-parameter selection model (McShane, 2016). Results of the trim-fill procedure showed that the estimated number of missing effects was zero with zero additional studies being "filled" in, suggesting little risk for publication bias. Because the PET regression intercept was not significantly greater than zero ($g = –0.036$, $p = 0.96$), this suggests little publication bias. The results of the three-step parameter selection model results revealed no significant evidence of publication bias (LRT = 0.0015, $p = 0.97$). Moreover, the true

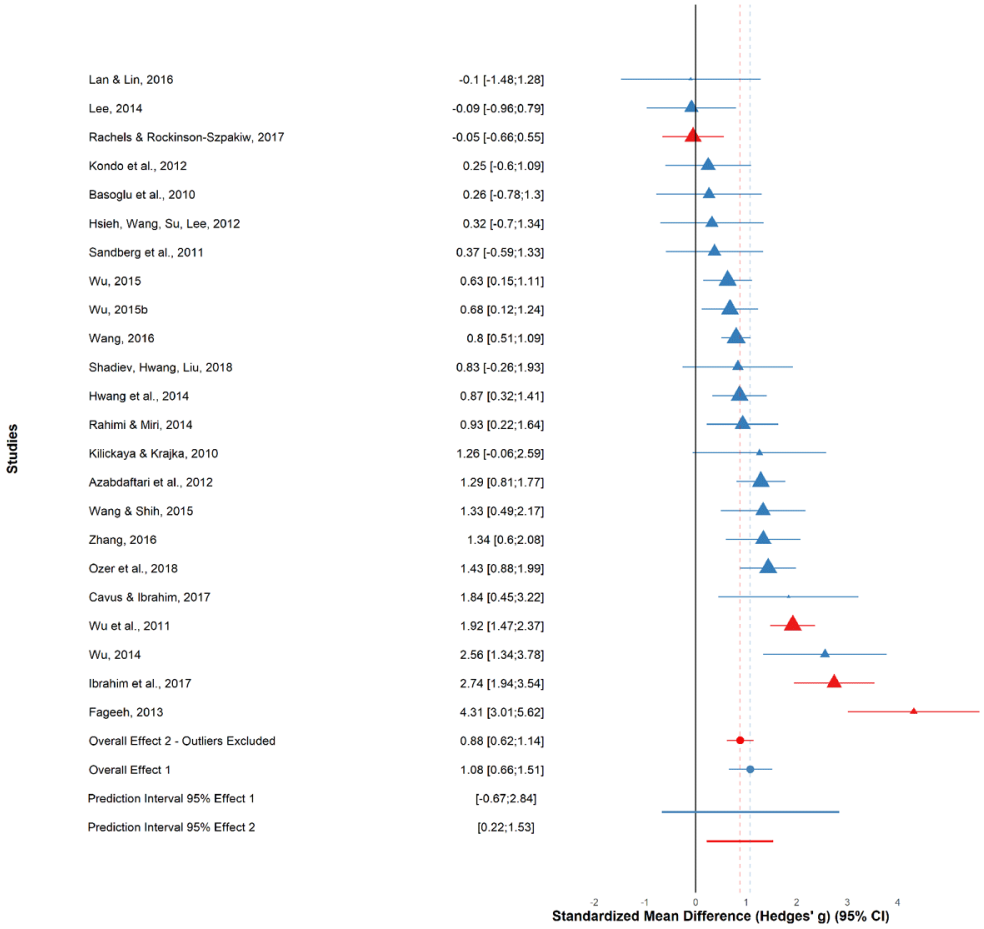
effect size estimate of $g = 1.07$ ($95\%CI : 0.50, 1.65$), which is nearly identical to our overall pooled estimate under the random-effects model ($g = 1.08$). This indicates our meta-analysis was not biased by a lower selection probability of non-significant results.

Overall Effects for MALL-application interventions

The summary of our findings is shown in Table 2. A total of 23 effect sizes were computed in this meta-analysis and pooled under the random-effects model with Knapp-Hartung adjustment (2003) to obtain the overall effect of MALL-application on L2 learning achievement (Borenstein et al., 2009) between experimental groups who utilized the MALL-application and control groups which used traditional pen-paper approaches. Overall, there is a large effect of $g = 1.08$ (95% CI: 0.66, 1.51, Figure 9) under the random effects model suggesting that MALL-application use facilitates L2 learning in the treatment group who used the application in question compared to the control group who learned via traditional approaches. The Q statistic was significant and suggests the overall sample is highly heterogenous ($Q = 106.10, p < 0.001, df = 22$). This further supported by the I^2 statistic of 79%, suggesting a 79% chance of results being due to heterogeneity rather than chance. A medium-to-large τ^2 of 0.67 indicates medium-to-high between-study variance. As each measure of heterogeneity is limited due to sensitivity to statistical power and precision, we additionally consider the prediction interval for a more robust estimate (Harrer et al., 2021). The prediction interval ($95\%PI : -0.67, 2.84$) falls slightly below zero, meaning we cannot be entirely certain that the strong positive effect we observe is robust in every sense.

Figure 9

Forest plot of all studies and overall effects



Note. Forest plot of all studies included in the meta-analysis. Individual effect sizes of all studies included in this meta-analysis with their respective weight in the analysis and confidence intervals represented by the triangles. Larger triangles indicate more weight in the random effects model. Red triangles denote the outliers identified in the analysis. The x-axis represents the standardized mean difference effect size (Hedges' g) and the y-axis is each individual study with its 95% confidence interval. The overall effect sizes are represented by the red and blue circles at the bottom. The 95% PI denotes the 95% prediction intervals for the overall effect of all studies (blue) and for the overall effect with outliers removed (red), indexed by the thick red and blue lines.

Outlier Analysis

To further investigate the contributions to significant heterogeneity ($Q = 106.10$, $p < 0.001$, $I^2 = 79\%$) in our sample, we ran outlier and influence analysis to detect and eliminate potential outliers in our sample that could be contributing to the between-study variance. A total of four influential cases (A2, A7, A13, A20; see Supplementary, Figure 10) were detected out of our total 23 studies.

A new overall effect size is recalculated with the weights of the four identified outliers set to zero in the random-effects model. The new overall effect size now stands at lower, but still strong $g = 0.88$ (95% CI: 0.62–1.14, 95% PI: 0.22, 1.53; see Figure 9). The Q statistic is still significant ($Q = 32.72, df = 18, p = 0.02$), however markedly reduced indicating a substantial reduction in heterogeneity. Further, I^2 (45%) suggests a low to medium chance of the results being due to heterogeneity rather than chance and a minimal τ^2 (0.083) suggests virtually non-existent between-study variance. The greatly reduced τ^2 value after outlier removal additionally suggests that a great deal of the heterogeneity was caused by outliers. The positive prediction interval ($95\%PI : 0.22, 1.53$) suggests that we can be relatively confident that future studies would find a similar effect (Harrer et al., 2021), a confidence we did not observe with the prediction interval of the initial model ($g = 1.08, 95\%PI : \breve{}0.67, 2.84$). Overall, results with outliers removed suggest a decreased, yet still moderate-to-strong, effect size with low heterogeneity and a positive prediction interval.

Subgroup Analysis

We conducted subgroup analysis for school level of learners (elementary school, middle school, secondary or high school, university) learning focus (vocabulary, reading, writing, grammar, or a mix of language learning skills), duration of reported intervention, type of application used (pre-existing or developed by authors) and learning principles. Analyses were conducted with the four outliers removed. Medium-to-strong effects were seen across all subgroups, indicating a beneficial impact of MALL-application across the

board (see Table 2 for summary of findings). No significant differences were observed between any subgroups. Inter-rater reliability scores for all subgroups were above 0.70, indicating good inter-rater agreement (see Supplementary, Table 4).

Follow-up Effects

Follow-up measurements between experimental and control groups were conducted in only two studies. In Kilickaya & Krajka (2010), the effect size for the follow-up post-test conducted 3 months after the intervention was $g = 0.20$ (95% CI: –1.07, 1.48), indicating a small effect of continued learning. Lee (2014) also reported a delayed post-test one week after the MALL-application intervention characterized by a small effect size ($g = 0.13$; 95% CI: –0.75, 1.00). Overall, these results show a weak effect of MALL-application versus control group at the delayed post-test, suggesting some sustained effects of MALL-application benefit after a delay, although there is not enough evidence currently to say for certain.

Exploratory Analysis: Learning Principles

We explored whether MALL-applications contained elements of learning principles (retrieval practice, feedback, distributed learning, multisensory learning) in the way they controlled learning content. Overall, all applications or learning system included some type of learning principles. Distributed learning was found present in every paper and was thus excluded from further statistical analysis. Strong effect sizes were also observed for all learning principles (see summary of findings, Table 2). The second most common learning principle was multimodal learning ($g = 0.87$, $N = 14$), followed by retrieval practice ($g = 0.95, N = 12$), and feedback ($g = 0.89, N = 9$). No significant differences between were observed between using a learning principle versus no principles used. Combinations of learning principles in one article were not investigated due low number of articles.

3.4.5 Quality of Evidence

To ascertain the quality of evidence in our included studies, we performed GRADE assessments across the domains of our subgroups: school level, learning focus, intervention

duration, learning app type (available on OSF). Outcomes were assessed across the five GRADE domains: risk of bias, inconsistency, indirectness of evidence and imprecision. The overall quality of evidence was poor, with "low" and "very low" being the most frequent GRADE assessment awarded for each outcome. The primary reasons for low certainty assessments were: a) high risk of bias across papers, b) high degree of inconsistency of results as indicated by I^2 estimates of over 40% in some cases) relatively imprecise measures as indicated by wide confidence intervals around each effect size for each subgroup. Indirectness of evidence was sometimes judged as "not serious" due to measurements being directly related to the outcome of interest (i.e., target word exams as a direct measure of vocabulary learning). Inconsistency was judged as "not serious" in some cases, as indexed by low I^2 values in some subgroups. Taken together, this analysis suggests that future studies might impact the overall effects and their confidence intervals found in the different outcomes assessed observed in the current work.

3.5 Discussion

This meta-analysis examined the effects of utilizing MALL-application in experimental groups over control groups who used traditional pen-paper classroom methods on L2 learning achievement. Our analysis revealed a strong overall effect in favor of MALL-application ($g = 1.08, N = 23$) over traditional-approach control groups on L2 learning. After outlier exclusion, we observed a moderate-to-strong effect ($g = 0.88, N = 19$). Publication bias was not detected in our sample. Subgroup analysis revealed moderate-to-strong effects across all moderator variables. In terms of overall effects sizes, our results seem to offer positive effectiveness for MALL-applications on L2 acquisition, however these results should be approached with caution as our results also revealed high risk of bias and overall low quality of evidence across all articles and outcomes in the meta-analysis.

Nearly all articles in our meta-analysis revealed positive effect sizes in favor of MALL-application in comparison to traditional approaches. These findings echo results

from previous meta-analyses that report positive medium-sized effects for L2 learning using MALL more generally (Chen et al., 2020; Cho et al., 2018; Mahdi, 2018; Sung et al., 2015, 2016; Taj et al., 2016). That is, the current finding extends beyond prior literature because it elucidates that MALL-application specifically provides a benefit over traditional learning approaches, as opposed to general MALL technology on learning previously conducted. Moreover, the observed moderate-to-strong in comparison to previously found moderate effects suggests that MALL-application might be slightly more beneficial than the more general MALL approach.

Both SSE and publication bias measures detected little to no publication bias in the current sample. This finding corroborates previous literature as far as SSE are concerned (e.g., Cho et al., 2018). However, our analysis went a step further by analyzing publication bias with multiple approaches not included in prior reviews on the MALL topic. It is important to note that the results of the SSE and publication bias approaches described in section 3.2.1 are with their limitations and weaknesses. For example, the PET-PEESE is prone to over-adjusting effects and leading to underestimation of the true effect size (Carter et al., 2019) and the trim-fill method is known to under-correct for publication bias (Harrer et al., 2021). Additionally, both the trim-fill and PET-PEESE methods are not robust when the heterogeneity is high, as is the case here. The three-parameter selection mode has been found to be more reliable than other methods (McShane et al., 2016), however it can be difficult to interpret (Harrer et al., 2021). Moreover, it is important to note that publication bias can be caused by a multitude of other factors, such as between-study heterogeneity and high risk of bias within studies (Harrer et al., 2021). Thus, although we found no quantifiable publication bias, the high risk of bias (see section 3.1.2) may be an alternate explanation for potential publication bias.

Only 23 studies fulfilled the inclusion criteria of our meta-analysis on MALL-application, which is surprising given the popularity of mobile devices and learning applications used today. One explanation for this low sample size is that a plethora of

MALL-applications exist and continue to be utilized in educational contexts or for individual L2 learning, but remain experimentally unvalidated for learning outcomes. That is, students and teachers may be using various MALL-applications without awareness as to their actual effects on learning. This resounds Burston's (2015) review, where it was found that over 40% of all articles published on MALL are unrelated to MALL-applications more specifically and an overwhelming majority of articles lack quantifiable learning outcomes. Given our finding that MALL-application might be more effective than general MALL on L2 learning, these findings thus highlight the importance of using MALL applications that are both specifically designed for learning and have been experimentally validated for learning outcomes. Translating this to practical terms, it is advisable for educators and students to be sparing in terms of the applications they utilize in the classroom and limit their use to only validated MALL-applications. At the same time, pre-existing or newly developed MALL and MALL-applications should be experimentally validated to ensure quantifiable learning outcomes prior to use in classrooms or on the market for individual use.

It has long been deemed necessary to integrate learning principles from fundamental memory research within mobile learning tools to enhance learning (Parsons & Ryu, 2006; Reber & Rothen, 2018; Zydney & Warner, 2016). We attempted to address this need by exploring whether the four learning principles of retrieval practice, feedback, distributed learning, and multimodal learning are utilized in MALL-application to boost L2 learning. Our analysis revealed that all MALL-applications used in all included articles featured distributed learning and included one or more learning principles. However, due to the small number of studies in the different subgroups, we were not able to identify differential effects. Critically, none of the learning principles were directly manipulated by authors in the studies. It will thus be an interesting endeavor for future studies to incorporate and manipulate learning principles to directly assess the relative contribution of each principle on learning and their interactions on L2 learning with a MALL-application. Such studies

would be interesting in their own rights for the field of memory research more generally because our current knowledge on the interaction of different learning principles is extremely limited (Belardi et al., 2021; Weinstein et al., 2018). Given how easy it is to collect large amounts of data across extended time periods with MALL-application, such studies therefore offer a unique opportunity to advance current knowledge in the field of basic memory research beyond mere L2 learning.

Only three studies in our included batch administered a delayed post-test following the intervention, with follow-up effects. This finding suggests some sustained positive benefit on L2 learning with MALL-application over the long-term, however given the small number of studies it is premature to draw conclusion on potential long-term benefit. Elucidating the long-term value of MALL-application beyond the intervention period requires a data-driven approach as some students might continue to use the MALL-application beyond the experimental intervention period (Kondo et al., 2012; Sandberg et al., 2011). In addition, because most studies investigated only short intervention periods, it might be the case that the effects of MALL-application interventions diminish over the time (Sung et al., 2016).

Besides the overall finding that MALL-application is likely to be beneficial for L2 learning achievement, we also identified several potential risks. Due to the applied nature of the research, most studies lacked proper randomization procedures in assigning participants to experimental and control groups (i.e., risk of bias). Additionally, articles lacked transparency when it came to how outcome variables were assessed. It is thus recommended for future articles on MALL-application interventions to provide scoring or coding schemes for how post-test questions were graded, as well as justification for why certain target words were chosen as the target words to assess learning following the MALL-application intervention. Recommendations to follow best practice guidelines when it comes to addressing missing outcome data, clearly explaining statistical analyses, and considering a priori power analyses are also encouraged in order to improve methodological

rigor and increase the replicability and reproducibility of the benefits of MALL-application on learning.

3.6 Conclusion and Future Directions

This meta-analysis identified 23 studies which systematically assessed L2 learning achievement outcomes by means of a MALL-application intervention in comparison to a traditional pen-paper learning control group. Based on these studies, we found a moderate-to-strong benefit of $g = 0.88$ of using MALL-application on L2 learning achievement over traditional classroom approaches. This suggests that MALL-applications themselves are an effective way to boost L2 learning, and such experimentally validated MALL-applications should be considered for L2 learning. All included studies showed evidence of implementing learning principles in their MALL-applications, however none manipulated or compared these principles directly and the overall sample size of the included papers is small. Therefore, it is hard to determine their individual contribution to learning. Future studies should examine which pedagogical learning principles are most conducive for L2 achievement in a MALL-application setting. The beneficial effects of MALL-application on learning are not, however, without risks. Limitations revealed in the current work were mainly related to short intervention durations, missing follow-up measurements after the actual intervention, lack of randomization, and unclear measurement of the outcome variable. Such risks should motivate future research to utilize best research practices in order to produce replicable and valid effects. Taken together, MALL-application appear to be beneficial for L2 learning achievement, however the low number of studies in combination with the observed risks and limitations and the missing manipulation of learning principles require further research efforts to determine the impact of MALL-application in educational contexts and memory.

Data Accessibility Statement The data and methods reported here are available at the Open Science Framework: https://osf.io/htybd/.

Additional File. The additional file for this article can be found as follows:

Supplementary Material: Criteria Definitions for Subgroup Analysis. DOI: https://doi.org/10.5334/pb.1146.s1

Acknowledgements This research was supported by the service of higher education of the Canton of Valais, Switzerland (project: "School of Tomorrow"). The authors would like to thank Eleonora Balbi for her efforts in the literature search of this meta-analysis.

Competing Interests. The authors have no competing interests to declare.

Author Contributions Mariela Mihaylova: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. Simon Gorin: Software, Methodology, Data curation, Validation, Writing – review & editing. Thomas Reber: Conceptualization, Methodology, Writing – review & editing, Funding Acquisition. Nicolas Rothen: Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding Acquisition.

Table 1

Characteristics of included studies

CODE	AUTHOR	TOTAL N	AGE	DURATION	FOCUS	TARGET LANGUAGE	APPLICATION	DESIGN	ORIGIN COUNTRY	TYPE	SOURCE
A1	Lan & Lin, 2016	34	17-23	4 weeks	Communication	Chinese	Mobile Seamless System (MOSE)	Between-subjects, Mixed Methods	Taiwan	Journal	jstor.org/stable/10.2307/jeductechsoci.19.3.335
A2	Rachels & Rockinson-Szapkiw, 2017	167	3rd and 4th graders	12 weeks	Vocabulary & Grammar	Spanish	Duolingo	Quasi-Experimental	United States	Journal	https://doi.org/10.1080/09588221.2017.1382536
A3	Sandberg, Maris, & Geus, 2011	75	8-10	1 day	Vocabulary	English	MEL Application	Quasi-Experimental	Netherlands	Journal	https://doi.org/10.1016/j.compedu.2011.01.015
A4	Basoglu & Akdemir, 2010	60	17-24	6 weeks	Vocabulary	English	ETACO mobile flashcards	Between-Subject, Mixed Methods	Turkey	Journal	ERIC Number: EJ898010 https://eric.ed.gov/?id=EJ898010
A5	Cavus & Ibrahim, 2017	37	10-13	4 weeks	Listening, Vocabulary, Comprehension Pronunciation	English	Near East University Children's Story Teller (NEUCST)	Between-Subjects	Turkey	Journal	https://doi.org/10.1111/bjet.12427
A6	Azabdaftari & Mozaheb, 2012	80	21	7 weeks	Vocabulary	English	Spaced Repetition System (SRS)	Between-Subject, Mixed Methods	Iran	Journal	ERIC Number: EJ1064983 https://eric.ed.gov/?id=EJ1064983
A7	Ibrahim, Chee & Yahaya, 2017	53	Primary school	4 weeks	Reading	Chinese	Learn Chinese Mandarin App	Quasi-Experimental	Malaysia	Journal	https://doi.org/10.1504/IJMLO.2017.10005992
A8	Ozer & Kilic, 2018	63	18-22	6 weeks	Grammar	English	Variety of mobile apps	Not Explicit, Between-Subjects	Turkey	Journal	https://doi.org/10.29333/ejmste/90992
A9	Kilickaya & Krajka, 2010	38	17-19	5 weeks	Vocabulary	English	WordChamp	Between-Subjects	Turkey	Journal	ERIC Number: EJ898003 https://eric.ed.gov/?id=EJ898003
A10	Wu, 2014	50	20-23	1 semester	Vocabulary	English	Word Learning	Not Explicit, Between-Subjects	China	Grey	https://doi.org/10.1016/j.sbspro.2014.07.409
A11	Wu, 2015a	70	20-23	8 weeks	Vocabulary	English	Word Learning-CET6	Not Explicit, Between-Subjects	China	Journal	https://doi.org/10.1016/j.compedu.2015.02.013
A12	Wu, 2015b	199	20-22	1 semester	Vocabulary	English	Word Learning-CET4	Not Explicit, Between-Subjects	China	Journal	https://doi.org/10.1371/journal.pone.0128762
A13	Fageeh, 2013	58	University students	1 semester	Vocabulary	English	Android Online Dictionary	Not Explicit, Between-Subjects	Saudi Arabia	Journal	web.a.ebscohost.com
A14	Rahimi & Miri, 2014	34	University students	16 sessions	Vocabulary	English	Longman Mobile Dictionary	Not Explicit, Between-Subjects	Iran	Grey	https://doi.org/10.1016/j.sbspro.2014.03.567
A15	Kondo et al., 2012	88	University students	April – July semester	Listening & Reading	English	TOEIC	Not Explicit, Between-Subjects, Mixed Methods	Japan	Journal	https://doi.org/10.1017/S0958344012000055,
A16	Wang, 2016	196	University students	1 semester	Reading	English	Learn English Audio & Video	Between-Subjects	Taiwan	Journal	https://doi.org/10.1080/10494820.2015.1131170
A17	Wang & Shih, 2015	93	19	1 semester	Vocabulary	English	The Most Important 2000 TOEIC Words	Not Explicit, Between-Subjects	Taiwan	Journal	https://doi.org/10.1504/IJMC.2015.070060
A28	Hwang, Chen, Shadiev, Huang & Chen, 2014	59	6th grade	3 classes per week for 1.5 months	Writing	English	Situated Writing System	Quasi-Experimental	Taiwan	Journal	https://doi.org/10.1080/09588221.2012.733711
A19	Shadiev, Hwang, Liu, 2018	53	13–14	3 weeks	Grammar, Writing, Reading	English	Mobile Multimedia Learning System (MMLS)	Quasi-Experimental	Taiwan	Journal	https://doi.org/10.1007/s11423-018-9590-1
A20	Wu, Sung, Huang, Yang, Yang, 2011	113	University students	7 weeks	Reading	English	Ubiquitous English-Reading Learning System	Not Explicit, Between-Subjects	Taiwan	Journal	www.jstor.org/stable/10.2307/jeductechsoci.14.4.164
A21	Hsieh, Wang, Su, Lee, 2012	60	University students	4 months	Reading & Vocabulary	English	Fuzzy, Logic-Based Personalized Learning System	Between-Subjects	Taiwan	Journal	www.jstor.org/stable/10.2307/jeductechsoci.15.1.273
A22	Zhang, 2016	120	University students	10 weeks	Listening Comprehension	English	Keke English & Easy IELTS	Between-Subjects	China	Grey	https://doi.org/10.12783/dtssehs/icaem2016/4290
A23	Lee, 2014	120	15-21	20 classes	Vocabulary	English	Teacher-Created Application	Not Explicit, Between-Subjects	Taiwan	Grey	https://doi.org/10.1109/MCSoC.2014.24

Table 2

Summary of meta-analytic findings

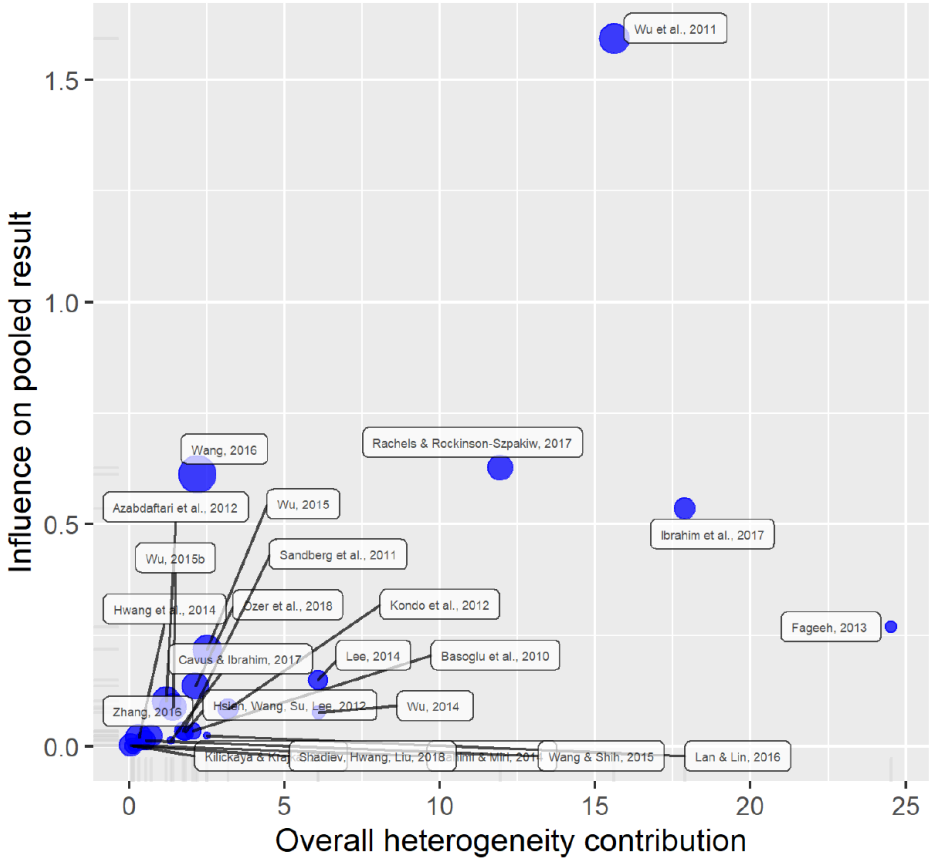
INTERVENTION (MALL-APPLICATION VS. CONTROL)	N	G	95% CI	TAU²	I²	Q	95% CI
Overall	23	1.08	0.66–1.51	0.67	79.3%	106.10***	<0.0001
Overall, outliers removed	19	0.88	0.62–1.14	0.08	45%	32.72*	0.018
INTERVENTION (MALL-APPLICATION VS. CONTROL)	**N**	**G**	**95% CI**			**$Q_{BETWEEN-GROUPS}$**	**P-VALUE**
School Level							
University	14	0.87	0.51–1.23	0.16	56.3%		
High School	1	0.83	0.26–1.93	NA	NA	1.44	0.84
Middle School	2	1.13	−4.39–6.66	0.18	39.1%		
Elementary School	1	0.37	−0.59–1.33	NA	NA		
Learning focus							
Vocabulary	10	0.87	0.41–1.33	0.17	55.1%		
Reading	1	0.80	0.51–1.09	NA	NA		
Writing	1	0.87	0.32–1.41	NA	NA	7.99	0.24
Grammar	1	1.43	0.88–1.99	NA	NA		
Communication	1	−0.10	−1.48–1.28	NA	NA		
Listening	1	1.34	0.60–2.08	NA	NA		
Mix of language skills	4	0.65	−0.38–1.67	0.06	28.6%		
Intervention Duration							
1 day	1	0.37	−0.59–1.33	NA	NA		
2–6 weeks	6	0.98	0.27–1.70	0.17	38.6%		
7–10 weeks	4	0.99	0.45–1.54	0.05	35.6%	2.34	0.67
1 semester	6	0.89	0.14–1.65	0.23	58.9%		
10–16 sessions (undefined weeks)	2	0.46	−5.97–6.89	0.35	67.8%		
App Type							
Pre-existing App	10	0.98	0.68–1.28	0.05	33.8%	0.63	0.43
Developed App	9	0.76	0.20–1.32	0.19	52.5%		
Learning Principles							
Retrieval Practice	12	0.95	0.56–1.34	0.13	51.1%	2.57	0.28
No-Retrieval Practice	6	0.73	0.35–1.11	0.006	28.1%		
Feedback	9	0.89	0.58–1.20	0.02	26.6%	0.01	0.91
No-Feedback	10	0.86	0.37–1.35	0.22	58.7%		
Multimodal	14	0.87	0.53–1.22	0.12	51.8%	0.03	0.87
No-Multimodal	5	0.91	0.33–1.49	0.07	23.1%		

Note. "*" = $p < 0.01$, "**" = $p < 0.001$, "***" = $p < 0.0001$. Q-between-groups denotes the Q statistic for between-groups comparison, NA denotes unavailable values due to small number of studies.

3.7 Supplementary

Figure 10

Outliers detected in meta-analysis



Note. Results of outlier analysis showing articles A2, A7, A13, and A20 as strong contributors to heterogeneity (x-axis) and influencing the overall pooled result (y-axis). The majority of the studies are clustered in the lower left corner, indicating lower heterogeneity contribution.

Criteria Definitions for Subgroup Analysis

School level: To be included in this subgroup, the paper should explicitly state that the age and school level of the participants. If only "university students" is mentioned, then classify into university level. If only ages are mentioned but no school level, we designated 18 years as the cut-off age for inclusion in university learners group; younger ages in the young learners group. Younger learners were further divided into secondary or high school (grades 9-12), middle school (grades 6-9) and primary school (grades 1-5).

Learning focus: Depends on the focus of the task performed with the application in the paper. If the focus was word learning such as with flashcards or mobile dictionary, it is primarily vocabulary unless otherwise mentioned in the paper. If performed a variety of learning topics, such as grammar, writing, reading, classify into other which includes multiple learning areas.

Type of app: Unless otherwise stated in the paper (such as when authors state they have created or developed the application or learning system used in that study), safe to assume the application exists already.

Duration of intervention: Classify based on what explicitly mentioned in article regarding how long participants used the app to learn before the post-test. If exact weeks not specified and only a semester is specified, assume 15 weeks per semester. If participants could use the device in their free time in addition to intervention, classify based on controlled intervention portion of study only for which duration info available.

Learning principles: Inclusion/exclusion criteria listed in Table 3. Efforts should be made to stick to what is explicitly described in the paper.

Table 3

Classification criteria for learning principles

Learning Principle	Inclusion	Exclusion
Retrieval	Participants were prompted to actively recall study material, also during learning phase	Materials were re-read or re-studied passively or not explicitly mentioned in article, multiple choice test performed
Distributed Learning	Learning/intervention took place over time, several different sessions	Intervention/learning took place during one session
Feedback	Feedback was given to participants directly in the application	No feedback given to participants during learning/intervention or not explicitly mentioned in article
Multisensory Learning	Learning material was presented in multiple modes during intervention (i.e – audio and visual)	Learning material was presented in one method only (i.e –text only)

Note. The inclusion criteria raters followed when rating each article for learning principles.

Table 4

Inter-rater reliability scores

Subgroup	Fleiss Kappa before discussion
Learning Stage	0.95
Learning Focus	1.00
Learning Mode	0.95
Duration	1.00
Learning Principles – RP	0.70
Learning Principles – FB	0.88
Learning Principles – MM	0.93

Note. Fleiss Kappa values before rater deliberation. RP = retrieval practice, FB = feedback, MM = multimodal.

4. Study 2: Does retrieval practice protect memory against stress?

4.1 Abstract

Stressors like test anxiety are known to decrease memory retrieval, whereas retrieval practice is the phenomenon that actively recalling information from memory enhances memory. Evidence suggests retrieval practice can protect memory against the negative effects of stress on memory (A. M. Smith & Thomas, 2018a; A. M. Smith et al., 2016), however the findings are mixed (e.g., Yang et al., 2020). Evaluating the effectiveness of retrieval practice in maintaining memory performance under stress could transform memory resilience and lead to new cognitive interventions. This raises the need for a meta-analytic summary of the literature to understand the effects of retrieval practice on memory in relation to stressors. In this registered report, we conducted a meta-analysis (k = 10, N = 856) on the impact of retrieval stress on memory following learning with retrieval practice using databases and other information sources from 2006-2024. We found a positive effect of retrieval practice versus restudy in stress conditions, Hedge's g = 0.45, 95% CI [0.19, 0.71], suggesting retrieval practice is more beneficial than restudy during stress. Significant heterogeneity was observed along with moderate risk of bias, yet little evidence for publication bias. We did not find significant moderators in this sample nor a significant effect of the stressor on memory performance. Future studies should be conducted with sufficiently powered samples and using other types of control, or non-retrieval practice, strategies to better understand protective effects of retrieval practice on memory following stress. We registered our meta-analysis, with datafile, code and supplementary here: https://osf.io/jwx4f/.

4.2 Introduction

Stress is defined as any event that is perceived as threatening (Dedovic et al., 2009) and encompasses situation-specific and socio-evaluative psychological stressors such as test anxiety, also known as exam-related stress (Cassady & Johnson, 2002; Dickerson &

Kemeny, 2004). A wealth of evidence suggests that retrieval stress, or stress occurring before memory recall, decreases memory subsequent learning (Gagnon et al., 2019a; Kuhlmann, 2005; McEwen & Gianaros, 2011; Schwabe & Wolf, 2010b; Shields, Sazma, et al., 2017; Vogel & Schwabe, 2016). At the same time, learning strategies such as retrieval practice (i.e., the act of actively recalling information from memory) are consistently shown to enhance memory retrieval (Roediger & Karpicke, 2006; Rowland, 2014). Recent evidence suggests that retrieval practice may have protective effects on memory against stress via memory strengthening mechanisms (Smith et al., 2016). These findings might suggest that memory may be made less sensitive against the detrimental effects of retrieval stress using an easy-to-use learning strategy. To date, the overall effects of retrieval stress on memory after learning with retrieval practice have not been examined in a meta-analytic approach. In this meta-analysis, we aim to explore the protective mechanisms of retrieval practice in the context of retrieval stress.

Stress can be experienced in many different forms. The one of interest for the current meta-analysis is psychological stress, which involves uncontrollable situations or events characterized by socio-evaluative threat, such as one's performance being evaluated or judged negatively by others (Dickerson & Kemeny, 2004). Examples of these situations include: evaluative situations such as exams and test anxiety (Cassady & Johnson, 2002; Hembree, 1988); the Tier Social Stress Test (TSST), which consists of socially evaluative situations such as making a speech in front of others and being judged (Kirschbaum et al., 1993); or via instruction sets which mention that performance will be judged (Almazrouei et al., 2022). Stress can also be induced through procedures such as the Cold Pressor Test, where participants place their hands in cold water for a specific time, coupled with socio-evaluative elements (Schwabe & Schächinger, 2018; Schwabe et al., 2008). The abovementioned measures to induce stress are typically associated with increased cortisol levels or state anxiety responses which signal a stress response. And, importantly, a plethora of studies suggest that stress induced through these methods is associated with

decreases in memory performance and memory retrieval (de Quervain et al., 2000; Kuhlmann, 2005; Kuhlmann, Kirschbaum, & Wolf, 2005; Schwabe & Wolf, 2010a).

Retrieval practice is a learning strategy where one actively recalls information from memory. In a classic experiment, Roediger and Karpicke (2006) presented participants with two passages to read for 7 minutes and then either restudy (reread) the passage or take a short test where they wrote down as much as they could remember from the passage. After a retention interval of 5 minutes, 2 days or one week, participants were asked to recall as much as they could from the initial passages. Results revealed that for the longer retention intervals of 2 days and 1 week, participants in the testing, or retrieval practice, condition performed significantly better than the restudy condition (Roediger & Karpicke, 2006), highlighting the effectiveness of this strategy for long-term learning. Since then, the benefits of retrieval practice have been shown for a wide range of learning materials and retention intervals (Karpicke, 2017; Rowland, 2014; Schwieren et al., 2017).

Critically, retrieval practice is shown to be more effective than commonly used learning strategies such as restudying, highlighting, note-taking or elaborative techniques such as drawing concept maps (Moreira et al., 2019). The benefits of retrieval practice are thought to occur via an episodic context account. Under the episodic context account, contextual cues become bound to memory traces and are reinstated each time an item is retrieved from memory, thereby strengthening the memory traces learned with each retrieval (Karpicke et al., 2014; Karpicke, 2017). When memory is strengthened as such, it could become less sensitive to the effects of stress, and thereby also to the contextual shifts that might occur due to stress (Smith & Thomas, 2018). Although this is only one of the possible mechanisms under which retrieval practice is presumed to be effective, it can explain why retrieval practice might protect memory against stress. However, despite its effectiveness, the benefits of retrieval practice are rarely examined in the face of situations where memory is likely to fail, such as during stressful situations.

That is, until A. M. Smith et al. (2016) examined the protective effects of retrieval

practice against retrieval stress for the first time. In their study, 120 participants were split into either a retrieval practice group, who learned material using the retrieval practice strategy, or a study practice group, who re-read the material. Twenty-four hours later, 30 participants from both the retrieval practice and study practice groups underwent TSST stress induction, and the other half underwent a non-stressful control task. At five minutes and 20 minutes into the stress induction, a memory test was administered to observe the immediate and delayed effects of the stress respectively. Results showed that participants who learned via retrieval practice and were exposed to stress outperformed those who restudied and were also exposed to stress (A. M. Smith et al., 2016). This study suggests that retrieval practice might protect memory from the otherwise detrimental effects of stress on memory, as well as carry over to other stressors such as during testing situations.

Other evidence comes from Agarwal and colleagues (2014), who administered surveys to students in classes involved in a school-wide retrieval practice learning program. When asked if retrieval practice made students more or less nervous for tests and exams, 72% reported that retrieval practice made them feel less nervous for upcoming tests. When asked if they experienced more or less test anxiety for classes in which they underwent the retrieval practice intervention compared to classes where they did not use retrieval practice, only 19% indicated feeling more test anxiety, and over half of students (54%) reported that retrieval practice reduced their test anxiety. Taken together, these results suggest that retrieval practice can help protect memory against stressors.

However, other studies present contradictory findings regarding the protective role of retrieval practice on memory following stress exposure. For example, a recent study by Yang et al. (2020) investigated whether learning with retrieval practice is modulated by individual differences such as test anxiety levels. Students filled out test anxiety questionnaires and engaged in a learning session where they learned word lists with either a retrieval practice or restudy strategy. Results showed that test anxiety scores did not significantly correlate with memory performance, suggesting that test anxiety does not

significantly modulate retrieval practice effects. However, other evidence from Clark and colleagues (2018) suggests a positive relationship between test anxiety and using retrieval practice when external incentives are applied, suggesting memory can be protected by retrieval practice in the face of stressors like test anxiety.

The above literature suggests mixed evidence for the protective effects of retrieval practice on memory following stress exposure. Further investigation is needed using a meta-analytic approach to determine the strength of the cumulative evidence for the protective effects of retrieval practice on memory following retrieval stress. Addressing this question is critical as it could imply that using non-invasive learning strategies might alleviate the memory impairment induced by stress. Such an investigation has the potential to challenge some of the major theories of stress, as it would suggest that there is a way to make memory less sensitive—and potentially protected—against what would be a stress-induced memory impairment. In terms of real-world value, the findings of this meta-analysis would additionally have major implications for designing learning-based interventions in applied settings such as schools and other learning environments.

To summarize, a wealth of evidence suggests that psychological stressors include situation-specific, socio-evaluative situations where individuals' performance is likely to be judged or evaluated such as test anxiety. Stressors experienced at retrieval decrease memory and learning. Retrieval practice has been consistently shown to boost memory and learning, however its protective effects in the face of retrieval stress are mixed. In this meta-analysis, we aim to answer the question of whether retrieval practice can make memory less sensitive to the detrimental effect of stress and potentially protect memory in the context of retrieval stress.

Retrieval Practice Main Effects

In line with existing literature showing the negative impact of retrieval stress on memory performance (Shields et al., 2017), our primary aim is to investigate whether retrieval practice can make memory less sensitive to the negative effects of acute stress. To

do this, we will perform a systematic literature search to identify studies which investigated the potential of retrieval practice to protect memory against acute stress (A. M. Smith et al., 2016). Based on the retrieved studies, our main research question will then be investigated via four primary hypotheses.

First, based on previous literature showcasing the detrimental effects of retrieval stress on memory (Shields et al., 2017), we anticipate that stress induction will lead to a decline in memory performance when no specific strategies are employed (H1). This hypothesis will be investigated by comparing the control learning strategies in a stress versus non-stress condition. Second, prior evidence suggests that retrieval practice benefits memory more so than other typically used strategies such as re-reading or highlighting (e.g., Moreira et al., 2019) under non-stressful conditions. We thus hypothesize that retrieval practice will yield memory benefits even in the absence of stress (H2). This hypothesis will be tested by comparing the effects of learning with retrieval practice versus a control strategy in non-stress conditions.

Third, using retrieval practice may make memory less sensitive against the detrimental impact of stress on memory (A. M. Smith et al., 2016). Thus, we expect that retrieval practice will outperform other control strategies in mitigating the memory impairments induced by stress (H3). This hypothesis will be tested by comparing the effects of learning with retrieval practice versus a control strategy in groups that underwent stress induction. And, in a second step, if H3 is confirmed, we will further explore this benefit by comparing the effects of retrieval practice on memory in a stress versus non-stress condition. Here, we expect relatively equal performance when using retrieval practice in a stress versus non-stress condition (H4), as the protective benefit of retrieval practice in the stress condition should make it equivalent with the benefit of the strategy already experienced in the non-stress condition (A. M. Smith et al., 2016).

Confirmatory Moderators

Stressor Type. When focusing on testing situations (i.e., memory retrieval), the

literature points to mixed effects for different types of stressors. Namely, stress induced via protocols in laboratory settings such as TSST leads to memory impairments (Shields, Sazma, et al., 2017), whereas test anxiety does not always have an effect (Clark et al., 2018; Yang et al., 2020). To further explore these differences, we coded this moderator according to the type of stressor: TSST or TA. Based on our current understanding of the literature, these are the two main types of stressor tasks. However, additional types of stressor tasks may be added at Stage 2 when we conduct the literature search.

We predict that all types of retrieval stressors will lead to negative effects on memory performance in groups who did not learn with retrieval practice but will not have a negative impact when learning with retrieval practice.

Other Strategies. Previous evidence suggests that retrieval practice benefits memory more so than other typically used strategies such as re-reading or highlighting (e.g., Moreira et al., 2019). Thus, we wanted to explore how different control strategies used in comparison with retrieval practice could moderate overall effects. This moderator was coded by categorizing each other strategy used (ie., restudy, highlighting, drawing diagrams).

We predict that strategies used other than retrieval practice would be less beneficial than retrieval practice.

Delay. Previous studies examining the impact of stress and retrieval practice on memory performance were conducted with varying lengths of retention intervals between the initial learning session and the final memory assessment. We wanted to explore whether retention interval impacts memory performance following learning with retrieval practice. This moderator was coded by different potential delay periods following learning to memory test (e.g., 1 day, 2 days, 1 week, etc.).

Because retrieval practice is shown to have sustained long-term benefits (Roediger & Karpicke, 2006), we expect its protective factor to continue even after a long-term delay (e.g., 1 week) following initial learning.

Task Type. Previous studies utilized different types of learning material when measuring the impact of retrieval practice on memory performance. For example, the classic study by Roediger and Karpicke (2006) had participants learn educational reading passages while A.M. Smith and colleagues (2016) asked participants to remember word lists. However, meta-analyses on the benefits of retrieval practice (e.g., Rowland, 2014) show consistently positive effects regardless of this learning strategy on different types of materials. Thus, we wanted to explore whether the type of learning material used impacts the overall effectiveness of retrieval practice. This moderator was coded as the different types of materials used (e.g., reading passages, word lists, questions, etc.).

We expect retrieval practice to have a positive effect on all types of learning material used.

4.3 Methods

We shared all procedures, materials, datasets, articles, and code on Open Science Framework (https://osf.io/nye73). There are no other unreported/unlinked pre-registrations for this meta-analysis project. The templates on OSF and the template for Stage 1 Registered Reports used in this meta-analysis have been adapted from the resources developed by Feldman (2019a, 2019b) and (Yeung & Feldman, 2022). We made all efforts recommended by the field to enhance reproducibility, openness, and transparency (Maassen et al., 2020; Schwab et al., 2022).

4.3.1 Literature Search

An unstructured literature search was first performed on these databases in April 2022 during the conceptualization stage of the current work to test and refine our search terms. During this initial probe, the articles were not systematically searched.

To find articles relevant on the topic, we used the following databases: PsycInfo, PubMed, JSTOR, and Web of Science. The following search terms were applied on all databases: ("testing effect*" OR "retrieval practice*") AND ("stress*" OR "test anxiety*") using the appropriate search syntax terms for each database. We used Boolean operators

such as “OR” and “AND” in the search pattern to connect test anxiety with stress and retrieval practice or the testing effect. These terms are similar to other reviews on the topic (Rowland et al., 2014; Schwieren et al., 2017) with the addition of the term “test anxiety.” We selected experimental studies published in peer-reviewed journals in English between the years of 2006 – 2022. This range was subsequently extended to the month and year in which Stage 1 acceptance was obtained (April 2024), thus making our literature search range from 2006 – 2024. The year 2006 was selected as the start date as that is the year Roediger & Karpicke (2006) published the initial findings regarding the benefits of retrieval practice, which has since then led to an explosion of research in that area. Grey literature was searched on pre-print archives (e.g., OSF Pre-Prints) and featured unpublished studies and theses databases (e.g., Thesis Commons, ProQuest) using the same search terms as the database search. We reran the searches at least twice to ensure all literature was up to date. The date last searched was April 28, 2024. The total outcome was 6,147 prospective articles. Following deletion of duplicates, we had a total of 6,028 articles (Figure 11).

After that, a search for relevant papers not listed in the primary database search was conducted, by manually searching for papers listed under the “related articles” and “cited by” features in Google Scholar (Walters, 2007) using the identified list of articles. This allowed us to find articles that were not detected in the keywords search process. Additionally, we also conducted one additional round of search by skimming the reference sections of identified articles from our primary search. The date last searched was May 30, 2024. The outcome was a total of 16 additional articles.

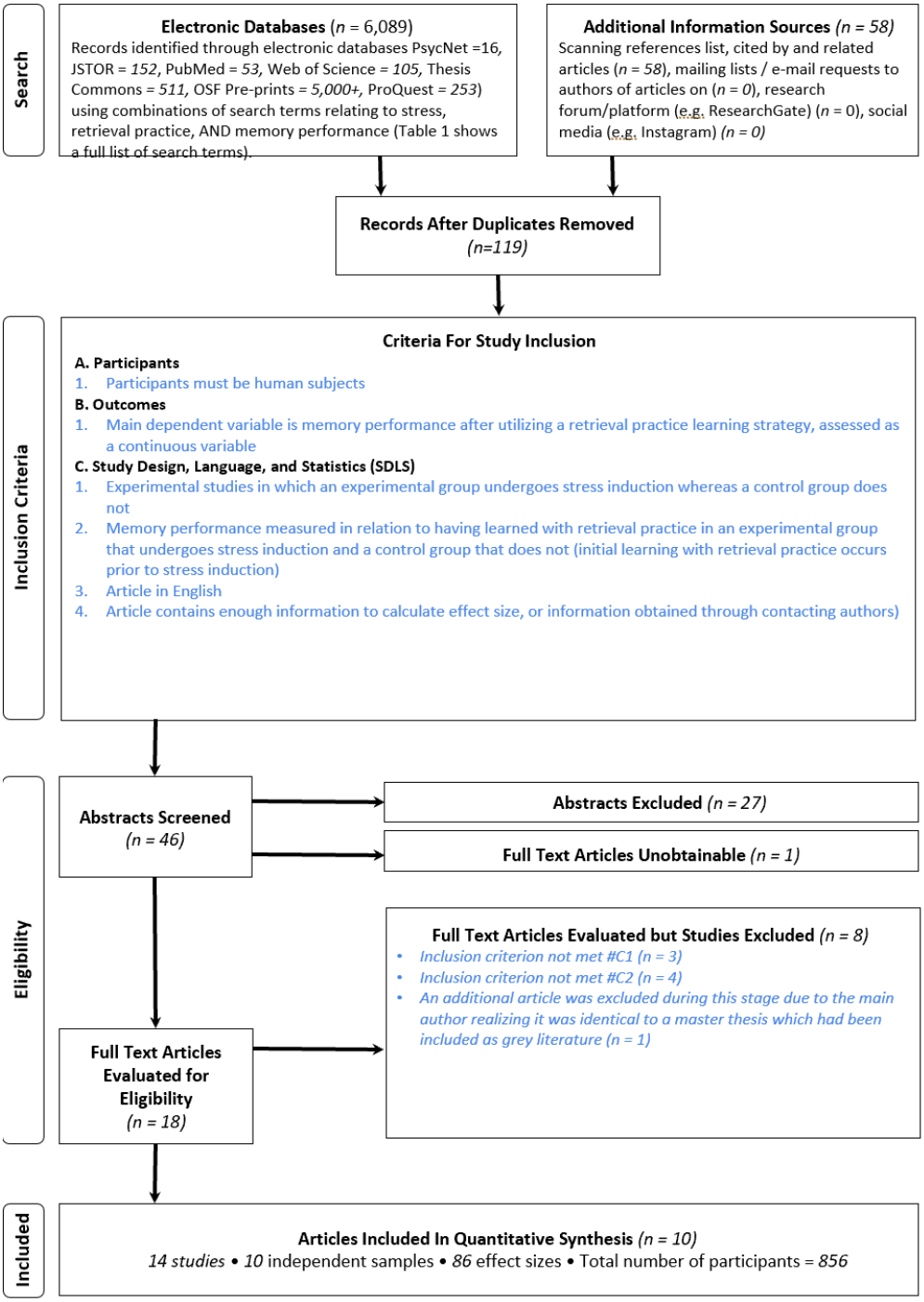
Furthermore, we identified authors in the field of the stress and memory literature along with authors of other identified articles and searched through their related publications. This ensured full coverage and maximized access to unpublished data and/or manuscripts that are also relevant. This is an essential part of a meta-analysis process, as it may reduce the effect of publication bias and may help prevent overestimated effect sizes (Feltz & May, 2017). In total, we contacted nine authors, three of which replied. None of

those who replied had any additional articles to provide. Lastly, we issued a call for unpublished findings on online forums, research platforms, and social media (e.g., ResearchGate, Meta) throughout the month of May 2024. This did not yield any additional articles. We did not contact authors for missing statistics as we were able to extract them from the papers.

After the above search procedures, MM and MZ scanned all abstracts, tables, and method sections to identify the relevance of the sources (see Screening section). If the articles indicated relevance for our analysis, MM and MZ read more of the articles to determine whether they met the inclusion criteria or whether articles had to be excluded based on our search criteria (see next section). Disagreements were resolved via discussion rounds at regular update meetings and reliability scores were performed at each step of the screening process to ensure consistency. A second scan round enabled us to exclude 6 articles, reducing our sample of studies to 12 articles with a total of 1,046 participants. At Stage 2, we further excluded two additional studies according to exclusion criteria, ending up with a final 10 articles with a total of 856 participants (see Supplementary, Table 6). We listed all the excluded articles in the Full Coding Sheet (available on OSF).

Figure 11

Systematic literature search flow diagram



Note. The above template is adapted from Moher et al. (2009) preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement (www.prisma-statement.org), as well as Moreau and Gamble (2022). Meta-analysis templates and materials Template 2 Search Flow Diagram (osf.io/q8stz). It has been used/adapted from Yeung and Feldman (2022).

4.3.2 Inclusion and Exclusion Criteria

Meta-analysis is meant to integrate similar or comparable studies (Higgins, 2003). Since the aim of our meta-analysis was to determine whether learning with retrieval practice protected memory following stress exposure, we established strict inclusion and exclusion criteria.

First, the main dependent variable in each article needs to assess the impact of the stressor vs. non stressor on memory in relation to having learned with retrieval practice. In the current work, stress induction is defined to mean undergoing a procedure for stress induction prior to retrieval. These procedures can include, but are not limited to, standard procedures for stress induction: the Cold Pressor Test or variations of the TSST (see Kirschbaum et al., 1993). As such, we included studies that feature an experimental vs. control group design where the experimental group undergoes stress induction procedure, and the control group does not. Additionally, stress induction is extended to the induction of test anxiety. For these studies, the "stressor" corresponds to taking a test, being placed in an evaluative situation, or otherwise inducing test anxiety or evaluative threat. In such cases, individuals in the test anxiety group are considered as the stress group and those in the non-test anxiety are the control group. Likewise, retrieval practice is defined to mean the activity of actively recalling information from memory or engaging in the testing effect (Roediger & Karpicke, 2006). Papers that included a learning session that is performed with retrieval practice versus a control strategy (i.e., restudy), or several other strategies, were also accepted. Papers which featured a non-retrieval practice group were also accepted as a viable comparison group.

Second, the experimental studies we focused on had to include adequate statistical information for computing the effect size for the effects of retrieval practice on memory following stress induction. Namely, the article needs to report means, standard deviations, and sample sizes for both the experimental and control groups who learned with retrieval practice versus another strategy (if included) after undergoing the stress procedure.

Alternatively, articles need to include the effect size that represents the magnitude of having learned with retrieval practice in a group that underwent a stressor versus a control or for the interaction effects between retrieval practice and a control strategy between stress and control groups. In cases of missing statistical data, we first attempted to contact the authors (Polanin et al., 2019) if the statistics were not reported or unable to be extracted from plots with WebPlotDigitizer (Rohatgi, 2020) or metaDigitise (Pick et al., 2018). If we were not able to obtain the required statistics, we excluded the articles even if the articles met all other search criteria. We excluded all correlational studies and other non-experimental studies.

Third, we excluded articles not written in English, unless we obtained all necessary data and information for coding in English, or we obtained such data and information from the authors. Fourth, we excluded retracted studies if the retraction is due to problems of data collection and data analysis (Fanelli et al., 2022).

4.3.3 Screening

Studies collected through database searches and through contacting authors were assessed for their eligibility based on their titles, abstracts, and contents. Titles were first scanned to identify the relevance of the sources. Relevant titles were rated with a "1" and irrelevant titles with a "0" by two independent raters (MM and MZ). If the titles seemed relevant for our analysis, the full articles underwent abstract inspection. Abstract screening followed the procedures suggested by Polanin and colleagues (2019). We looked for relevant key words (i.e., "stress," "retrieval practice") and words indicating an experimental design. Relevant titles were also cross-checked using the automated tool in the litsearchr package in R (Grames et al., 2019, v0.1.0), though the searches were of manageable size and didn't require automation. All relevant abstracts identified in the abstract screening then underwent eligibility screening via methods inspection. During this round of screening, the methods section of all identified articles was carefully inspected to ensure they met all inclusion criteria and were eligible for inclusion.

The methods sections were independently checked by MM and MZ using separate spreadsheets. Raters met regularly to update on progress and discuss any disagreements at each stage. Disagreements were resolved through deliberations with a senior member. Inter-rater reliability scores were assessed before and after rater deliberation. All decisions for inclusion and exclusion were documented clearly, transparently, and systematically in the excel spreadsheet Full Coding Sheet spreadsheet and Literature Searches (tab name "Article List Inclusion and Exclusion Criteria," see OSF https://osf.io/nye73, based on https://osf.io/4cdzx). We saved all preliminary references of the studies in the total search into a list available on OSF. The open-access full-texts are accessible on OSF.

4.3.4 Coding and Pre-testing

We developed a data coding sheet (tab name "Coding" in the Full Coding Sheet) and a Codebook (available on OSF) to keep a clear record of our decisions at different stages and enhance reproducibility (Arslan, 2019; Obels et al., 2020; Siddaway et al., 2019). Before we began with the coding process, we pilot-tested 2 randomly selected studies in two stages and refined it accordingly in every stage. Authors MM and MZ completed the coding process to ensure a higher inter-rater reliability. We documented gaps and reported decisions in detail in the "Article and Decision" tab of the Full Coding Sheet (see OSF https://osf.io/nye73). As the main contributor, MM then verified the coding sheet and adjusted any discrepancies if necessary following deliberations.

4.3.5 Included Studies Coding

Once we completed the article selection procedures, pre-test coding, and confirmed the included studies, MM and MZ coded the studies independently. A codebook with instructions (see OSF) on how to code each column was provided to all coders during the coding process. Both coders coded half the studies individually, then checked each other's work for the other half of studies. Inter-rater agreements were checked with Cohen's Kappa and intra-class correlation coefficient (Hohn et al., 2019; Siddaway et al., 2019). If, during the coding process, the article was found not suitable for meta-analytic inclusion, all

reasons were clearly stated in the "Excluded Studies" tab.

4.3.6 Confirmatory Analyses

We used RStudio v4.1.3 (R Core Team, 2020) for the statistical analyses with packages for meta-analysis such as metafor (Viechtbauer, 2010, v3.8-1). We used the analysis templates adapted from Yeung and Feldman (2022) for meta-analyses in psychology. We also followed the guidebook laid out by Harrer and colleagues (Harrer et al., 2021) to conduct the meta-analysis.

We converted all effect sizes into Hedges' g during analysis to facilitate comparison. Multiple effect sizes (i.e., different measures within the same study) were handled by computing a separate effect size for each different relevant scenario described in the article (Appelbaum et al., 2018). For missing data (e.g., effect size missing, but M and SD reported), we calculated using packages such as esc (Lüdecke, 2019, v0.5.1) or compute.es (Re, 2020, v0.2-5). Calculation or coding procedures, as well as all packages and functions used, were documented in the Full Coding Sheet ("Included Studies Effect Coding" tab).

Whenever standardized effect sizes were not available, we used either descriptive statistics or inferential statistics, such as means and standard deviations. We also verified statistical results from articles using statcheck (Nuijten & Polanin, 2020) to confirm internal consistency. If the original article did not directly report mean and standard deviation but simply provided graphs, we used WebPlotDigitizer (Rohatgi, 2020) and/or metaDigitise (Pick et al., 2018) to extract the necessary values. We documented all conversions and coding decisions. We included the original quotes and/or table/page numbers from the original articles into the "Included Studies Effect Coding" tab to facilitate reproducibility.

For main-effects, we analyzed the data with a two-level random-effects model (Borenstein et al., 2009; Slaney et al., 2018). This model was adopted because it assumes that studies stem from different populations, thus resulting in a distribution of effect sizes rather than one true effect (Harrer et al., 2021). This model seemed applicable for our

research question because it is unlikely that the selected studies will be completely homogeneous. The random-effects model produces an overall effect size, along with heterogeneity measures.

Importantly, because our main hypotheses look at the effects of retrieval practice compared to control strategies in stressed and non-stressed conditions, as well as of control strategies in stressed versus non-stressed conditions, we conducted the random-effects model for all primary hypotheses (H1, H2, H3, H4). For H1, we compared the effects of learning with control strategies in stress versus control or non-stress conditions. For H2, we compared the effects of learning with retrieval practice versus a control strategy in a control or non-stress condition. For H3, we compared the effects of learning with retrieval practice versus other strategies in stress conditions. Lastly, for H4, we compared the effects of learning with retrieval practice in stress versus non-stress conditions. This breakdown allowed us to isolate and compare the effects of retrieval practice versus other strategies on memory performance. Because we are running models on all four scenarios, we decided not to conduct multivariate models, which are typically used to assess multiple correlated outcomes within the same study. However, we ran sensitivity analysis with three-level models at Stage 2 for each H to account for the nested structure of the data (i.e., effect sizes within experiments, and experiments within studies). We also included the robust variance estimator (RVE) at Stage 2 as another sensitivity analysis to account for potential dependencies among effect sizes that might not be fully addressed by the multi-level structure.

To determine the impact of the stressor on memory performance, we conducted additional analysis by applying meta-regression using participant's scores on the stress manipulation checks as moderators weighed on memory performance scores in the stress condition. This analysis was conducted as a sanity check to verify that the stress procedure was successful in included studies. The stress scores were extracted from each study and reflected participant's self-reported stress score on a stress or anxiety measure taken before

and after the stressor in both stress and control groups. This analysis will only be conducted if the stress measurements extracted from studies are sufficiently comparable at Stage 2.

We plotted forest plots presenting the effect size of each study. We presented the effect size with confidence intervals and sample size of each study. Statistical heterogeneity between studies was determined using the Q statistic and quantified with I^2 (Huedo-Medina et al., 2006). This global meta-analysis yielded a point estimate, confidence interval, and p-value, along with statistics for heterogeneity. We determined a threshold of I^2 of over 50% and a significant Q statistic as an indicator to perform subsequent moderator analysis (Harrer et al., 2021). If we obtain such results, we can assume that there are sources of variation other than sampling error in our sample, thus warranting further investigation. If there was indeed meaningful heterogeneity, we investigated and explored potential moderators.

For moderator analysis, we used two-level plural models for contrasting moderator categories. These models combine the fixed-effects model to assess differences in true effect sizes between fixed subgroup levels and the random-effects model to account for potential heterogeneity within and among subgroups (Harrer et al. 2021). Because moderator analysis is heavily dependent on statistical power (Harrer et al., 2021), we controlled for the low power issue by using the MetaForest package (van Lissa, 2020, v0.1.3). This procedure uses bootstrapping techniques to overcome the low power issues in moderator analyses. It provides a ranking of moderators in terms of variable importance.

Publication bias was assessed by first evaluating "small study effects" (Harrer et al., 2021). Small study effects (SSEs) refer to the phenomenon where studies with smaller sample sizes tend to show larger and more extreme effects compared to studies with larger sample sizes. Thus, small studies are more likely to get published while studies with non-significant results are more likely to be unpublished, creating skewed evidence. To assess small study effects, we first plotted the effect sizes and standard errors of each study,

visually depicted in a funnel plot. Egger's Test of the Intercept (Egger et al., 1997; Sterne & Egger, 2005) was then used to calculate whether asymmetry exists in the funnel plot. If Egger's Test is significant, this may be due to missing studies. To check this, we then applied the trim-and-fill procedure, which corrects for this asymmetry by filling in missing studies (Duval and Tweedy, 2000). We also conducted the Rank correlation test (Begg & Mazumdar, 1994) which assesses the association between effect sizes and their standard errors. The Rank test produces a measure of association with Kendall's tau, where significant correlations suggest publication bias.

To check for publication bias, we applied the PET-PEESE method (Stanley & Doucouliagos, 2014). In the PET method, the effect of small studies is controlled by including the standard error as a predictor in a weighted regression model where the study's effect size is regressed on its standard error (Harrer et al., 2021). Similarly, the PEESE method uses the squared standard error as a predictor. If the regression intercept calculated by PET is significantly larger than zero, the PEESE is used as the true effect estimate. If the PET intercept is not significantly larger than zero, the PET is used as the true effect estimate (Harrer et al., 2021). We also conducted a three-parameter selection model (Iyengar & Greenhouse, 1988). This model uses three parameters to assess publication bias: the effect size parameter, the heterogeneity parameter (τ^2), and the likelihood of selection. Selection models predict how likely it is that a study is published (i.e., "selected) based on its results (i.e., it's p-value). The model then "removes" the assumed bias due to selected publication and derives a corrected estimate of the true effect (Harrer et al, 2021).

The above publication bias methods are our preferred methods based on simulations of false positives, statistical power, and recommendations from the field (Carter et al., 2019). We acknowledge that there are many different approaches to publication bias correction. We also acknowledge that heterogeneity and publication bias are closely intertwined, and some measures of publication bias can be sensitive to underlying study

heterogeneity. This sensitivity could affect the reliability and interpretation of our findings. One way in which we will disentangle the two in the current work involves conducting Egger's test to assess the presence of publication bias, while also evaluating heterogeneity using methods such as Cochran's Q or I^2 statistic, as outlined above. Significant heterogeneity may indicate that studies are estimating different underlying effects, whereas significant results from Egger's test could suggest publication bias. Moreover, we will perform sensitivity analysis using the leave-on-out method (Harrer et al., 2021) where effect sizes are recalculated with one study removed each time to assess the robustness of findings and identify potential outliers. Additionally, we will consider the sample size and quality of included studies when interpreting results, recognizing that small sample sizes and low-quality studies are more vulnerable to biases and spurious results (Brysbaert, 2019), which may influence our understanding of potential publication bias. Additionally, we will also assess study level power to check whether publication bias is likely (Quintana, 2023).

4.3.7 Power Analysis

A priori power analysis was conducted prior to beginning the current work. We expected the effect size of retrieval practice following stress exposure to be $d = 0.61$, as previously demonstrated in experimental results for memory performance in a stressed group of participants that learned with retrieval practice (A. M. Smith et al., 2016). Because our understanding of the literature is that the current field is still emerging, we expected to include 10 studies. We expected the average sample size per study and condition to be 25 and we expected moderate-to-high heterogeneity. We conducted a priori-power calculation with dmetar v0.0.9000 package (Harrer et al., 2019, available on OSF), which yielded a power estimate of 99.91%. We also conducted sensitivity power analysis by conducting a simulation with the same parameters, but assuming an effect of $d = 0.4$, the smallest effect size needed for real world application in psychological research (Brysbaert, 2019). The estimated power for these parameters was 92.88%. Both analyses suggested we had viable power to conduct the meta-analysis assuming those parameters.

Post-hoc power analysis was performed at Stage 2 by re-running our initial power analysis script above with the actual effect sizes obtained from our meta-analysis. We also used the upper and lower bounds of the confidence intervals for each H overall effect, as suggested during Stage 2 review. As a complementary approach, we also applied the metameta package (Quintana, 2023). The metameta package serves as a versatile tool for conducting post-hoc power analysis in meta-analysis, enabling researchers to determine the range of effect sizes reliably detectable within a body of studies. By utilizing data extracted from meta-analysis forest plots and tables, metameta calculates study-level statistical power and median statistical power based on published effect-size and variance data.

4.3.8 Risk of Bias

Risk of bias of individual studies included in our meta-analysis was assessed with Cochrane Risk of Bias 2 tool (Sterne et al., 2019). Risk of bias is essential to perform in a meta-analysis to assess and weigh the relative bias risk each study poses. Risk of bias is assessed across the following domains: the randomization process, deviations from intended interventions, missing outcome data, measurement of outcome, selection of the reported results. Judgments regarding the risk of bias for each domain are based on answers to signaling questions, which are rated on the basis of "yes," "probably yes," "no," "probably no," or "no information." The resulting judgments of "low," "some concerns," or "high" risk of bias are outputted by the risk of bias algorithm in the tool. Risk of bias judgments were performed by two independent raters (MM and MZ). Risk of bias ratings are available on OSF (https://osf.io/v32b9).

4.4 Results

In the following results section, we first present the retrieval practice main effect findings, followed by publication bias, moderator analysis, post-hoc power analysis, and risk of bias. Note that the forest plot for H3 only is provided in the main paper (see Supplementary for H1, H2, H4 forest plots). Also note that $k = 9$ for H1-H3, as one study (Hupbach & Feiman, 2012) only had effects for H4 and is thus only factored into H4.

4.4.1 Retrieval Practice Main Effects

Random-Effects Two-Level Model for H1 (other strategy in a stress vs. non-stress condition)

We first examined the overall effect of having learned with a control learning strategy in a stress compared to non-stress condition. The mean effect was negative, and the *p*-value was not significant, $k = 9$ with 21 effect sizes, g = -0.05, CI [-0.15, 0.05], p = 0.34. Thus, we found no support for H1. Heterogeniety for H1 was not significant ($Q(df = 17) = 15.99$, $p = 0.72$, $I^2 = 0\%$). Sensitivity analysis with a three-level model and robust variance estimation accounting for effect size dependencies showed similar results as the main model (see Supplementary).

Random-Effects Two-Level Model for H2 (retrieval practice vs. control strategy in a non-stress condition)

Next, we examined the overall effect of having learned with retrieval practice compared to a control strategy in a non-stress condition. The mean effect was positive, and the *p*-value was significant, $k = 9$ with 21 effect sizes, $g = 0.37$, CI [0.09, 0.66], $p = 0.01$. Thus, we found support for H2. Additionally, there was significant heterogeniety ($Q(df = 20) = 115.38$, $p < 0.01$, $I^2 = 86\%$). Sensitivity analysis with a three-level model and robust variance estimation accounting for effect size dependencies showed similar results as the main model (see Supplementary).

Random-Effects Two-Level Model for H3 (retrieval practice vs. other strategy in a stress condition)

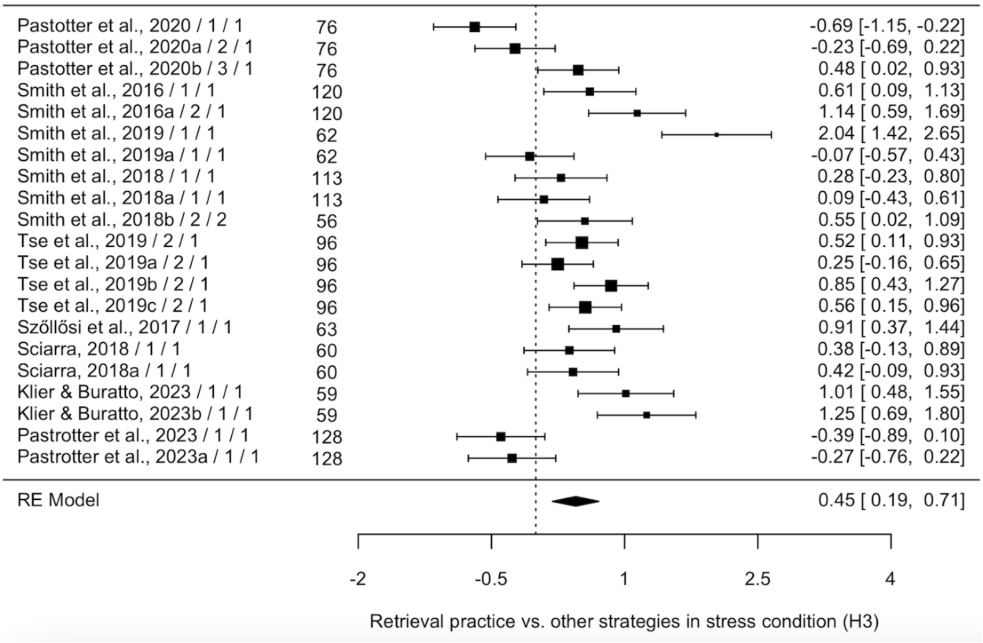
We then examined the overall effect having learned with retrieval practice versus a control strategy on memory performance in a stress condition, our main hypothesis of interest. The mean effect was positive, and the *p*-value was significant, $k = 9$ with 21 effect sizes, $g = 0.45$, CI [0.19, 0.71], $p < 0.01$ (Figure 12). Thus, we found support for H3. Additionally, there was significant heterogeniety ($Q(df = 20) = 109.20$, $p < 0.01$, $I^2 = 83\%$). Sensitivity analysis with a three-level model

and robust variance estimation accounting for effect size dependencies showed similar results as the main model (see Supplementary).

Random-Effects Two-Level Model for H4 (retrieval practice in a stress vs. non-stress condition)

Lastly, we examined the overall effect having learned with retrieval practice in a stress versus non-stress condition. The mean effect was positive with a negative confidence interval, and the p-value was not significant, $k = 10$ with 23 effect sizes, $g = 0.08$ CI [-0.02, 0.19], $p = 0.11$. Thus, we found no support for H4. Additionally, heterogeniety was not significant ($Q(df = 21) = 23.65$, $p = 0.37$, $I^2 = 6.47\%$). Sensitivity analysis with a three-level model and robust variance estimation accounting for effect size dependencies showed similar results as the main model (see Supplementary).

Figure 12
Forest plot for H3



Note. Forest plot of all studies in H3 (with their coded study and experiment number) included in this meta-analysis with their respective sample size, effect size (represented by the squares), weight, and 95% confidence intervals. Larger squares indicate more weight in the random effects model. The x-axis represents the standardized mean difference effect size (Hedges' g), and the y-axis is each individual study. RE model at the bottom represents the overall effect size and its 95% confidence interval using the random-effects model. Article name (e.g., Pastötter 2020), followed by study number, experiment number within the study, and sample number (e.g., Pastötter 2020 / 1 / 2 / 1). Letters (e.g., "a", "b") indicate multiple studies within the same article.

4.4.2 Effect of Stressors

To examine the effect of the stress manipulation on memory performance, we submitted participant's subjective stress scores to a meta-regression. Subjective stress scores were used as the stress measure because they were consistently reported across most studies. One study (Klier & Buratto, 2023) was not included in this analysis as they did not report these scores. The test for residual heterogeneity ($QE(df = 6) = 4.57$, $p = 0.60$)

indicated no significant unaccounted variability in the model. The test of moderators yielded a QM statistic of 0.38 ($p = 0.54$), suggesting a nonsignificant influence of stress scores on memory performance overall.

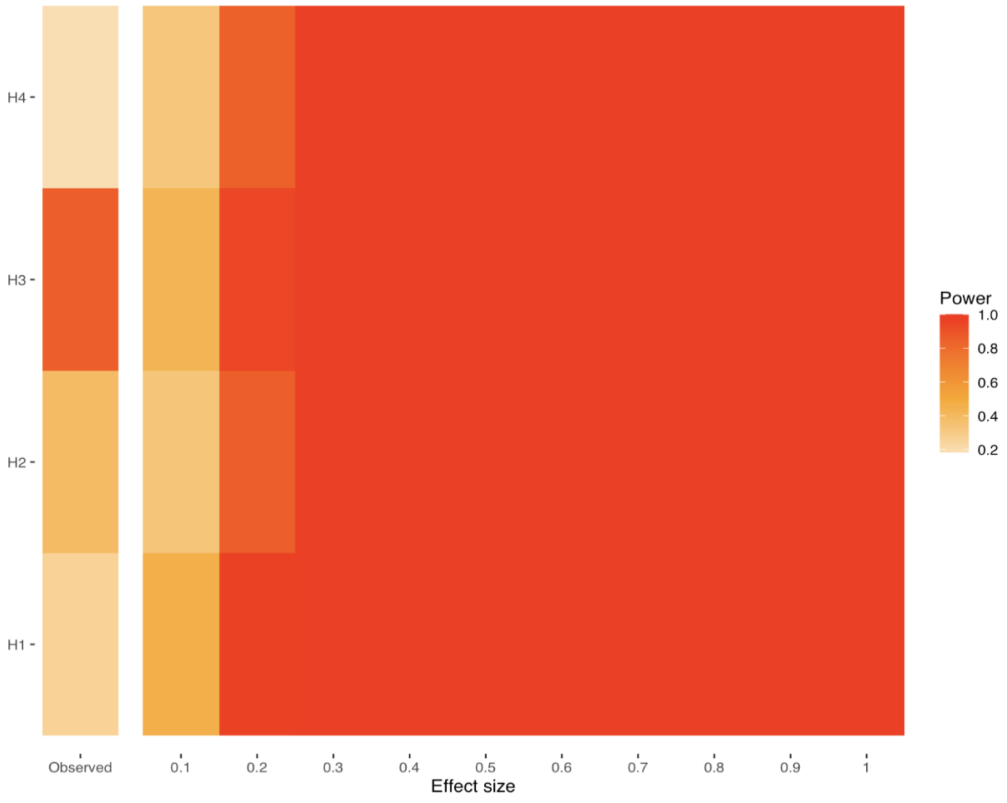
4.4.3 Statistical Power

Post-hoc power analysis was conducted to determine the adequacy of our study design under the assumption of a random effects model with moderate heterogeneity. The analysis was done by using the upper and lower bounds of the confidence intervals, as suggested during Stage 2 review, from the main effect of each H with 10 studies and 30 participants per group. For H1, power estimates ranged from 8.29% - 43.74%; for H2, estimates ranged from 12.91% - 99.98%; for H3, estimates ranged from 37.53% - 100%; and for H4, from 5.52% - 52.13%. Overall, H3 was the most well-powered of all hypotheses. However, none of the hypotheses demonstrated consistently sufficient power across the full range of estimates, particularly at the lower end of the confidence intervals.

We also conducted post-hoc power analysis by calculating study-level statistical power and median statistical power based on published effect-size and variance data from our meta-analysis using the metameta package (Quintana, 2023). We visualized median power in a Firepower plot (Figure 13), which summarizes the median statistical power across multiple meta-analyses (i.e., in our case, the four Hs) and illustrates the ability of studies within each H to detect the overall effects found per H. Results showed that larger effect sizes are more likely to be detected, however the overall power across the studies is relatively low, as they are unable to detect effects smaller than 0.3.

Figure 13

Fireplot visualizing power across all Hs



Note. Fire plot indicating statistical power for all hypotheses (Hs) in the meta-analysis. The x-axis represents the total power ranging from 0.1 (10% or low power) to 1.0 (100% or high power) and the y-axis represents the hypotheses. Colors indicate the median power of the conducted meta-analyses. Higher color saturation indicates higher statistical power for effect sizes (see Power bar).

4.4.4 Publication Bias

Null findings are less likely to be published (Begg & Berlin, 1988; Duval & Tweedie, 2000), resulting in biased published literature and a possible overestimation of an effect. We looked at small-study effects (SSEs) and correction methods, and employed six different statistical approaches to examine publication bias according to recommendations from the field (Harrer et al., 2021). A summary of publication bias analyses is provided in

Table 5 for H3 and in Supplementary (Table 7) for H1, H2, and H4. There were discrepancies in the methods. For instance, for H3, no missing studies were identified using the Trim-Fill, however Egger's Test and Kendall's *tau* were both significant, indicating funnel plot asymmetry, though visual inspection of the funnel plot did not confirm this (Figure 14). Correction methods using PET-PEESE and the selection model were not significant, showing no strong evidence of publication bias. Significant publication bias was found for H1 and H4 with the Trim-Fill imputing missing studies (Figure 15) in both, as well as significant PET-PEESE intercepts suggesting a positive effect. However, as neither H1 nor H4 are supported, this more likely indicates no effects for this hypothesis. For H2, Egger's test suggested funnel plot asymmetry, but this was not accompanied by any other bias indications (i.e., PET was not significant). Additionally, the Likelihood ratio tests from the selection model were insignificant across all Hs, indicating the pooled effects were not distorted by selective reporting.

We also conducted a sensitivity analysis using the leave-one-out method (Harrer et al., 2021) to disentangle the effects of heterogeneity and publication bias, recognizing that some measures of publication bias are sensitive to underlying heterogeneity. This analysis revealed that the overall effect size was robust, with stable estimates and confidence intervals when individual studies were excluded. Despite the robustness, there was significant heterogeneity in the dataset, as indicated by high Q statistics and I^2 values, suggesting that there is considerable variability in effect sizes across the studies included in the meta-analysis.

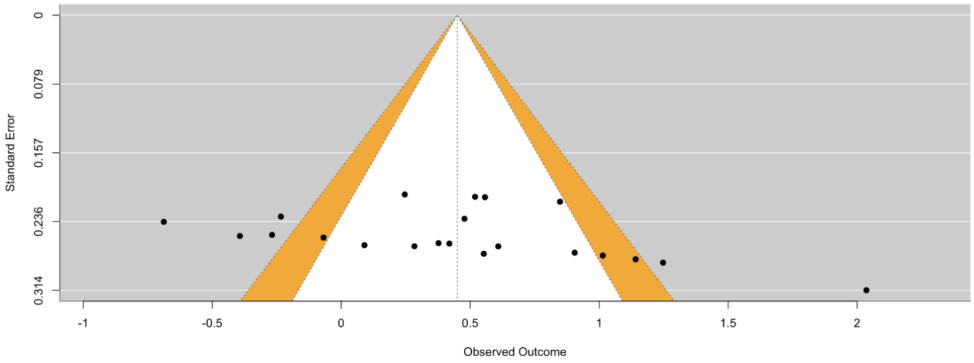
Table 5

Summary of publication bias for H3

Publication bias analysis method	Results and adjusted models
Small Study Effects	
Trim and fill funnel plot asymmetry	0 missing on the left side. Adjusted model: g = 0.45, 95% CI [0.19, 0.71]
Egger's regression test	z = 2.34, p = 0.02
Rank correlation test	Kendall's *tau* = 0.45, p = 0.004
Publication Bias Correction	
Three-parameter selection model	Likelihood Ratio Test: 0.69, p = 0.4 Adjusted Model: g = 0.31, 95% CI [-0.10, 0.72]
PET	b = -1.41 [-3.63, 0.81], p = 0.10
PEESE	b = -0.61 [-1.73, 0.50], p = 0.06

Note. Values in parentheses indicate 95% confidence intervals [lower bound, upper bound].

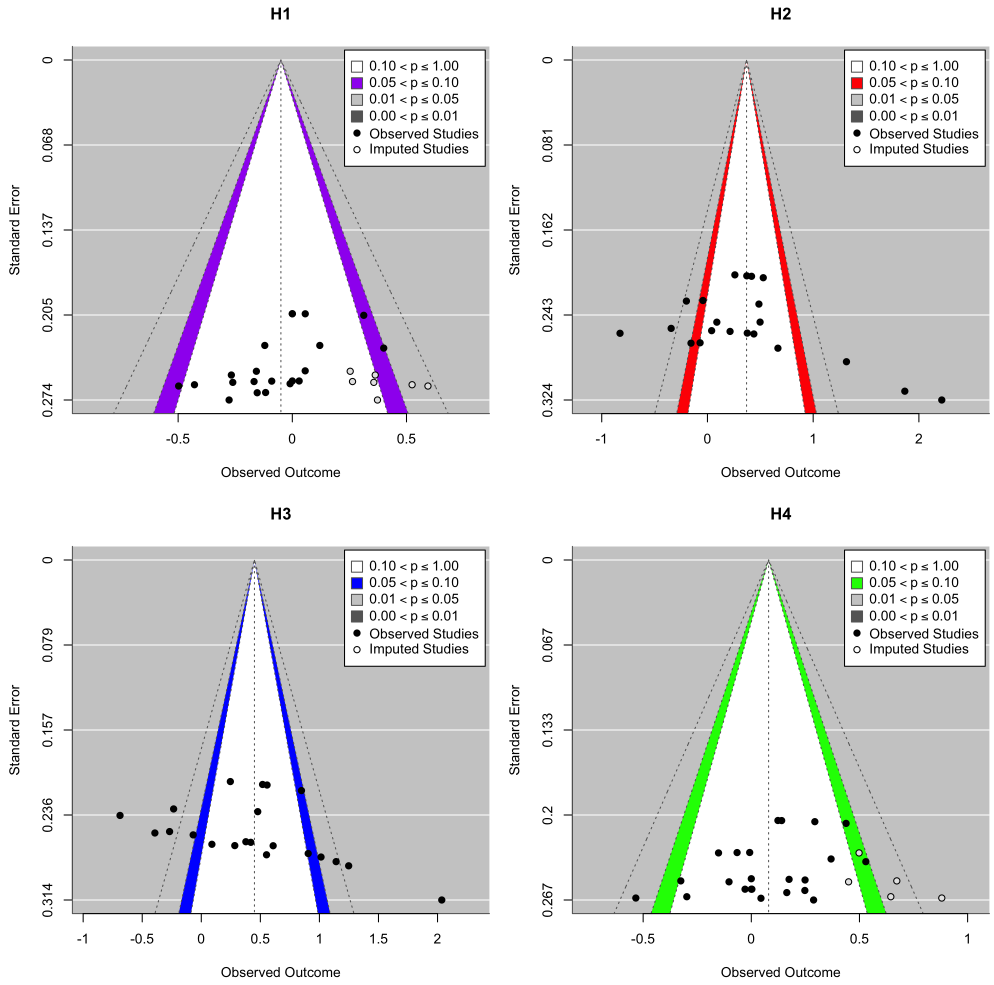
Figure 14
Funnel plot for H3



Note. Funnel plot of H3 showing each effect size (x-axis) plotted against its standard error (y-axis). Studies with higher precision are at the top and studies with lower precision (i.e., larger standard errors) congregate at the bottom of the funnel and signify publication bias. The current funnel plot shows studies homogeneously spread out, effects are generally consistent, though not many studies are at the top.

Figure 15

Funnel plots for H1, H2, H3, H4



Note. Funnel plot of all studies per hypothesis (H) showing each effect size (x-axis) plotted against its standard error (y-axis). Studies with higher precision are at the top and studies with lower precision (i.e., larger standard errors) congregate at the bottom of the funnel and signify publication bias. The dotted line in the center represents the overall mean effect. Significant publication bias is seen in H1 and H4, indicated by the seven and five imputed studies from the Trim-Fill procedure.

4.4.5 MetaForest Moderator Analyses

To address the problem of limited studies and lack of statistical power without risk overfitting, we adopted MetaForest (van Lissa, 2017, v0.1.3). MetaForest uses "random forests," a machine learning technique, and bootstrapping to examine several possible moderators. The main model indicator, R^2 out-of-bag (R^2-OOB), was 0.23, suggesting that the included moderators explained approximately 23% of the variance in effect sizes. This reflects a moderate level of explanatory power and indicates that the moderators contributed somewhat meaningfully to accounting for variability in the data. Delay was the most important moderator, followed by Task Type and Stressor Type. Other Strategies had a negligible impact.

4.4.6 Moderator Analyses

Statistical heterogeneity was determined using Cochran's Q statistic and quantified with I^2 (Higgins & Thompson, 2002). The Q statistic was significant ($Q = 109.20, p < 0.01, df = 20$), suggesting the overall sample for H3 was highly heterogeneous. This is further supported by an I^2 value of 83.92%, suggesting a nearly 84% chance of the results being due to heterogeneity rather than chance. To examine the sources of heterogeniety, we examined three possible theoretical and methodological moderators according to a pre-registered criteria: Stressor Type (TSST vs. TA stress), Other Strategies used (restudy only), Delay (immediate, 1 day, 2 days, 1 week), and learning Task Type (words/lists, facts/short answer, educational texts). Results of moderator analyses are summarized in Figure 16 for H3, our main H of interest. Moderator analyses for H1, H2 and H4 are presented in Supplementary. Because all studies employed the same control strategy (restudy), the Other Strategies moderator lacked variability and could not be meaningfully analyzed. Therefore, we omitted it at Stage 2 for clarity and parsimony across all Hs.

As suggested during Stage 2 review of this registered report, we also ran Delay as a continuous moderator. This analysis found a significant effect ($QM = 7.75$, $p = 0.005$),

indicating longer delays were associated with increased memory performance for H3. The results for the other Hs are presented in Supplementary.

Stressor Type. Seventeen effect sizes for TSST stress had an effect size of $g = 0.43$, CI [0.10, 0.75], $p = 0.01$. Four effect sizes for TA stress had an effect size of $g = 0.54$, CI [0.30, 0.78], $p < 0.001$. We used a fixed-effects contrast and MetaForest to test if there is a meaningful moderating effect. We did not find a significant difference between TSST and TA type stressors.

Delay. Seven effect sizes for a delay of less than one day (i.e., immediate) had an effect size of $g = 0.05$, CI [-0.38, 0.47], $p = 0.83$. Four effect sizes for a delay of one day had an effect size of $g = 0.52$, CI [0.08, 0.97], $p = 0.02$. Six effect sizes for a delay of one week had an effect size of $g = 0.93$, CI [0.38, 1.49], $p = 0.001$. Four effect sizes for a delay of two days had an effect size $g = 0.40$ CI [0.17, 0.62], $p < 0.001$. We used a fixed-effects two-level model and MetaForest to test if there is a meaningful moderating effect. We did not find a significant difference between the moderators.

Task Type. Fifteen effect sizes for word or word lists tasks had an effect size of $g = 0.43$, CI [0.06, 0.80], $p = 0.02$. Four effect sizes for facts/short answer tasks had an effect size of $g = 0.54$, CI [0.30, 0.78], $p < 0.001$. Two effect sizes for educational texts had an effect size of $g = 0.40$ CI [0.04, 0.76], $p = 0.03$. We used a fixed-effects two-level model and MetaForest to test if there is a meaningful moderating effect. We did not find a significant difference between the moderators.

Figure 16

Summary of moderator analyses for H3

Moderator	*k*	*Q*	*df*	*g*	95% CI	*Tau²*	*I²*	Diff	*p*	*Categories*
					Stressor Type					
TSST	17	102	16	0.43	0.10-0.75	0.39	85.25%	0.29	0.59	TSST vs. TA
TA	4	4	3	0.54	0.30-0.78	0.02	27.70%			
					Delay					
Immediate (<1 day)	7	38	6	0.05	-0.38-0.47	0.28	83.92%	2.32	0.13	Immediate vs. 1 day
1 day	4	8	3	0.52	0.08-0.97	0.13	65.29%	1.27	0.26	1 day vs. 1 week
1 week	6	30	5	0.93	0.38-1.49	0.41	84.16	3.05	0.08	1 week vs. 2 days
2 days	4	1	3	0.40	0.17-0.62	0.00	0.00%			
					Type of Task					
Words/word list	15	102	14	0.43	0.06-0.80	0.46	87.14%	0.22	0.64	Words vs. Short answer
Facts/short answer	4	4	3	0.54	0.30-0.78	0.02	27.69%	0.40	0.53	Short answer vs. texts
Texts	2	0.01	1	0.40	0.04-0.76	0.00	0.00%			

Note. k = number of studies (i.e., effects from each study); g = Hedge's g effect size, CI = lower and upper limits of 95% confidence interval, tau^2 = tau squared value, I^2 = I-squared value, * $p < .05$, ** $p < .01$, *** $p < .001$, (all two-tailed); $Q = Q$ statistic indicating heterogeneity (decimals only reported if $Q < 1$), $Diff = QM$ statistic, test of moderator effects showing any differences between the moderators.

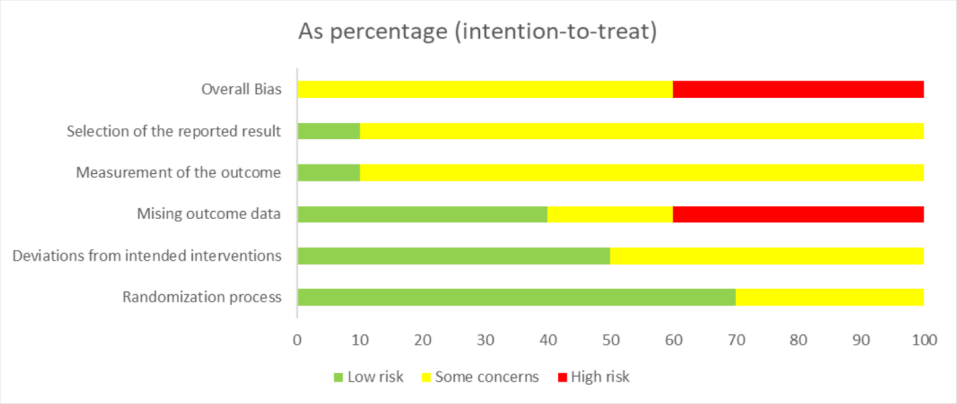
4.4.7 Risk of Bias

Risk of bias was assessed with the Cochrane Risk of Bias 2 tool (Sterne et al., 2019). The total risk of bias was "some concerns" across all studies (Figure 17). Most studies (70%) stated participants were randomly selected into stress or control groups, although they did not elaborate on the method of randomization (e.g., random number generator). The other half (30%) were rated as "some concerns" in this category for not mentioning any randomization or opting for quasi-randomization methods. For deviations from intended interventions, exactly half of studies were rated as "some concerns" due to participants and those delivering the intervention likely being aware of their assignment to the stress or control group or not having enough information to determine if they were. In terms of missing outcome data, 40% were at "high risk" due to having excluded enough participants to make a difference in the outcome or not mentioning reasons for participant

exclusion from analysis, and no sensitivity analyses being done to test how the removed data points impacted overall results. Nearly all studies (90%) had "some concerns" when it came to measurement of the outcome variable for assessors likely being aware of participants' assignment to experimental conditions and there not being sufficient information reported in the papers to determine if that knowledge could have made a difference or not in the measurement. All studies had "some concerns" when it came to measurement of the reported outcome due to not reporting a study protocol or analysis plan prior to having conducted the intervention.

Figure 17

Risk of bias across studies



Note. Summary table of the risk of bias in all included studies overall and across each of the five domains: overall bias (some concerns), selection of the reported result (some concerns), measurement of the outcome (largely some concerns), missing outcome data (a mix of low and high risk), deviations from intended interventions (a mix of low risk and some concerns), and randomization process (mainly low risk, but some concerns). The bias domain is seen on the y-axis, and the score out of 100 is illustrated on the x-axis.

4.5 Discussion

This meta-analysis examined whether retrieval practice can make memory less sensitive to the detrimental effects of stress. We probed the literature for studies examining

memory performance in participants who underwent stress exposure versus a control group after learning with retrieval practice. We found a positive overall effect for retrieval practice versus a control strategy (i.e., restudy) on memory performance in stress conditions, $k = 9$ with 21 effect sizes, $g = 0.45$, CI [0.19, 0.71]. These results support our primary hypothesis (H3) and suggest retrieval practice is more effective than restudy for memory during stressful situations. However, meta-regression analysis revealed a nonsignificant impact of the stress induction on memory performance. We did not detect significant publication bias in this sample nor any significant moderators.

Retrieval practice main effects

The current work found retrieval practice significantly improves memory performance compared to other strategies under stress conditions, supporting our H3. This aligns with previous findings that retrieval practice is effective in the context of stressors (A. M. Smith et al., 2016) and with studies showing its benefit for other stressors like test anxiety (Yang, Shanks, et al., 2023). This also aligns with the context-dependent explanation of stress, where the repeated binding of contextual cues to the memory trace strengthens the memory and makes it easier to access, thereby potentially making it less vulnerable to the detrimental effects of stress (Karpicke, 2017; Karpicke et al., 2014).

We then explored the benefit of retrieval practice by comparing its effects on memory in stress versus non-stress conditions (H4). Results showed null findings of $g =$ 0.08 CI [-0.02, 0.19], indicating no statistically significant effect of retrieval practice between stress and non-stress conditions. These findings align to some extent with those of A. M. Smith et al. (2016) and Szőllősi et al. (2017), who observed similar memory performance levels between stressed and non-stressed participants who used retrieval practice. It is also possible that here, the repeated binding of contextual cues during retrieval contributes to strengthening the memory trace and making it robust during stress situations (Lehman et al., 2014), but further research is needed to substantiate this explanation.

The evidence for our other hypotheses was inconclusive. First, we found null effect when only restudy was used in a stress versus non-stress setting, g = -0.05, CI [-0.15, 0.05], failing to support H1. This means there was no significant difference in memory performance between stress and non-stress conditions when participants used restudy. This does not align with other research showing the detrimental effects of retrieval stress on memory when no specific strategies are used (Schwabe & Wolf, 2010a; Shields, Sazma, et al., 2017). We also found a positive effect for retrieval practice versus restudy in non-stress conditions, g = 0.37, CI [0.09, 0.66], consistent with the robust testing effect (Agarwal et al., 2021; Rowland, 2014; Schwieren et al., 2017; Yang et al., 2021).

Together, these results suggest that retrieval practice is more effective when compared to control strategies in both non-stress (H2) and stress conditions (H3). However, there is no evidence to suggest that stress impacts performance when only retrieval practice (H4) or control strategies (H1) are used. Thus, our findings seem to suggest that retrieval practice may offer a stable benefit compared to control strategies regardless of stress, rather than specifically protecting against the stress. This is further supported by our meta-regression analyses which did not reveal a significant effect of stress scores on memory performance. We note that these findings do not confirm the absence of an effect for stress, but rather that no detectable effects were seen here. We would further like to note that the meta-regression results should also be approached with caution due to the variability of subjective stress scores measured in each study and the low number of studies in the analysis.

Publication bias

In terms of publication bias, the current work showed mixed results, consistent with the publication bias results of previous reviews on retrieval practice (e.g., Rowland, 2014; Yang et al., 2021). For H3, there was some evidence of publication bias as shown by Egger's Test and Kendall's *tau*, but the Trim-Fill did not find missing studies and the selection model did not confirm selective reporting. Although PET-PEESE were not

significant, their corrected estimates showed the effect may be weaker than initially observed. For H2, there was little to no evidence of publication bias across all measures. This means the benefits of retrieval practice compared to restudy in stress and non-stress conditions hold up and are not likely influenced by publication bias.

For H1, both Egger's Test and Kendall's *tau* revealed significant funnel plot asymmetry, and the Trim-Fill imputed seven "missing" studies. Moreover, both PET and PEESE yielded significant positive intercepts, indicating that after correcting for SSEs, the true effects for H1 might be positive, meaning publication bias may have obscured true effects here. H4 showed similar results, with five imputed studies and the PET-PEESE being marginally significant $p = 0.05$, and suggesting positive corrected effects. However, due to the null findings for both H1 and H4, the presence of bias does not necessarily indicate hidden effects but rather suggests the effects, if any, are negligible and spotlights the need for cautious interpretation. In these cases, it is important to note the limitations of these measures. For instance, PET-PEESE and the Trim-Fill methods are not robust when heterogeneity is high (H2 and H3) and when the power is low (all Hs) (Harrer et al., 2021).

Publication bias can also be explained by other factors such as heterogeneity and risk of bias (Harrer et al., 2021). Indeed, this review observed high levels of heterogeneity indexed by a significant Cochrane's Q for H2 and H3. Additionally, heterogeneity remained high during sensitivity analysis using the leave-one-out analysis, meaning that the effect observed is not driven by any single study but is consistent across the included studies. All studies in our sample included some levels of bias as indexed by our risk of bias assessment. Specifically, issues with randomization, awareness of group assignment, missing outcome data, and lack of analysis protocols may introduce biases that could overestimate the positive effects of retrieval practice on memory performance under stress found in the individual studies. Finally, we note that most studies contained smaller sample sizes than recommended for their designs, making them vulnerable to spurious results (Brysbaert,

2019). All these methods together confirm that bias is present, potentially inflating overall effects, and reinforce the need to examine bias from multiple angles for robustness.

Moderator analyses

Moderator analyses were also a mixed bag. No significant moderators were found for H3, suggesting that the positive effects of retrieval practice versus restudy are consistent and apparently unaffected by various study characteristics and contexts. This aligns with previous studies showing the effectiveness of retrieval practice regardless of various factors (Adesope et al., 2017). In H4, 2-day delays showed stronger effects than 1-week delays, suggesting a benefit for shorter retention intervals when retrieval practice is used in stress compared to non-stress conditions. This is partly in line with previous research where the benefits of retrieval practice become stronger in the next day or so (e.g., Karpicke & Roediger, 2007), but not with the classic retrieval practice paradigm where its effects are strongest after one week (Roediger & Karpicke, 2006). This pattern was reversed in H2, which showed a significant difference for 1-week delays compared to 1-day, consistent with the classic testing effect, so this pattern warrants further investigation during stressful situations. In both H1 and H4, effects for facts or short answer learning material were stronger compared to educational texts (H1) and word lists (H4). This suggests restudy seems to benefit shorter learning material more than more complex texts, and that retrieval practice may specifically benefit facts or short answer material over simpler material like words in stress versus non-stress conditions. These results need further investigation in stress settings, but they generally align to the literature showing retrieval practice benefits this type of material (Agarwal et al., 2021; Moreira et al., 2019; Yang et al., 2021).

In line with our meta-regression findings, we did not find a significant moderating effect of type of stressor on memory performance across all Hs. This is surprising given the large body of literature showing that stress induced through TSST procedures impacts memory (Schwabe et al., 2008; Shields, Sazma, et al., 2017), but is partially in line with research suggesting test anxiety type stressors do not always impact memory performance

with retrieval practice (Clark et al., 2018; Hinze & Rapp, 2014; Yang et al., 2020). One possible explanation is that the benefits of retrieval practice might be robust enough to negate the effects of the stressor altogether. This has already been shown in studies where retrieval practice successfully decreased test anxiety stressors (Piroozmanesh & Imanipour, 2018; Szpunar et al., 2013) but requires further investigation using different types of stressors.

Implications

The results found here suggest retrieval practice provides a stable benefit compared to restudy in both stress and non-stress situations. Having a noninvasive strategy to mitigate the potential detrimental effects of stress on memory is a promising step in the memory and learning field. In educational settings, where students often experience stress during assessments (Vogel & Schwabe, 2016), incorporating retrieval practice into study routines could help alleviate this stress and improve performance through individuals' own efforts. For instance, Yang and colleagues (2023) found that testing can reduce test anxiety, highlighting its value in real-world scenarios. In high-stress professions, such as healthcare or emergency services, using retrieval practice to remember medical symptoms or protocols could enhance memory retention and accuracy, potentially leading to better decision-making and outcomes under pressure. In daily life, individuals facing stress-related memory challenges—whether due to work demands or caregiving responsibilities—could benefit from retrieval-based strategies to enhance memory resilience for daily life tasks like remembering a grocery list.

At the same time, high heterogeneity in the overall results suggests that methodological variations across the included studies, such as differences in retrieval practice implementation, could have played a role. For instance, some studies allowed only a few seconds for encoding or recall (A. M. Smith et al., 2019; Tse et al., 2019), while others gave participants more extensive study and retrieval opportunities (Szőllősi et al., 2017). These inconsistencies could have obscured true effects, and point to the need for more

standardized and well-controlled implementations of retrieval practice in future research.

Limitations and future directions

Restudy was the only control learning strategy utilized across all the papers, highlighting a limitation in the field and in our results. It would be informative to examine how other "strong" learning strategies fare in comparison to retrieval practice, for instance, feedback or spaced learning (Reber & Rothen, 2018) to determine if they could also have potentially protective effects on memory during stress. Most studies also exclusively utilized word lists for learning material, which did not show consistently strong effects in our moderator analyses. This opens questions on why and how different types of learning material, such as more complex educational material, might moderate learning with retrieval practice in stressful settings. Previous work has found differing effects for test format on memory (i.e., cued recall vs. recognition tests) during stress (Shields, Sazma, et al., 2017), opening doors for future studies to further investigate how different types of memory assessments interact with stress to influence memory performance. Although included studies reported that stress induction was successful, we did not find stress induction to contribute meaningfully to differences in memory performance across the studies, at least at the subjective level. Future studies should further investigate other factors that may influence memory performance under stress, such as the type and duration of stress, individual differences in stress response, and the context in which stress and retrieval practice are applied.

In addition, the general lack of power in individual studies suggests that their findings may be less reliable and could contribute to variability in the overall results. Post-hoc power analysis revealed no hypothesis was sufficiently powered on its own and our meta-analysis was not powered enough to detect effects smaller than 0.3. This means that meta-analytic estimates are more sensitive to the inclusion or exclusion of individual studies, as studies have more of an impact since they are so few. It also highlights the need for caution when interpreting the results, as limited statistical power increases the risk of

both false positives and false negatives. This is especially dangerous for budding fields such as the current, as early, potentially inflated, effects could disappear as better powered studies get conducted in the future. Thus, current and future studies must aim for adequately powered studies by following recommendations from the field (e.g., Brysbaert, 2019) to enhance the reliability and value of future meta-analytic endeavors.

4.6 Conclusion

This meta-analysis explored whether retrieval practice can make memory less sensitive to the detrimental effects of stress. We found a medium effect size for retrieval practice versus restudy in stress conditions, $g = 0.45$, CI [0.19, 0.71], and a similar effect size for retrieval practice versus restudy in non-stress conditions, $g = 0.37$, CI [0.09, 0.66], suggesting retrieval practice is more beneficial than restudy in both stress and non-stress conditions. Null findings for both restudy and retrieval practice alone in stress versus non-stress conditions suggest that neither strategy was particularly vulnerable to the effects of stress when used alone. Moderator analyses across the hypotheses suggest factors like delay and type of learning task may influence outcomes but not across all situations. Trends in publication bias analyses were mixed but suggested possible overestimation of effects. Additionally, the inadequately powered studies in the meta-analysis as a whole suggest the observed effects should be approached with caution. Although retrieval practice is more effective than restudy in both stress and non-stress conditions, the lack of impact of the stressor on memory performance makes any takeaways preliminary and highlights that this is an emerging field requiring more sufficiently powered studies to clarify the conditions under which retrieval practice is or isn't protective.

4.7 Supplementary

In this section, we present sensitivity analyses for our main effects and moderator analyses. Our included studies are shown in Table 6, forest plots for H1, H2 and H4 in Figure 18, publication bias summary for H1, H2 and H4 in Table 7, and moderator results for H1, H2 and H4 in Tables 8-10.

4.7.1 Sensitivity Analysis: Main Effects

We conducted three-level meta-analytic models for all hypotheses to account for the hierarchical structure of the data (i.e., multiple effect sizes nested within experiments and studies) and followed this up with robust variance estimation (RVE). RVE was conducted with the 'clubSandwich' R package (Pustejovsky & Tipton, 2018) to address potential dependencies among effect sizes and provide small-sample corrections. All sensitivity analyses showed the same pattern of results as the main effects two-level models as did RVE measures.

H1. A three-level model was conducted to examine the overall effect, accounting for dependencies among effect sizes nested within experiments and studies ($k = 21$ effects). The overall effect was not statistically significant, $g = –0.05$, $p = 0.35$, CI [–0.15, 0.05]. The test for heterogeneity was also non-significant, $Q(20) = 15.99$, $p = 0.72$, indicating no substantial variability in effect sizes beyond sampling error.

The robust variance estimator model showed no significant effect for restudy (intercept = -0.05, $SE = 0.05$, $t(19.1) = -1.04$, $p = 0.31$), suggesting no impact on memory performance.

H2. A three-level model was conducted to account for dependencies among effect sizes nested within experiments and studies ($k = 21$ effects). The overall effect was statistically significant, $g = 0.37$, $p = 0.01$, CI [0.09, 0.66]. Heterogeneity was significant, $Q(20) = 115.38$, $p < 0.001$, suggesting substantial variability in effect sizes beyond sampling error.

The robust variance estimator model for H2 indicated a significant effect for retrieval

practice (intercept = 0.37, $SE = 0.14$, $t(20) = 2.57$, $p = 0.02$) on memory performance.

H3. A three-level model was conducted for H3 to account for dependencies among effect sizes nested within experiments and studies ($k = 21$ effects). The overall effect was significant, $g = 0.45$, $p < 0.01$, CI [0.19, 0.71]. The test for heterogeneity was significant, $Q(20) = 109.20$, $p < 0.001$, indicating substantial variability across effect sizes beyond chance.

The robust variance estimator model showed a significant positive effect of retrieval practice (intercept = 0.449, $SE = 0.13$, $t(20) = 3.38$, $p = 0.003$) on memory performance.

H4. A three-level model was conducted for H4 with 23 effect sizes to account for dependencies among multiple outcomes nested within experiments and studies. The overall effect was non-significant, $g = 0.08$, $p = 0.11$, CI [-0.02, 0.18]. Although the point estimate was positive, the confidence interval included zero, indicating no clear evidence for an effect. The test for heterogeneity was not significant, $Q(22) = 23.64$, $p = 0.37$, suggesting low variability in effect sizes.

The robust variance estimator model showed no significant effect (intercept = 0.08, $SE = 0.05$, $t(21.3) = 1.62$, $p = 0.12$), indicating that stress did not significantly impact memory performance under retrieval practice in this analysis.

4.7.2 Moderator Analyses

We conducted moderator analysis for H1, H2 and H4 in order to compare the effects of the different moderating variables for each H. These results are presented in Tables 8-10.

We used a fixed-effects contrast and MetaForest to test if there is a meaningful moderating effect between moderators. We found significant effect for short answer/facts versus texts learning material in H1, as well as for short answer/facts versus word lists in H4. We also found a significant moderating effect for a delay of 1 week versus 1 day in H2 and 2 days versus 1 week in H4.

Following suggestions received at Stage 2, we also ran Delay as a continuous moderator. H2 showed a similar pattern of results as the categorical moderator and as H3.

H1 and H4 did not show significant differences.

In H1, there was no significant effect of delay ($QM = 0.52$, $p = 0.47$), indicating that length of delay did not predict differences in effect sizes, with minimal residual heterogeneity ($I^2 = 1.02$).

For H2, there was a significant effect of delay ($QM = 10.32$, $p = 0.001$), indicating that memory performance increased with longer delays. Residual heterogeniety remained high ($I^2 = 79\%$).

For H4, there was no significant effect of delay ($QM = 1.89$, $p = 0.17$), meaning length of delay did not impact memory performance. Heterogeniety was minimal ($I^2 = 1.04$).

4.7.3 Supplementary Figures

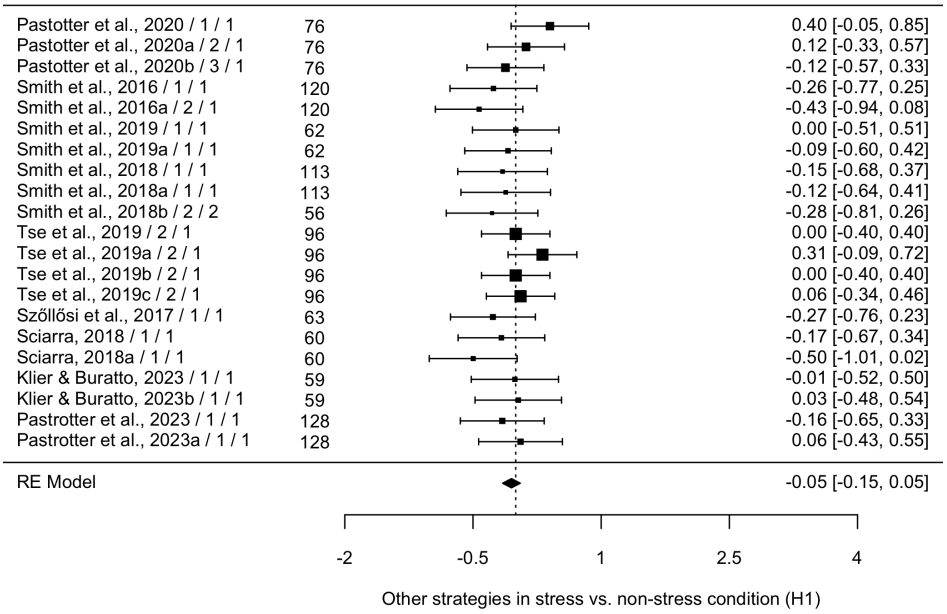
Table 6

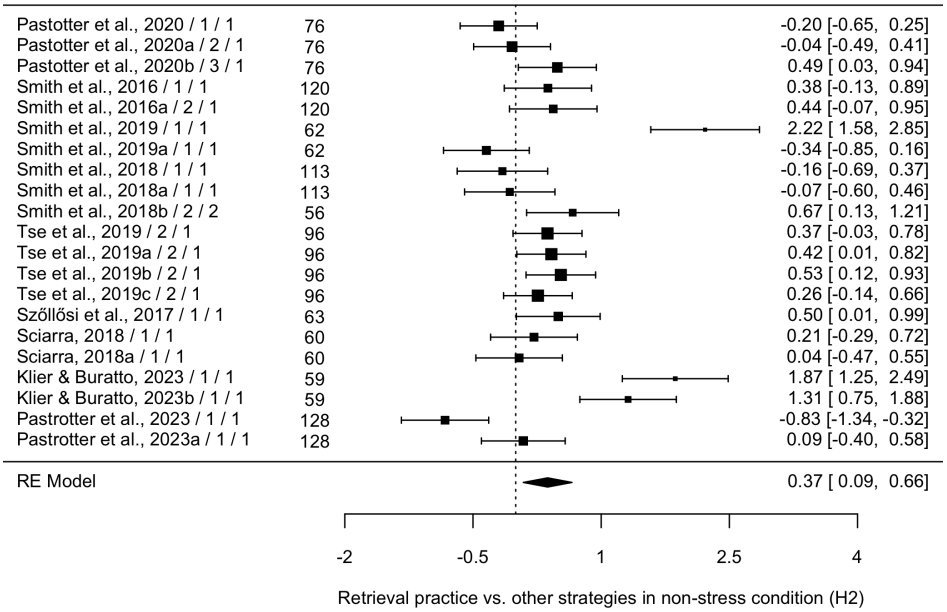
List of included studies

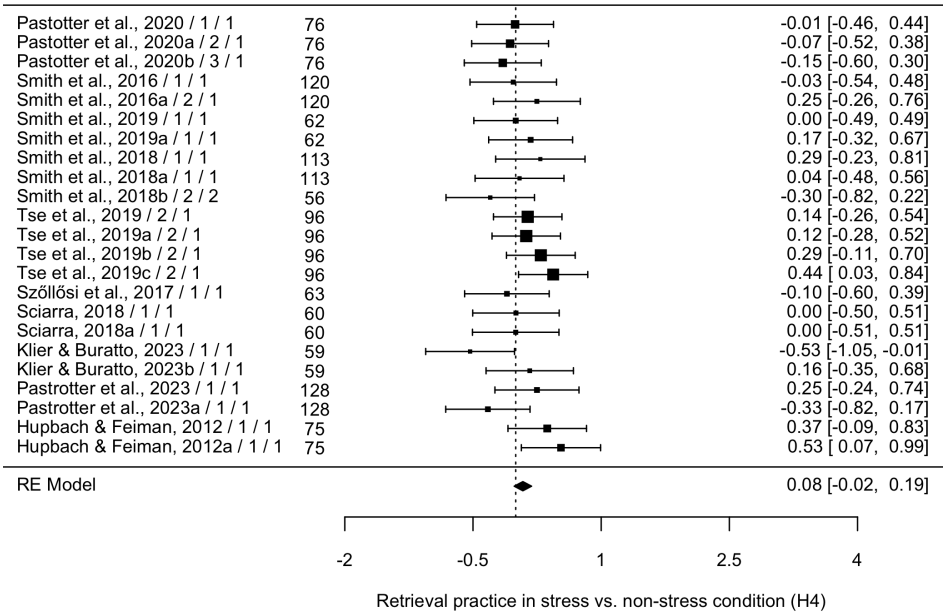
#	Study	N	Country	Sample	Design	Published
1	Pastotter et al., 2020	76	Germany	Students	Mixed	Yes
2	Smith et al., 2016	120	USA	Students	Between	Yes
3	Smith et al., 2019	62	USA	Students	Mixed	Yes
4	Smith et al., 2018	113	USA	Students	Between	Yes
5	Tse et al., 2019	96	China	Students	Mixed	Yes
6	Szőllősi et al., 2017	63	Hungary	Students	Mixed	Yes
7	Sciarra, 2018	60	Canada	Students	Between	No
8	Klier & Buratto, 2023	59	Brazil	Students	Mixed	Yes
9	Pastotter et al., 2023	128	Germany	Students	Between	Yes
10	Hupbach & Feiman, 2012	75	USA	Students	Between	Yes

Figure 18

Forest plots for H1, H2, H4







Note. Forest plots for H1, H2, and H4 (with their coded study and experiment number) with their respective sample size, effect size (represented by the squares), weight, and 95% confidence intervals. Larger squares indicate more weight in the random effects model. The x-axis represents the standardized mean difference effect size (Hedges' g), and the y-axis is each individual study. RE model at the bottom represents the overall effect size and its 95% confidence interval using the random-effects model. Article name (e.g., Pastötter 2020), followed by study number, experiment number within the study, and sample number (e.g., Pastötter 2020 / 1 / 2 / 1). Letters (e.g., "a", "b") indicate multiple studies within the same article.

Table 7

Publication bias summary for H1, H2, H4

Hypothesis	**Trim & Fill**	**Egger's Regression Test**	**Rank Correlation Test**	**Three Parameter Selection Model**	**PET-PEESE**
H1	7 missing studies Adjusted model: g = 0.04 [-0.05, 0.14]	Egger's Test of the Intercept: z = -2.37, p = 0.02	Kendall's *tau* = -0.49, p = 0.002	Likelihood ratio test = 0.68, p = 0.4 Adjusted Model: g = -0.04, 95% CI [-0.15, 0.07]	PET: b = 1.18, [0.37, 1.99], p = 0.005 PEESE: b=0.59, [0.17, 1.06], p = 0.004
H2	0 missing studies Adjusted model: g = 0.37 [0.08, 0.66]	Egger's Test of the Intercept: z = 2.65, p = 0.008	Kendall's *tau* = 0.18, p = 0.27	Likelihood ratio test = 0.49, p = 0.5 Adjusted Model: g = 0.51, 95% CI [0.02, 0.99]	PET: b = -1.46, [-3.49, 0.57], p = 0.08 PEESE: b = -0.68, [-1.68, 0.32], p = 0.04
H4	5 missing studies, Adjusted model: g = 0.17, [0.06, 0.28]	Egger's Test of the Intercept: z = -1.97, p = 0.05	Kendall's tau = -0.15, p = 0.34	Likelihood ratio test = 0.57, p = 0.5 Adjusted Model: g = 0.06, 95% CI [-0.05, 0.18]	PET: b = 1.15, [0.05, 2.24], p = 0.05 PEESE: b = 0.63, [0.06, 1.21], p = 0.05

Table 8

Moderator analysis for H1

Moderator	*k*	*Q*	*df*	*g*	95% CI	*Tau2*	I^2	Diff	*p*	*Categories*
					Stressor Type					
TSST	17	11	16	-0.10	-0.22, 0.02	0.00	0.00%	3.40	0.06	TSST vs. TA stressor
TA	4	1	3	0.09	-0.10, 0.29	0.00	0.00%			
					Delay					
Immediate	7	3	6	0.04	-0.12, 0.22	0.00	0.00%	3.29	0.07	Immediate vs.1 day
1 day	4	0.85	3	-0.24	-0.50, -0.02	0.00	0.00%	0.70	0.40	1 day vs. 1 week
1 week	6	1	4	-0.10	-0.31, 0.11	0.00	0.00%	0.08	0.78	1 week vs. 2 days
2 days	4	6	3	-0.04	-0.38, 0.29	0.06	53.75%			
					Type of Task					
Words/word lists	15	9	14	-0.07	-0.20, 0.05	0.00	0.00%	1.83	0.18	Words vs. Facts/short answer
Facts/short answer	4	1	3	0.09	-0.11, 0.29	0.00	0.00%	4.80	0.04*	Short answer vs. Texts
Texts	2	0.80	1	-0.33	-0.69, 0.03	0.00	0.00%			

Note. k = number of studies (i.e., number of effects); g = Hedge's g effect size, CI = lower and upper limits of 95% confidence interval, tau2 = tau squared value, $I2$ = I-squared value, * $p <$ .05, ** $p < .01$, *** $p <$.001, (all two-tailed); $Q = Q$ statistic indicating heterogeneity (decimals only reported if $Q < 1$), $Diff = QM$ statistic, test of moderator effects showing any differences between the moderators.

Table 9

Moderator analysis for H2

Moderator	*k*	*Q*	*df*	*g*	95% CI	*Tau2*	I^2	Diff	*p*	*Categories*
					Stressor Type					
TSST	17	113	16	0.37	0.01, 0.73	0.51	88.10%	0.01	0.92	TSST vs. TA stressor
TA	4	0.87	3	0.39	0.19, 0.59	0.00	0.00%			
					Delay					
Immediate	7	23	6	0.07	-0.27, 0.41	0.16	75.59%	0.13	0.72	Immediate vs. 1 day
1 day	4	3	3	0.15	-0.14, 0.45	0.02	23.43%	4.37	0.04*	1 days vs. 1 week
1 week	6	54	5	1.02	0.26, 1.78	0.81	91.10%	3.62	0.06	1 week vs. 2 days
2 days	4	1	3	0.25	0.03, 0.48	0.00	0.00%			
					Type of Task					
Words/word lists	15	112	14	0.41	-0.002, 0.82	0.58	89.53%	0.004	0.95	Words vs. Facts/short answer
Facts/short answer	4	0.87	3	0.39	0.19, 0.59	0.00	0.00%	1.16	0.20	Short answer vs. Texts
Texts	2	0.22	1	0.13	-0.23, 0.48	0.00	0.00%			

Note. k = number of studies (i.e., number of effects); g = Hedge's g effect size, CI = lower and upper limits of 95% confidence interval, tau2 = tau squared value, $I2$ = I-squared value, * $p <$ .05, ** $p < .01$, *** $p <$.001, (all two-tailed); $Q = Q$ statistic indicating heterogeneity (decimals only reported if $Q < 1$), $Diff = QM$ statistic, test of moderator effects showing any differences between the moderators.

Table 10

Moderator analysis for H4

Moderator	*k*	*Q*	*df*	*g*	95% CI	*Tau2*	I^2	Diff	*p*	*Categories*
					Stressor Type					
TSST	19	18	18	0.03	-0.08, 0.15	0.003	4.57%	3.27	0.07	TSST vs. TA stressor
TA	4	1	3	0.25	0.05, 0.45	0.00	0.00%			
					Delay					
Immediate	7	5	6	0.03	-0.13, -0.20	0.00	0.00%	1.19	0.27	Immediate vs. 1 day
1 day	5	1	4	0.19	-0.03, 0.42	0.00	0.00%	3.19	0.07	1 day vs. 1 week
1 week	6	5	5	-0.09	-0.31, 0.12	0.01	8.87%	4.60	0.03*	1 week vs. 2 days
2 days	5	4	4	0.24	0.03, 0.45	0.00	8.12%			
					Type of Task					
Words/word lists	15	11	14	-0.02	-0.15, 0.10	0.00	0.00%	5.03	0.02*	Words vs. Facts/short answer
Facts/short answer	4	1	3	0.25	0.05, 0.45	0.00	0.00%	<0.01	0.97	Short answer vs. Texts
Texts	4	3	3	0.24	-0.02, 0.50	0.01	14.32%			

Note. k = number of studies (i.e., number of effects); g = Hedge's g effect size, CI = lower and upper limits of 95% confidence interval, tau2 = tau squared value, $I2$ = I-squared value, * $p <$ .05, ** $p <$ .01, *** $p <$.001, (all two-tailed); $Q = Q$ statistic indicating heterogeneity (decimals only reported if $Q < 1$), $Diff = QM$ statistic, test of moderator effects showing any differences between the moderators.

5. Study 3: Does retrieval practice protect memory against the detrimental effects of test anxiety?

5.1 Abstract

Test anxiety is characterized by acute stress and worry in evaluative situations and leads to memory retrieval deficits. Retrieval practice is a learning strategy in which one actively recalls information from memory and is consistently associated with memory enhancements. Recent research suggests retrieval practice can reduce the negative effects of stress on memory. Currently, no studies have investigated whether retrieval practice can protect memory against test anxiety induction. The present study aims to investigate whether learning with retrieval practice can overcome the detrimental effects of test anxiety on memory. Participants learned two texts using a retrieval practice strategy for one text and a restudy learning strategy for the second text in an online setting. One week later, participants performed the recall session where they were instructed to recall as many items as possible from both texts. Prior to memory recall, participants were randomly placed in a test-like condition via evaluative, test anxiety-inducing instructions (evaluative group) or a neutral testing condition via control instructions (control group). Retrieval practice improved memory performance across both evaluative and control conditions and remained effective regardless of participants' anxiety levels. These findings support the robustness of the testing effect in online settings and highlight retrieval practice as a stable learning strategy across varying psychological contexts. Implications for research-based learning interventions in educational settings are discussed.

5.2 Introduction

"I totally blanked during the test!" This is something we have all experienced at one point or another. Test anxiety is a type of stressor that occurs during evaluative (i.e., testing) situations and leads to memory retrieval deficits (Cassady & Johnson, 2002; Hembree, 1988; Vogel & Schwabe, 2016). While recent models suggest that these deficits

could be due to the neurobiological stress response (Vogel & Schwabe, 2016), that may not be the only explanation. For instance, earlier work has suggested that test anxiety-related memory performance deficits might be attributed to poor study skills (Naveh-Benjamin et al., 1981; Wittmaier, 1972). Evidence suggests that retrieval practice, one of the most powerful learning strategies where learning material is retrieved from memory instead of just re-studied (Roediger & Karpicke, 2006), protects memory from the detrimental effects of stress (A. M. Smith et al., 2016) and is robust against test anxiety more generally (Yang et al., 2020). However, up until now, the protective benefits of learning with retrieval practice against stressors like test anxiety have not been examined in an experimental setting. The current study endeavored to fill this gap by investigating whether learning with retrieval practice can protect memory from the negative effects of test anxiety on memory.

In educational settings, test anxiety is referred to as "academic stress" and occurs during evaluative, or testing, situations (Cassady & Johnson, 2002). Test anxiety encompasses a wide range of physiological and psychological symptoms, such as worry, disruptive thoughts, racing heart, sweating, and rapid breathing, among others (Cassady & Johnson, 2002; Hembree, 1988). Test anxiety has also been linked with a plethora of unfavorable outcomes, including poor exam performance, decreased academic achievement, poor memory and learning abilities, and increased vulnerability to other types of anxiety disorders and stress (von der Embse et al., 2018). It is estimated that between 15-22% of students experience high levels of test anxiety (Thomas et al., 2018), and between 10% and 40% of all students experience some level of test anxiety that can surface as early as age seven (von der Embse et al., 2013). Recent estimates suggest that this number can go up to 75% before an actual exam takes place (Thiriveedhi et al., 2023).

Our theoretical understanding of the psychological factors of test anxiety has evolved over the last few decades. Early on, test anxiety was conceptualized through a cognitive interference model, in which high degrees of emotionality and worry were thought

to disrupt information recall (Liebert & Morris, 1967; Wine, 1971). Then, theorists, such as Wittmaier (1972) and Culler and Holahan (1980), proposed a skills-deficits model. This model supposes that students suffer from test anxiety because they do not employ effective study habits during learning. At the same time, the information processing model was put forth, positing that students have difficulties both in learning and organizing material (i.e., encoding) and retrieving it during test (Naveh-Benjamin et al., 1981).

More recent models have incorporated findings from the neurobiology of stress to conceptualize test anxiety as a physiological stressor occurring in educational settings (Vogel & Schwabe, 2016). When various classroom situations, such as difficult exams or academic pressures, are perceived as threatening, they disrupt our body's homeostasis, leading us to feel stress (McEwen & Gianaros, 2011; Vogel & Schwabe, 2016). Stress triggers a well-described neurobiological cascade in the body, comprising the fast-acting autonomous nervous system (ANS) and the slower hypothalamus-pituitary-adrenal (HPA) axis (Shields, Sazma, et al., 2017). These systems work together to produce changes in the body and brain, which ultimately alter memory processes. Although the effects of stress on memory vary depending on the timing of the stressor and the memory process investigated, a wealth of evidence agrees that stress experienced prior to retrieval, called retrieval stress, decreases memory (Gagnon et al., 2019a; Kuhlmann, Kirschbaum, & Wolf, 2005; Shields, Sazma, et al., 2017; Vogel & Schwabe, 2016). Indeed, studies have found that increased physiological stress reactions are associated with decreased exam performance (M. Cohen & Khalaila, 2014; Schoofs et al., 2008).

Moreover, stress, including that experienced in relation to exams, has been hypothesized to induce a shift in context-dependent memory processes (Shields, Sazma, et al., 2017; Vogel & Schwabe, 2016). Context-dependent memory is the well-replicated finding that items are better retrieved if the circumstances in which they were initially learned are reinstated (S. M. Smith & Vela, 2001). These circumstances can be manipulated externally to induce a disruption in context-dependent memory processes.

Such external changes can include, for instance, when students go from being calm or neutral to highly stressed because of an exam (A. M. Smith et al., 2019).

Context-dependent shifts have been shown to impair memory (Schwabe & Wolf, 2009; Shields, Sazma, et al., 2017). For example, Schwabe and Wolf (2009) exposed participants to a stress induction or a control task prior to learning in a room scented with vanilla. Participants then completed a memory test in either the same context (the room scented with vanilla) or a different context (a different room with no vanilla scent). Results revealed that stress impaired memory when it was learned in the different context, but it did not impair memory in the same context. These findings suggest that stress creates a context mismatch strong enough to impair memory when it is accompanied by an external manipulation. Importantly, providing enough contextual support during retrieval (i.e., the room scented with vanilla) alleviated this issue, suggesting that this impairment can be overcome by providing enough contextual support.

Taken together, the evidence suggests that the psychological and physiological mechanisms of stress in combination with poor learning strategies result in poor memory performance. These factors together might further impair context-dependent processes, which can be subverted if enough contextual information is provided. One potential way to overcome the underlying issues of test anxiety is through strategies that address both the poor learning strategies aspect and the stress aspect. For example, if test anxiety disrupts study habits during learning, as proposed by the skills-deficits model and information-process models, applying effective learning strategies could counteract the detrimental effects of test anxiety on memory.

Retrieval practice is a learning strategy in which information is actively recalled from memory, leading to the well-replicated memory-enhancement phenomenon known as the testing effect (Roediger & Karpicke, 2006). Retrieval practice is robust across a wide range of learning materials, school levels, and even in individuals with learning disabilities (Rowland, 2014; Sumowski et al., 2010a). Retrieval practice is also more effective than

commonly used learning strategies, such as re-reading, highlighting, and note-taking (Moreira et al., 2019). Importantly, when these commonly used learning strategies are employed by individuals with test anxiety as a way to alleviate the stressful symptoms of test anxiety and improve performance (Lent & Russell, 1978), they fail to do either (Huntley et al., 2019).

Retrieval practice is thought to increase memory for learned items during encoding through contextual reinstatement mechanisms, which strengthens memory traces continually at each retrieval (Karpicke, 2017; Lehman et al., 2014). Similarly, retrieval practice may enhance memory because it creates a match between the conditions of retrieval and the conditions of the eventual test (Morris et al., 1977; Veltre et al., 2015). Because retrieval practice might create equivalent conditions between learning and recall, and may further increase contextual cues available to participants at test, it might help avoid a potential stress-induced shift in context-dependent memory. This would mean that memory does not get impaired due to stress. Additionally, retrieval practice is shown to improve memory performance in the classroom, where test anxiety is most likely to affect learners, to a medium extent (Yang, Li, et al., 2023), indicating that there is good reason to believe it protects memory against the detrimental effects of test anxiety on memory.

This protective factor of retrieval practice against the detrimental effects of acute stress on memory was exemplified in A. M. Smith et al. (2016). In their study, participants learned word lists with both a retrieval practice and a restudy strategy. One day later, one group of participants underwent a laboratory-based stress induction procedure prior to memory retrieval, and another group underwent a non-stressful control procedure prior to retrieval. Results showed that despite being stressed, which led to a decrease in memory retrieval in the control group, participants who learned with retrieval practice showed better memory than non-stressed participants who also learned with retrieval practice. In other words, retrieval practice protected memory against the negative effects of stress on memory.

However, other evidence on the protective benefit of retrieval practice against test anxiety is scarce and mixed. First, researchers surveyed 1408 middle and high school students about their reactions to a larger retrieval practice-based intervention that had been ongoing for several years (Agarwal et al., 2012). The surveyed students engaged in frequent, low or no-stakes retrieval practice through classroom-based tests and quizzes, primarily administered via clicker response systems with immediate feedback. Importantly, 72% of students reported feeling less test anxiety following the use of retrieval practice in the classroom, suggesting that using this strategy makes students feel less nervous about upcoming exams (Agarwal, D'Antonio, Roediger III, et al., 2014). At the same time, correlational evidence suggests a weak relationship between using retrieval practice and self-reported test anxiety (Tse & Pu, 2012; Yang et al., 2020). Taken together, these studies suggest that while retrieval practice may reduce feelings of test anxiety, it remains unclear whether this translates into better memory performance under test-related stress.

The goal of the current study was to fill this gap by investigating whether learning with retrieval practice can protect memory from the negative effects of test anxiety on memory. Test anxiety was induced in one group of participants via a set of evaluative instructions (evaluative group) and a second group served as the control with non-evaluative instructions. In Part 1, both groups learned with both retrieval practice and restudy prior to undergoing retrieval stress via the instructions in Part 2 one week later. State anxiety was measured throughout the study, as well as trait anxiety, general anxiety, and depression at the end.

In the present study, we formulated three hypotheses. First, we hypothesized that participants will demonstrate better memory performance for the learned text learned with retrieval practice compared to the one learned with restudy, replicating the established testing effect (Roediger & Karpicke, 2006). Second, since we induce test anxiety in the evaluative group through the instructions, we expected higher levels of state anxiety in the evaluative group relative to the control after reading the instructions. Third, because

evidence suggests that retrieval practice may reduce the detrimental effects of acute stress on memory (A. M. Smith et al., 2016), we expected memory performance to be higher in the evaluative group compared to the control group for the text learned with retrieval practice versus restudy. Finally, we explored whether the differential effects of retrieval practice on memory performance in evaluative situations are specifically related to trait-test anxiety or if they extend to general anxiety.

5.3 Method

This study was pre-registered on Open Science Framework (OSF link: https://osf.io/68aw5). All analysis scripts and data are available in the OSF repository (https://osf.io/aefy4/?view_only=).

5.3.1 Participants

A total of $N = 280$ university students participated in the study. French and German-speaking participants were recruited from the UniDistance Suisse online participant pool and the University of Geneva's bachelor's course Psychology of Cognitive Aging. Participants were healthy volunteers without severe depression and were remunerated with course credits. A total of 74 participants were excluded for the following reasons: having an incomplete dataset (e.g., abandoning the study in the middle of the experiment, $N = 49$); not returning for part 2 or completing the experiment on the wrong day ($N = 13$); having a learning, cognitive, or attention disorder ($N = 8$); having a time lapse of more than 30 minutes on either session during the study ($N = 2$); and not providing written responses during the recall task ($N = 2$). The current analyses were performed on $N = 206$ ($Mean$ age = 33.72, $SD = 13.83$, $N = 35$ male), with $N = 99$ participants randomly assigned to the evaluative condition and $N = 107$ participants randomly assigned to the control condition. Of these participants, $N = 167$ performed the study online (N evaluative condition = 79), and $N = 39$ performed the study in an in-person laboratory setting (N evaluative condition = 20). The experiment was performed in exactly the same way (i.e., performed online on the computer) for both the lab sample

and the online sample, except that the lab sample completed the task in a laboratory setting with the experimenter in the room. Ethics approval was obtained from both UniDistance Suisse (Approval #2021-12-00002) and the University of Geneva (Approval #2022-12-138).

5.3.2 Materials

Learning Materials. The learning materials consisted of two educational reading texts used in previous studies titled "The Sun" and "Sea Otters" (Emmerdinger & Kuhbandner, 2019; Roediger & Karpicke, 2006). The texts consisted of 249 and 267 words, respectively. The German versions of both texts were taken from Emmerdinger and Kuhbandner (2019). As no French translation of the texts is currently known to exist, the texts were translated into French by the authors of this paper. Each text contains 30 idea units according to the scoring rubric used in previous work (Roediger & Karpicke, 2006).

Raven's Progressive Matrices. Raven's Standard Progressive Matrices (Raven, 1941) and Raven's Advanced Progressive Matrices (Raven, 1965) were used as the matrix task for both groups. Both the standard and advanced matrices were used in order to ensure that the items were challenging enough.

Questionnaires. Test anxiety was measured using two different questionnaires: the Cognitive Test Anxiety Scale (CTAS, see Cassady & Johnson, 2002, for the English version and Stefan et al., 2020, for the German version) and the Test Anxiety Inventory (TAI, see Spielberger, 1980, for the original English version and Ringeisen et al., 2010, for the German version). The CTAS (Cassady & Johnson, 2002; Stefan et al., 2020) consists of 27 statements measuring test anxiety on a scale of 1 ("not at all like me") to 4 ("very typical of me"). The TAI (Ringeisen et al., 2010) consists of 21 items on a scale of 1 ("almost never") to 4 ("almost always"). Each questionnaire was scored according to its standard scoring criteria. Both test anxiety questionnaires were translated into French by the authors.

Both state and trait anxiety were measured using the State-Trait Anxiety Inventory (STAI, see Spielberger, 1970). We used the validated French version from Gauthier and

Bouchard (1993) and the validated German version from Grimm (2009). This questionnaire contains 40 items, with 20 items measuring state anxiety and 20 items measuring trait anxiety. The state anxiety questionnaire measures anxiety felt in the current moment on a scale of 1 ("almost never") to 8 ("almost always"), while the trait anxiety questionnaire measures anxiety felt in general on a scale of 1 ("almost never") to 8 ("almost always"). State anxiety was measured at six time points throughout both sessions, resulting in six different state anxiety scores. The trait anxiety items were measured only once at the end of the study, resulting in a single trait anxiety score. The items were scored by reverse-coding answers on certain questions and adding them together to obtain the total. Each questionnaire was scored according to its standard scoring criteria.

General anxiety was measured via the Generalized Anxiety Diagnostic (GAD, see Micoulaud-Franchi et al., 2016, for the French version and Spitzer et al., 2006, for the English and German versions). The GAD is a seven-item questionnaire measuring one's propensity to generalized anxiety disorder over the last two weeks on a scale of 0 ("not at all") to 3 ("nearly every day"). The GAD is scored by adding the total scores for the seven questions (Micoulaud-Franchi et al., 2016; Spitzer et al., 2006).

Depression was measured via the Beck Depression Inventory II (BDI, see Beck et al., 1996, for the English and French versions and Hautzinger et al., 2006 for the German version). The BDI is a 21-item questionnaire that measures one's depressive symptoms over the last two weeks. The BDI was scored based on standard scoring criteria. Participants with BDI scores representing severe depression (BDI $>$ 29) were excluded in order to avoid co-morbidities with other disorders (Beck et al., 1996; Hautzinger et al., 2006).

5.3.3 Design and Power Analysis

This study followed a 2 x 2 mixed factorial design. The within-subject factor was *Strategy* (retrieval practice vs. restudy), and the between-subject factor was *Instruction* (evaluative vs. control). The dependent variable was memory performance, measured by accurately recalling idea units from both educational reading texts. Test anxiety, trait

anxiety, and general anxiety were covariates.

A-priori power analysis and effect size estimation were based on the recommendations from the field (Brysbaert, 2019). To observe an interaction between two variables assuming an effect of $d = 0.4$ and an alpha level of 0.05 with 80% power, a minimum of 200 participants was required.

5.3.4 Procedure

The procedure was identical for the participants who completed the study online and those who were tested in the laboratory. Participants received the link to the online experiment and a personalized code that they had to type in on the first page to begin the experiment. At the beginning of Part 1, which was the learning session, the participants first provided their informed consent. They then responded to basic demographic questions (i.e., age, gender, and neurological disorders) and filled out the STAI.

During the learning session, participants read both texts for 7 minutes each, separated by a 1-minute distractor math task. After a longer distractor math task of 5 minutes, participants engaged in both restudy and retrieval practice learning strategies. This paradigm is similar to the one reported in Roediger and Karpicke (2006), except for the following differences, which were implemented due to the online nature of the current study. First, in order to tally the number of times each text was read, the participants clicked a button on the screen each time they finished reading the text during the initial learning phase and during the restudy learning strategy. Second, participants recorded their responses by typing them into a text box rather than writing them on paper. Third, the 7 minute timer was measured by a countdown clock on the screen rather than a physical timer. In the restudy strategy, participants simply re-read one of the texts for 7 minutes. In the retrieval practice strategy, they were instructed to recall as much information from the other text as they could remember for 7 minutes by typing the information into the text box on the screen.

One week after the learning session, participants returned for Part 2, the recall

session. Participants were given either an evaluative or a control set of instructions (available on OSF and in Supplementary). The participants in the evaluative group were told that the relationship between their intelligence and memory abilities would be assessed via an IQ test and a memory test. This instruction emphasized that the task would evaluate their personal abilities, which is a socio-evaluative manipulation known to lead to stress (Almazrouei et al., 2022). The word "test" was repeated throughout the study in order to convey a sense of pressure and high stakes, which are also known to increase stress (Hinze & Rapp, 2014). In the control group, the participants were told that they would perform a puzzle task and a memory task. They were also told to simply try their best and not worry if the task was difficult for them. After reading the critical instructions, the participants filled out the STAI to measure their in-the-moment anxiety levels. Then, they engaged in Raven's Standard Progressive Matrices (Raven, 1938, 1962) and Raven's Advanced Progressive Matrices (Raven, 1965) for 20 minutes.

After performing the matrix task, participants filled out the STAI again and completed the memory recall task. During memory recall, they were asked to type in as many ideas or phrases they could recall from the texts learned in the learning session during the 7 minutes for each text. Because the experimental manipulation hinged on the successful comprehension of the instruction sets, we implemented a sanity check at the beginning of the recall task to ensure that the participants had read the instructions carefully. Specifically, participants were asked to press "w" instead of the "Space" bar to continue through to the experiment, although they could still continue if they pressed the Space bar.

At the end of both the learning and recall tasks, participants were asked to measure their heart rates via a pulse-counting task performed by the participants themselves (Laskowski, 2018). They were instructed to extend their left arm, place the second and third fingers of their right hand on their left wrist, and to gently press on their left wrist below the thumb until they could feel their pulse. They counted the number of times they

felt their pulse during a 15-second countdown window. After the 15-second window, the participants typed in the number of times they felt their pulse while counting and responded to two follow-up questions. In the first question, they were asked to rate how certain they were of their response on a scale of 1 ("not at all") to 5 ("very certain"). In the second question, they rated how difficult the task was for them on a scale of 1 ("not hard at all") to 5 ("very hard").

The presentation order of the learning material and the assignment of texts to each learning strategy were counterbalanced across the participants. To ensure consistency, the text that was read first was kept constant for each participant across both learning and recall sessions.

5.3.5 Reliability and Coding

Reliability. Reliability measures were calculated for the responses to each of the questionnaires (CTAS, TAI, Trait Anxiety, and GAD). The R package corrplot (Wei et al., 2017) was used to calculate pairwise Pearson correlation coefficients between the questionnaire response scores. The coefficients were then visualized in a color-coded correlation matrix to examine the relationships between the questionnaire responses and to identify potential sources of redundancy or ambiguity.

Idea units coding. Participants' retrieval practice responses in the learning and recall sessions were coded according to the scoring rubric used in previous work (Roediger & Karpicke, 2006). Each text contained a maximum of 30 idea units, representing the amount of successfully recalled items from each text. In the recall session, each participant produced two retrieval practice responses, one for each text learned in Part 1. Each participant's retrieval practice responses were first translated into English using DeepL so the main investigator (MM) could understand them, and then they were coded by labeling the successfully recalled idea units. This coding process also enhanced the study's reproducibility by making it easier for other researchers to accurately compare and replicate results. The accurately recalled idea units were then added up and converted into

proportions to obtain an overall memory performance score.

To establish inter-rater reliability (IRR) of the idea units coding scheme, a second rater independently coded a subset of 40 participant responses using the same procedure. Both raters then discussed the rating criteria and underwent several rounds of deliberations to discuss any discrepancies in their ratings. IRR was calculated using Pearson correlation before and after the rater deliberations (Sedgwick, 2012).

5.3.6 Pre-registered Data Processing

In our pre-registration, we first planned to perform sanity checks on the data to ensure that we only included participants who followed the instructions. Because our experiment relied on careful reading of the instructions, we limited our initial analyses to participants who pressed the w key instead of the Space bar during Part 2 of the study. However, we later also included the responses from those who pressed Space (see "Deviations from pre-registration" section).

We then checked that participants' retrieval practice responses were adequate (i.e., were completed and not left blank). We also verified that they had performed Part 1 and Part 2 of the studies on the designated dates for which they were registered to participate, as Part 2 had to be completed exactly one week after Part 1. Those who completed the study +/- 1 day within the week were included (see "Deviations from pre-registration" section).

Next, we ensured that participants did not report engaging in behaviors that gave them an advantage over other participants in remembering the texts, like copying down the texts or taking photos of them during either part of the study. Additionally, we verified that the participants performed all parts of both sessions and that the data were complete and did not have any significant (>30 minutes) time lags at any point during the study.

5.3.7 Pre-registered Main Analyses

The full planned analyses for this study can be found on OSF (https://osf.io/68aw5). Deviations from our pre-registered analysis plan are reported

transparently in the results section.

Memory performance was assessed using a 2 (strategy: retrieval practice vs. restudy) x 2 (instruction: evaluative vs. control) mixed-design analysis of variance (ANOVA) to determine the effect of *Strategy* (within-subject factor) and *Instruction* (between-subject factor) on memory performance. We conducted post-hoc pairwise comparisons to examine significant interactions.

To investigate the role of test anxiety and generalized anxiety on memory performance, we added the test anxiety scores as the covariate in a 2 x 2 mixed model analysis of covariance (ANCOVA). To test the robustness of the results, we conducted a multiverse approach and used the scores from the CTAS in a first analysis and the TAI in a second analysis. This analysis was then repeated with the Trait Anxiety and GAD scores as covariates in the ANCOVA.

To confirm the effectiveness of the experimental manipulation, we investigated the effect of the evaluative and control instructions on state anxiety scores measured by the STAI. In the learning session, we performed a 2 (instruction: evaluative vs. control) x 2 (state anxiety time point measurement: before reading text 1 and after performing the second learning strategy) mixed-design ANOVA. In the recall session, we performed a 2 (instruction: evaluative vs. control) x 4 (state anxiety time point measurement: after instructions, after the matrix task, after memory recall task 1, and after memory recall task 2) mixed-design ANOVA. We also investigated how test anxiety was induced via our physiological proxy (i.e., the heart rate measurements) by performing a 2 (instruction: evaluative vs. control) x 2 (session: learning vs. recall) mixed-design ANOVA. Because heart rate cannot be assessed directly, we acknowledge that the results of all heart rate analyses should be interpreted with caution.

As a follow-up, we performed a sensitivity analysis of memory performance using an extreme groups approach to ensure that our results were not due to poor induction of the evaluative context. To do this, we repeated the memory performance ANOVA after

including the participants from the evaluative group who had a high score on the STAI (i.e., the upper quartile for that group) and the participants from the control group with a low score on the STAI (i.e., the lower quartile for that group).

We checked for and reported all major assumptions made in the ANOVAs (i.e., normality and homogeneity of variances and covariances). For all the analyses, we used the standard $p < 0.05$ criteria for determining significant differences. The effect sizes for the main effects and interactions in our statistical models are reported as partial eta-squared (η_p^2), with standard interpretations for small ($\eta_p^2 = 0.01$), medium ($\eta_p^2 = 0.06$), and large ($\eta_p^2 = 0.14$) effects. The effect sizes for the post-hoc t-tests are reported as Cohen's d, with the standard effect size interpretations for small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effects (J. Cohen, 1988).

Outliers in the memory performance and pulse-counting tasks were identified by establishing an initial cut-off of 3 standard deviations from the mean, followed by 2.5, 2, 1.5, and 1 standard deviations. To examine the robustness of the findings, the memory performance and heart rate analyses were rerun with each of the outlier cut-offs. For the pulse-counting task, any physically unrealistic values (i.e., BPM < 20 or > 400) were removed. See the pre-registration for the full list of exclusion criteria (https://osf.io/aefy4/?view_only=).

5.4 Deviations from Pre-Registration

We deviated from the pre-registration by incorporating an in-lab validation sample. Although we originally planned to collect all the data online, difficulties in recruiting online participants led us to incorporate additional in-lab data collection. This change in our methodology provided an opportunity to compare the results between the online and lab samples. This also led to the addition of a third factor for *Sample* (online vs. lab) in our initial ANOVAs and ANCOVAs.

Additionally, we initially planned to analyze the results from only the participants who pressed w in Part 2 of the study, which implied that the sample only included those

who read the instructions carefully. To avoid losing valid data, we compared the responses from the w group with those who pressed the Space bar instead of w in the recall session. The results revealed similar patterns in the responses from both the w and Space groups, suggesting that there were no differences whether participants pressed w or not (see Supplementary). Therefore, we applied a second deviation from the pre-registration by merging these participants into the full sample and analyzing them together.

We also initially planned to exclude participants who completed the study on the wrong date. To avoid losing more quality data, we applied a third deviation by extending this cut-off to include participants who completed the experiment +/- 1 day from the appointed dates after ensuring that their data met all other inclusion criteria.

Lastly, we examined the relationship between test anxiety and the testing effect. More specifically, we examined how the test anxiety scores were related to the testing effect across different groups and samples by conducting exploratory analyses using a linear regression model.

5.5 Results

The preparatory analyses, including reliability, coding and the number of times each text was read, are presented in Appendix I. In the results section, we present our findings with the inclusion of the third factor, *Sample* (online vs. lab). This factor was added to the study to examine the potential differences in the participants' responses between the online and lab setting. Additional results for the online versus lab sample are presented in Appendix II. The results from the full sample excluding the *Sample* factor are presented in Appendix III. Reference tables for the full ANOVA/ANCOVA main effects and interaction results are provided unless the results are summarized directly in the text.

The Shapiro-Wilk test indicated a significant departure from normality ($p > 0.05$). Given the potential limitations of this test (Field, 2013), skewness and kurtosis statistics were examined for both the full sample and the online versus lab sample. These statistics revealed a slight positive skewness (0.36) and peakedness similar to a normal distribution

(Kurtosis value 2.59). Subsequent visual inspection of QQ plots confirmed the approximate normality of the data. Further tests for normality, including Levene's Test (homogeneity of variances) and Box's M (homogeneity of covariances), supported the parametricity of the data without violating any assumptions ($p > 0.05$ for all analyses).

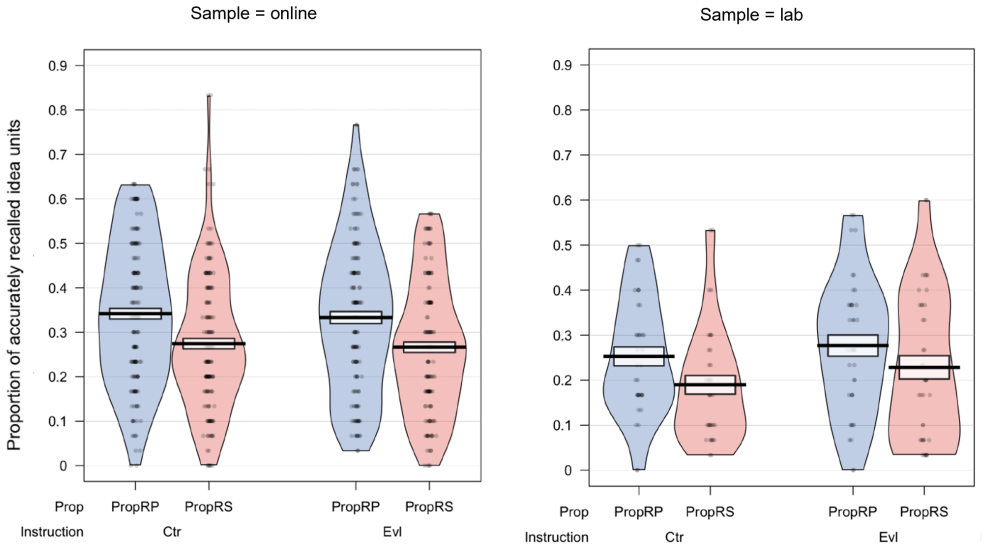
5.5.1 Memory Performance

Descriptive statistics of memory performance, presented as the proportion of accurately recalled idea units, are shown in Figure 19. To examine memory performance, we performed a 2 (strategy: retrieval practice vs. restudy, within-subject factor) x 2 (instruction: evaluative vs. control, between-subject factor) x 2 (sample: online vs. lab, between-subject factor) mixed-design ANOVA (Table 11). The results revealed a significant main effect of *Sample* ($F(1, 202) = 8.05$, $p = 0.005$, $\eta_p^2 = 0.04$), indicating that memory performance was higher in the online sample than in the lab sample (Cohen's $d = 0.25$, 95 % CI [0.06, 0.45]).

This ANOVA also revealed a significant main effect of *Strategy* on memory performance ($F(1, 202) = 17.56$, $p < 0.01$, $\eta_p^2 = 0.08$), indicating that participants had higher scores when they used retrieval practice instead of restudy (Cohen's $d = 0.31$, 95% CI [0.11, 0.50]). No other main effects or interactions were observed (Table 11). Rerunning the ANOVA with the trimming procedure also showed significant effects of *Sample* and *Strategy* on memory performance (see Appendix II). This significant effect of *Strategy* was also seen in the full sample (see Appendix III).

Figure 19

Memory performance in the evaluative and control groups across online and lab samples



Note. The x-axis represents learning strategy, retrieval practice (RP), and restudy (RS). The y-axis represents the proportions of idea units. Ctr is the control group, and Evl is the evaluative group. The black line in the center is the average, and the white box denotes the standard errors (SE). The grey dots represent individual data points.

Table 11

Summary of ANOVA for memory performance in the online versus lab sample

Effect	**F(1, 202)**	**p**	η_p^2
Sample*	8.05	0.005	0.04
Strategy*	17.56	< 0.01	0.08
Instruction	0.24	0.63	< 0.01
Instruction × Strategy	0.07	0.79	< 0.01
Instruction × Sample	0.71	0.40	< 0.01
Sample × Strategy	0.15	0.70	< 0.01
Instruction × Sample × Strategy	0.06	0.81	< 0.01

Note. The table shows the results of the memory performance ANOVA in the online versus lab sample, with a significant effect of *Strategy* and *Sample* indicated by the "*". *F*-values, *p*-values, and η_p^2 values are also reported.

5.5.2 Memory Performance and Anxiety

To investigate the role of test anxiety and generalized anxiety on memory performance, we added the test anxiety and anxiety scores as the covariates in a 2 x 2 x 2 ANCOVA, which was rerun with the results of each anxiety questionnaire. We detected violations of assumptions related to interactions between the grouping variables and the covariates. To address this issue, we implemented a centering procedure, consisting of subtracting the mean of the covariate from each individual score, which aims to reduce multicollinearity and enhance the interpretability of the regression coefficients (Schneider et al., 2015).

The results of all covariates for the four different anxiety questionnaires are summarized in Table 12 for both the online versus lab and the full samples. The CTAS and Trait Anxiety covariates were marginally significant ($p = 0.06$ and $p = 0.05$) in the online versus lab sample. Both test anxiety measures (CTAS and TAI) and Trait Anxiety emerged as significant predictors in the full sample ($p's < 0.01$).

Sample was a significant predictor of memory performance in the online versus lab

sample, mirroring the main 2 x 2 x 2 ANOVA on memory performance. All ANCOVAs for both the online versus lab and full samples showed significant effects of *Strategy*, mirroring the main memory performance ANOVAs for both samples. Across all ANCOVAs, the consistently small effect sizes suggested that the covariates had limited impact on memory performance (see Appendix II and III for the full ANCOVA results for the online versus lab and full samples).

Table 12

Summary of ANCOVA results for covariates in the online versus lab sample and the full sample

Effect	DF	F	p	η_p^2
Online vs. lab sample				
CTAS+	(1, 815)	3.70	0.06	0.01
TAI	(1, 815)	1.22	0.27	< 0.01
Trait Anxiety+	(1, 815)	3.73	0.05	0.01
GAD	(1, 815)	0.19	0.66	< 0.01
Full sample				
CTAS*	(1, 819)	8.29	< 0.01	0.01
TAI*	(1, 819)	8.95	< 0.01	0.01
Trait Anxiety*	(1, 819)	13.04	< 0.01	0.02
GAD	(1, 819)	2.69	0.10	< 0.01

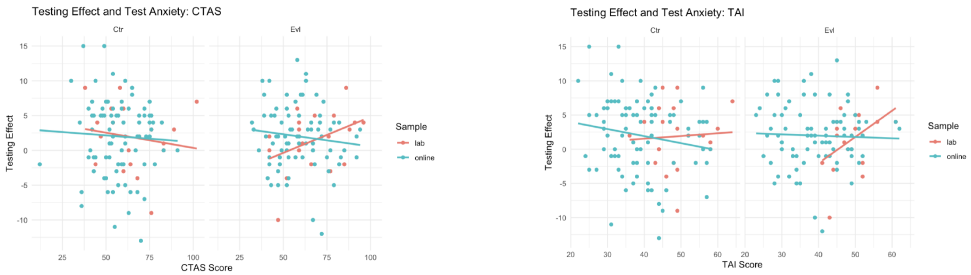
Note. The table shows the results for the covariates in both the online versus lab and full samples. CTAS, TAI and Trait Anxiety were significant in the full sample, indicated by the "*" next to the covariates. CTAS and Trait Anxiety were marginally significant in the online versus lab sample, indicated by the "+." F-values, p-values, and η_p^2 values are also reported.

Interestingly, our exploratory analyses using linear regression revealed a positive relationship between the testing effect and having higher levels of test anxiety in the evaluative group of the lab sample (Figure 20). This pattern remained consistent for both CTAS and TAI, with the exception of the lab sample in the control group for CTAS where the relationship was slightly negative and for the control group lab sample for TAI where the relationship was positive. No differences, or slightly negative relationships, were observed in the online samples for both CTAS and TAI. This analysis did not reveal any

significant patterns in the full sample (see Appendix III).

Figure 20

Testing effect as a function of test anxiety, group, and sample in the online versus lab sample



Note. The x-axis represents the scores from the CTAS and TAI questionnaires. The y-axis represents the testing effect, calculated as the difference in the recalled idea units between the retrieval practice and restudy conditions. Ctr is the control group, and Evl is the evaluative group. The lines in the graph are regression lines, representing the relationship among the dots for the lab and online samples. The colored dots represent individual data points. We expected a positive relationship between the testing effect and test anxiety in the evaluative group, such that a higher testing effect would be accompanied by higher levels of test anxiety because retrieval practice would have a protective role in these situations.

5.5.3 State Anxiety

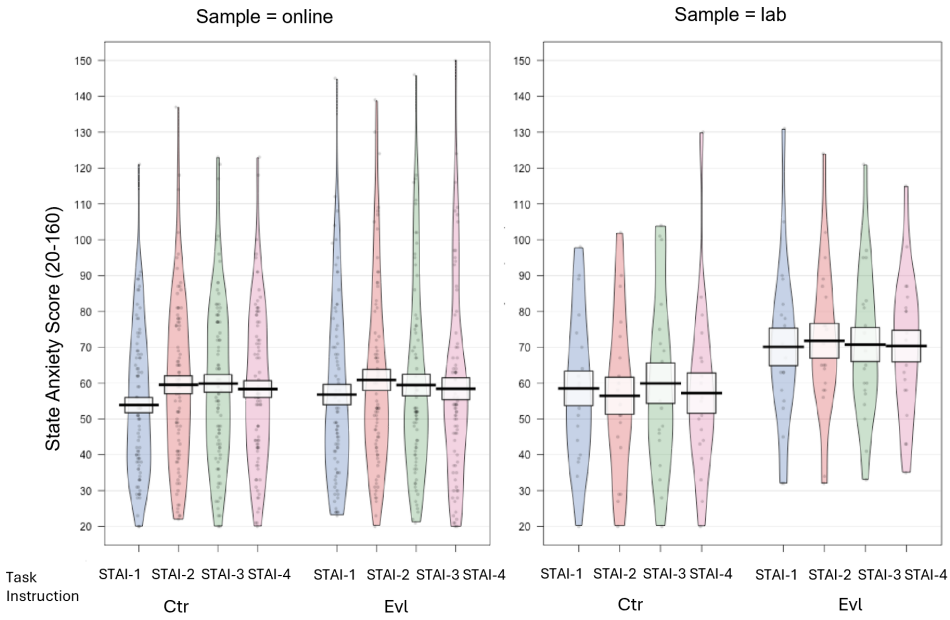
To examine the effect of the instructions to induce test anxiety during the recall session, we conducted a 2 (instruction: evaluative vs. control, between-subject factor) x 2 (sample: online vs. lab, between-subject factor) x 4 (task: STAI1, STAI2, STAI3, and STAI4, within-subject factor) mixed-design ANOVA (see Figure 21 for descriptive statistics). The results revealed numerically higher levels of state anxiety in the evaluative group in the lab sample, but this difference was not statistically significant compared to the control group ($p = 0.08$, Cohen's $d = 0.17$, 95% CI [-0.02, 0.36]). No other main effects or interactions were observed (Table 13).

In the full sample, the same analysis showed a main effect for *Task*. In this case,

participants in both groups were more stressed at time points 2 and 3 compared to time point 1 (see Appendix III).

Figure 21

State anxiety levels during recall sessions in the online versus lab sample



Note. The x-axis represents STAI1, STAI2, STAI3, and STAI4, which are the time points when state anxiety was measured during the recall session (i.e., after reading the instructions, after Raven's matrices or the puzzle task, after memory test 1, and after memory test 2). The y-axis represents the state anxiety scores. Ctr is the control group, and Evl is the evaluative group. The black line in the center is the average, and the white box denotes the standard errors (SE). The grey dots represent individual data points.

To check whether participants in the evaluative and control groups experienced similar stress levels during the initial learning session, we then performed a 2 (instruction: evaluative vs. control, between-subject factor) x 2 (task: before reading text 1 and after performing the second learning strategy, within-subject factor) x 2 (sample: online vs. lab, between-subject factor) mixed-design ANOVA. The results revealed a significant main effect of *Sample* ($F(1, 201) = 4.91$, $p = 0.03$, $\eta_p^2 = 0.02$), showing that the participants in the lab sample had higher STAI scores (M= 67.1, SE= 3.50) in Part 1 of the study than the online sample (M= 58.5, SE = 1.70, Cohen's $d = 0.20$, 95% CI [0.004, 0.39]). There were no other main effects or interactions (Table 14).

Table 13

Summary of ANOVA results for STAI scores during the recall session in the online versus lab sample

Effect	**DF**	**F**	**p**	$\boldsymbol{\eta_p^2}$
Instruction+	1, 201	2.99	0.08	0.01
Sample	1, 201	2.29	0.13	0.01
Task	2, 540	1.93	0.13	0.01
Instruction × Sample	1, 201	2.20	0.14	0.01
Instruction × Task	2, 540	0.57	0.62	< 0.01
Sample × Task	2, 540	1.49	0.22	< 0.01
Instruction × Sample × Task	2, 540	0.45	0.70	< 0.01

Note. The table shows the results from the state anxiety ANOVA in the online versus lab sample in the recall session. *Instruction* was marginally significant, indicated by the "+." F-values, p-values, and η_p^2 values are also reported. One participant was excluded from this analysis due to technical reasons.

Table 14

Summary of ANOVA results for baseline STAI scores in the online versus lab sample

Effect	DF	F	p	η_p^2
Instruction	1, 201	1.25	0.27	< 0.01
Sample*	1, 201	4.91	0.03	0.02
Task	1, 201	0.35	0.55	< 0.01
Instruction × Sample	1, 201	0.45	0.50	< 0.01
Instruction × Task	1, 201	2.62	0.11	0.10
Sample × Task	1, 201	2.26	0.13	< 0.01
Instruction × Sample × Task	1, 201	0.80	0.37	< 0.01

Note. The table shows the results from the state anxiety ANOVA in the online versus lab sample during the learning session, with a significant effect of *Sample*, indicated by the "*". *F*-values, *p*-values, and η_p^2 values are also reported. One participant was excluded from this analysis due to technical reasons.

5.5.4 Heart Rate

To investigate how test anxiety was induced via our physiological proxy, heart rate was analyzed using a 2 (instruction: evaluative vs. control, between-subject factor) x 2 (session: learning vs. recall, within-subject factor) x 2 (sample: online vs. lab, between-subject factor) mixed-design ANOVA (see Appendix II for descriptive statistics). Any physically unlikely values (BPM < 20 or > 400) were excluded. This analysis did not reveal any significant main effects or interactions (Table 15).

Rerunning this ANOVA with the trimming procedure revealed significant differences for *Sample* in all standard deviation cut-offs (see Appendix II). Marginal significance was seen for the *Sample* by *Session* interaction ($p = 0.05$) in the 1 standard deviation cut-off. Given that the sample sizes of the trimmed dataset was smaller and the heart rate measure was not our primary variable of interest, we remained cautious and refrained from interpreting these results further and drawing conclusions from them.

Table 15

Summary of ANOVA results for heart rate in the online versus lab sample

Effect	DF	F	p	η_p^2
Instruction	1, 201	0.24	0.62	< 0.01
Sample	1, 201	1.04	0.31	< 0.01
Session	1, 201	0.26	0.61	< 0.01
Instruction × Sample	1, 201	0.09	0.77	< 0.01
Instruction × Session	1, 201	0.01	0.94	< 0.01
Sample × Session	1, 201	1.36	0.25	< 0.01
Instruction × Sample × Session	1, 201	0.00	0.96	< 0.01

Note. The table shows the results from the heart rate ANOVA in the online versus lab sample, where no significant main effects or interactions were observed. F-values, p-values, and η_p^2 values are also reported. One participant was excluded from this analysis due to technical reasons.

5.5.5 Extreme Groups

Since we did not observe a significant interaction between *Instruction* and *Strategy* in our ANOVA of memory performance (Table 11), we investigated whether this could be because the evaluative context was not well-induced. To examine this, we used an extreme group approach. We repeated the 2 x 2 x 2 memory performance ANOVA by including participants in the evaluative group with a high score on the STAI (i.e., the upper quartile) and participants in the control group with a low score on the STAI (i.e., lower quartile). This analysis mirrored the main 2 x 2 x 2 ANOVA and showed a marginal main effect for *Sample* ($F(1, 68) = 4.01$, $p = 0.049$, $\eta_p^2 = 0.06$), indicating that the online sample performed better than the lab sample (Cohen's $d = 0.31$, 95% CI [-0.20, 0.65]). This ANOVA also showed a main effect for *Strategy* ($F(1, 68) = 4.34$, $p = 0.04$, $\eta_p^2 = 0.06$), indicating higher scores for the text learned with retrieval practice (Cohen's $d = 0.27$, 95 % CI [-0.07, 0.61]).

5.6 Discussion

The current study investigated whether learning with retrieval practice can protect memory against the detrimental effects of stressors like test anxiety. We found that

retrieval practice consistently benefits memory more than restudy over the long-term, replicating the testing effect (McDaniel et al., 2007; Roediger & Karpicke, 2006; Rowland, 2014; Schwieren et al., 2017) and confirming our first hypothesis. Previous studies have largely used pen-and-paper approaches or computerized tests to examine the testing effect in educational settings (Agarwal et al., 2012; Butler et al., 2014; Cadaret & Yates, 2018; Roediger & Butler, 2011; Rowland, 2014; Yang et al., 2018). To our knowledge, this makes the current study the first to replicate the testing effect using a fully online design. This finding is crucial because it suggests that the testing effect is robust despite today's rapidly evolving digital landscape. This not only expands the validity of the testing effect but also highlights the effectiveness of retrieval practice in modern online learning environments.

In Part 2 of the study, participants in the full sample reported higher stress levels in both the evaluative and control groups after completing the matrix task and after the first memory test compared to right after reading the instructions. Previous studies have shown that tasks like Raven's Matrices can elicit higher levels of anxiety on their own, especially if the tasks are timed (Gonthier, 2023; Kumari & Corr, 1998). This suggests that the matrices were effective enough to trigger stress regardless of whether the tasks were framed as an IQ test or a puzzle task. Thus, it is possible that the matrices' difficulty and associated cognitive load contributed to the increased STAI scores in both evaluative and control groups, independent of the evaluative or neutral instructions.

Participants in the evaluative group showed numerically higher, although non-significant ($p = 0.08$), STAI ratings than those in the control group in the lab setting, suggesting a potential effect of the instruction manipulation in the in-person setting. This means that the evaluative instructions were perhaps not strong enough to elicit considerable differences in this group and points to the need for future research to optimize instruction-based stress induction methods.

Participants in the lab reported significantly higher baseline stress levels compared to the online sample, possibly due to the social component that is present in this setting.

This aligns with our current understanding that stressor efficacy depends on the intensity of the stressor and its relevance, which can be amplified by the physical proximity of the participant to the stressor (Brosch et al., 2010; Dickerson & Kemeny, 2004; Sakaki et al., 2012). Even without formally inducing test anxiety in Part 1, the lab setting itself may have elevated the participants' stress levels nonetheless. These findings suggest that the lab environment alone could increase stress levels, potentially impacting performance, and begs the question of whether some tasks like stress induction are more effective in in-person settings.

We expected that learning with retrieval practice would counteract the detrimental effects of test anxiety on memory (i.e., memory performance would be higher in the evaluative group compared to the control group for the text learned with retrieval practice versus restudy), but this effect was not seen overall. Thus, the current findings do not provide sufficient evidence that retrieval practice protects memory against test anxiety, thereby not confirming our third hypothesis and not replicating the study by A. M. Smith et al. (2016). One possible reason for this is that our test anxiety manipulation may not have been intense or prolonged enough to trigger high stress and impair memory retrieval. In A. M. Smith et al. (2016), the TSST procedure lasted a total of around 6 minutes (i.e., giving speeches and solving math problems in front of others) and was accompanied by higher levels of cortisol in the stress group. In other studies, stress procedures lasted more than 1h, including pre- and post-procedure wait times (Birkett, 2011; Dickerson & Kemeny, 2004). In contrast, our study featured a brief anxiety exposure throughout the study, which focused on the evaluative component rather than negative judgment. Additionally, self-reported stress levels remained in the low range, suggesting that stress was likely insufficient to cause memory impairment. As a result, the amount of stress exposure in the current study may not have provided an opportunity for the stress response to be triggered and therefore, for retrieval practice to exert a protective influence.

In the follow-up extreme groups analysis, the online sample performed marginally

better than the lab sample, mirroring the main ANOVA on memory performance. This could be due to reduced social evaluative pressure in online settings, which may benefit performance when the participants are under stress (Dickerson & Kemeny, 2004; von der Embse et al., 2018). Additionally, retrieval practice again outperformed restudy in this analysis, demonstrating its benefit regardless of the testing environment, instruction, or anxiety level. Concurrently, the exploratory regression analysis indicated that the testing effect increased as test anxiety scores rose, particularly in evaluative lab-based settings. This finding could mean that retrieval practice may be especially beneficial for individuals who experience high test anxiety and in in-person testing conditions. This exploratory result also hints that retrieval practice may serve a buffering role in high-anxiety situations, though it should be interpreted cautiously due to the absence of significant interaction effects in the main analyses.

In terms of covariates, our study found that test anxiety and trait anxiety were significant predictors of memory performance when all other variables were held constant in the full sample. Test anxiety as measured by the CTAS questionnaire and trait anxiety also saw marginal significance in the online versus lab sample. This aligns with established findings in the literature showing that test anxiety and varying levels of trait anxiety influence memory (Cassady, 2004; Eysenck et al., 2007a). Yet, across all ANCOVA analyses, retrieval practice continued to outperform restudy when all other variables were held constant. This suggests that its benefits may be partially independent of anxiety. Our observations in this regard agree with the findings by Yang and colleagues (2020), who found that retrieval practice is minimally affected by factors like test anxiety, and is in line with the episodic context theory (Lehman et al., 2014).

At the same time, stressors like test anxiety may not always result in memory performance deficits. For example, Jerrim (2023) did not find a clear link between test anxiety and grade outcomes, and Sommer and Arendasy (2015) also did not find a causal relationship between test anxiety and performance in high-stakes contexts. Similarly,

Howard (2020) suggested that, after accounting for ability, elevated levels of test anxiety only lead to minor performance declines. One speculative explanation could be related to the Yerkes-Dodson Law, which posits that moderate arousal can improve performance, but excessive anxiety diminishes it (Yerkes, Dodson, et al., 1908). Although test anxiety does not always correspond to physiological arousal (Leininger & Skeel, 2012), moderate anxiety may still enhance performance, as observed in the inverted-U relationship between statistics anxiety and exam performance in Keeley et al. (2008). Participants in our study may also have experienced this optimal arousal, potentially improving cognitive performance during recall. This pattern is also supported by studies in animals and humans demonstrating memory enhancements under moderately acute stress (Goldfarb, 2019; Shields, Sazma, et al., 2017).

The current study contains several limitations. First, it lacks physiological measurements of stress, such as cortisol levels obtained from saliva samples, which is a common practice in similar studies (e.g., A. M. Smith et al., 2016). Due to the online nature of our study, we opted for an alternative proxy measure for stress by having participants self-report their heart rate. However, self-reported heart rates may lack the precision and objectivity associated with direct physiological measures. The second limitation is related to statistical power. Although the sample was sufficiently powered for the initial 2 x 2 design, the online versus lab comparison may not have been sufficiently powered for our analysis plan due to the addition of a third factor (Brysbaert, 2019). This could have led to an overestimation of the effects. Third, although the online portion of the study contained ample manipulation checks and stringent exclusion criteria, it is still possible that some participants did not carefully engage with the assigned tasks or provided responses inattentively. This potential lack of adherence to the study requirements could introduce biases or compromise internal validity.

The results from the current study have important implications for our understanding of learning strategies in both applied and educational settings. The

successful replication of the testing effect in an online setting means retrieval practice is an effective method to enhance long-term memory retention outside traditional classroom environments and can be reliably used in digital learning contexts. We also demonstrated that retrieval practice benefits memory in both the evaluative and control conditions, though the evaluative manipulation may have been too weak to fully detect a difference between these contexts. This means that learners might be able to benefit from retrieval practice regardless of whether they are in a high-pressure evaluative or a low-pressure setting in classrooms and other learning contexts, but further research is needed to confirm the effects of the evaluative condition. Retrieval practice consistently outperformed restudy, but evidence that it buffers against stress or test anxiety was inconclusive, likely due to the mild anxiety induction. These findings underscore the need for further research to refine stress induction protocols in online settings when assessing the interaction with retrieval practice.

5.7 Conclusion

The current work successfully replicated the testing effect in an online study. Test anxiety induced through instructions was not effective enough to considerably increase stress levels in the evaluative group in the online versus lab sample. The increase in state anxiety scores in both groups in the full sample might be more attributed to the matrix task rather than the instructions themselves. Retrieval practice was not protective for memory performance in the evaluative group overall. However, exploratory analyses suggested that the testing effect may benefit individuals with higher levels of test anxiety in lab settings more than online samples. Future research should refine the magnitude and nature of the stress induction when using retrieval practice.

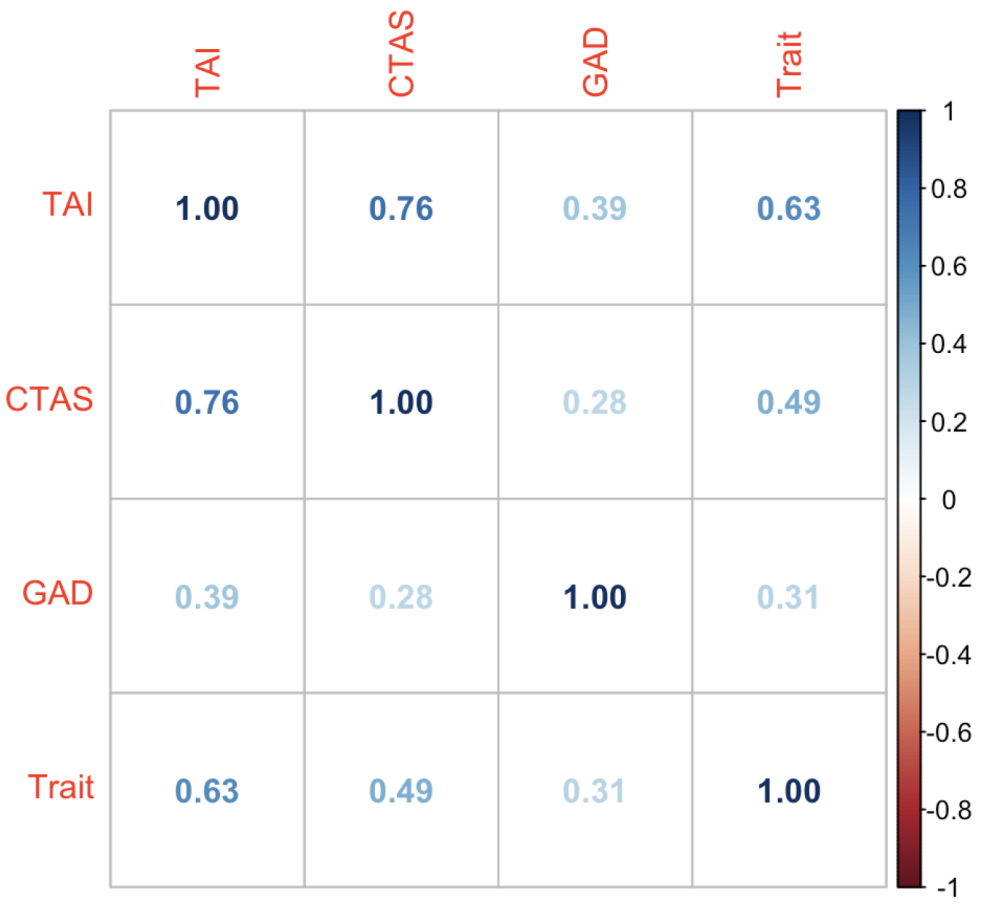
5.8 Appendix I: Preparatory Analyses

5.8.1 Reliability

We explored the correlations between the different anxiety measures assessed through our four questionnaires: CTAS, TAI, Trait Anxiety, and GAD (see Figure 22). A strong positive correlation ($r = 0.76$) was found between scores on the CTAS and TAI questionnaires, indicating a robust linear relationship between these measures. This suggests that CTAS and TAI measure similar underlying test anxiety traits and that individuals scoring high on one measure are likely to score high on the other as well. Moderately positive correlations were observed between Trait Anxiety and TAI ($r = 0.63$) and Trait Anxiety and CTAS ($r = 0.49$), suggesting that individuals with higher levels of trait anxiety tend to experience greater test anxiety. In contrast, a weak positive correlation was observed between the scores on the GAD and Trait Anxiety ($r = 0.31$), CTAS ($r = 0.28$), and TAI ($r = 0.39$). These weak correlations indicate that although individuals with higher levels of generalized anxiety disorder may exhibit higher trait and test anxiety, it is likely that GAD measures a broader or different aspect of anxiety. Overall, while these concepts are interconnected, they represent different dimensions of anxiety, with trait and test anxiety being more closely connected than other anxiety disorders.

Figure 22

Correlation matrix between anxiety variables



Note. Correlation matrix illustrating the relationships between all anxiety measures on a scale of -1 (low correlation) to 1 (high correlation). The numbers in dark blue indicate high correlation, and the numbers in light blue indicate lower correlations.

5.8.2 Inter-rater Reliability

Inter-rater reliability was calculated by having a second rater rate a subset of 40 responses and comparing the ratings of the two raters. The Pearson correlation between the two ratings indicated substantial agreement between the two raters for both "The Sun" ($r = 0.89$) and "Sea Otters" ($r = 0.88$) texts. This also reflects the strong agreement in the coding of idea units between both raters, suggesting that the coding procedure is highly reliable and replicable.

5.8.3 Times read

To ensure that our results were not biased by differences in the number of times participants read and restudied the two learning texts during Part 1, we computed a three-way factorial ANOVA examining the effects of learning *Strategy* (read vs. restudy, within-subject factor), *Sample* (lab vs. online, between-subject factor), and *Instruction* (control vs. evaluative, between-subject factor) on the number of times each text was read (see Table 16 for descriptive statistics). The analysis did not reveal any significant results.

Table 16

Descriptive statistics for the number of times each text was read in both samples

Instruction	Sample	Strategy	M	SD
Ctr	Online	Read	5.40	2.88
Ctr	Online	Restudy	6.71	9.13
Ctr	Lab	Read	6.11	3.05
Ctr	Lab	Restudy	4.78	2.17
Evl	Online	Read	5.15	2.17
Evl	Online	Restudy	8.47	25.70
Evl	Lab	Read	5.11	1.88
Evl	Lab	Restudy	4.22	1.64

Note. The table shows the number of times each text was read and restudied in both the evaluative ("Evl") and control ("Ctr") groups of both the online and the lab samples. M is the mean of the times each texts were read and SD is the standard deviation of M. No differences were observed in the number of times the texts were read between the samples and groups with a three-way factorial ANOVA.

5.9 Appendix II: Additional Results for Online vs. Lab Sample

5.9.1 Memory Performance

To examine the potential effects of outliers on memory performance, we reran the 2 x 2 x 2 ANOVA using a trimming procedure with 3, 2.5, 2, 1.5, and 1 standard deviations from the mean. The results are summarized in Table 17. Results showed a pattern similar to the main ANOVA, with a significant main effect of *Sample* and *Strategy* on memory performance ($p < 0.01$ for all analyses).

Table 17

Results of the trimming procedure for memory performance in the online versus lab sample

SD	Effect	DF	F	p	η_p^2
3	Instruction	(1, 202)	0.24	0.63	< 0.01
	Sample*	(1, 202)	8.05	0.005	0.04
	Instruction × Sample	(1, 202)	0.71	0.40	< 0.01
	Strategy*	(1, 202)	17.56	< 0.01	0.08
	Instruction × Strategy	(1, 202)	0.07	0.79	< 0.01
	Sample × Strategy	(1, 202)	0.15	0.70	< 0.01
	Instruction × Sample × Strategy	(1, 202)	0.06	0.81	< 0.01
2.5	Instruction	(1, 202)	0.24	0.63	< 0.01
	Sample*	(1, 202)	8.05	0.005	0.04
	Instruction × Sample	(1, 202)	0.71	0.40	< 0.01
	Strategy*	(1, 202)	17.56	< 0.01	0.08
	Instruction × Strategy	(1, 202)	0.07	0.79	< 0.01
	Sample × Strategy	(1, 202)	0.15	0.70	< 0.01
	Instruction × Sample × Strategy	(1, 202)	0.06	0.81	< 0.01
2	Instruction	(1, 201)	0.37	0.54	< 0.01
	Sample*	(1, 201)	7.83	< 0.01	0.04
	Instruction × Sample	(1, 201)	0.57	0.45	< 0.01
	Strategy*	(1, 201)	18.13	< 0.01	0.08
	Instruction × Strategy	(1, 201)	0.10	0.75	< 0.01
	Sample × Strategy	(1, 201)	0.19	0.66	< 0.01
	Instruction × Sample × Strategy	(1, 201)	0.04	0.85	< 0.01
1.5	Instruction	(1, 196)	0.05	0.82	< 0.01
	Sample*	(1, 196)	7.22	< 0.01	0.04
	Instruction × Sample	(1, 196)	1.47	0.23	< 0.01
	Strategy*	(1, 196)	17.72	< 0.01	0.08
	Instruction × Strategy	(1, 196)	0.11	0.74	< 0.01
	Sample × Strategy	(1, 196)	0.20	0.65	< 0.01
	Instruction × Sample × Strategy	(1, 196)	0.03	0.86	< 0.01
1	Instruction	(1, 177)	2.03	0.16	0.01
	Sample*	(1, 177)	9.73	0.02	0.05
	Instruction × Sample	(1, 177)	2.62	0.11	0.01
	Strategy*	(1, 177)	19.57	< 0.01	0.10
	Instruction × Strategy	(1, 177)	0.02	0.88	< 0.01
	Sample × Strategy	(1, 177)	0.02	0.90	< 0.01
	Instruction × Sample × Strategy	(1, 177)	0.22	0.64	< 0.01

Note. The table shows the ANOVA results on memory performance across all trimming parameters. Significant effects for *Sample* and *Strategy* are indicated by the "*". *F*-values, *p*-values, and η_p^2 values are also reported.

5.9.2 Memory Performance and Anxiety

The full ANCOVA results (Table 18) revealed a similar pattern across all questionnaires, with the *Strategy* and *Sample* variables statistically significant across the board. Marginal significance was seen for the CTAS covariate ($p = 0.06$) and the Trait Anxiety covariate ($p = 0.05$). The effect sizes were small for all variables.

Table 18

Full ANCOVA results for CTAS, TAI, Trait Anxiety, and GAD in the online versus lab sample

Variable	Effect	F(1, 815)	p	η_p^2
CTAS	CTAS+ (covariate)	3.70	0.06	< 0.01
	Strategy*	36.03	< 0.001	4.04
	Instruction	0.01	0.942	< 0.01
	Sample*	18.60	< 0.001	0.02
	Strategy × Instruction	0.03	0.87	< 0.01
	Strategy × Sample	0.17	0.68	< 0.01
	Instruction × Sample	2.12	0.15	< 0.01
	Strategy × Instruction × Sample	0.07	0.80	< 0.01
TAI	TAI (covariate)	1.22	0.27	< 0.01
	Strategy*	35.92	< 0.001	0.04
	Instruction	< 0.01	0.95	< 0.01
	Sample*	15.47	< 0.01	0.02
	Strategy × Instruction	0.03	0.87	< 0.01
	Strategy × Sample	0.17	0.68	< 0.01
	Instruction × Sample	2.07	0.15	< 0.01
	Strategy × Instruction × Sample	0.07	0.80	< 0.01
Trait Anxiety	Trait Anxiety+ (covariate)	3.73	0.05	< 0.01
	Strategy*	36.03	< 0.001	0.04
	Instruction	0.07	0.80	< 0.01
	Sample*	13.67	< 0.01	0.02
	Strategy × Instruction	0.03	0.87	< 0.01
	Strategy × Sample	0.17	0.68	< 0.01
	Instruction × Sample	2.31	0.13	< 0.01
	Strategy × Instruction × Sample	0.07	0.80	< 0.01
GAD	GAD (covariate)	0.19	0.66	< 0.01
	Strategy*	35.87	< 0.01	0.04
	Instruction	< 0.01	0.96	< 0.01
	Sample*	20.78	< 0.01	0.02
	Strategy × Instruction	0.03	0.87	< 0.01
	Strategy × Sample	0.17	0.68	< 0.01
	Instruction × Sample	2.04	0.15	< 0.01
	Strategy × Instruction × Sample	0.07	0.79	< 0.01

Note. The table shows the full ANCOVA results for the CTAS, TAI, Trait Anxiety, and GAD questionnaires for the online versus lab sample. *Strategy* and *Sample* are significant across all variables, indicated by the "*". Marginal significance for CTAS and Trait Anxiety covariates is indicated by the "+". F-values, p-values, and η_p^2 values are also reported.

5.9.3 Heart Rate

The full heart rate analysis results are presented in Tables 19 (descriptive statistics) and 20 (trimming procedure). The results showed a significant effect of *Sample* at all cut-offs and a marginally significant *Sample* by *Session* interaction at the 1 standard deviation cut-off ($p = 0.05$). Because the sample size was small and this was not our main variable of interest, we refrain from making a heavy interpretation of these effects.

Table 19

Descriptive statistics for heart rate in both samples

Instruction	Session	Sample	Mean Pulse	Median Pulse	Std Dev Pulse	Min Pulse	Max Pulse
Control	Learning	lab	71.79	72	12.29	36	92
Control	Learning	online	65.73	64	11.21	36	96
Control	Recall	lab	70.53	72	16.48	28	100
Control	Recall	online	69.36	64	27.91	44	308
Evaluative	Learning	lab	72.40	74	13.16	48	92
Evaluative	Learning	online	68.15	68	13.08	40	108
Evaluative	Recall	lab	71.00	68	12.20	52	96
Evaluative	Recall	online	71.29	68	22.71	40	240

Note. The table shows the descriptive results for heart rate in the control and evaluative groups across both learning conditions in both samples. The mean, median, min, and max values refer to the pulse rates.

Table 20

Results of the trimming procedure for heart rate in the online versus lab sample

SD	Effect	DF	F	p	η_p^2
3	Instruction	1, 199	0.52	0.47	0.003
	Sample*	1, 199	4.35	0.04	0.021
	Instruction × Sample	1, 199	0.20	0.66	< 0.01
	Session	1, 199	0.02	0.89	< 0.01
	Instruction × Session	1, 199	0.00	0.95	< 0.01
	Sample × Session	1, 199	1.48	0.22	0.01
	Instruction × Sample × Session	1, 199	0.02	0.90	< 0.01
2.5	Instruction	1, 199	0.52	0.47	0.003
	Sample*	1, 199	4.35	0.04	0.021
	Instruction × Sample	1, 199	0.20	0.66	< 0.01
	Session	1, 199	0.02	0.89	< 0.01
	Instruction × Session	1, 199	0.00	0.95	< 0.01
	Sample × Session	1, 199	1.48	0.22	0.01
	Instruction × Sample × Session	1, 199	0.02	0.90	< 0.01
2	Instruction	1, 197	0.00	0.95	< 0.01
	Sample*	1, 197	8.16	0.005	0.04
	Instruction × Sample	1, 197	0.86	0.35	< 0.01
	Session	1, 197	0.00	0.99	< 0.01
	Instruction × Session	1, 197	0.00	0.98	< 0.01
	Sample × Session	1, 197	1.37	0.24	0.01
	Instruction × Sample × Session	1, 197	0.08	0.78	< 0.01
1.5	Instruction	1, 189	0.06	0.80	< 0.01
	Sample*	1, 189	7.64	0.01	0.04
	Instruction × Sample	1, 189	0.38	0.54	< 0.01
	Session	1, 189	0.22	0.64	< 0.01
	Instruction × Session	1, 189	0.42	0.52	< 0.01
	Sample × Session	1, 189	2.19	0.14	0.01
	Instruction × Sample × Session	1, 189	0.06	0.80	< 0.01
1	Instruction	1, 162	2.07	0.15	0.01
	Sample*	1, 162	4.73	0.03	0.03
	Instruction × Sample	1, 162	0.15	0.70	< 0.01
	Session	1, 162	3.31	0.07	0.02
	Instruction × Session	1, 162	0.02	0.89	< 0.01
	Sample × Session+	1, 162	3.83	0.05	0.02
	Instruction × Sample × Session	1, 162	0.18	0.67	< 0.01

Note. The table shows the ANOVA results on heart rate analysis across all trimming parameters. Significant effects for *Sample* are indicated by "*" and marginally significant results by "+." *F*-values, *p*-values, and η_p^2 values are also reported.

5.10 Appendix III: Full Sample Results

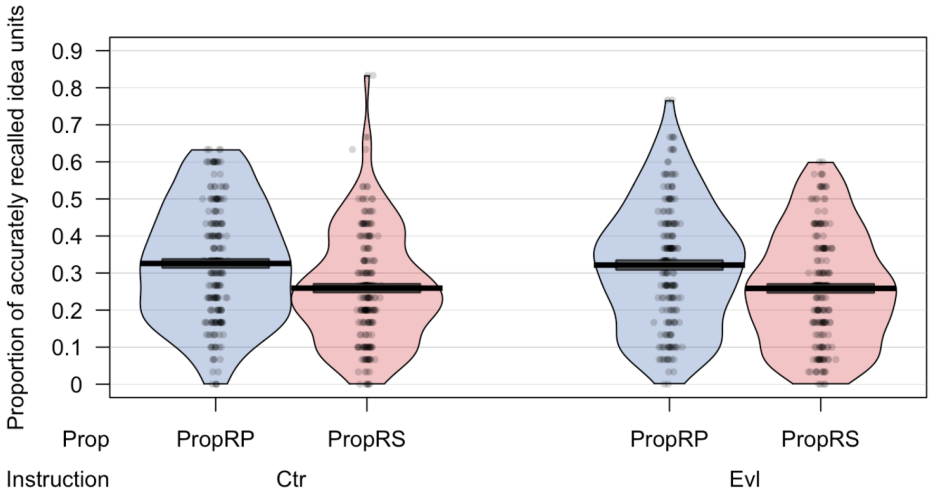
5.10.1 Memory Performance

To examine memory performance, we performed a 2 (strategy: retrieval practice vs. restudy, within-subject factor) x 2 (instruction: evaluative vs. control, between-subject factor) mixed-design ANOVA (see Figure 23 for descriptive statistics). The results showed a significant main effect of *Strategy* ($F(1, 204) = 32.14, p < 0.001, \eta_p^2 = 0.14$), with follow-up effects showing that the retrieval practice scores were higher than the restudy scores (Cohen's $d = 0.41$, 95% CI [0.22, 0.61]). The main effect of *Instruction* was not significant ($F(1, 204) = 0.02, p = 0.90, \eta_p^2 < 0.01$) nor was the interaction effect between *Instruction* and *Strategy* ($F(1, 204) = 0.03, p = 0.87, \eta_p^2 < 0.01$).

Rerunning the ANOVAs with the trimming procedure (Table 21) showed significant effects of *Strategy* on memory performance ($p < 0.01$ for all analyses). No other significant main effects or interactions were observed.

Figure 23

Memory performance of the evaluative and control groups in the full sample



Note. The x-axis represents learning strategy, retrieval practice (RP), and restudy (RS). The y-axis represents the proportions of idea units. Ctr is the control group, and Evl is the evaluative group. The black line in the center is the average, and the white box denotes the standard errors (SE). The grey dots represent individual data points.

Table 21

Results of the trimming procedure for memory performance in the full sample

SD	Effect	DF	F	p	η_p^2
3	Instruction	(1, 204)	0.02	0.90	< 0.01
	Strategy*	(1, 204)	32.14	< 0.01	0.14
	Instruction× Strategy	(1, 204)	0.03	0.87	< 0.01
2.5	Instruction	(1, 204)	0.02	0.90	< 0.01
	Strategy*	(1, 204)	32.14	< 0.01	0.14
	Instruction × Strategy	(1, 204)	0.03	0.87	< 0.01
2	Instruction	(1, 203)	0.01	0.93	< 0.01
	Strategy*	(1, 203)	33.53	< 0.01	0.14
	Instruction × Strategy	(1, 203)	0.07	0.79	< 0.01
1.5	Instruction	(1, 198)	0.54	0.46	< 0.01
	Strategy*	(1, 198)	32.16	< 0.01	0.14
	Instruction × Strategy	(1, 198)	0.09	0.76	< 0.01
1	Instruction	(1, 179)	0.15	0.70	< 0.01
	Strategy*	(1, 179)	32.46	< 0.01	0.15
	Instruction × Strategy	(1, 179)	0.03	0.87	< 0.01

Note. The table shows the results of the memory performance ANOVA across all trimming parameters. Significant effects are seen for *Strategy*, indicated by the "*". F-values, p-values, and η_p^2 values are also reported.

5.10.2 Memory Performance and Anxiety

To investigate the effects of test anxiety and general anxiety on memory performance, we added the test anxiety and anxiety scores as the covariates in a 2 x 2 ANCOVA, which was rerun with the results of each anxiety questionnaire. We detected violations of assumptions underlying these ANCOVAs related to interactions between the grouping variables and the covariates and homogeneity of variances. To address this, we implemented the same centering procedure as in the factor analysis. *Strategy* consistently emerged as significant in all ANCOVAs ($p < 0.05$ in all analyses). Test anxiety, as measured by the CTAS and TAI questionnaires, and Trait Anxiety emerged as significant predictors (Table 22). Across all ANCOVAs, consistent small effect sizes were observed. In

addition, the exploratory regression analyses did not reveal any meaningful patterns between test anxiety and the testing effect (Figure 24).

Table 22

Full ANCOVA results for CTAS, TAI, Trait Anxiety, and GAD in the full sample

Variable	Effect	F(1, 819)	p	η_p^2
CTAS	CTAS (covariate)*	8.29	< 0.01	0.01
	Strategy*	35.30	< 0.01	0.04
	Instruction	< 0.01	0.10	< 0.01
	Strategy × Instruction	0.03	0.87	< 0.01
TAI	TAI (covariate)*	8.95	< 0.01	0.01
	Strategy*	35.32	< 0.01	0.04
	Instruction	0.02	0.88	< 0.01
	Strategy × Instruction	0.03	0.86	< 0.01
Trait Anxiety	Trait (covariate)*	13.04	< 0.01	0.02
	Strategy*	35.50	< 0.01	0.04
	Instruction	0.25	0.61	< 0.01
	Strategy × Instruction	0.03	0.86	< 0.01
GAD	GAD (covariate)	2.69	0.10	< 0.01
	Strategy*	35.06	< 0.01	0.04
	Instruction	0.02	0.89	< 0.01
	Strategy × Instruction	0.03	0.87	< 0.01

Note. The table shows the full ANCOVA results for the CTAS, TAI, Trait Anxiety, and GAD questionnaires in the full sample. *Strategy* is significant across all measures, and the covariates CTAS, TAI, and Trait Anxiety are also significant predictors of memory performance, indicated by the "*". *F*-values, *p*-values, and η_p^2 values are also reported.

Figure 24

Testing effect as a function of test anxiety and instruction in the full sample

Note. The x-axis represents the scores on the CTAS and TAI questionnaires. The y-axis represents the testing effect, calculated as the difference in the recalled idea units between retrieval practice and restudy conditions. Ctr is the control group, and Evl is the evaluative group. The lines in the graph are regression lines, representing the relationship among the dots. The colored dots represent individual data points. We expected a positive relationship between the testing effect and test anxiety in the evaluative group, such that a higher testing effect would be accompanied by higher levels of test anxiety because retrieval practice would have a protective role in these situations.

5.10.3 State Anxiety

To investigate the impact of the test anxiety induction on state anxiety levels, we performed a 2 (instruction: evaluative vs. control, between-subject factor) x 4 (task: STAI1, STAI2, STAI3, and STAI4, within-subject factor) mixed-design ANOVA (see Figure 25 for descriptive statistics). There was no main effect of *Group* ($F(1, 203) = 1.17$, $p = 0.28$, $\eta_p^2 = 0.01$) nor a significant *Group* by *Task* interaction ($F(2, 546) = 0.95$,p = 0.41,$\eta_p^2 = 0.01$).

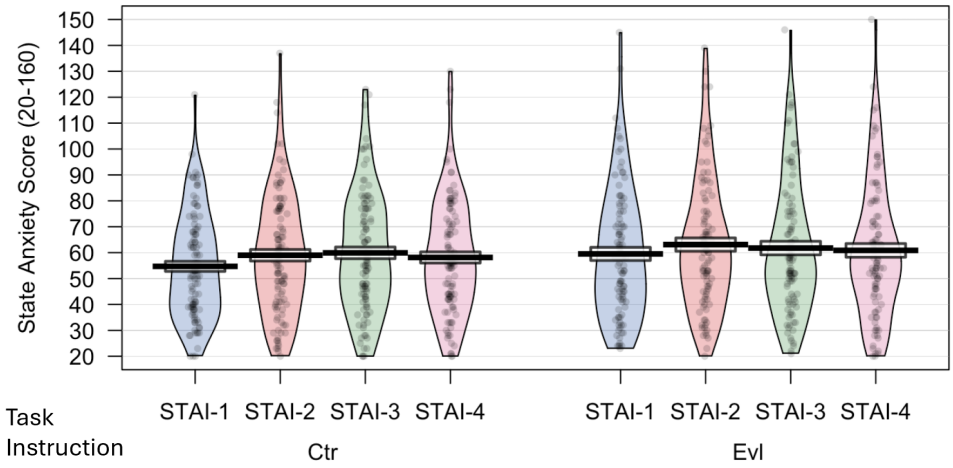
Results showed a significant effect of *Task* ($F(2, 546) = 7.00$, $p < 0.01$, $\eta_p^2 = 0.03$). We then performed paired sample *t*-tests to examine the differences among the different time points for the STAI scores. The results showed a significant increase in the state anxiety levels from the STAI1 to STAI2 time points ($t(203) = -3.82$, $p < 0.01$, *Cohen's d* = 0.17, 95% CI [-0.02, 0.36]) and from STAI1 to STAI3 time points

($t(203) = -3.62$, $p < 0.01$, *Cohen's* $d = 0.15$, 95% CI [-0.04, 0.35]). STAI1 to STAI4 was also marginally significant ($t(203) = -2.58$, $p = 0.051$, *Cohen's* $d = 0.10$, 95% CI [-0.09, 0.29]).

To check whether participants in the evaluative and control groups had similar stress levels during the initial learning session, we then performed a 2 (instruction: evaluative vs. control, between-subject factor) x 2 (task: before reading text 1 and after performing the second learning strategy, within-subject factor) mixed-design ANOVA. We found no main effect of *Instruction* ($F(1, 203) = 0.65$, $p = 0.42$, $\eta_p^2 < 0.01$) and no main effect of *Task* ($F(1, 203) = 0.21$, $p = 0.65$, $\eta_p^2 < 0.01$). The interaction between *Instruction* and *Task* was also not significant ($F(1, 203) = 1.70$, $p = 0.19$, $\eta_p^2 < 0.01$).

Figure 25

State anxiety levels during the recall session in the full sample



Note. The x-axis represents STAI1, STAI2, STAI3, and STAI4. These are the time points at which state anxiety was measured during the recall session (i.e., after reading the instructions, after Raven's matrices or the puzzle task, after memory test 1, and after memory test 2). The y-axis represents the State Anxiety scores. Ctr is the control group, and Evl is the evaluative group. The black line in the center is the average, and the white box denotes the standard errors (SE). The grey dots represent individual data points.

5.10.4 Heart Rate

To investigate how test anxiety is induced via our physiological proxy, heart rate was analyzed as a 2 (instruction: evaluative vs. control, between-subject factor) x 2 (session: learning vs. recall, within-subject factor) mixed-design ANOVA (see Table 19 for descriptive statistics). Any physically unlikely values (BPM < 20 or > 400) were excluded.

There was no significant effect of *Instruction* ($F(1, 203) = 0.81$, $p = 0.37$, $\eta_p^2 < 0.01$) or *Session* ($F(1, 203) = 2.45$, $p = 0.12$, $\eta_p^2 = 0.01$). The interaction effect between *Instruction* and *Session* was also not significant ($F(1, 203) = 0.03$, $p = 0.86$, $\eta_p^2 < 0.01$). Additionally, applying the trimming procedure did not reveal any differences (Table 23).

Table 23
Results of the trimming procedure for heart rate in the full sample

SD	Effect	DF	F	p	η_p^2
3	Instruction	(1, 201)	1.74	0.19	0.01
	Session	(1, 201)	0.59	0.44	< 0.01
	Instruction × Session	(1, 201)	0.02	0.89	< 0.01
2.5	Instruction	(1, 201)	1.74	0.19	0.01
	Session	(1, 201)	0.59	0.44	< 0.01
	Instruction × Session	(1, 201)	0.02	0.89	< 0.01
2	Instruction	(1, 199)	0.87	0.35	< 0.01
	Session	(1, 199)	0.87	0.35	< 0.01
	Instruction × Session	(1, 199)	0.04	0.84	< 0.01
1.5	Instruction	(1, 191)	0.92	0.34	< 0.01
	Session	(1, 191)	0.29	0.59	< 0.01
	Instruction × Session	(1, 191)	0.87	0.35	< 0.01
1	Instruction	(1, 164)	2.55	0.11	0.01
	Session	(1, 164)	0.56	0.45	< 0.01
	Instruction × Session	(1, 164)	0.21	0.64	< 0.01

Note. The table shows the results for heart rate across all trimming parameters. No significant effects or interactions were found. F-values, p-values, and partial η^2 values are also reported.

5.11 Supplementary: w and S Group Results

We initially planned to conduct analyses only for participants who pressed the w during Part 2 to ensure participants followed instructions. Subsequently, we compared the results from this subset against those who did not press w, but pressed the Space bar (S group) to not relinquish quality data. In this Supplementary section to Paper 3, we report the results of the w and S groups on all main analyses to check they are sufficiently similar. Both groups showed broadly similar patterns of results on the main analyses (memory performance, anxiety, and state anxiety). As such, they were collapsed together to form the full sample for analysis reported in the main paper.

Only significant results are reported. All assumptions for normality were checked for. Deviations in normality using Shapiro Wilk test were resolved by calculating Skew-Kurtosis values. Skewness and Kurtosis values were within range (Skew of 0.39 and Kurtosis value of 2.81 for the w group; Skew of 0.22 and Kurtosis of 2.12 in the S group) and visual inspection with QQ plots suggested data were within normal range. No other deviations from normality were found.

5.11.1 Memory Performance

Memory performance was analyzed as a 2 (strategy: retrieval practice vs. restudy) x 2 (instruction: evaluative vs. control) mixed ANOVA. Descriptive statistics can be seen in Figure 26 for the w group and Figure 27 for the S group. In the w group, this analysis revealed a significant main effect of *Strategy* ($F(1, 114) = 19.43, p < 0.001, \eta_p^2 = 0.15$) on memory performance, suggesting that learning with retrieval practice resulted in better performance than using restudy (Cohen's $d = 0.45$ [95% CI: 0.19-0.71]). This analysis was repeated with the trimming procedure, which showed no differences from the initial ANOVA (Table 24).

Similarly, the S group showed a significant main effect of *Strategy* ($F(1, 49) = 6.86, p = 0.01, \eta_p^2 = 0.12$) on memory performance (Figure 27), suggesting better memory for retrieval practice than restudy (Cohen's $d = 0.38$ [95% CI: -0.01-0.77]),

mirroring the results of the w group. This analysis was repeated with the trimming procedure. A significant effect of *Instruction* emerged between the 2 and 1.5 SD cut-offs, however the effect disappeared at the 1 SD cut-off, suggesting it was not robust (Table 25).

The memory performance 2 x 2 ANOVA was repeated using an extreme groups approach by including participants in the evaluative group with a high score on the STAI (i.e., the upper quartile) and participants in the control group with a low score on the STAI (i.e., lower quartile). Results revealed a significant main effect of *Strategy* ($F(1, 38) = 7.86, p = 0.008, \eta_p^2 = 0.17$ in the w group and also in the S group ($F(1, 16) = 8.75, p = 0.009, \eta_p^2 = 0.35$).

5.11.2 Memory Performance and Anxiety

In the w group, we found violations of Levene's Test ($p's < 0.01$), as well as significant interactions between the grouping variables and covariates in the S group. As such, we performed the same centering procedure as the main analysis for both w and S groups to aid interpretability.

We conducted four 2-way ANCOVAs to determine the effect of *Strategy* (retrieval practice vs. restudy) and *Instruction* (evaluative vs. control) while controlling for scores on all four anxiety measures (CTAS, TAI, Trait Anxiety, GAD). For both w and S groups, *Strategy* consistently emerged as a strong and significant predictor of memory performance for all variables (all $p's < 0.05$, see Table 26), signifying a similar pattern of results.

In the w group, CTAS emerged as a marginally significant predictor ($p = 0.05$), meaning test anxiety levels as measured by the CTAS questionnaire may have moderated performance. In the S group, the covariates TAI, Trait and General Anxiety were all significant predictors (Table 26), suggesting test and trait anxiety levels impacted performance here as well. In the S group, *Instruction* also showed a significant main effect, with the control group scoring higher overall ($M = 0.31$, $SE = 0.02$) than the evaluative group ($M = 0.25$, $SE = 0.02$). Given the the small sample size of this sample, we refrain from in-depth interpretation.

5.11.3 State Anxiety

State anxiety was assessed in the recall session by performing a 2 (instruction: evaluative vs. control) x 4 (task: STAI1, STAI2, STAI3, STAI4) mixed ANOVA. Descriptive statistics are shown in Figure 28 for the w group and Figure 29 for the S group. For the w group, this analysis revealed a significant main effect of *Task* ($F(2, 297) = 5.41$, $p < 0.01$, $\eta_p^2 = 0.04$). Post-hoc t-tests revealed that participants were more stressed at the second time point measurement compared to the first time point measurement, ($t(114) = -3.16, p = 0.02$, Cohen's $d = 0.17$, [-0.43, 0.08]); and at the third time point measurement compared to the first time point measurement ($t(114) = -3.16, p = 0.01$, Cohen's $d = 0.17$, [-0.43, 0.08]).

For the S group, this analysis also showed a marginal main effect of *Task* ($F(2, 116) = 2.91$, $p = 0.048$, $\eta_p^2 = 0.06$). Follow-up analyses showed a similar pattern as the w group, where participants were more stressed at the second time point measurement compared to the first, however the t-test did not reach significance ($t(48) = 2.67, p = 0.06$).

To ensure that differences didn't emerge from participants being more stressed during the initial learning session, we performed a 2 (instruction: evaluative vs. control) by 2 (task: before reading texts vs. after reading texts in session 1) mixed ANOVA. No significant main effects or interactions were found (all $p's > 0.05$) for both w and S groups, suggesting no differences in baseline state anxiety scores depending on whether participants ended up in evaluative or control groups.

5.11.4 Heart Rate

No significant main effects or interactions were found for heart rate (all $p's > 0.05$) for both the w and the S groups, suggesting no differences in the physiological proxy between sessions and conditions (see Table 27 for descriptive statistics). These analysis were also repeated with a trimming procedure consisting of re-running the analysis by restricting from 3 SD to 1 SD from the mean. Results of this trimming procedure did not reveal any significant differences for either group.

5.11.5 Times Read

Descriptive statistics for the amount of times texts were read in the w or S groups are shown in Table 28. Analysis with *t*-tests showed no differences in the amount of times the texts were read across both reading conditions between evaluative and control groups.

5.11.6 Supplementary Tables and Figures

Table 24

Memory performance trimming procedure: w group

SD	Effect	DF	F	p	η_p^2
3	Instruction	(1, 113)	0.91	0.34	< 0.01
	Strategy*	(1, 113)	20.88	< 0.001	0.15
	Instruction × Strategy	(1, 113)	0.13	0.72	< 0.01
2.5	Instruction	(1, 112)	0.56	0.46	< 0.01
	Strategy*	(1, 112)	19.62	< 0.001	0.15
	Instruction × Strategy	(1, 112)	0.26	0.61	< 0.01
2	Instruction	(1, 107)	0.26	0.61	< 0.01
	Strategy*	(1, 107)	22.69	< 0.001	0.17
	Instruction × Strategy	(1, 107)	1.85	0.18	0.01
1.5	Instruction	(1, 86)	0.21	0.65	< 0.05
	Strategy*	(1, 86)	10.88	0.001	0.11
	Instruction × Strategy	(1, 86)	1.01	0.32	0.01
1	Instruction	(1, 58)	0.14	0.71	0.01
	Strategy*	(1, 58)	14.04	< 0.001	0.19
	Instruction × Strategy	(1, 58)	0.23	0.63	< 0.01

Note. Table shows ANOVA results for memory performance at each standard deviation cut-off in the w group. Significant effects for *Strategy* are indicated by "*". F, p, and η_p^2 are also reported.

Table 25

Memory performance trimming procedure: S group

SD	Effect	DF	F	p	η_p^2
3	Instruction	(1, 49)	2.79	0.10	0.05
	Strategy*	(1, 49)	6.86	0.01	0.12
	Instruction × Strategy	(1, 49)	0.06	0.81	< 0.01
2.5	Instruction	(1, 49)	2.79	0.10	0.05
	Strategy*	(1, 49)	6.86	0.01	0.12
	Instruction × Strategy	(1, 49)	0.06	0.81	< 0.01
2	Instruction+	(1, 47)	4.08	0.05	0.08
	Strategy*	(1, 47)	5.48	0.02	0.10
	Instruction × Strategy	(1, 47)	0.20	0.65	< 0.01
1.5	Instruction*	(1, 37)	4.88	0.03	0.12
	Strategy	(1, 37)	3.34	0.08	0.08
	Instruction × Strategy	(1, 37)	1.16	0.29	0.03
1	Instruction	(1, 17)	1.07	0.32	0.06
	Strategy	(1, 17)	0.01	0.91	< 0.01
	Instruction × Strategy	(1, 17)	0.36	0.56	0.02

Note. Table shows ANOVA results for memory performance at each standard deviation cut-off in the S group. Significant effects for *Strategy* and *Instruction* are indicated by "*" and marginally significant effects by the "+". F, p, and η_p^2 are also reported.

Table 26

Summary of ANCOVA results in the w and S groups

Effect	df	F	p	η_p^2
w group				
CTAS (covariate)+	1, 459	3.96	0.05	0.01
Strategy*	1, 459	23.78	< 0.01	0.05
Instruction	1, 459	1.19	0.28	< 0.00
Strategy × Instruction	1, 459	0.06	0.81	< 0.00
TAI (covariate)	1, 459	0.50	0.48	< 0.00
Strategy*	1, 459	23.61	< 0.001	0.05
Instruction	1, 459	1.15	0.28	< 0.00
Strategy × Instruction	1, 459	0.06	0.81	< 0.00
Trait Anxiety (covariate)	1, 459	1.05	0.31	< 0.00
Strategy*	1, 459	23.64	< 0.001	0.05
Group	1, 459	0.61	0.44	< 0.00
Strategy × Instruction	1, 459	0.06	0.81	< 0.00
GAD (covariate)	1, 459	0.08	0.78	< 0.00
Strategy*	1, 459	23.58	< 0.001	0.05
Instruction	1, 459	1.08	0.30	< 0.00
Strategy × Instruction	1, 459	0.06	0.82	< 0.00
S group				
CTAS (covariate)	1, 199	0.23	0.64	< 0.01
Strategy*	1, 199	7.12	0.01	0.04
Instruction*	1, 199	7.75	0.01	0.04
Strategy × Instruction	1, 199	0.06	0.80	0.00
TAI (covariate)*	1, 199	20.09	< 0.01	0.09
Strategy*	1, 199	7.83	0.01	0.04
Instruction*	1, 199	4.33	0.04	0.02
Strategy × Instruction	1, 199	0.07	0.79	< 0.00
Trait Anxiety (covariate)*	1, 199	6.43	0.01	0.03
Strategy*	1, 199	7.34	0.01	0.04
Instruction*	1, 199	7.52	0.01	0.04
Strategy × Instruction	1, 199	0.07	0.80	< 0.00
GAD (covariate)*	1, 199	4.44	0.04	0.02
Strategy*	1, 199	7.27	0.01	0.04
Instruction*	1, 199	7.66	0.01	0.04
Strategy × Instruction	1, 199	0.07	0.80	< 0.00

Note. Table shows ANCOVA results for each anxiety questionnaire for both w and S groups. *Strategy* is significant across all measures and *Instruction* is significant in the S group, indicated by the "*". CTAS showed marginal significance in the w group, indicated by the "+". TAI, Trait Anxiety and GAD were all significant predictors in the S group. F, p, and η_p^2 are also reported.

Table 27

Descriptive statistics for heart rate in the w and S groups

Group	Instruction	Session	Mean Pulse	Std Dev Pulse	Min Pulse	Max Pulse
w Group	Control	Learning	65.97	10.58	36	96
	Control	Recall	70.31	32.85	44	308
	Evaluative	Learning	69.00	12.73	44	108
	Evaluative	Recall	69.47	11.15	44	104
S Group	Control	Learning	65.57	12.54	40	84
	Control	Recall	67.45	12.91	48	100
	Evaluative	Learning	66.00	13.76	40	92
	Evaluative	Recall	76.00	38.87	40	240

Note. Table shows mean, median, minimum and maximum pulse rates for both the w and S groups. The mean pulse rate is reported, along with the standard deviation of the mean, and the minimum and maximum pulse rates.

Table 28

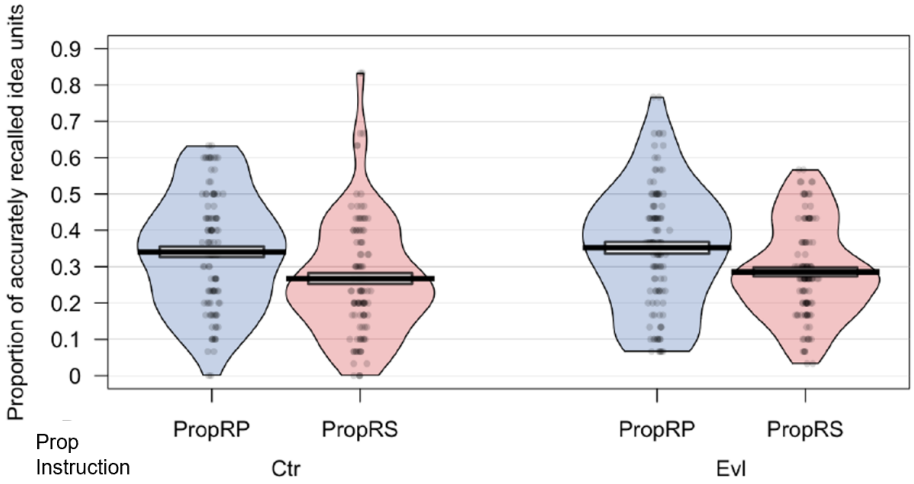
Descriptive statistics for times texts were read in w and S groups

Sample	Instruction	Strategy	M	SD
w Group	Control	Read	5.31	2.23
	Control	Restudy	6.16	8.58
	Evaluative	Read	5.11	1.90
	Evaluative	Restudy	6.53	3.44
S Group	Control	Read	5.66	3.61
	Control	Restudy	6.54	6.66
	Evaluative	Read	5.70	2.39
	Evaluative	Restudy	12.30	37.81

Note. Table shows number of click counts during the initial learning and restudy phase for both evaluative and control groups in the w and S groups. *T*-test analyses showed no significant differences.

Figure 26

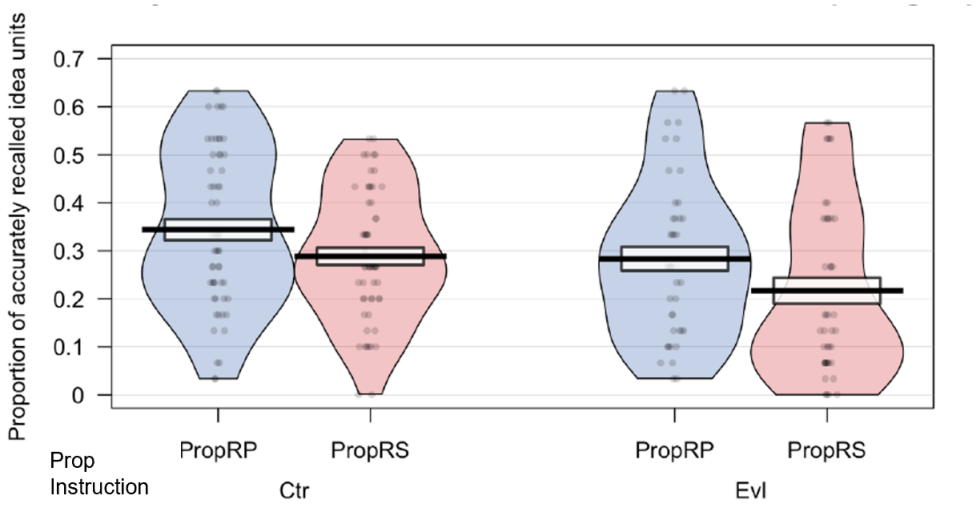
Memory performance between evaluative and control groups in w group



Note. The x-axis represents PropRP and PropRS, proportions of retrieval practice (RP) and restudy (RS). The y-axis represents the proportions of idea units. Ctr is the control group, and Evl is the evaluative group. The black line in the center is the average, the white box denotes standard error (SE). The grey dots represent individual data points.

Figure 27

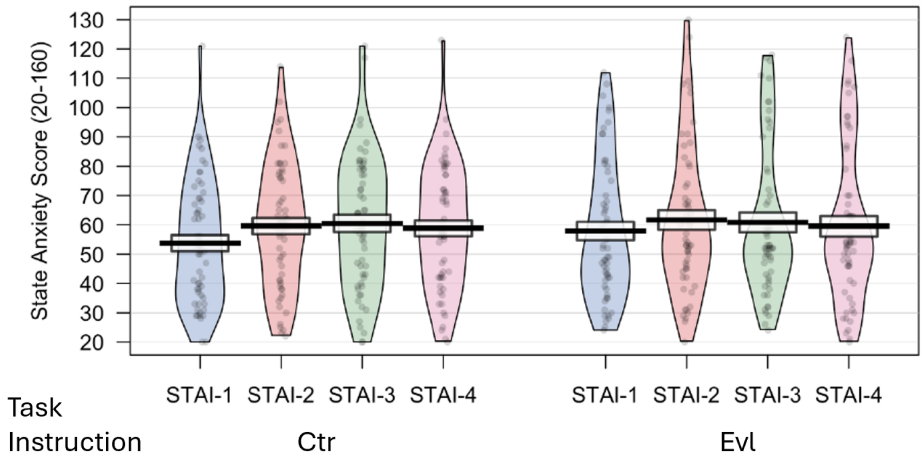
Memory performance between evaluative and control groups in S group



Note. The x-axis represents PropRP and PropRS, proportions of retrieval practice (RP) and restudy (RS). The y-axis represents the proportions of idea units. Ctr is the control group, and Evl is the evaluative group. The black line in the center is the average, the white box denotes standard error (SE). The grey dots represent individual data points.

Figure 28

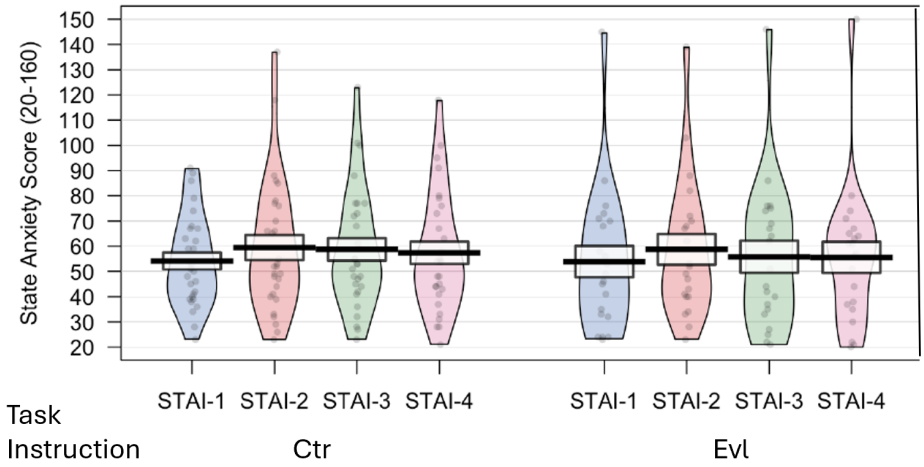
Recall session state anxiety levels in the w group



Note. STAI1, STAI2, STAI3, and STAI4 represent the time points at which state anxiety was measured in the recall session (i.e., after reading the instructions, after the Raven's matrices or the puzzle task, after memory test 1, after memory test 2, respectively). Ctr is the control group, and Evl is the evaluative group. The black line in the center is the average, the white box denotes standard error (SE). The grey dots represent individual data points.

Figure 29

Recall session state anxiety levels in the S group



Note. STAI1, STAI2, STAI3, and STAI4 represent the time points at which state anxiety was measured in the recall session (i.e., after reading the instructions, after the Raven's matrices or the puzzle task, after memory test 1, after memory test 2, respectively). Ctr is the control group, and Evl is the evaluative group. The black line in the center is the average, the white box denotes standard error (SE). The grey dots represent individual data points.

5.11.7 Instructions

Evaluative Group Instructions

General. In this session, we will test your intelligence and your memory abilities. That is, we want to know how intelligent you are and how good your memory is. Because we want to study the relationship between intelligence and memory, it is extremely important that you complete all tasks to the very best of your abilities.

IQ test. The next task is the intelligence test which results in an IQ (intelligence quotient) reflecting how intelligent you are. Because we are interested in your intelligence, it is extremely important that you do not rush through the test, but only select a response when you are absolutely certain that the response is correct. Continue to the next screen to start the intelligence test.

Memory test. In the following task, we will test your memory abilities. Specifically, we are interested in how many facts you can remember from the two texts you read in the first session of this study. You will be given an entire 7 minutes for the recall to make sure we are genuinely testing nothing but your memory abilities without the limitations of time pressure. Because we are interested in your memory abilities, it is extremely important that you use the entire 7 minutes to the very best of your abilities to recall as many facts as possible from the required text. On the next page you will be prompted to recall a text from the first session. The title on the next page indicates which text you must recall. A timer indicating how much time you have left will be presented on the upper right hand side of the screen. Continue to the next screen to start the memory test.

Control Group Instructions

General. In this session, you will do a puzzle task and a memory task. That is, you will have to search for missing pieces and you will have to remember things. Because we want to study the relationship between problem solving and memory, it is extremely important that you try to avoid guessing.

Puzzle task. The next task is the puzzle task which is used to assess your problem-solving strategies. Because we are interested in problem solving strategies, it is extremely important that you do not rush through the test, but prioritize accuracy over speed. Continue to the next screen as soon as you are ready to continue with the puzzle task.

Memory task. In the following task, we will ask you to remember things. Specifically, we are interested in how much you can remember from the two texts you read in the first session of this study. You will be given 7 minutes for the recall to make sure you have enough time to write down everything you remember. Do not worry, if you find this task difficult. We know that it is not an easy task. Nevertheless, because we are interested in how much can be recalled when people try to the best of their ability , it is extremely important that you use the entire 7 minutes to recall as much as possible from the required text. On the next page you will be asked to recall a text from the first session. The title on the next page indicates which text you must recall. A timer indicating how much time you have left will be presented on the upper right hand side of the screen. Continue to the next screen to start the memory task.

6. General Discussion

This thesis investigated the potential of scientifically-backed learning strategies–feedback, spacing, retrieval practice, and multisensory learning–to optimize memory and learning in two under-explored, real-world contexts, namely mobile learning and stress. Study 1 found that all of these strategies improved memory when mobile learning applications were used, but retrieval practice had the greatest impact on performance. In Studies 2 and 3, we then zoomed in to determine whether retrieval practice can act as a buffer against stressful situations. As such, the current discussion section focuses mainly on the effects of retrieval practice. The cumulative evidence in Study 2 showed for the first time that retrieval practice may provide an advantage over restudy when learning under stressful conditions. Likewise, Study 3 demonstrated that retrieval practice is effective in online learning settings and remained effective regardless of evaluative or control settings. This section discusses the research questions (section 6.1), integrates their findings (section 6.2), and raises important questions and recommendations for future research (section 6.3).

6.1 Discussion of Research Questions

6.1.1 Study 1: What is the effectiveness of mobile applications built solely for learning purposes, and how do established learning strategies contribute to their effectiveness?

Study 1 conducted a meta-analysis to evaluate the effectiveness of mobile language learning applications and found that they had a moderate-to-strong effect on memory and learning performance ($g = 0.88$, 95% CI [0.62–1.14]). Unlike prior research, this study assessed applications specifically designed for learning, which allowed their impact to be isolated from other features found in digital tools (Abdulrahaman et al., 2020; Bano et al., 2018), and highlighted their use of learning strategies like retrieval practice, feedback, spaced learning, and multisensory learning. These results provided clearer evidence of the

educational value of MALL-applications compared to traditional methods.

A key objective of Study 1 was to examine whether scientifically-backed learning strategies were implemented in MALL-applications. Retrieval practice emerged as the front-runner, with the highest effect size ($g = 0.95$, 95% CI [0.56, 1.34]) for learning, suggesting a strong benefit in mobile learning contexts. High effect sizes have been observed in mobile applications that feature self-testing and quizzing through flashcards and vocabulary learning (Abdollapour & Maleki, 2012; Sandberg et al., 2011; Q. Wu, 2014). This aligns with a plethora of empirical evidence highlighting the benefits of retrieval practice for learning (Agarwal, D'Antonio, Roediger, et al., 2014; Agarwal et al., 2012; Butler et al., 2014). Spaced learning was also widely implemented in all applications, suggesting that it is a key concern when designing mobile learning applications. Most learning applications explored in Study 1 presented the same vocabulary words in-between other words (spaced) rather than presenting the same words consecutively (massed) (Q. Wu, 2015). They also implemented features, like daily practice reminders, scheduled review sessions, and progressive unlocking of new content (Chen & Chung, 2008; Zou & Xie, 2018), all of which naturally incorporate spacing to encourage users to revisit material over time rather than cramming in a single session. Feedback and multisensory strategies also showed strong effects on learning outcomes ($g > 0.80$), suggesting that they are also drivers of performance in mobile applications. Adaptive feedback systems and multisensory elements enhance engagement and cognitive processing and accommodate diverse learning styles, improving retention and focus in mobile contexts (Shadiev et al., 2018; Troussas et al., 2022).

However, despite their promise, none of the studies actually manipulated the learning strategies themselves. In other words, the strategies were used by chance in the applications and were not the main focus of the applications. So far, a direct manipulation of these learning strategies has only been attempted in (Belardi et al., 2020), who evaluated the impact of all four learning strategies on vocabulary learning outcomes using

a self-developed web application. They found significant effects for spacing and feedback, as well as for a combination of spacing, corrective feedback, and testing, compared to massed learning and testing without corrective feedback. Similarly, Rüth et al. (2021) examined the effects of two feedback types—standard corrective feedback and enriched feedback with additional information—in quiz applications in lab and mobile settings, and they found that both feedback types improved cognitive outcomes in the short and long term. These studies were conducted on consumer-grade computers using a web-based or classroom interface rather than a mobile learning context, making their contributions relevant but still limited in addressing the specific context and unique challenges of mobile learning environments.

Overall, Study 1 showed that learning strategies can improve performance in mobile learning contexts and suggested that retrieval practice offers the most benefit compared to spacing, feedback, and multisensory strategies. This finding highlights the potential of retrieval practice as a key mechanism in optimizing learning in mobile contexts and the need to further explore and integrate this strategy into other contexts where it could also be beneficial.

6.1.2 Study 2: Is there evidence supporting the cumulative benefits of retrieval practice on memory in stressful contexts, and what is the effect size of the protective benefits of retrieval practice on memory in these settings?

As seen in Study 1, retrieval practice is a promising strategy for investigation into other underexplored contexts, such as stress. Retrieval stress is known to decrease memory, leading to memory impairments (Klier & Buratto, 2020; Roozendaal, 2002; Shields, Sazma, et al., 2017), whereas retrieval practice has been shown to protect memory from this impairment (A. M. Smith et al., 2016). Study 2 builds on A. M. Smith et al. (2016) by examining whether there other evidence in the field to support a benefit of retrieval practice in stressful settings.

The primary hypothesis in Study 2 was that participants who engaged in retrieval

practice would perform better under stress than those who were stressed but did not use retrieval practice. This hypothesis was supported by a medium-sized positive effect (g = 0.45, 95% CI [0.19, 0.71]), suggesting that retrieval practice is more effective than other strategies (in this case, only restudy) under stressful conditions. Additionally, Study 2 confirmed that the testing effect was still robust with a small-to-medium effect of $g = 0.37$, 95% CI [0.09, 0.66]. These results suggest a cumulative benefit of retrieval practice over restudy during stressful and non-stressful situations.

Our meta-analysis did not find a significant difference in the effectiveness of retrieval practice between stress and non-stress situations (g = -0.08, 95% CI [-0.02, 0.19]). This means that although retrieval practice is more effective than restudy under stress, memory performance is not different between stress and non-stress conditions when retrieval is used. These results are similar to those observed in Szőllősi et al. (2017), where relatively equal performance in stress and non-stress conditions for retrieval practice was observed, meaning that stress did not decrease its effectiveness. Interestingly, we observed a similar pattern of results when examining the effects of control strategies, or restudy, during stress and non-stress conditions. This analysis showed another null effect of g = -0.05, 95% CI [-0.15, 0.05], which suggests that there was no significant difference in memory performance between stress and non-stress conditions for participants who learned using a control strategy.

We did not observe a significant effect of the stressor via our regression or moderator analysis. This was unexpected given the large number of past studies showing that stress impairs memory, especially when no specific strategies were used (de Quervain et al., 1998, 2000; Gagnon & Wagner, 2016; Gagnon et al., 2019b; Kuhlmann, Piel, & Wolf, 2005; Roozendaal, 2002; Shields, Sazma, et al., 2017). This discrepancy may be due to the reliance in our study on self-reported stress scores, which varied across the different questionnaires and study timings, and the absence of cortisol measurements, which could have reduced the sensitivity of these results. Moreover, different types of stressors may have

varying effects on memory. Notably, test anxiety does not always lead to decreased memory (Clark et al., 2018; Hinze & Rapp, 2014). Finally, we acknowledge that the absence of observed stress effects in this study does not imply the nonexistence of such effects.

Overall, the results obtained in Study 2 suggest that retrieval practice is more beneficial than restudy under both stress and non-stress conditions. This means that retrieval practice provides a more stable and consistent benefit to memory performance, as opposed to protecting it. Importantly, the magnitude of the stressor has yet to be clarified because the stress scores did not have a significant effect, and moderator analyses showed no significant effects for the type of stressor induced. Lastly, insufficient statistical power could have limited the ability to detect or overestimated certain effects in this meta-analysis. Study-level heterogeneity, including differences in retrieval practice protocols and stress procedures, insufficient power across the studies, and the mixed publication bias findings likely contributed to noise, making it difficult to confidently confirm the effects.

6.1.3 Study 3: Can retrieval practice protect memory against stressors like test anxiety?

One takeaway from Study 2 was that higher-powered studies are required to reliably determine the protective effects of retrieval practice on memory and stress. Study 3 went a step further by using a sufficiently powered sample in the full sample to examine whether retrieval practice improved memory after inducing test anxiety.

An important finding in Study 3 was that the testing effect can be replicated successfully in an online setting, thereby confirming its value for memory and learning in online situations. Moreover, participants in the online portion of the study performed better than those in the lab when sample was added as a factor in the online versus lab sample analysis, further supporting retrieval practice as an effective tool in these environments.

Study 3 also suggested that test anxiety induced by evaluative instructions may be somewhat more effective in in-person settings than in online settings, although the

difference was not statistically significant in the online versus lab factor analysis. Notably, the stress levels increased in both the evaluative and control groups in the full sample following the matrix task. These results contrast with Almazrouei et al. (2022), who successfully induced stress in an online setting. Their task used instructions paired with explicit negative feedback (e.g., "WRONG" in red or "TIME OUT!") and negative social comparisons. Our study relied solely on the evaluative aspect of "taking a test" without negative feedback or comparisons. Our stressor may have been mild enough that it affected participants regardless of evaluative or control condition, possibly due to the difficulty of the subsequent matrix task and resulting cognitive load rather than the evaluative instructions alone. Importantly, we note that although the state anxiety scores increased in both groups in the full sample, they still remained in the low range, suggesting that the induced anxiety was mild and may not have been sufficient to produce memory impairments.

Another key finding was that retrieval practice improved memory under both evaluative and control conditions, meaning we cannot conclusively determine whether retrieval practice protects memory against test anxiety, especially given the mild impact of evaluative instructions on state anxiety levels. This matches the findings of Szőllősi et al. (2017) and Pastötter et al. (2023), who found that retrieval practice improved memory regardless of the stress condition. However, it contrasts with Hinze and Rapp (2014), who reported a reversal of the testing effect under high-pressure conditions. Methodological differences may explain these discrepancies. For instance, Hinze and Rapp (2014) manipulated high- and low-stakes testing during initial learning rather than during retrieval and used multiple-choice tests, which are less sensitive to retrieval practice benefits (Larsen, 2018) than the free recall test used in our study.

6.2 Integrative Discussion

To summarize, Study 1 identified retrieval practice as the most effective strategy to enhance performance in mobile learning contexts. Studies 2 and 3 then focused further on

elucidating the benefits of retrieval practice in other types of non-traditional contexts, namely stressful contexts. Study 2 demonstrated that retrieval practice is robust in the face of stressors, and Study 3 showed its efficacy in online learning environments and in both evaluative and control conditions. In this integrative discussion, I explore overarching themes between the studies, revisit mechanisms of action, and identify directions for future research.

6.2.1 Common Themes Across Studies

Theme 1: Adaptability of retrieval practice. This thesis builds on earlier studies and demonstrates the effectiveness of retrieval practice in mobile learning and stressful contexts, both of which are underexplored domains. The effectiveness of retrieval practice in these areas highlights its adaptability and applicability across diverse learning environments. Study 1 found a large effect size for mobile applications featuring retrieval-based learning activities, indicating that retrieval practice can optimize learning outcomes in fast-paced digital settings. Study 3 then confirmed the effectiveness of retrieval practice in online settings, adding to a growing body of evidence supporting its use for online learning (Cadaret & Yates, 2018; Simone et al., 2023). Both Studies 2 and 3 demonstrated the robustness of retrieval practice under stressful and evaluative conditions. All these results align with the broader literature consistently showing strong effects for retrieval practice across a variety of settings and across different learning materials (Adesope et al., 2017; Agarwal et al., 2021; Moreira et al., 2019; Schwieren et al., 2017).

Theme 2: Retrieval practice as a stable optimizer. This thesis was inspired by the findings of A. M. Smith et al. (2016) and further investigated whether retrieval practice could protect memory against the detrimental effects of stress. Only a subset of the results in Study 3 went in this direction, where a positive relationship between having high levels of test anxiety and the testing effect was observed in lab and evaluative situations. However, the exploratory nature of these results precluded us from confirming the protective effects of retrieval practice.

The rest of the results in this thesis point to retrieval practice being more of what I'm calling a "stable optimizer" of memory rather than protector. By "stable optimizer," I mean that retrieval has consistent beneficial effects on memory regardless of other factors. In Studies 2 and 3, retrieval practice consistently outperformed restudy under non-stress conditions (Study 2, H2), stress conditions (Study 2, H3), and both evaluative and control conditions (Study 3). Moreover, retrieval practice was unaffected by the various moderators in Study 2 and remained effective when test anxiety and general anxiety were controlled using ANCOVA in Study 3. A key caveat was that the effects of the stressors could not be confirmed in either study, though this does not mean that no effect exists. These findings indicate that retrieval practice provides consistent benefits under various emotional conditions and settings rather than selectively buffering against stress or anxiety, even though our attempts to manipulate stress may not have strongly impacted performance.

Although Study 1 did not directly examine stress or anxiety, mobile learning environments could involve stress-inducing factors, such as time constraints, interruptions, and evaluative contexts (Gonthier, 2023; Kumari & Corr, 1998). Given its potential effectiveness in mobile settings observed in Study 1, retrieval practice may continue to enhance memory performance even when mobile learning conditions involve stressful or evaluative situations. This possibility remains to be tested in future research, as discussed later.

Theme 3: Impact of stressors. Another theme encountered in the studies of this thesis is the nuanced interplay of learning strategies, stress, and anxiety. First, we did not find strong evidence in Study 3 that evaluative instructions had a considerable impact on participant stress levels, which contrasts with the results in (Almazrouei et al., 2022) and suggests that the intensity of the stressor may be a key factor. Participants in the lab sample in Study 3 showed numerically, but not significantly, higher stress ratings in the evaluative condition. This finding is partially consistent with the well-established TSST procedure, which is typically performed in in-person settings where it elicits effects likely

due to physical proximity and higher perceived intensity (Dickerson & Kemeny, 2004; Kirschbaum et al., 1993). In Study 2, although most of the included studies featured a variant of the TSST protocol and reported it to be successful, the moderating effect of stress on memory was not evident overall. Individual differences, such as cortisol reactivity, may have influenced stress sensitivity here (Klier & Buratto, 2023; Szőllősi et al., 2017).

The effect of the stressor is further complicated when learning strategies are considered. When control strategies (in our case, restudy) were used, stress did not significantly impact memory (Study 2, H1). This finding is inconsistent with those from the general stress literature that show retrieval stress negatively impacts memory (Klier & Buratto, 2020; Kuhlmann, 2005; Oei et al., 2006; Roozendaal, 2002; Smeets, 2011). In those studies, learning materials are not necessarily systematically studied or restudied (i.e., participants may only see the learning materials once), while in our Study 2, participants from the included studies were explicitly instructed to restudy the materials, sometimes over multiple cycles. Studies have shown that when learners have metacognitive control of the items they need to restudy (like when they are aware they are using a restudy strategy), they perform better (Tullis & Benjamin, 2012). Therefore, repeated restudy, while less effective than retrieval practice overall, may still offer a baseline level of memory support by reinforcing encoding through multiple exposures, an effect that has been shown to be effective compared to little or no study (Storm et al., 2008).

Similarly, in Study 2 (H4), there was also no significant difference in memory performance between stress and non-stress conditions when retrieval practice was used. Therefore using retrieval practice yields similar outcomes regardless of whether stress was present. These results align with the growing consensus that retrieval practice could be robust regardless of stress: Szőllősi et al. (2017) found that participants performed equally well when they used retrieval practice under both stress and control conditions; Pastötter et al. (2023) showed that the forward testing effect is unaffected by stress; and a review by Yang et. al (2020) argued that retrieval practice is independent of stressors like test

anxiety. In addition, these results align with the findings in Study 3 showing that retrieval practice was effective under both evaluative and control conditions, which means that its effects can be generalized to evaluative situations.

Together, these results raise the question of how much stress actually impairs memory when learning strategies are used. Specifically, they suggest that the effectiveness of the strategy employed (namely, retrieval practice) may be a more critical determinant of learning outcomes than the presence or absence of stress.

Theme 4: Impact of learning material. A fourth theme emerging from these studies is the interplay between retrieval practice and the type of learning material. Most of the mobile application learning materials featured vocabulary words in Study 1, and most of the articles examined in Study 2 featured word lists or short answer questions. These stimuli showed small-to-medium positive effect sizes in both studies, which aligns with other evidence demonstrating that retrieval practice benefits this type of learning material (Klier & Buratto, 2023; Pastötter & Frings, 2019; A. M. Smith & Thomas, 2018b; A. M. Smith et al., 2016).

Inconclusive results were observed for longer texts and educational reading materials. These materials showed a clear testing effect in Study 3, aligning with previous research (Emmerdinger & Kuhbandner, 2019; Roediger & Karpicke, 2006), but smaller effects in all moderator analyses for all hypotheses in Study 2. A. M. Smith (2018) suggested that in the presence of stress, the effectiveness of retrieval practice may decrease as a function of stimulus difficulty, particularly for stimuli that require the binding of information (i.e., texts) as opposed to those that can be learned as individual units (i.e., words). Individual unit stimuli, especially when learned over multiple retrieval trials, can lead to automation, which increases retention (Racsmány et al., 2020), whereas materials that require binding demand content integration, a process that A. M. Smith (2018) suggests becomes disturbed during stress.

However, this explanation is not consistent with the episodic context theory

(Lehman et al., 2014). It is also inconsistent with evidence showing that retrieval practice *improves* recall of more difficult items during stress (Klier & Buratto, 2023) and with findings showing that making multiple retrieval attempts can eliminate the detrimental impact that retrieval stress typically has on memory retrieval (A. M. Smith et al., 2018) (although these items were, again, word lists). Moreover, reading texts and other complex materials is closer to what we do in real life (i.e., reading emails, articles online, ingredient lists of food items in stores), with evidence showing that this type of material is not subject to the memory-impairing effects of stress (Stock & Merz, 2018). In addition, experts and scholars have suggested that materials requiring more cognitive engagement (e.g., essays and longer texts), rather than word lists or individual items, appear to be the most effective way to use retrieval practice. This is because more complex materials require learners to construct and retrieve more information (Larsen, 2018), which then improves memory consolidation and retention (Antony et al., 2017).

Theme 5: Impact of retention interval. Retrieval practice becomes more effective after longer periods (Roediger & Karpicke, 2006), aligning with the results observed in Study 3. Study 2, however, showed conflicting results in this regard. Namely, a significant difference was found between a two-day and a one-week delay between learning and final recall when retrieval practice was used in stress versus non-stress conditions (H4). In contrast, when retrieval practice was compared against restudy in non-stress situations (H2), the typical one-week advantage for retrieval practice returned. A. M. Smith (2018) also observed that retrieval practice may be more beneficial for memory in the short term rather than the long term when stress is involved. This would support the moderator pattern in Study 2 but clashes with Study 3, where retrieval practice continued to be effective even after a one-week delay regardless of the evaluative condition. Furthermore, using delay as a continuous moderator in Study 2 revealed that memory performance increased with longer delays when retrieval practice was compared against restudy in both non-stress and stress conditions (H2 and H3). However, the same trend was not observed

when retrieval practice alone was compared in stress versus non-stress conditions (H4). Therefore, its benefits may not actually decrease in the presence of stress as A. M. Smith stated. This is also in line with meta-analytic work showing that retrieval practice is effective regardless of the delay to the final test (Adesope et al., 2017).

6.2.2 The Testing Effect Revisited

The results of Studies 2 and 3 align with the broader retrieval practice literature replicating the testing effect (Karpicke, 2017; Roediger & Karpicke, 2006; Rowland, 2014). Testing effects are typically more noticeable in cued and free recall tests (Karpicke et al., 2014), as we observed in Study 3, likely because retrieval practice strengthens the recollection processes (Weinstein et al., 2010). Building on this literature, we replicated the testing effect in stress and anxiety situations and demonstrated its robustness in these settings.

The observed testing effect in Study 2 ($g = 0.37$) was slightly lower than typical estimates, such as $g = 0.499$ reported by Yang et al. (2021) in a large-scale meta-analysis involving over 48,000 participants. Our testing effect also showed high heterogeneity and was not consistently replicated across individual studies, indicating variability at the study level in Study 2. One reason behind the heterogeneity may be that certain studies in Study 2 used recognition tests (e.g., A.M. Smith et al., 2019), where the testing effect tends to be weaker (Karpicke et al., 2014) because recognition relies more on familiarity (Weinstein et al., 2010). Another reason may be due to the inconsistent use of retrieval practice paradigms, altering the number of retrieval trials, spacing, or study-test intervals.

For instance, some of the studies in Study 2 included multiple retrieval cycles but allowed participants only a few seconds to encode the material. This short encoding time potentially limited encoding and resulted in unclear testing effects following stress (A. M. Smith et al., 2018, 2019). In Klier and Buratto (2023), participants alternated between restudy and retrieval practice with short encoding and recall windows (6 seconds per word pair, 4.5 seconds for retrieval), which led to a testing effect only for difficult

items. Similarly, Tse et al. (2019) found that retrieval and restudy produced similar performance when participants studied facts for 11 seconds but were tested with 8 seconds for recall and 3 seconds of feedback. A. M. Smith et al. (2019) found no difference in source memory between retrieval practice and restudy when participants had 2 seconds to study each word, followed by either two additional 2-minute restudy sessions or two 2-minute free-recall tests. Then, in cases when participants had 5 seconds to learn and 8 seconds to either restudy or retrieve but did this over 6 cycles like in Szőllősi et al. (2017), a clear testing effect emerged.

Consequently, if retrieval practice is not implemented correctly or consistently, it may not offer significant advantages over restudy. In fact, the testing effect can be reversed when the initial learning criterion is too low (Racsmány et al., 2020; Storm et al., 2014) and, in certain situations, repeated, spaced restudying can even outperform retrieval practice (Higham et al., 2022, 2023). This trend was also observed in certain analyses included in Study 2 (e.g., A.M. Smith et al., 2019 and Tse et al., 2019) where participants showed higher recall for items learned with restudy.

These findings highlight two key takeaways: (1) when retrieval practice is applied correctly, it yields strong and stable long-term retention, but (2) if alternative retrieval practice protocols are used, they may not always be more beneficial than restudy. In this latter case, it may be advisable to enhance these protocols with other evidence-based learning strategies to maximize their effectiveness. One such strategy could be successive relearning, which combines repeated retrieval with spacing and corrective feedback. Successive relearning has been shown to improve retention over time (Rawson & Dunlosky, 2011a, 2013, 2022). Research shows that integrating retrieval with these other strategies strengthens memory by helping to preserve and enhance the testing effect, even under conditions where retrieval alone may fall short (Racsmány et al., 2020; Rawson & Dunlosky, 2022). Scholars also encourage the use of retrieval practice together with spacing and feedback to improve retention and help learners retain information for longer periods

(Larsen, 2018).

6.2.3 Mechanisms of Action Revisited

This thesis argues that episodic context and context-shift theories explain the benefits of retrieval practice. Retrieval practice strengthens memory traces through contextual binding, making them less susceptible to contextual disruptions, such as stress (Lehman et al., 2014; Shields, Sazma, et al., 2017). These theories can explain the benefits of retrieval practice in both stress and non-stress contexts, as observed in Studies 2 and 3. However, because the effect of the stressor was not fully supported in either study, this theory does not fully explain our findings.

Our findings could alternatively be explained by the transfer-appropriate processing theory (TAP). According to TAP, retrieval practice is effective because it mimics the cognitive processing that is needed for the final test. This is supported by a meta-analysis showing that the testing effect is more robust when the format of the practice test matches the format of the final test (Adesope et al., 2017). Study 3 further supports this theory, as a clear testing effect was observed when participants engaged in free recall during both the initial learning and the final test. Similarly, the results reported in the individual studies in Study 2 were similar when the same format was used for learning and final testing (e.g., Klier & Buratto, 2023). The TAP framework could also explain the findings in Study 1, as participants were often tested in a format similar to how they learned the materials on the mobile application.

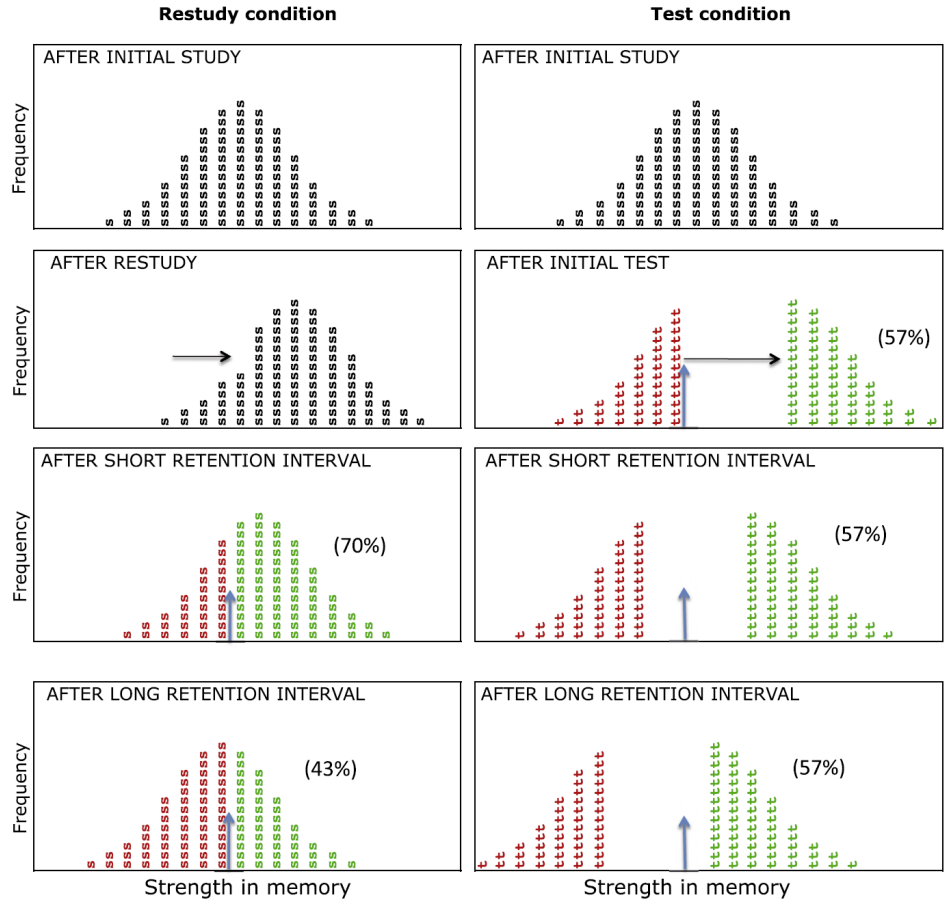
In situations where no testing effects are observed, the bifurcation model (Figure 30) could offer an explanation. In this model, retrieval practice strengthens memory by selectively enhancing successfully recalled items, while items that fail to be recalled remain weaker, similar to unpracticed items (Kornell et al., 2011). In Study 2, if some retrieval attempts were unsuccessful in the individual studies (for example, due to insufficient encoding), the items unsuccessfully recalled would not have been strengthened and might have remained at a lower level of memory strength, which could explain why some studies

didn't find a testing effect. Moreover, not all studies included feedback during learning, which means the retrieval attempts would not have been corrected, limiting the benefits of testing to the few items that are successfully retrieved (Mundt et al., 2020; Racsmány et al., 2020). This further suggests the added benefit of incorporating feedback and spacing with retrieval practice.

In terms of stress, the results seen in this thesis can be explained by the transactional model of stress (Lazarus, 1984; Lazarus & Folkman, 1987). This model emphasizes that stress responses depend on how individuals appraise a situation and their perceived ability to cope with it. Since stress varies by context and individual perception, it may explain some of the findings we saw in our Studies 2 and 3. For instance, some studies included in Study 2 used the Cold Pressor Test to induce stress (Hupbach & Fieman, 2012), which, while physically uncomfortable, can still be appraised by participants as manageable, thus reducing its stress impact. Also, the evaluative manipulation in Study 3 may not have been perceived as something participants couldn't cope with, reflecting the generally low levels of state anxiety ratings we found. Similarly, participants in Study 3 may not have felt like the experimental procedure was out of their control or like they were judged too harshly, both known triggers of psychological stress (Dickerson & Kemeny, 2004). Thus, overall, the relatively low stress responses observed in Studies 2 and 3 may be attributed to participants' appraisals of the situations as non-threatening or manageable.

Figure 30

The bifurcation model of retrieval practice



Note. The bifurcation model for items previously restudied (left column) or previously tested (right column). Initially, both distributions are identical and normal. After intervention, restudied items uniformly increase in strength, while tested items show a bifurcated pattern where retrieved items gain more strength and unretrieved items show no gain. The vertical line marks the recall threshold. After short and long delays, all items decay at the same rate, but the bifurcated test distribution maintains more items above threshold, preserving recall performance. Note and figure adapted from Kornell et al. (2011).

6.3 Implications and Future Directions

As discussed in earlier sections of this chapter, the interplay among mobile contexts, stress, learning strategies, and memory is complex and leaves open questions. In this section, I outline key theoretical and practical implications, suggest future research directions, and make methodological recommendations.

6.3.1 For Retrieval Practice and Mobile Learning

Study 1 showed the importance of integrating key learning strategies into mobile applications to enhance long-term retention. Although each learning strategy is individually effective, how they interact remains poorly understood. Therefore, it will be worth investigating how these learning strategies work together both in lab settings and within mobile learning environments. For example, retrieval-based learning and spaced repetition are both well-established strategies (Carpenter et al., 2022; Weinstein et al., 2018), but their combined effects in a mobile context remain underexplored. Similarly, quizzing platforms like Kahoot!, Quizizz, and Quizlet enhance retrieval practice by allowing self-paced engagement, immediate feedback, and varied question formats, which support learning achievement (Johnson, 2018). However, the impact of different feedback types, such as immediate versus delayed and explanatory versus simple correctness feedback, is not well-understood in mobile learning. And, although feedback is generally effective, its benefits may not always be amplified by retrieval practice (Roediger & Butler, 2011) so further investigation into these interactions is warranted.

Similarly, retrieval practice appeared especially helpful in mobile learning, likely because it promotes self-regulated learning. Yang et al. (2017) found that retrieval practice helped learners maintain sustained study effort in self-paced settings; without interim tests, study time declined across materials, whereas testing maintained engagement. Sustained self-regulation and motivation are key for mobile learning (Viberg & Grönlund, 2012), and future studies should investigate how retrieval practice can be tailored to maximize these features in mobile and on-the-go learning environments. Also, checking if these

self-regulation and motivation carry over to the backwards testing effect to determine if the same rules apply would also be worthwhile.

Large-scale data collected over long periods through MALL and mobile learning presents a unique opportunity to explore these aforementioned interactions comprehensively (Belardi et al., 2020; Weinstein et al., 2018). Mobile learning platforms generate vast amounts of user data, including engagement patterns, quiz performance, time spent on tasks, and frequency of application usage, which can easily be analyzed for insights using machine learning models. This data could be leveraged to understand how learning strategies could fare during stressful situations in mobile applications, opening yet another area of research.

For example, mobile learning applications could leverage AI-driven analytics to track user performance under varying levels of stress, such as before an important exam or during workplace training involving tight deadlines. Machine learning models could then detect or predict when stress negatively impacts learning and adapt the learning experience accordingly. This could involve adjusting retrieval difficulty levels and providing tailored feedback (e.g., explanatory versus correctness-based feedback) before high-stakes learning situations. It would be interesting to see which learning strategies remain effective under stress so that we could explore how mobile applications could best support learners in high-stress environments. Insights from such studies may help educators and developers create more personalized learning experiences that respond dynamically to user needs.

6.3.2 For Retrieval Practice and Stress

The results of Studies 2 and 3 suggest that retrieval practice is more effective than restudy in stress and evaluative settings. This has direct implications for classroom environments, where students frequently experience stress, like during exams, presentations, or performance evaluations. Yet, we know that most students still prefer to use weaker strategies like highlighting and restudying even when they see better results using retrieval practice (Karpicke, 2012). Thus, two things need to happen in the future: 1)

retrieval practice will need to be integrated into classrooms not just as a general learning tool but also during evaluative situations; and 2) awareness of the advantages of retrieval-based strategies will need to be increased to encourage their use in academic and real-world learning contexts. This awareness-raising could be achieved by training teachers and improving scientific communication to the public.

Study 3 observed a benefit of retrieval practice for individuals with high levels of test anxiety compared to those with low levels specifically in lab settings. Though exploratory, this finding means retrieval practice could help "level the playing field" in learning environments by offering a way to support learners who would be most vulnerable to stress-related performance impairments. This potential has already been observed in lower-ability students, where retrieval practice has been shown to benefit their memory outcomes (Agarwal et al., 2017; Yang et al., 2020). As prior studies have also demonstrated, retrieval practice can help reduce test anxiety (Agarwal, D'Antonio, Roediger III, et al., 2014; Brown & Tallon, 2015), but further research is needed to clarify its effects on memory and performance in more extreme or at-risk populations.

Different retrieval practice paradigms were used in Study 2, which produced inconsistent testing effects. Thus, an important avenue for future research is to fine-tune and standardize retrieval practice protocols for both stress and non-stress situations. Based on the results from Study 3, we know that the classic retrieval practice paradigm (Roediger & Karpicke, 2006) produces the testing effect when individuals experience evaluative pressure. Cued recall tests like those described in Szőllősi et al. (2017) are also a possibility. Thus, replicating these studies will help improve the reliability of these effects.

Similarly, although not explored in detail in this thesis, retrieval practice exhibits a dose-dependent relationship whereby after three successful retrievals, additional repetitions yield diminishing returns, particularly when the tests are closely spaced (Rawson & Dunlosky, 2011b). Examining how this relationship functions during stressful situations in a standardized retrieval practice paradigm is now necessary to determine how much

retrieval is actually needed and whether this amount varies in stressful situations. Then, extending this set-up to examine highly anxious participants would allow us to know whether the same amount of retrieval practice is needed for these populations.

Although validated stress protocols were used in the papers included in Study 2 and used in Study 3, their effects on memory were not confirmed, raising questions about how different stressors interact with retrieval practice, if at all. This opens up a need for a more detailed look at the intensity and duration of stress required to impact memory when retrieval practice is used. For example, a replication of our Study 3 with a more pronounced stress procedure (such as that in Almazrouei et al. (2022) may yield clearer insights about how intense the stressor would need to be to have a full effect in an online sample. While stress enhances the backward testing effect in that it improves long-term retention after a delay (A.M. Smith et al., 2016), it does not significantly influence the forward testing effect (Pastötter et al., 2023). Thus, future studies need to better investigate how both of these retrieval practice effects might play out with different types of stressors (i.e., test anxiety) and learning materials.

Despite its robustness, retrieval practice is a mixed bag when it comes to individual differences. For example, individual differences like cortisol reactivity, anxiety levels, working memory capacity, and hormone usage, can all influence stress reactivity (Shields, Sazma, et al., 2017). However, such differences do not always seem to influence retrieval practice (Bertilsson et al., 2021; de Lima & Buratto, 2024; Yang, Li, et al., 2023; Yang et al., 2020). It would be interesting for future studies to take a closer look at the interplay among these factors and their impact on memory performance in both stress and non-stress situations. Moreover, certain studies have shown that up to 30% fail to demonstrate any benefit of retrieval practice (Brewer & Unsworth, 2012; Minear et al., 2018), with this subsample performing even better on various cognitive tasks than participants who showed a testing effect (Minear et al., 2018). It would be important for the field to understand in which conditions participants do *not* benefit from retrieval

practice and how people under those conditions fare in terms of memory during stress.

Retrieval practice has been shown to improve memory deficits in individuals following traumatic brain injury (Coyne et al., 2015; Pastötter et al., 2013), as well as in patients with multiple sclerosis (Sumowski et al., 2010b) and even in patients with schizophrenia (Jantzi et al., 2019). Given our preliminary findings that retrieval practice can support memory during situations where stress or evaluation is expected, a key next step is to investigate whether these benefits can apply to other conditions that impair memory. For example, aging is associated with memory declines, and research suggests that certain encoding strategies could be used to help aging memory (Hering et al., 2014). However, research into retrieval practice's effects on aging is still scarce. Alzheimer's Disease is another area where memory decline is prominent (Grober et al., 2008), offering opportunities to explore whether retrieval practice can mitigate memory impairments or slow cognitive decline.

Along a similar vein, retrieval practice is beginning to be investigated in populations with attentional and developmental disorders, both of which are highly prevalent in educational settings (van Bergen et al., 2025). For example, retrieval practice has shown a promising benefit for students with attention deficit hyperactivity disorder, or ADHD (Knouse et al., 2016). Minear et al. (2023) found that students with ADHD tend to use fewer deep encoding strategies, suggesting that retrieval practice alone may be insufficient and should be combined with targeted support to improve encoding processes. Similarly, individuals with dyslexia exhibit poorer retrieval performance when responding via typing but perform comparably to typical learners when responding by speech, indicating that processing delays rather than memory deficits underlie their difficulties (Wilschut et al., 2025). Future research should focus on adapting retrieval practice to optimize learning for neurodivergent populations, especially in classroom settings where tailored approaches are essential.

Lastly, another possibility emerging from this thesis is that retrieval practice may be

even more beneficial when it is used with other scientifically-backed learning strategies, like spacing and feedback, in the form of successive relearning (Carpenter et al., 2022; Rawson et al., 2018; Vaughn et al., 2016). Studies show that successive relearning can also reduce anxiety (Higham et al., 2023), meaning that there is reason to believe that these strategies together could further protect memory against the negative effects of stress. Experimental research will be needed to improve our understanding of the combined effects of retrieval with other strategies on memory and whether these combinations perform better than retrieval practice alone in stress and non-stress situations.

6.3.3 For Methodology

This thesis featured two meta-analyses, which provides a unique opportunity to recommend best practices for conducting meta-analyses and individual studies. Accurate and reliable meta-analytic findings have the power to change our understanding of entire fields, potentially leading to significant shifts in policy, decision making, and practice, so their rigorous execution is critical.

For meta-analyses, assessing the risk of bias in the included studies is essential to have a better understanding of the sources of bias within each study and how they might impact results. Including and interpreting confidence intervals is also essential to ensure that interpretations are valid and any uncertainty is taken into account. Careful attention must be paid to statistical power because studies with low power can lead to inflated or inconsistent effect sizes. I recommend conducting power analysis simulations and post-hoc power analyses, as described in Study 2. Assessing issues regarding heterogeneity and publication bias in the meta-analyses via a multiverse approach and sensitivity analysis is also needed to have a better understanding of the variances because the sources of these variances are rarely ever clear-cut. It should go without saying that transparency when reporting the methods, inclusion/exclusion criteria, and results is vital. I encourage everyone to conduct a registered report when it comes to this in order to clearly define in advance how the study will be conducted, reducing researcher bias and increasing the

credibility and reproducibility of the findings.

Meta-analyses are only as good as the studies they include, so I urge researchers to plan accordingly when they conduct their studies and make sure their studies are "meta-analysis ready". This means having an adequate sample size for the anticipated design and effects, reporting all outcomes, and sticking to best research practices throughout (for a review, see Schwabe et al., 2022). Studies should also be transparent in their results reporting and provide clear effect size metrics and confidence intervals so that they can be accurately synthesized in meta-analyses. Lastly, ensuring that raw data files are accessible in open access repositories makes it easier to recalculate effects and retrieve more statistical information if needed.

7. Conclusion

This thesis started out by examining the four most effective learning strategies known today–retrieval practice, spacing, feedback and multisensory learning–in two underexplored contexts critical for everyday life, mobile learning and stressful situations. Study 1 showed that mobile learning applications improve learning compared to traditional methods and that scientifically backed learning strategies are implemented, but not experimentally. Moreover, retrieval practice emerged as the strongest strategy used in these contexts. Coupled with evidence from the stress literature that retrieval practice can also protect memory against stressors, we proceeded to investigate whether it can also improve and protect memory in another underexplored context, stressful situations. We saw that retrieval practice is more effective than restudy in both stress and non-stress conditions (Study 2), as well as in evaluative and control conditions (Study 3). However, the relative impact of the stressor in both Studies 2 and 3 was unconfirmed overall, suggesting that retrieval practice benefits memory consistently regardless of whether stress levels are meaningfully elevated, and highlighting the need for further research to clarify how different stressors influence memory when retrieval practice is concerned. Based on the various sources of bias observed in Study 2, we also considered the possibility that retrieval practice can be fortified using other strong learning strategies for maximal benefit. As such, we return to the beginning aspects of this thesis and conclude that likely a combination of learning strategies, such as retrieval practice, spacing, and feedback, is necessary to fully optimize memory.

8. References

Abdollapour, Z., & Maleki, N. A. (2012). Second Language Vocabulary Acquisition in CALL and MALL Environments and Their Effect on L2 Vocabulary Retention: A Comparative Study. *6*(9), 109–118.

Abdulrahaman, M., Faruk, N., Oloyede, A., Surajudeen-Bakinde, N. T., Olawoyin, L. A., Mejabi, O. V., Imam-Fulani, Y. O., Fahm, A., & Azeez, A. L. (2020). Multimedia tools in the teaching and learning processes: A systematic review. *Heliyon*, *6*(11).

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of educational research*, *87*(3), 659–701.

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The Value of Applied Research: Retrieval Practice Improves Classroom Learning and Recommendations from a Teacher, a Principal, and a Scientist. *Educ Psychol Rev*, *24*(3), 437–448. https://doi.org/10.1007/s10648-012-9210-2

Agarwal, P. K., D'Antonio, L., Roediger, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*(3), 131–139. https://doi.org/10.1016/j.jarmac.2014.07.002

Agarwal, P. K., D'Antonio, L., Roediger III, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*(3), 131–139.

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

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval Practice Consistently Benefits Student Learning: A Systematic Review of Applied Research in Schools

and Classrooms. *Educ Psychol Rev*, *33*(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Ajayi, L. (2012). How teachers deploy multimodal textbooks to enhance english language learning. *Tesol Journal*, *6*(1), 16–35.

Akan, M., Stanley, S. E., & Benjamin, A. S. (2018). Testing enhances memory for context. *Journal of Memory and Language*, *103*, 19–27. https://doi.org/10.1016/j.jml.2018.07.003

Akinleke, W., & Adeaga, T. (2014). Contributions of test anxiety, study habits and locus of control to academic performance. *British Journal of Psychology Research*, *2*(1), 14–24.

Almazrouei, M. A., Morgan, R. M., & Dror, I. E. (2022). A method to induce stress in human subjects in online research environments. *Behav Res*. https://doi.org/10.3758/s13428-022-01915-3

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in cognitive sciences*, *21*(8), 573–576.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The apa publications and communications board task force report. *American Psychologist*, *73*(1), 3.

Arslan, R. C. (2019). How to automatically document data with the codebook package to facilitate data reuse. *Advances in Methods and Practices in Psychological Science*, *2*(2), 169–187.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*(4), 566–577.

Bano, M., Zowghi, D., Kearney, M., Schuck, S., & Aubusson, P. (2018). Mobile learning for science and mathematics school education: A systematic review of empirical

evidence. *Computers & Education*, *121*, 30–58. https://doi.org/10.1016/j.compedu.2018.02.006

Baran, E. (2014). A review of research on mobile learning in teacher education. *Journal of Educational Technology & Society*, *17*(4), 17–32.

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck depression inventory–ii. *Psychological assessment*.

Begg, C. B., & Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias [Publisher: [Wiley, International Biometric Society]]. *Biometrics*, *50*(4), 1088–1101. https://doi.org/10.2307/2533446

Beilock, S. L., & Carr, T. H. (2016). When High-Powered People Fail: Working Memory and "Choking Under Pressure" in Math [Publisher: SAGE PublicationsSage CA: Los Angeles, CA]. *Psychological Science*. Retrieved July 29, 2020, from https://journals.sagepub.com/doi/10.1111/j.0956-7976.2005.00789.x

Belardi, A., Pedrett, S., Rothen, N., & Reber, T. (2020). Spacing, feedback, and testing boost vocabulary learning in a web application.

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

Birkett, M. A. (2011). The trier social stress test protocol for inducing psychological stress. *JoVE (Journal of Visualized Experiments)*, (56), e3238.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied research in Memory and Cognition*, *9*(4), 475.

Borenstein, M., Hedges, L., V., Higgins, J. P., & Rothstein, H. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons.

Bozdoğan, D. (2015). MALL Revisited: Current Trends and Pedagogical Implications. *Procedia - Social and Behavioral Sciences*, *195*, 932–939. https://doi.org/10.1016/j.sbspro.2015.06.373

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*(3), 407–415.

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and emotion*, *24*(3), 377–400.

Brown, M. J., & Tallon, J. (2015). The effects of pre-lecture quizzes on test anxiety and performance in a statistics course. *Education*, *135*(3), 346–350.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables [Place: United Kingdom Publisher: Ubiquity Press]. *Journal of Cognition*, *2*. https://doi.org/10.5334/joc.72

Buchanan, T. W., & Lovallo, W. R. (2001). Enhanced memory for emotional material following stress-level cortisol treatment in humans. *Psychoneuroendocrinology*, *26*(3), 307–317.

Butler, A. C., Marsh, E. J., Slavinsky, J. P., & Baraniuk, R. G. (2014). Integrating Cognitive Science and Technology Improves Learning in a STEM Classroom. *Educ Psychol Rev*, *26*(2), 331–340. https://doi.org/10.1007/s10648-014-9256-4

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & cognition*, *36*(3), 604–616.

Cabeza, R., & Nyberg, L. (2000). Neural bases of learning and memory: Functional neuroimaging evidence. *Current opinion in neurology*, *13*(4), 415–421.

Cadaret, C. N., & Yates, D. T. (2018). Retrieval practice in the form of online homework improved information retention more when spaced 5 days rather than 1 day after class in two physiology courses. *Advances in physiology education*, *42*(2), 305–310.

Cardozo, L. T., Azevedo, M. A. R. d., Carvalho, M. S. M., Costa, R., de Lima, P. O., & Marcondes, F. K. (2020). Effect of an active learning methodology combined with formative assessments on performance, test anxiety, and stress of university students. *Advances in Physiology Education*, *44*(4), 744–751.

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, *28*(2), 353–375.

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*(9), 496–511.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic bulletin & review*, *13*(5), 826–830.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144.

Cassady, J. C. (2004). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *18*(3), 311–325.

Cassady, J. C. (2010). Test anxiety: Contemporary theories and implications for learning. *Anxiety in schools: The causes, consequences, and solutions for academic anxieties*, 7–26.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary educational psychology*, *27*(2), 270–295.

Cavus, N. (2016). Development of an Intellegent Mobile Application for Teaching English Pronunciation. *Procedia Computer Science*, *102*, 365–369. https://doi.org/10.1016/j.procs.2016.09.413

Cavus, N., & Ibrahim, D. (2017). Learning English using children's stories in mobile devices: Children's stories in mobile devices. *Br J Educ Technol*, *48*(2), 625–641. https://doi.org/10.1111/bjet.12427

Chen, C.-M., & Chung, C.-J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education*, *51*(2), 624–645. https://doi.org/10.1016/j.compedu.2007.06.011

Chen, C.-M., & Li, Y.-L. (2010). Personalised context-aware ubiquitous learning system for supporting effective English vocabulary learning. *Interactive Learning Environments*, *18*(4), 341–364. https://doi.org/10.1080/10494820802602329

Cho, H., Ryu, S., Noh, J., & Lee, J. (2016). The Effectiveness of Daily Mindful Breathing Practices on Test Anxiety of Students [Publisher: Public Library of Science]. *PLOS ONE*, *11*(10), e0164822. https://doi.org/10.1371/journal.pone.0164822

Cho, K., Lee, S., Joo, M.-H., & Becker, B. (2018). The Effects of Using Mobile Devices on Student Achievement in Language Learning: A Meta-Analysis. *Education Sciences*, *8*(3), 105. https://doi.org/10.3390/educsci8030105

Clark, D. A., Crandall, J. R., & Robinson, D. H. (2018). Incentives and test anxiety may moderate the effect of retrieval on learning. *Learning and Individual Differences*, *63*, 70–77. https://doi.org/10.1016/j.lindif.2018.03.001

Cohen, J. (1988). Statistical power analysis for the behavioural sciences (2'eu.) *HilIsUale, NJ: Lawrence Eribaum Associates*.

Cohen, M., & Khalaila, R. (2014). Saliva pH as a biomarker of exam stress and a predictor of exam performance. *Journal of Psychosomatic Research*, *77*(5), 420–425. https://doi.org/10.1016/j.jpsychores.2014.07.003

Coyne, J. H., Borg, J. M., DeLuca, J., Glass, L., & Sumowski, J. F. (2015). Retrieval practice as an effective memory strategy in children and adolescents with traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, *96*(4), 742–745.

Culler, R. E., & Holahan, C. J. (1980). Test anxiety and academic performance: The effects of study-related behaviors [Place: US Publisher: American Psychological Association]. *Journal of Educational Psychology*, *72*(1), 16–20. https://doi.org/10.1037/0022-0663.72.1.16

Darmi, R., & Albion, P. (2014). A Review of Integrating Mobile Phones for Language Learning. *International Association for Development of the Information Society*, 8.

Dedovic, K., Duchesne, A., Andrews, J., Engert, V., & Pruessner, J. C. (2009). The brain and the stress axis: The neural correlates of cortisol regulation in response to stress. *NeuroImage*, *47*(3), 864–871. https://doi.org/10.1016/j.neuroimage.2009.05.074

de Lima, M. F. R., & Buratto, L. G. (2024). Retrieval practice effect and individual differences: Current status and future directions. *Journal of Cognitive Psychology*, *36*(4), 443–456.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*(2), 162–170. https://doi.org/10.1037/0022-0663.79.2.162

de Quervain, D. J.-F., Roozendaal, B., & McGaugh, J. L. (1998). Stress and glucocorticoids impair retrieval of long-term spatial memory [Place: United Kingdom Publisher: Nature Publishing Group]. *Nature*, *394*(6695), 787–790. https://doi.org/10.1038/29542

de Quervain, D. J.-F., Roozendaal, B., Nitsch, R. M., McGaugh, J. L., & Hock, C. (2000). Acute cortisone administration impairs retrieval of long-term declarative memory in humans [Number: 4 Publisher: Nature Publishing Group]. *Nature Neuroscience*, *3*(4), 313–314. https://doi.org/10.1038/73873

Dickerson, S. S., & Kemeny, M. E. (2004). Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *130*, 355–391. https://doi.org/10.1037/0033-2909.130.3.355

Drigas, A. S., & Pappas, M. A. (2015). A review of mobile learning applications for mathematics. *International Journal of Interactive Mobile Technologies*, *9*(3).

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.

Duval, S., & Tweedie, R. (2000). Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.2000.00455.x]. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of neurosciences*, *20*(4), 155.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Emmerdinger, K. J., & Kuhbandner, C. (2019). C. *Memory*, *27*(8), 1043–1053. https://doi.org/10.1080/09658211.2019.1618339

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007a). Anxiety and cognitive performance: Attentional control theory [Place: US Publisher: American Psychological Association]. *Emotion*, *7*, 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007b). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336.

Fageeh, A. A. I. (2013). Effects of MALL Applications on Vocabulary Acquisition and Motivation. *Arab World English Journal*, *4*(4), 28.

Fanelli, D., Wong, J., & Moher, D. (2022). What difference might retractions make? an estimate of the potential epistemic cost of retractions on meta-analyses. *Accountability in Research*, *29*(7), 442–459.

Feltz, A., & May, J. (2017). The means/side-effect distinction in moral cognition: A meta-analysis. *Cognition*, *166*, 314–327.

Field, A. (2013). *Discovering statistics using ibm spss statistics*. sage.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815–836.

Fryer, L. K., & Leenknecht, M. J. (2023). Toward an organising theoretical model for teacher clarity, feedback and self-efficacy in the classroom. *Educational Psychology Review*, *35*(3), 68.

Gagnon, S. A., & Wagner, A. D. (2016). Acute stress and episodic memory retrieval: Neurobiological mechanisms and behavioral consequences. *Annals of the New York Academy of Sciences*, *1369*(1), 55–75.

Gagnon, S. A., Waskom, M. L., Brown, T. I., & Wagner, A. D. (2019a). Stress Impairs Episodic Retrieval by Disrupting Hippocampal and Cortical Mechanisms of Remembering. *Cerebral Cortex*, *29*(7), 2947–2964. https://doi.org/10.1093/cercor/bhy162

Gagnon, S. A., Waskom, M. L., Brown, T. I., & Wagner, A. D. (2019b). Stress impairs episodic retrieval by disrupting hippocampal and cortical mechanisms of remembering. *Cerebral Cortex*, *29*(7), 2947–2964.

Garzón, J., Lampropoulos, G., & Burgos, D. (2023). Effects of mobile learning in english language learning: A meta-analysis and research synthesis. *Electronics*, *12*(7), 1595.

Gauthier, J., & Bouchard, S. (1993). French-canadian adaptation of the revised form of spielberger's state–trait anxiety inventory. *Canadian Journal of Behavioral Science/Revue canadienne des sciences du behaviour*, *25*(4), 559.

Gettinger, M., & Seibert, J. K. (2002). Contributions of study skills to academic competence. *School Psychology Review*, *31*(3), 350–365.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*(2), 95–112.

Goldfarb, E. V. (2019). Enhancing memory with stress: Progress, challenges, and opportunities. *Brain and cognition*, *133*, 94–105.

Gonthier, C. (2023). Should intelligence tests be speeded or unspeeded? a brief review of the effects of time pressure on response processes and an experimental study with raven's matrices. *Journal of Intelligence*, *11*(6), 120.

Goossens, N. A., Camp, G., Verkoeijen, P. P., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology*, *28*(1), 135–142.

Grimm, J. (2009). State-trait-anxiety inventory nach spielberger. deutsche lang-und kurzversion.-methodenforum der universität wien.

Grober, E., Hall, C. B., Lipton, R. B., Zonderman, A. B., Resnick, S. M., & Kawas, C. (2008). Memory impairment, executive dysfunction, and intellectual decline in preclinical alzheimer's disease. *Journal of the International Neuropsychological Society*, *14*(2), 266–278.

Guenzel, F. M., Wolf, O. T., & Schwabe, L. (2013). Stress disrupts response memory retrieval. *Psychoneuroendocrinology*, *38*(8), 1460–1465.

Güler, M., Bütüner, S. Ö., Danişman, Ş., & Gürsoy, K. (2022). A meta-analysis of the impact of mobile learning on mathematics achievement. *Education and Information Technologies*, 1–21.

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2019). Dmetar: Companion R Package For The Guide 'Doing Meta-Analysis in R'. http://dmetar.protectlab.org

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81–112.

Hautzinger, M., Keller, F., & Kühner, C. (2006). *Beck depressions-inventar (bdi-ii)*. Harcourt Test Services.

Hembree, R. (1988). Correlates, Causes, Effects, and Treatment of Test Anxiety [Publisher: American Educational Research Association]. *Review of Educational Research*, *58*(1), 47–77. https://doi.org/10.3102/00346543058001047

Hering, A., Rendell, P. G., Rose, N. S., Schnitzspahn, K. M., & Kliegel, M. (2014). Prospective memory training in older adults and its relevance for successful aging. *Psychological research*, *78*, 892–904.

Higgins. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Higgins, J. P., Savović, J., Page, M. J., Elbers, R. G., & Sterne, J. A. (2019). Assessing risk of bias in a randomized trial. *Cochrane handbook for systematic reviews of interventions*, 205–228.

Higham, P. A., Fastrich, G. M., Potts, R., Murayama, K., Pickering, J. S., & Hadwin, J. A. (2023). Spaced retrieval practice: Can restudying trump retrieval? *Educational Psychology Review*, *35*(4), 98.

Higham, P. A., Zengel, B., Bartlett, L. K., & Hadwin, J. A. (2022). The benefits of successive relearning on multiple learning outcomes. *Journal of Educational Psychology*, *114*(5), 928.

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology*, *28*(4), 597–606.

Hohn, R. E., Slaney, K. L., & Tafreshi, D. (2019). Primary study quality in psychological meta-analyses: An empirical assessment of recent practice. *Frontiers in Psychology*, *9*, 2667.

Howard, E. (2020). A review of the literature concerning anxiety for educational assessments. *Research and analysis, Coventry, UK: Ofqual*.

Hsieh, T.-C., Wang, T.-I., Su, C.-Y., & Lee, M.-C. (2012). A fuzzy logic-based personalized learning system for supporting adaptive english learning. *Journal of Educational Technology & Society*, *15*(1), 273–288.

Hsu, C.-K., Hwang, G.-J., & Chang, C.-K. (2013). A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. *Computers & Education*, *63*, 327–336. https://doi.org/10.1016/j.compedu.2012.12.004

Huberty, T. J. (2009). Test and performance anxiety. *Principal leadership*, *10*(1), 12–16.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods*, *11*(2), 193–206. https://doi.org/10.1037/1082-989X.11.2.193

Huntley, C. D., Young, B., Temple, J., Longworth, M., Smith, C. T., Jha, V., & Fisher, P. L. (2019). The efficacy of interventions for test-anxious university students: A meta-analysis of randomized controlled trials. *Journal of Anxiety Disorders*, *63*, 36–50. https://doi.org/10.1016/j.janxdis.2019.01.007

Hupbach, A., & Fieman, R. (2012). Moderate stress enhances immediate and delayed retrieval of educationally relevant material in healthy young men. *Behavioral Neuroscience*, *126*(6), 819.

Hwang, G.-J., & Chang, H.-F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, *56*(4), 1023–1031. https://doi.org/10.1016/j.compedu.2010.12.002

Hwang, W.-Y., Chen, H. S., Shadiev, R., Huang, R. Y.-M., & Chen, C.-Y. (2014). Improving English as a foreign language writing in elementary schools using mobile devices in familiar situational contexts. *Computer Assisted Language Learning*, *27*(5), 359–378. https://doi.org/10.1080/09588221.2012.733711

Ibrahim, N. H., Chee, K. N., & Yahaya, N. (2017). Effectiveness of mobile learning application in improving reading skills in Chinese language and towards post-attitudes. *IJMLO*, *11*(3), 210. https://doi.org/10.1504/IJMLO.2017.10005992

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 109–117.

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory & cognition*, *41*, 625–637.

Jantzi, C., Mengin, A. C., Serfaty, D., Bacon, E., Elowe, J., Severac, F., Meyer, N., Berna, F., & Vidailhet, P. (2019). Retrieval practice improves memory in patients with schizophrenia: New perspectives for cognitive remediation. *BMC psychiatry*, *19*, 1–12.

Jerrim, J. (2023). Test anxiety: Is it associated with performance in high-stakes examinations? *Oxford Review of Education*, *49*(3), 321–341.

Joëls, M., & Baram, T. Z. (2009). The neuro-symphony of stress [Number: 6 Publisher: Nature Publishing Group]. *Nat Rev Neurosci*, *10*(6), 459–466. https://doi.org/10.1038/nrn2632

Johnson, R. (2018). Supporting retrieval practice with quizzing technology. *Technology and the Curriculum: Summer 2018*.

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

Kang, S. H. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 12–19.

Kang, S. H., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic bulletin & review*, *20*(6), 1259–1265.

Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning. *Learning and Instruction*, *36*, 38–45.

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157–163.

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. *Grantee Submission*.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775.

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

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

Kast, M., Baschera, G.-M., Gross, M., Jäncke, L., & Meyer, M. (2011). Computer-based learning of spelling skills in children with and without dyslexia. *Annals of dyslexia*, *61*, 177–200.

Keeley, J., Zayac, R., & Correia, C. (2008). Curvilinear relationships between statistics anxiety and performance among undergraduate students: Evidence for optimal anxiety. *Statistics Education Research Journal*, *7*(1), 4–15.

Khodabandeh, F., Soleimani, H., et al. (2017). The effect of mall-based tasks on efl learners' grammar learning. *Teaching English with Technology*, *17*(2), 29–41.

Kiliçkaya, F., & Krajka, J. (2010). COMPARATIVE USEFULNESS OF ONLINE AND TRADITIONAL VOCABULARY LEARNING. *The Turkish Online Journal of Educational Technology*, *9*(2), 9.

Kim, H.-Y., & Kim, E.-Y. (2023). Effects of medical education program using virtual reality: A systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*, *20*(5), 3895.

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting [Publisher: Karger Publishers]. *NPS*, *28*(1-2), 76–81. https://doi.org/10.1159/000119004

Kleijn, W. C., van der Ploeg, H. M., & Topman, R. M. (1994). Cognition, study habits, test anxiety, and academic performance. *Psychological Reports*, *75*(3), 1219–1226.

Klier, C., & Buratto, L. G. (2020). Stress and long-term memory retrieval: A systematic review. *Trends in psychiatry and psychotherapy*, *42*(3), 284–291.

Klier, C., & Buratto, L. G. (2023). The benefit of retrieval practice on cued recall under stress depends on item difficulty. *Neuroscience Letters*, *797*, 137066.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in medicine*, *22*(17), 2693–2710.

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

Kondo, M., Ishikawa, Y., Smith, C., Sakamoto, K., Shimomura, H., & Wada, N. (2012). Mobile Assisted Language Learning in university EFL courses in Japan: Developing attitudes and skills for self-regulated learning. *ReCALL*, *24*(2), 169–187. https://doi.org/10.1017/S0958344012000055

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

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

Kuhlmann, S. (2005). Impaired Memory Retrieval after Psychosocial Stress in Healthy Young Men. *Journal of Neuroscience*, *25*(11), 2977–2982. https://doi.org/10.1523/JNEUROSCI.5139-04.2005

Kuhlmann, S., Kirschbaum, C., & Wolf, O. T. (2005). Effects of oral cortisol treatment in healthy young women on memory retrieval of negative and neutral words. *Neurobiology of Learning and Memory*, 5.

Kuhlmann, S., Piel, M., & Wolf, O. T. (2005). Impaired memory retrieval after psychosocial stress in healthy young men. *Journal of Neuroscience*, *25*(11), 2977–2982.

Kukulska-Hulme, A., & Bull, S. (2009). Theory-based Support for Mobile Language Learning: Noticing and Recording. *Int. J. Interact. Mob. Technol.*, *3*(2), pp. 12–18. https://doi.org/10.3991/ijim.v3i2.740

Kumari, V., & Corr, P. J. (1998). Trait anxiety, stress and the menstrual cycle: Effects on raven's standard progressive matrices test. *Personality and Individual Differences*, *24*(5), 615–623.

Larsen, D. P. (2018). Planning education for long-term retention: The cognitive science and implementation of retrieval practice. *Seminars in neurology*, *38*(04), 449–456.

Laskowski, E. (2018, August). *What's a normal resting heart rate? expert opinion* [Mayo Clinic]. https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979#:~:text=To%20check%20your%20pulse%20at,calculate%20your%20beats%20per%20minute

Lazarus, R. S. (1984). *Stress, appraisal, and coping* (Vol. 464). Springer.

Lazarus, R. S., & Folkman, S. (1987). Transactional theory and research on emotions and coping. *European Journal of personality*, *1*(3), 141–169.

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.

Leininger, S., & Skeel, R. (2012). Cortisol and self-report measures of anxiety as predictors of neuropsychological performance. *Archives of Clinical Neuropsychology*, *27*(3), 318–328.

Lent, R. W., & Russell, R. K. (1978). Treatment of test anxiety by cue-controlled desensitization and study-skills training [Place: US Publisher: American Psychological Association]. *Journal of Counseling Psychology*, *25*(3), 217–224. https://doi.org/10.1037/0022-0167.25.3.217

Li, Z., & Hegelheimer, V. (2013). Mobile-Assisted Grammar Exercises: Effects on Self-Editing in L2 Writing. *Language Learning & Technology*, *17*(3), 135–156.

Liebert, R. M., & Morris, L. W. (1967). Cognitive and Emotional Components of Test Anxiety: A Distinction and Some Initial Data [Publisher: SAGE Publications Inc]. *Psychol Rep*, *20*(3), 975–978. https://doi.org/10.2466/pr0.1967.20.3.975

Lin, J.-J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis [Publisher: Routledge _eprint: https://doi.org/10.1080/09588221.2018.1541359]. *Computer Assisted Language Learning*, *32*(8), 878–919. https://doi.org/10.1080/09588221.2018.1541359

Lüdecke, D. (2019). _Esc: Effect Size Computation for Meta Analysis (Version 0.5.1). https://doi.org/10.5281/zenodo.1249218

Maassen, E., Assen, M. A. L. M. v., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology [Publisher: Public Library of Science]. *PLOS ONE*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107

Mahdi, H. S. (2018). Effectiveness of mobile devices on vocabulary learning: A meta-analysis. *Journal of educational computing research*, *56*(1), 134–154.

Martin, F., & Ertzberger, J. (2013). Here and now mobile learning: An experimental study on the use of mobile technology. *Computers & Education*, *68*, 76–85.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom [Publisher: Routledge _eprint: https://doi.org/10.1080/09541440701326154]. *European Journal of Cognitive Psychology*, *19*(4-5), 494–513. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., Fadler, C. L., & Pashler, H. (2013). Effects of spaced versus massed training in function learning [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1417–1432. https://doi.org/10.1037/a0032184

McEwen, B. S., & Gianaros, P. J. (2011). Stress- and Allostasis-Induced Brain Plasticity. *Annu Rev Med*, *62*, 431–445. https://doi.org/10.1146/annurev-med-052209-100430

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes [Publisher: SAGE Publications Inc]. *Perspect Psychol Sci*, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

Mendezabal, M. J. N. (2013). Study habits and attitudes: The road to academic success. *Open Science Repository Education*, (open-access), e70081928.

Metcalfe, J. (2017). Learning from Errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Micoulaud-Franchi, J.-A., Lagarde, S., Barkate, G., Dufournet, B., Besancon, C., Trébuchon-Da Fonseca, A., Gavaret, M., Bartolomei, F., Bonini, F., & McGonigal, A. (2016). Rapid detection of generalized anxiety disorder and major depression in epilepsy: Validation of the gad-7 as a complementary tool to the nddi-e in a french sample. *Epilepsy & Behavior*, *57*, 211–216.

Mihaylova, M., & Rothen, N. (2025). Overcoming test anxiety through learning principles: A science-based intervention. *In prep*.

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.

Minear, M., Coane, J. H., Cooney, L. H., Boland, S. C., & Serrano, J. W. (2023). Is practice good enough? retrieval benefits students with adhd but does not compensate for poor encoding in unmedicated students. *Frontiers in Psychology*, *14*, 1186566.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement [Publisher: Public Library of Science]. *PLOS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Monteiro, C. F. d. S., Freitas, J. F. d. M., & Ribeiro, A. A. P. (2007). Estresse no cotidiano acadêmico: O olhar dos alunos de enfermagem da universidade federal do piauí. *Escola Anna Nery*, *11*, 66–72.

Moreau, D., & Gamble, B. (2022). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychol Methods*, *27*(3), 426–432. https://doi.org/10.1037/met0000351

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Education*, *4*, 5.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of verbal learning and verbal behavior*, *16*(5), 519–533.

Mundt, D., Abel, R., & Hänze, M. (2020). Exploring the effect of testing on forgetting in vocabulary learning: An examination of the bifurcation model. *Journal of Cognitive Psychology*, *32*(2), 214–228.

Naveh-Benjamin, M., McKeachie, W. J., Lin, Y.-g., & Holinger, D. P. (1981). Test anxiety: Deficits in information processing [Place: US Publisher: American Psychological

Association]. *Journal of Educational Psychology*, *73*(6), 816–824. https://doi.org/10.1037/0022-0663.73.6.816

Nuijten, M. B., & Polanin, J. R. (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Res Synth Methods*, *11*(5), 574–579. https://doi.org/10.1002/jrsm.1408

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 229–237.

O'Day, G. M., & Karpicke, J. D. (2020). Comparing and combining retrieval practice and concept mapping. *Journal of Educational Psychology*.

Oei, N. Y., Everaerd, W. T., Elzinga, B. M., van Well, S., & Bermond, B. (2006). Psychosocial stress impairs working memory at high loads: An association with cortisol levels and memory retrieval. *Stress*, *9*(3), 133–141.

Oz, H. (2014). Prospective english teachers' ownership and usage of mobile devices as m-learning tools. *Procedia-Social and Behavioral Sciences*, *141*, 1031–1041.

Ozer, O., & Kılıç, F. (2018). The Effect of Mobile-Assisted Language Learning Environment on EFL Students' Academic Achievement, Cognitive Load and Acceptance of Mobile Learning Tools. *EURASIA J MATH SCI T*, *14*(7). https://doi.org/10.29333/ejmste/90992

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews [Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting]. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of experimental psychology: Learning, Memory, and Cognition*, *31*(1), 3.

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing [Publisher: Frontiers]. *Front. Psychol.*, *5*. https://doi.org/10.3389/fpsyg.2014.00286

Pastötter, B., & Frings, C. (2019). The forward testing effect is reliable and independent of learners' working memory capacity. *Journal of cognition*, *2*(1).

Pastötter, B., von Dawans, B., Domes, G., & Frings, C. (2023). The forward testing effect is resistant to acute psychosocial retrieval stress. *Experimental Psychology*.

Pastötter, B., Weber, J., & Bäuml, K.-H. T. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology*, *27*(2), 280.

Piroozmanesh, A., & Imanipour, M. (2018). The effect of formative assessment on test anxiety of nursing students. *Journal of Medical Education Development*, *10*(28), 18–26.

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1354]. *Research Synthesis Methods*, *10*(3), 330–342. https://doi.org/10.1002/jrsm.1354

Pruett, S. B. (2001). Quantitative aspects of stress-induced immunomodulation. *International immunopharmacology*, *1*(3), 507–520.

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences [Place: US Publisher: John Wiley & Sons]. *Research Synthesis Methods*, *10*, 57–71. https://doi.org/10.1002/jrsm.1332

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, *36*(4), 672–683.

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.

Quintana, D. S. (2023). A guide for calculating study-level statistical power for meta-analyses. *Advances in Methods and Practices in Psychological Science*, *6*(1), 25152459221147260.

Rachels, J. R., & Rockinson-Szapkiw, A. J. (2018). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Computer Assisted Language Learning*, *31*(1-2), 72–89. https://doi.org/10.1080/09588221.2017.1382536

Racsmány, M., Szőllősi, Á., & Marián, M. (2020). Reversing the testing effect by feedback is a matter of performance criterion at practice. *Memory & Cognition*, *48*, 1161–1170.

Rahimi, M., & Miri, S. S. (2014). The Impact of Mobile Dictionary Use on Language Learning. *Procedia - Social and Behavioral Sciences*, *98*, 1469–1474. https://doi.org/10.1016/j.sbspro.2014.03.567

Raven, J. C. (1941). Standardization of progressive matrices, 1938 [Place: United Kingdom Publisher: British Psychological Society]. *British Journal of Medical Psychology*, *19*, 137–150. https://doi.org/10.1111/j.2044-8341.1941.tb00316.x

Raven, J. (1938). Progressive matrices: A perceptual test of intelligence. *London: HK Lewis*, *19*, 20.

Raven, J. (1962). Advanced progressive matrices: Sets i and ii. *(No Title)*.

Raven, J. (1965). Advanced progressive matrices. sets i and ii. london: Hk lewis & co.

Rawson, K. A., & Dunlosky, J. (2011a). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283.

Rawson, K. A., & Dunlosky, J. (2011b). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? [Place: US Publisher: American

Psychological Association]. *Journal of Experimental Psychology: General*, *140*(3), 283–302. https://doi.org/10.1037/a0023956

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*, *142*(4), 1113.

Rawson, K. A., & Dunlosky, J. (2022). Successive relearning: An underexplored but potent technique for obtaining and maintaining knowledge. *Current Directions in Psychological Science*, *31*(4), 362–368.

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology: Applied*, *24*(1), 57.

Reber, T. P., & Rothen, N. (2018). Educational App-Development needs to be informed by the Cognitive Neurosciences of Learning & Memory. *npj Science Learn*, *3*(1), 22. https://doi.org/10.1038/s41539-018-0039-4

Ringeisen, T., Buchwald, P., & Hodapp, V. (2010). Capturing the multidimensionality of test anxiety in cross-cultural research: An english adaptation of the german test anxiety inventory. *Cognition, Brain, Behavior*, *14*(4), 347.

Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2013). Retrieval practice, with or without mind mapping, boosts fact learning in primary school children. *PloS one*, *8*(11), e78976.

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

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

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied*

Research in Memory and Cognition, *1*(4), 242–248. https://doi.org/10.1016/j.jarmac.2012.09.002

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

Roozendaal, B. (2002). Stress and memory: Opposing effects of glucocorticoids on memory consolidation and memory retrieval. *Neurobiology of learning and memory*, *78*(3), 578–595.

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.

Rüth, M., Breuer, J., Zimmermann, D., & Kaspar, K. (2021). The effects of different feedback types on learning with mobile quiz apps. *Frontiers in Psychology*, *12*, 665144.

Sakaki, M., Niki, K., & Mather, M. (2012). Beyond arousal and valence: The importance of the biological versus social relevance of emotional stimuli. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(1), 115–139.

Salend, S. J. (2011). Addressing test anxiety. *Teaching exceptional children*, *44*(2), 58–68.

Sandberg, J., Maris, M., & de Geus, K. (2011). Mobile English learning: An evidence-based study with fifth graders. *Computers & Education*, *57*(1), 1334–1347. https://doi.org/10.1016/j.compedu.2011.01.015

Sanfilippo, F., Blazauskas, T., Salvietti, G., Ramos, I., Vert, S., Radianti, J., Majchrzak, T. A., & Oliveira, D. (2022). A perspective review on integrating vr/ar with haptics into stem education for multi-sensory learning. *Robotics*, *11*(2), 41.

Sarason, I. G. (1961). Test anxiety and the intellectual performance of college students [Place: US Publisher: American Psychological Association]. *Journal of Educational Psychology*, *52*, 201–206. https://doi.org/10.1037/h0049095

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *46*, 929–938. https://doi.org/10.1037/0022-3514.46.4.929

Schneider, B. A., Avivi-Reich, M., & Mozuraitis, M. (2015). A cautionary note on the use of the analysis of covariance (ancova) in classification designs with and without within-subject factors. *Frontiers in psychology*, *6*, 474.

Schoofs, D., Hartmann, R., & Wolf, O. (2008). Neuroendocrine stress responses to an oral academic examination: No strong influence of sex, repeated participation and personality traits. *Stress*, *11*(1), 52–61.

Schwab, S., Janiaud, P., Dayan, M., Amrhein, V., Panczak, R., Palagi, P. M., Hemkens, L. G., Ramon, M., Rothen, N., Senn, S., et al. (2022). Ten simple rules for good research practice. *PLoS computational biology*, *18*(6), e1010139.

Schwabe, L., Bohringer, A., Chatterjee, M., & Schachinger, H. (2008). Effects of pre-learning stress on memory for neutral, positive and negative words: Different roles of cortisol and autonomic arousal. *Neurobiology of Learning and Memory*, *90*(1), 44–53. https://doi.org/10.1016/j.nlm.2008.02.002

Schwabe, L., Böhringer, A., & Wolf, O. T. (2009). Stress disrupts context-dependent memory. *Learning & Memory*, *16*(2), 110–113.

Schwabe, L., & Schächinger, H. (2018). Ten years of research with the Socially Evaluated Cold Pressor Test: Data from the past and guidelines for the future. *Psychoneuroendocrinology*, *92*, 155–161. https://doi.org/10.1016/j.psyneuen.2018.03.010

Schwabe, L., & Wolf, O. T. (2009). The context counts: Congruent learning and testing environments prevent memory retrieval impairment following stress. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(3), 229–236.

Schwabe, L., & Wolf, O. T. (2010a). Learning under stress impairs memory formation. *Neurobiology of Learning and Memory*, *93*(2), 183–188. https://doi.org/10.1016/j.nlm.2009.09.009

Schwabe, L., & Wolf, O. T. (2010b). Stress impairs the reconsolidation of autobiographical memories. *Neurobiology of Learning and Memory*, *94*(2), 153–157.

Schwabe, L., & Wolf, O. T. (2013). Stress and multiple memory systems: From 'thinking'to 'doing'. *Trends in cognitive sciences*, *17*(2), 60–68.

Schwabe, L., & Wolf, O. T. (2014). Timing matters: Temporal dynamics of stress effects on memory retrieval. *Cognitive, Affective, & Behavioral Neuroscience*, *14*, 1041–1048.

Schwarzer, G., Carpenter, J. R., Rücker, G., Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). Fixed effect and random effects meta-analysis. *Meta-analysis with R*, 21–53.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing Effect in the Psychology Classroom: A Meta-Analytic Perspective [Publisher: SAGE Publications]. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Sedgwick. (2012). Pearson's correlation coefficient. *Bmj*, *345*.

Shadiev, R., Hwang, W.-Y., & Liu, T.-Y. (2018). Investigating the effectiveness of a learning activity supported by a mobile multimedia learning system to enhance autonomous EFL learning in authentic contexts. *Education Tech Research Dev*, *66*(4), 893–912. https://doi.org/10.1007/s11423-018-9590-1

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in cognitive sciences*, *12*(11), 411–417.

Shields, G. S., Doty, D., Shields, R. H., Gower, G., Slavich, G. M., & Yonelinas, A. P. (2017). Recent life stress exposure is associated with poorer long-term memory, working memory, and self-reported memory. *Stress*, *20*(6), 598–607.

Shields, G. S., Sazma, M. A., McCullough, A. M., & Yonelinas, A. P. (2017). The effects of acute stress on episodic memory: A meta-analysis and integrative review. *Psychological bulletin*, *143*(6), 636.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. *Annu Rev Psychol*, *70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

Simone, P. M., Whitfield, L. C., Bell, M. C., Kher, P., & Tamashiro, T. (2023). Shifting students toward testing: Impact of instruction and context on self-regulated learning. *Cognitive Research: Principles and Implications*, *8*(1), 14.

Slaney, K. L., Tafreshi, D., & Hohn, R. (2018). Random or fixed? an empirical examination of meta-analysis model choices. *Review of General Psychology*, *22*(3), 290–304.

Smeets, T. (2011). Acute stress impairs memory retrieval independent of time of day. *Psychoneuroendocrinology*, *36*(4), 495–501. https://doi.org/10.1016/j.psyneuen.2010.08.001

Smeets, T., Otgaar, H., Candel, I., & Wolf, O. T. (2008). True or false? memory is differentially affected by stress-induced cortisol elevations and sympathetic activity at consolidation and retrieval. *Psychoneuroendocrinology*, *33*(10), 1378–1386.

Smith, A. M., Davis, F. C., & Thomas, A. K. (2018). Criterial learning is not enough: Retrieval practice is necessary for improving post-stress memory accessibility. *Behavioral neuroscience*, *132*(3), 161.

Smith, A. M., Floerke, V. A., & Thomas, A. K. (2016). Retrieval practice protects memory against acute stress. *Science*, *354*(6315), 1046–1048.

Smith, A. M., Race, E., Davis, F. C., & Thomas, A. K. (2019). Retrieval practice improves item memory but not source memory in the context of stress. *Brain and cognition*, *133*, 24–32.

Smith, A. M., & Thomas, A. K. (2018a). Reducing the Consequences of Acute Stress on Memory Retrieval. *Journal of Applied Research in Memory and Cognition*, *7*(2), 219–229. https://doi.org/10.1016/j.jarmac.2017.09.007

Smith, A. M., & Thomas, A. K. (2018b). Reducing the consequences of acute stress on memory retrieval. *Journal of Applied Research in Memory and Cognition*, *7*(2), 219–229.

Smith, A. M. (2018). *Examining the role of retrieval practice in improving memory accessibility under stress* [Doctoral dissertation, Tufts University].

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1712–1725. https://doi.org/10.1037/a0033569

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review*, *8*(2), 203–220.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767. https://doi.org/10.1002/acp.1747

Sommer, M., & Arendasy, M. E. (2015). Further evidence for the deficit account of the test anxiety–test performance relationship from a high-stakes admission testing setting. *Intelligence*, *53*, 72–80.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch Intern Med*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias [Place: US Publisher: John Wiley & Sons]. *Research Synthesis Methods*, *5*, 60–78. https://doi.org/10.1002/jrsm.1095

Stefan, A., Berchtold, C. M., & Angstwurm, M. (2020). Translation of a scale measuring cognitive test anxiety (g-ctas) and its psychometric examination among medical students in germany. *GMS Journal for Medical Education*, *37*(5).

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials [Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting]. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

Stock, L.-M., & Merz, C. J. (2018). Memory retrieval of everyday information under stress. *Neurobiology of Learning and Memory*, *152*, 32–38.

Stojanović, N. M., Randjelović, P. J., Pavlović, D., Stojiljković, N. I., Jovanović, I., Sokolović, D., Radulović, N. S., et al. (2021). An impact of psychological stress on the interplay between salivary oxidative stress and the classic psychological stress-related parameters. *Oxidative Medicine and Cellular Longevity*, *2021*.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 230.

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

Sumowski, J. F., Chiaravalloti, N., & DeLuca, J. (2010a). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect [Place: US Publisher: American Psychological Association]. *Neuropsychology*, *24*(2), 267–272. https://doi.org/10.1037/a0017533

Sumowski, J. F., Chiaravalloti, N., & DeLuca, J. (2010b). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. *Neuropsychology*, *24*(2), 267.

Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275. https://doi.org/10.1016/j.compedu.2015.11.008

Sung, Y.-T., Chang, K.-E., & Yang, J.-M. (2015). How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review*, *16*, 68–84. https://doi.org/10.1016/j.edurev.2015.09.001

Szőllősi, Á., Keresztes, A., Novák, B., Szászi, B., Kéri, S., & Racsmány, M. (2017). The testing effect is preserved in stressful final testing environment. *Applied Cognitive Psychology*, *31*(6), 615–622.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313–6317.

Taj, I. H., Sulan, N., Sipra, M., & Ahmad, W. (2016). Impact of Mobile Assisted Language Learning (MALL) on EFL: A Meta-Analysis. *7*(2). Retrieved April 16, 2020, from https://papers.ssrn.com/abstract=2931654

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational research*, *81*(1), 4–28.

Thiriveedhi, S., Myla, A., Priya, C., Vuppuluri, K., Dulipala, P., & Vudathaneni, V. K. P. (2023). A study on the assessment of anxiety and its effects on students taking the national eligibility cum entrance test for undergraduates (neet-ug) 2020. *Cureus*, *15*(8).

Thomas, C. L., Cassady, J. C., & Finch, W. H. (2018). Identifying Severity Standards on the Cognitive Test Anxiety Scale: Cut Score Determination Using Latent Class and Cluster Analysis [Publisher: SAGE Publications Inc]. *Journal of Psychoeducational Assessment*, *36*(5), 492–508. https://doi.org/10.1177/0734282916686004

Thornton, P., & Houser, C. (2005). Using mobile phones in English education in Japan: Using mobile phones in English education in Japan. *Journal of Computer Assisted Learning*, *21*(3), 217–228. https://doi.org/10.1111/j.1365-2729.2005.00129.x

Tlili, A., Padilla-Zea, N., Garzón, J., Wang, Y., Kinshuk, K., & Burgos, D. (2023). The changing landscape of mobile learning pedagogy: A systematic literature review. *Interactive Learning Environments*, *31*(10), 6462–6479.

Trammell, J. P., & Clore, G. L. (2014). Does stress enhance or impair memory consolidation? *Cognition & Emotion*, *28*(2), 361–374.

Troussas, C., Krouska, A., & Sgouropoulou, C. (2022). Enriching mobile learning software with interactive activities and motivational feedback for advancing users' high-level cognitive skills. *Computers*, *11*(2), 18.

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

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

Tseng, C.-C., Lu, C.-H., & Hsu, W.-L. (2007). A Mobile Environment for Chinese Language Learning. In M. J. Smith & G. Salvendy (Eds.), *Human Interface and the Management of Information. Interacting in Information Environments* (pp. 485–489). Springer. https://doi.org/10.1007/978-3-540-73354-6_53

Tullis, J. G., & Benjamin, A. S. (2012). Consequences of restudy choices in younger and older learners. *Psychonomic Bulletin & Review*, *19*, 743–749.

Tulving, E. (1993). What is episodic memory? *Current directions in psychological science*, *2*(3), 67–70.

van Bergen, E., de Zeeuw, E. L., Hart, S. A., Boomsma, D. I., de Geus, E. J., & Kan, K.-J. (2025). * co-occurrence and causality among adhd, dyslexia, and dyscalculia. *Psychological Science*, *36*(3), 204–217.

Vaughn, K. E., Dunlosky, J., & Rawson, K. A. (2016). Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion? *Memory & cognition*, *44*, 897–909.

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, *23*(8), 1229–1237.

Viberg, O., & Grönlund, Å. (2012). Mobile Assisted Language Learning : A Literature Review. Retrieved October 7, 2022, from http://urn.kb.se/resolve?urn=urn:nbn:se:du-10659

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *36*(3), 1–48. http://www.jstatsoft.org/v36/i03/

Vogel, S., & Schwabe, L. (2016). Learning and memory under stress: Implications for the classroom [Number: 1 Publisher: Nature Publishing Group]. *npj Science Learn*, *1*(1), 1–10. https://doi.org/10.1038/npjscilearn.2016.11

von der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000–2010. *Psychology in the Schools*, *50*(1), 57–71.

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Vuogan, A., & Li, S. (2023). Examining the effectiveness of peer feedback in second language writing: A meta-analysis. *Tesol Quarterly*, *57*(4), 1115–1138.

Wahlheim, C. N., Maddox, G. B., & Jacoby, L. L. (2014). The role of reminding in the effects of spaced repetitions on cued recall: Sufficient but not necessary [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 94–105. https://doi.org/10.1037/a0034055

Walters, W. H. (2007). Google scholar coverage of a multidisciplinary field. *Information processing & management*, *43*(4), 1121–1132.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., Zemla, J., et al. (2017). Package 'corrplot'. *Statistician*, *56*(316), e24.

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications*, *3*(1), 2. https://doi.org/10.1186/s41235-017-0087-y

Weinstein, Y., McDermott, K. B., & Chan, J. C. (2010). True and false memories in the drm paradigm on a forced choice test. *Memory*, *18*(4), 375–384.

Wenzel, K., & Reinhard, M.-A. (2021). Learning with a double-edged sword? beneficial and detrimental effects of learning tests—taking a first look at linkages among tests, later learning outcomes, stress perceptions, and intelligence. *Frontiers in Psychology*, *12*, 693585.

Wilkinson, S. (2014). Accounting for the phenomenology and varieties of auditory verbal hallucination within a predictive processing framework. *Consciousness and Cognition*, *30*, 142–155.

Wilschut, T., Sense, F., & van Rijn, H. (2025). Modality matters: Evidence for the benefits of speech-based adaptive retrieval practice in learners with dyslexia. *Topics in Cognitive Science*, *17*(1), 57–72.

Wine, J. (1971). Test anxiety and direction of attention [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *76*, 92–104. https://doi.org/10.1037/h0031332

Wittmaier, B. C. (1972). Test Anxiety and Study Habits [Publisher: Routledge _eprint: https://doi.org/10.1080/00220671.1972.10884344]. *The Journal of Educational Research*, *65*(8), 352–354. https://doi.org/10.1080/00220671.1972.10884344

Wolf, O. T. (2017). Stress and memory retrieval: Mechanisms and consequences. *Current Opinion in Behavioral Sciences*, *14*, 40–46.

Wu, Q. (2014). Learning ESL Vocabulary with Smartphones. *Procedia - Social and Behavioral Sciences*, *143*, 302–307. https://doi.org/10.1016/j.sbspro.2014.07.409

Wu, Q. (2015). Designing a smartphone app to teach English (L2) vocabulary. *Computers & Education*, *85*, 170–179. https://doi.org/10.1016/j.compedu.2015.02.013

Wu, W.-H., Jim Wu, Y.-C., Chen, C.-Y., Kao, H.-Y., Lin, C.-H., & Huang, S.-H. (2012). Review of trends from mobile learning studies: A meta-analysis. *Computers & Education*, *59*(2), 817–827. https://doi.org/10.1016/j.compedu.2012.03.016

Yaman, İ., Şenel, M., & Yeşilel, D. B. A. (2015). Exploring the extent to which elt students utilise smartphones for language learning purposes. *South African Journal of Education*, *35*(4).

Yang, C., Li, J., Zhao, W., Luo, L., & Shanks, D. R. (2023). Do practice tests (quizzes) reduce or provoke test anxiety? a meta-analytic review. *Educational Psychology Review*, *35*(3), 87.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological bulletin*, *147*(4), 399.

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied*, *23*(3), 263.

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ science of learning*, *3*(1), 1–9.

Yang, C., Shanks, D. R., Zhao, W., Fan, T., & Luo, L. (2023). Frequent quizzing accelerates classroom learning. *In their own words: What scholars and teachers want you to know about why and how to apply the science of learning in your academic setting*, 190–199.

Yang, C., Sun, B., Potts, R., Yu, R., Luo, L., & Shanks, D. R. (2020). Do working memory capacity and test anxiety modulate the beneficial effects of testing on new learning? [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology: Applied*, *26*, 724–738. https://doi.org/10.1037/xap0000278

Yerkes, R. M., Dodson, J. D., et al. (1908). The relation of strength of stimulus to rapidity of habit-formation.

Yeung, S. K., & Feldman, G. (2022). Action-inaction asymmetries in emotions and counterfactual thoughts: Meta-analysis of the action effect [registered report stage 1]. *Open Science Framework*.

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

Zou, D., & Xie, H. (2018). Personalized word-learning based on technique feature analysis and learning analytics. *Journal of Educational Technology & Society*, *21*(2), 233–244.