sp open
swiss psychology open

Does Retrieval Practice Protect Memory Against Stress? A Meta-Analysis

META-ANALYSIS

MARIELA MIHAYLOVA
MARIE ZOCCA
MATTHIAS KLIEGEL
NICOLAS ROTHEN

*Author affiliations can be found in the back matter of this article

ubiquity press

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 (Agarwal et al., 2014; Smith & Thomas, 2018), however the findings are mixed (Yang et al., 2020). Understanding whether retrieval practice maintains 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 = 908) on the impact of stress following learning with retrieval practice from 2006–2024 using databases and other information sources. We found a positive effect of retrieval practice in stress conditions, Hedges' g = 0.45, 95% CI [0.17, 0.72] with a random-effects two-level model, suggesting retrieval practice is more beneficial than restudy during stress. Significant heterogeneity was observed along with a moderate risk of bias, yet little evidence for publication bias. We did not find significant moderators in this sample. Future studies should be conducted with sufficiently powered samples and using other types of control, or non-retrieval practice, strategies to better understand the protective effects of retrieval practice on memory following stress. We registered our meta-analysis, with datafile, code and supplementary here: https://osf.io/jwx4f/.

CORRESPONDING AUTHOR:
Mariela Mihaylova, PhD
Faculty of Psychology, UniDistance Suisse, Switzerland; Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland
marielamihay@gmail.com

KEYWORDS:
retrieval practice; meta-analysis; registered report; memory; test anxiety; stress

TO CITE THIS ARTICLE:
Mihaylova, M., Zocca, M., Kliegel, M., & Rothen, N. (2025). Does Retrieval Practice Protect Memory Against Stress? A Meta-Analysis. *Swiss Psychology Open,* 5(1): 6, pp. 1–24. DOI: https://doi.org/10.5334/spo.101

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 subsequent learning (Gagnon et al., 2019; Kuhlmann, 2005; McEwen & Gianaros, 2011; Schwabe & Wolf, 2010; Shields 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 (Agarwal et al., 2021; Roediger & Karpicke, 2006; Rowland, 2014; Schwieren et al., 2017; Yang et al., 2021). 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 to 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 Trier 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 et al., 2008; Schwabe & Schächinger, 2018). The above-mentioned 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 et al., 2005; Schwabe & Wolf, 2010).

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 seven 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 five minutes, two 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 two days and one 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; Roedgier & Butler, 2011; 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 Smith and colleagues (2016) explored 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 (Smith et al., 2016). This study suggested 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 about 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 demonstrate 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 and colleagues (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, meaning that test anxiety does not significantly modulate retrieval practice effects. However, other evidence from Clark and colleagues (2018) showed 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 memory impairments 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 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 decrease the detrimental effect of stress on memory performance 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 (e.g., 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-stress 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 to the detrimental impact of stress on memory (e.g., Smith et al., 2016). Thus, we expect that retrieval practice will outperform other control strategies in mitigating the memory impairment 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 (Smith et al., 2016).

CONFIRMATORY MODERATORS

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 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 first moderator according to the type of stressor: TSST or TA (test anxiety). 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.

> Stressor Type. 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.

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 (i.e., restudy, highlighting, drawing diagrams).

> Other Strategies. We predict that any strategies used other than retrieval practice would be less beneficial than retrieval practice.

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 performance measurement. Thus, we wanted to explore whether retention interval impacts memory performance following learning with retrieval practice. This moderator was coded by different potential delay periods between initial learning and memory test (e.g., one day, two days, one week, etc.).

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

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 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 the different types of materials used. 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.).

> Task Type. We expect retrieval practice to have a positive effect on any type of learning material used.

METHODS

OPEN SCIENCE DISCLOSURES

We shared all procedures, materials, datasets, articles, and code on Open Science Framework (OSF, here). There are no other unreported/unlinked pre-registrations for this meta-analysis project. See Open Science Disclosures in Supplementary for details (here). The templates on OSF and the template for Registered Reports used in this meta-analysis have been adapted from the resources developed by Feldman (2019a, 2019b) and Yeung and colleagues (2021) (Yeung & Feldman, 2021). We made all efforts recommended by the field to enhance reproducibility, openness, and transparency (Lindsay, 2020; Maassen et al., 2020; Moreau & Gamble, 2020b; Page, 2020; Polanin, 2020b).

LITERATURE SEARCH

An unstructured literature search was first performed 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 (see Table 1 in Supplementary for the full list). 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, 2014; Schwieren et al., 2017) with the addition of the term "test anxiety." We selected experimental articles 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 and 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 unpublished articles 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 outcome was a total of 6,147 prospective articles. Following deletion of duplicates, we had a total of 6,028 articles (Figure 1).

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 another 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. All these search procedures are documented on the Literature Search file on OSF.

Furthermore, we identified authors in the field along with authors of other identified articles and searched through their related publications before contacting them. This ensured full coverage and maximized access to unpublished data and/or manuscripts that could also be relevant (see Supplementary Materials – Template for Contacting Authors for Published and Unpublished Data for a mail-merge Word template sent). 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 were relevant for our analysis, MM and MZ read more to determine whether they met the inclusion criteria or whether the article 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 six articles, reducing our pool to 12 articles. At Stage 2, we further excluded two additional articles according to exclusion criteria, leaving us with a final 10 articles consisting of 13 different studies and a total of 908 participants. We listed all the excluded articles in Table 2 of Supplementary and in the Full Coding Sheet.

INCLUSION AND EXCLUSION CRITERIA

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

First, the main dependent variable in each article needs to assess the impact of the stressor versus 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 featured an experimental versus control group design where the experimental group underwent stress induction procedure, and the control group did not. Additionally, stress induction was 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 were 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 that featured a non-retrieval practice group were also accepted as a viable comparison group.

Second, the experimental studies 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 needed to report means (*M*), standard deviations (*SD*), and sample sizes (*n*) for both the experimental and control groups who learned with retrieval practice versus another strategy (if included) after undergoing the stress procedure. Alternatively, articles needed to include the effect size that represents the magnitude of having learned with retrieval practice in a group that underwent a stress versus a control task or for the interaction effects between the retrieval practice and control strategies for stress and control groups. In cases of missing statistical data, we first attempted to contact the authors (Polanin et al., 2020a) 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 could not 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.

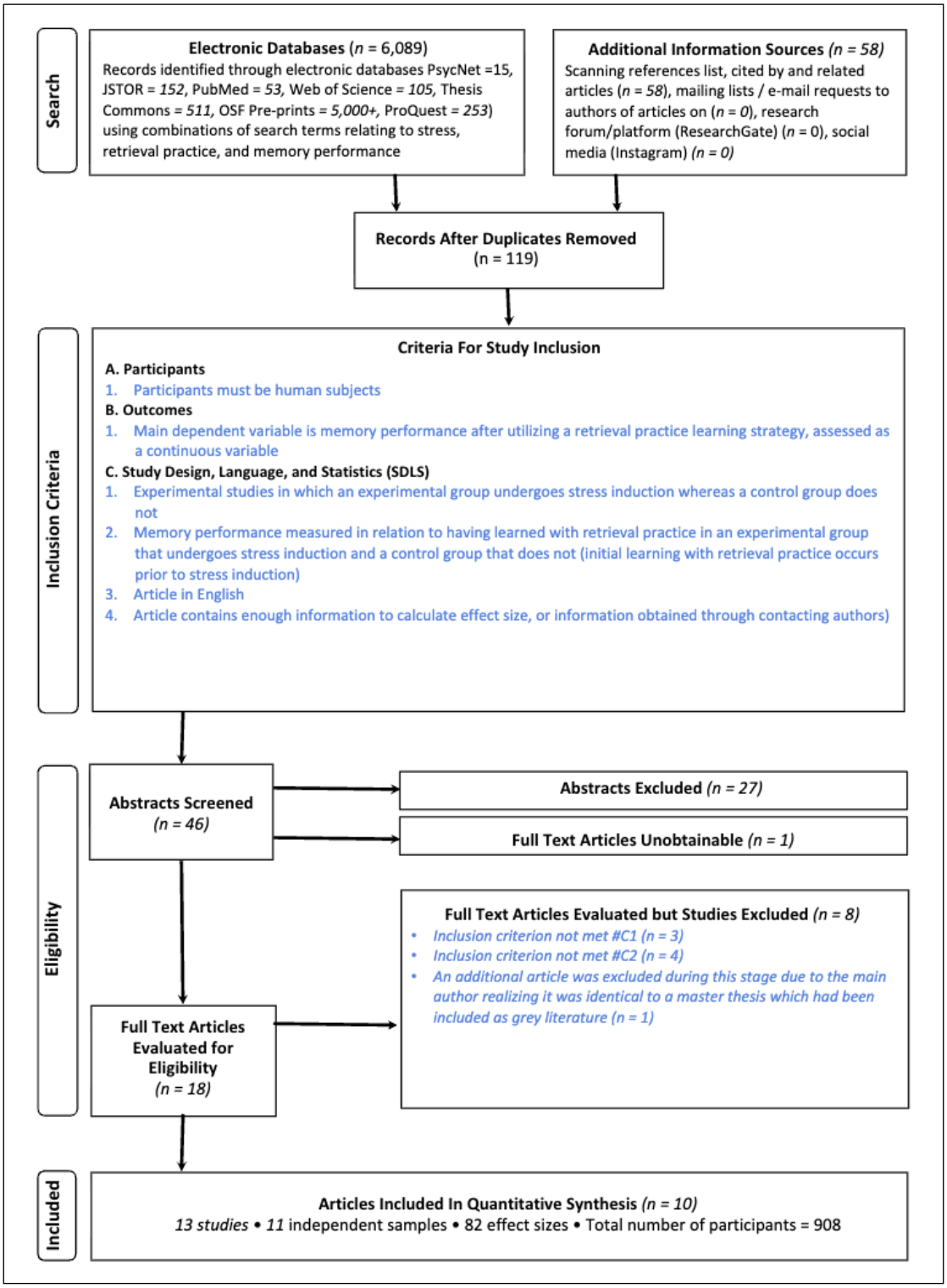


Figure 1 Systematic literature search flow diagram.

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) (Page, 2020), as well as Moreau and Gamble (2020a). Meta-analysis templates and materials Template 2 Search Flow Diagram (osf.io/q8stz). It has been used/adapted in Yeung et al. (2021) meta-analysis.

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., 2021).

SCREENING

Articles 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 discuss progress or any disagreements at each stage of coding. Disagreements were resolved through deliberations and discussion 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 Full Coding Sheet spreadsheet and Literature Searches (tab name "Article List Inclusion and Exclusion Criteria," in the Full Coding Sheet, based on the Literature Search). We saved all included references of the studies in the total search into a list available on OSF. The open-access full-texts are accessible on OSF.

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 the coding process, we pilot-tested two randomly selected articles in two stages and refined accordingly at every stage. MM and MZ both completed the coding process to ensure high inter-rater reliability. We documented gaps and reported decisions in detail in the "Article and Decision" tab of the Full Coding Sheet. As the main contributor, MM then verified the coding sheet and adjusted any discrepancies if necessary following deliberations.

INCLUDED STUDIES CODING

Once we completed the article selection procedures, pre-test coding, and confirmed the included articles, MM and MZ coded the studies independently using the Full Coding sheet. The Codebook with instructions (see OSF) on how to code each column were 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 articles. Inter-rater agreements were checked with Cohen's Kappa and intra-class correlation coefficient (Hohn et al., 2020; 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.

CONFIRMATORY ANALYSES

We used RStudio v4.1.3 (R Core Team, 2020) for statistical analyses with packages for meta-analysis such as 'metafor' (Viechtbauer, 2010, v3.8–1). We used the analysis templates adapted from Yeung and colleagues (2021) (Yeung & Feldman, 2021) 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* to facilitate comparison. Multiple effect sizes (i.e., different measures within the same article) 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* or *SD* reported), we calculated using packages like '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 *M* and *SD*. We also verified statistical results from articles using statcheck (Nuijten & Polanin, 2020) to confirm internal consistency. If the original article did not directly report *M* and *SD* 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., 2010; 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 homogenous. The random-effects model produces an overall effect size.

Importantly, because our main hypotheses look at the effects of retrieval practice compared to control strategies in stress and non-stress conditions, as well as of control strategies in stress versus non-stress 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 article (Cheung, 2014). 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 studies within articles). 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 participants' 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 participants' 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 representing 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 (Higgins & Thompson, 2002; Huedo-Medina et al., 2006). The global meta-analysis yielded an overall effect with 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 which warrant further investigation. If there was indeed meaningful heterogeneity, we investigated 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 (Borenstein & Higgins, 2013). 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, 2017, v0.1.3). This procedure uses bootstrapping techniques to overcome low power 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 this asymmetry by filling in missing studies (Duval and Tweedie, 2000). We also conducted the Rank correlation test (Begg and 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 a strong correlation suggests publication bias.

To correct 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 (tau^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., its *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; Harrer et al., 2021). 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 between-study heterogeneity. This sensitivity could affect the reliability and interpretation of our findings. One way in which we disentangled 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 performed sensitivity analysis using the leave-one-out method (Harrer et al., 2021) where effect sizes are recalculated with one effect removed each time to assess the robustness of findings and identify potential outliers. Additionally, we considered 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. Lastly, we also assessed study level power to check whether publication bias is likely (Quintana, 2023).

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 (Smith et al., 2016). Because our current understanding of the literature is that the field is still emerging, we expected to include 10 articles. 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 using the 'dmetar' package (Harrer et al., 2019, v0.0.9000, available on OSF). We estimated the power of the meta-analysis to be 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 also estimated to be 92.88%. Both analyses suggested we had viable power to conduct the meta-analysis with those parameters.

Post-hoc power analysis was performed at Stage 2 by re-running our initial power analysis script above with the actual values obtained from our meta-analysis, as well as using the upper and lower bounds of the confidence intervals obtained for each overall effect for each H 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-analyses, 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.

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 in a meta-analysis to assess and weigh the relative bias risk each article poses. Risk of bias was 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 judgements were performed by two independent raters (MM and MZ). Risk of bias ratings are available on OSF.

RESULTS

We summarized our findings in Table 1, publication bias in Table 2, and moderator analysis in Table 3 for H3 only. We provided the list of articles included in the meta-analysis in Table 4. We presented forest and funnel plots of the studies included in Figures 2 and 3. Post-hoc power analysis results are shown in Figure 4 and risk of bias in Figure 5. In the following, we first present the main-effect findings, followed by publication bias, moderator analysis, post-hoc power analysis, and risk of bias at the end. Moderator and publication bias analyses for H1, H2, and H4 can be found in in Supplementary. Note that H1-H3 have k = 9 due to one study (Hupbach & Feiman, 2012) only reporting effects for H4.

As suggested during Stage 2 review, running sensitivity analysis with a three-level model and robust variance estimation revealed similar results as the main model for all Hs (see Supplementary. Similarly, for null overall effects, we followed up with equivalence testing to determine whether the observed effects were statistically equivalent to zero within meaningful small-effect bounds (see Supplementary).

OVERALL RETRIEVAL PRACTICE EFFECTS

Random-effects two-level model for H1 (other strategies 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,

HYPOTHESES	KEY FINDINGS/THEORIES IN THE LITERATURE	FINDINGS IN THE META-ANALYSIS
Main Hypotheses		
H1	Stress induction leads to a decline in memory performance if no specific strategies are used (Shields et al., 2017).	Not supported
H2	Retrieval practice benefits memory more than other typically used strategies such as re-reading or highlighting (e.g., Moreira et al., 2019) in non-stressful conditions.	Supported
H3	Retrieval practice may make memory less sensitive to the detrimental impact of stress on memory (e.g., Smith et al., 2016).	Supported
H4	Retrieval practice equalizes memory performance 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 (Smith et al., 2016).	Not supported
Confirmatory Moderator Hypotheses		
Stressor Type	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.	Not supported
Other Strategies	Strategies used other than retrieval practice would be less beneficial than retrieval practice.	NA – re-reading only other strategy used
Delay	Because retrieval practice is shown to have sustained long-term benefit (Roediger & Karpicke, 2006), we expect its protective factor to continue even after a long-term delay (e.g., 1 week) following initial learning.	Not supported
Task Type	Retrieval practice will have a positive effect on any type of learning material used.	Not supported

Table 1 Summarized Results of the Meta-Analysis.

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 [0.17, 0.72]
Egger's regression test	z = 2.33, p = 0.02
Rank correlation test	Kendall's *tau* = 0.50, p = 0.0014
Publication Bias Correction	
Three-parameter selection model	Likelihood Ratio Test: 0.22, p = 0.6 Adjusted Model: g = 0.36, 95% CI [–0.07, 0.80]
PET	b = –1.46 [–3.76, 0.85], p = 0.10
PEESE	b = –0.65 [–1.82, 0.52], p = 0.07

Table 2 Publication bias analyses for H3.

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

k = 9 with 20 effect sizes, g = –0.05, CI [–0.15, 0.06], p = 0.39. Thus, we found no support for H1. Heterogeneity for H1 was not significant ($Q(df$ = 19) = 15.89, p = 0.66, I^2 = 0%), suggesting there was no substantial variability beyond what would be expected by chance.

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, k = 9 with 20 effect sizes, g = 0.37, CI [0.07, 0.67], p = 0.02. Thus, we found support for H2. Additionally, there was significant heterogeneity ($Q(df$ = 19) = 114.84, p < 0.01, I^2 = 86%), suggesting considerable variability across studies likely due to moderating factors.

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 another strategy on memory

MODERATOR	*k*	*Q*	*df*	*g*	95% CI	*Tau*²	I^2	Diff	*p*	CATEGORIES
Stressor Type										
TSST	16	102	15	0.42	0.08–0.77	0.42	85.25%	0.27	0.60	TSST vs. TA
TA	4	4	3	0.54	0.30–0.78	0.02	27.70%			
Delay										
Immediate (<1 day)	6	35	5	–0.03	–0.50–0.45	0.30	85.03%	2.74	0.09	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	14	102	13	0.43	0.03–0.83	0.50	87.91%	0.30	0.08	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%			

Table 3 Results of moderator analyses for H3.

Note. *k* = number of effect sizes, *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 differences between moderator groups).

	STUDY	*N*	COUNTRY	SAMPLE	DESIGN	PUBLISHED
1	Pastötter 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	Pastötter et al., 2023	128	Germany	Students	Between	Yes
10	Hupbach & Feiman, 2012	75	USA	Students	Between	Yes

Table 4 Studies included in the meta-analysis.

Note. Table shows included articles in meta-analysis, with their sample size, country of origine, type of sample, experimental design, and whether they are published. Experiments within each article are not listed here for clarity.

performance in a stress condition. The mean effect was positive, *k* = 9 with 20 effect sizes, *g* = 0.45, CI [0.17, 0.72], *p* < 0.01. Thus, we found support for H3. Additionally, there was significant heterogeneity (*Q*(*df* = 19) = 109.11, *p* < 0.01, I^2 = 84%), suggesting considerable variability across studies likely due to moderating factors.

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, *k* = 10 with 22 effect sizes, *g* = 0.09 CI [–0.01, 0.19], *p* = 0.07. Thus, we found no support for H4. Additionally, heterogeneity was not significant (*Q*(*df* = 21) = 22.53, *p* = 0.37, I^2 = 5.57%), suggesting there was no substantial variability beyond what would be expected by chance.

EFFECT OF STRESSORS

To examine the effect of stress manipulation on memory performance, we submitted participants' 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 article (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)

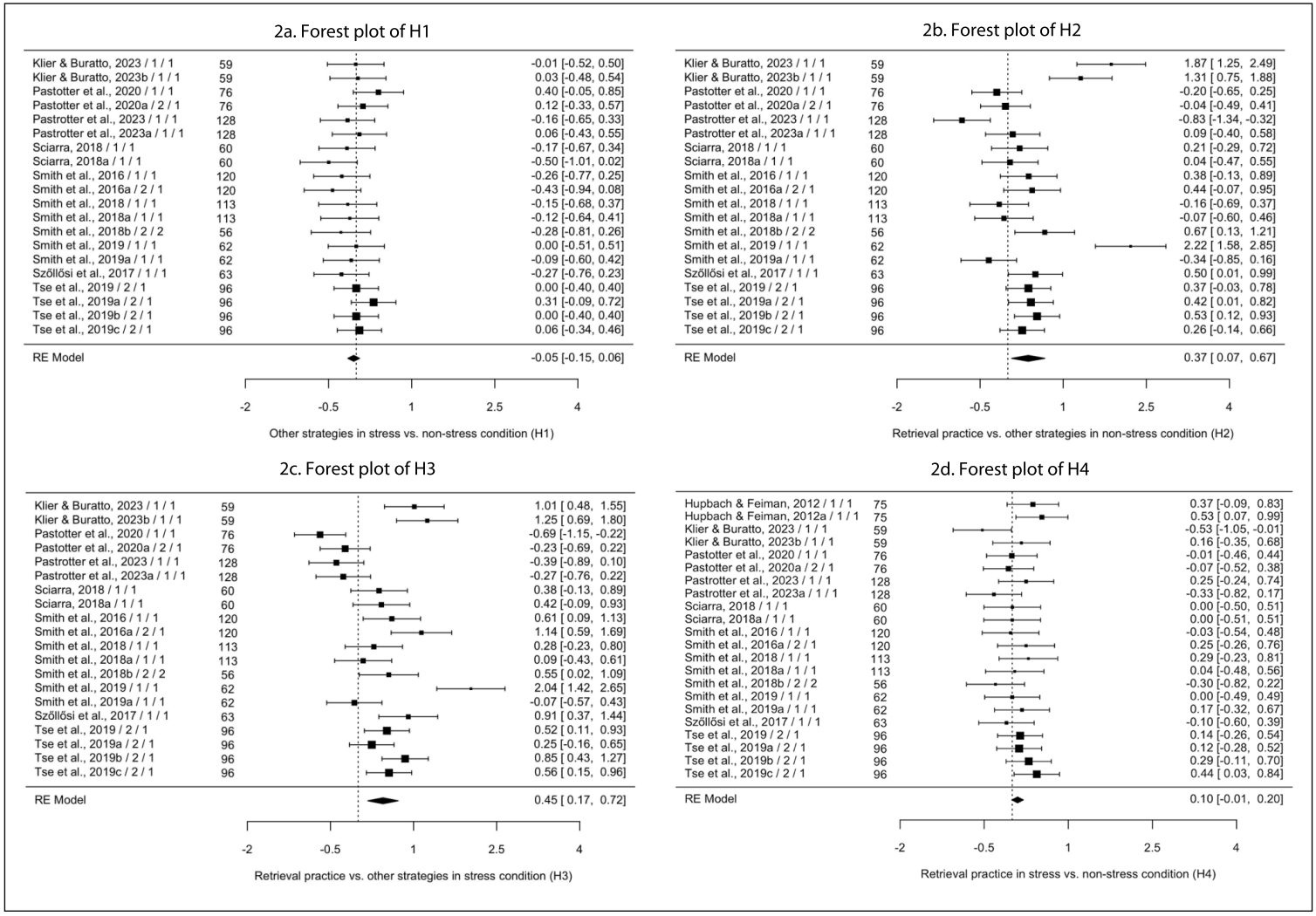


Figure 2 Forest plots of H1, H2, H3, H4.

Note for all plots. Forest plot of all articles (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 left represents the overall effect size and its 95% confidence interval under the random-effects model. Article name (e.g., Pastötter et al., 2020), followed by study number, experiment number within the study, and sample number (e.g., Pastötter et al., 2020 / 1 / 2 / 1). Letters (e.g., "a", "b") indicate multiple studies within the same article. The dotted line in the center represents the overall mean effect for each H. Note that some effect sizes (e.g., Smith et al., 2019) were calculated from reported standard errors and small rounding differences in SEs can substantially affect Hedges' *g*, so this estimate should be interpreted with caution. Note the overall effect for H4 is rounded up in the forest plot.

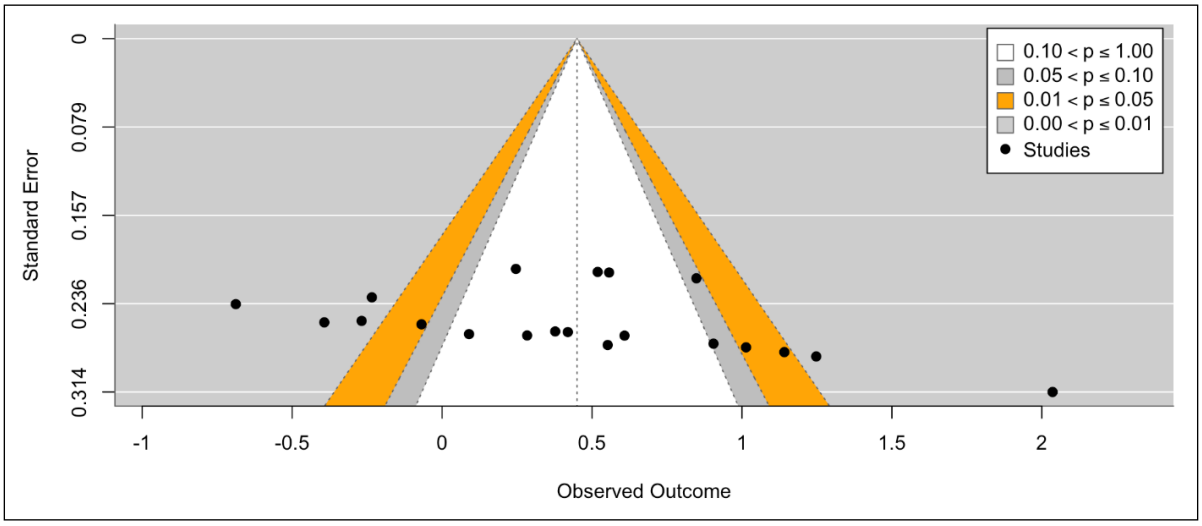


Figure 3a Funnel plot of H3.

Note. Funnel plot for H3 showing the effect sizes on the *x*-axis and standard errors on the *y*-axis. Studies are evenly distributed throughout the plot, though not many studies are seen at the top of the funnel, indicating a lack of studies with very small (i.e., precise) standard errors. The dotted line in the center represents the overall mean effect.

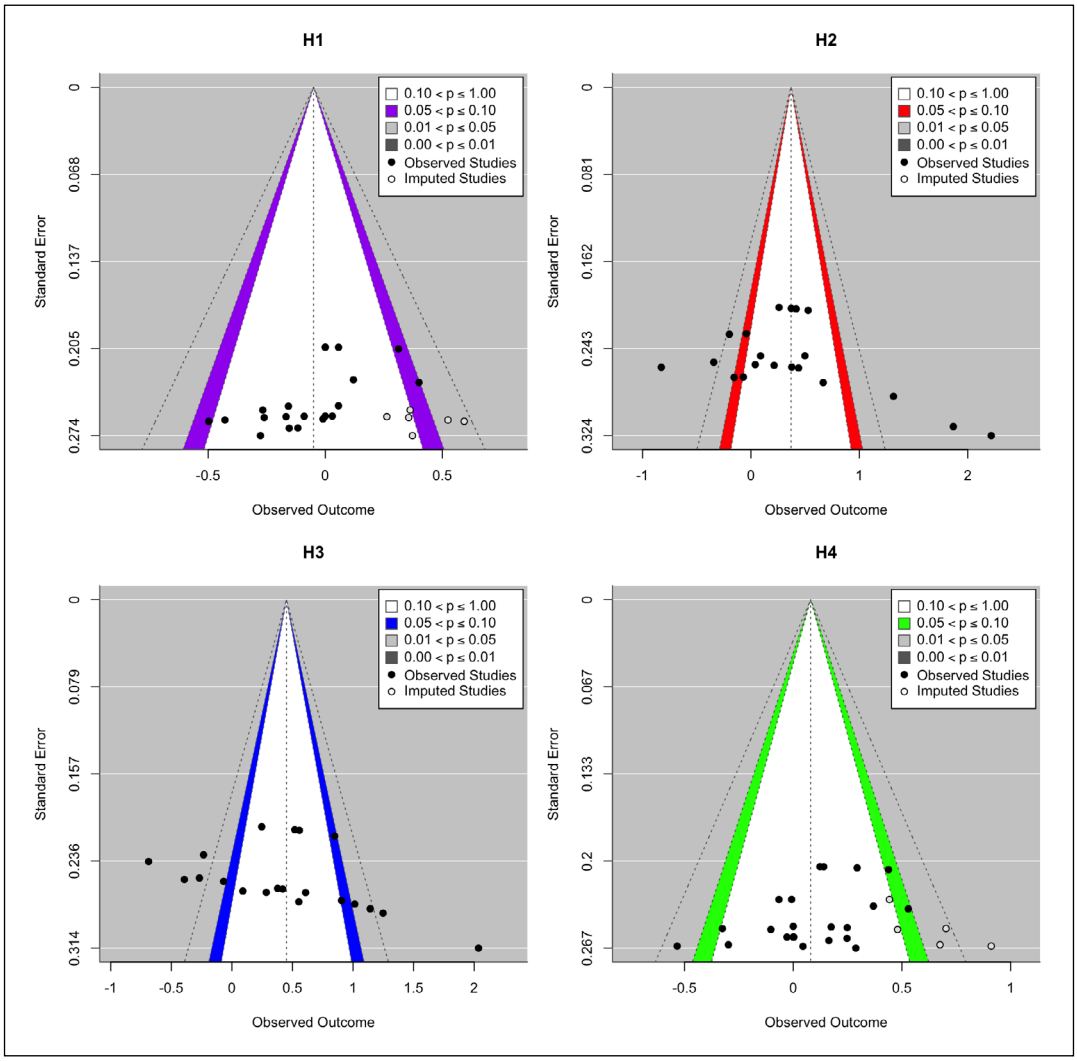


Figure 3b Funnel plots of studies in H1, H2, H3, and H4.

Note. Funnel plot of all studies per hypothesis 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 six and five imputed studies from the trim-and-fill procedure.

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.

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 9.28%–32.63%; for H2, estimates ranged from 8.86%–99.97%; for H3, estimates ranged from 28.62%–99.99%; and for H4, from 5.13%–52.13%. Overall, H3 was the most well-powered of all hypotheses, with several studies approaching high power. 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 4), which summarizes the median statistical power across multiple meta-analyses (i.e., in our case, the four Hs) and illustrates the ability of the studies within each H to detect the overall effect found for each 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.

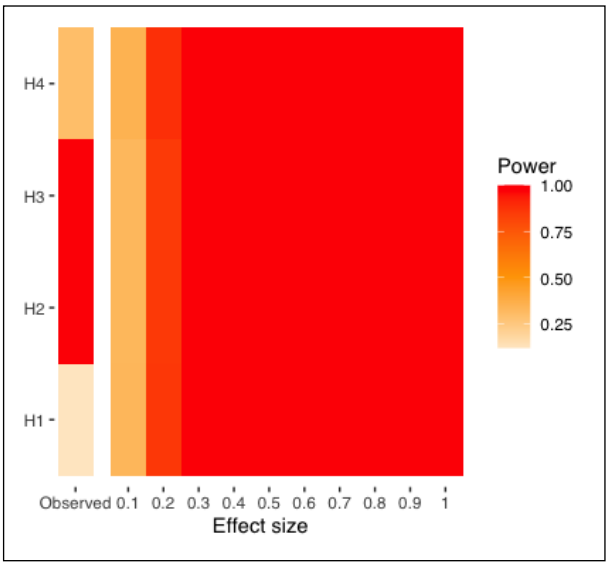


Figure 4 Fireplot visualizing power across all Hs.

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

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 and correction methods, employing six

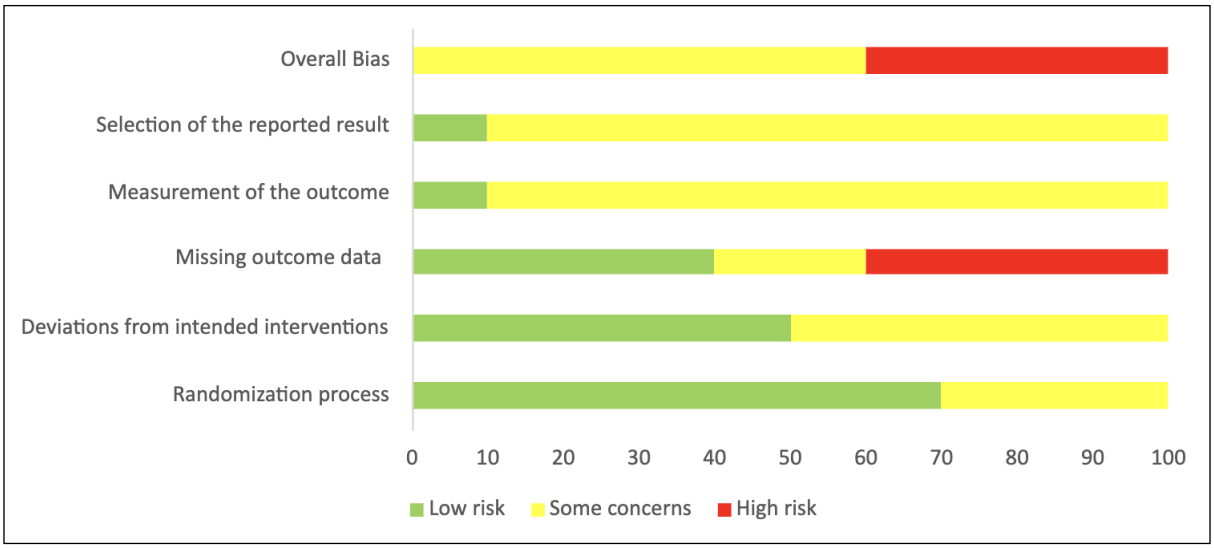


Figure 5 Risk of bias ratings across included articles.

Note. Summary table of the risk of bias in all included articles 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.

different statistical approaches to examine publication bias according to recommendations from the field (Carter et al., 2019; Harrer et al., 2021). A summary of publication bias analyses is provided in Table 2 for H3 and in Supplementary for H1, H2, and H4. There were discrepancies in the methods. For instance, for H3, no missing studies were identified using the trim-and-fill procedure, however Egger's test and Kendall's *tau* were both significant, indicating funnel plot asymmetry. Visual inspection of the funnel plot (Figure 3a) did not greatly suggest this, as most articles were clustered in the center, though studies with very small standard errors (i.e., standard errors < 0.1) were not observed. 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-and-fill imputing six and five missing studies (Figure 3b), as well as significant PET-PEESE intercepts suggesting a positive effect. However, as neither H1 nor H4 are supported, this more likely indicates no effect for these hypotheses. 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, and no extreme or influential cases. Despite the robustness, there was significant heterogeneity in the dataset, as indicated by high Q statistics and I^2 values, meaning that there was considerable variability in effect sizes across the studies included in the meta-analysis.

METAFOREST MODERATOR ANALYSES

To address the problem of low number of studies and lack of statistical power without risk overfitting, we adopted the 'MetaForest' package (van Lissa, 2017, v0.1.3). 'MetaForest' uses random forests, a machine learning technique, and bootstrapping to examine several possible moderators. Results suggested significant heterogeneity among the effect sizes. The main model indicator, R^2 out-of-bag (R^2-OOB), was 0.29, suggesting that the included moderators explained approximately 29% 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 Stressor Type and Type of Task. Other Strategies had a negligible impact.

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 = 109.11, p < 0.01, df = 19), suggesting the H3 sample is highly heterogenous. This was further supported by an I^2 value of 84%, suggesting an 84% chance of the results being due to heterogeneity rather than chance. The significant heterogeneity warrants a deeper look to examine the sources of heterogeneity. To this aim, we examined four theoretical and methodological moderators according to a pre-registered criteria in the Full Coding Sheet: Stressor Type (TSST vs. TA stress), Other Strategies (restudy only), Delay (immediate delay, one day, one week, two-day delay), and Type of Task (words/lists, facts/short answer, educational texts). The results of moderator analyses are summarized in Table 3 for the main hypothesis (H3) and in Supplementary for H1, H2, and H4. Because all studies employed the same 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.

Stressor Type. Sixteen effect sizes for TSST stress had an effect size of g = 0.42, CI [0.08, 0.77], 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. Six effect sizes for a delay of less than one day (i.e., immediate) had an effect size of g = –0.03, CI [–0.50, 0.45], p = 0.91. 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.

As suggested during Stage 2 review, we also ran delay as a continuous moderator. This analysis found a significant effect (QM = 7.87, p = 0.005) for H3, indicating longer delays were associated with increased memory performance. This model accounted for approximately 31% of the between-study heterogeneity (R^2 = 31.36%), though residual heterogeneity remained high (I^2 = 77.48%). The results for the other Hs are presented in Supplementary.

Task Type. Fourteen effect sizes for word or word lists had an effect size of g = 0.43, CI [0.03, 0.82], p = 0.03. 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.

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 5). 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.

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 on memory performance in stress conditions, k = 9 with 20 effect sizes, g = 0.45, CI [0.17, 0.72]. These results support our primary hypothesis (H3) and suggest retrieval practice is more effective than restudy for memory during stressful situations. Meta-regression analysis revealed a nonsignificant impact of stress induction on memory performance. We did not detect significant publication bias in this sample nor any significant moderator results.

RETRIEVAL PRACTICE MAIN EFFECTS

The current work found that retrieval practice significantly improved memory performance compared to other strategies under stress conditions, supporting H3. This aligns with previous findings that retrieval practice is effective in the context of stressors (Smith et al., 2016) and with studies showing its benefit for other stressors like test anxiety (Yang et al., 2023). This also aligns with the context-dependent explanation of retrieval practice, 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 et al., 2014; Karpicke, 2017).

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 = 0.09 CI [–0.01, 0.19], indicating no statistically significant effect of the effectiveness of retrieval practice between stress and non-stress conditions. These findings align to some extent with those of Smith and colleagues (2016) and Szőllősi and colleagues (2017), who observed relatively similar memory performance levels between stressed and non-stressed participants using retrieval practice. It is also possible that here, the repeated binding of contextual cues during retrieval (Karpicke et al., 2014; Karpicke, 2017) contributed to strengthening the memory trace and making it robust during stress situations, but further research is needed to substantiate this explanation.

The evidence for our other hypotheses was mixed. First, we found a null effect when only restudy was used in a stress versus non-stress setting, g = –0.05, CI [–0.15, 0.06], failing to support H1. This means there was no evidence for a difference in memory performance between stress and non-stress conditions for participants who learned with a control strategy. This does not align with other research showing the detrimental effects of retrieval stress on memory (Schwabe & Wolf, 2010; Shields et al., 2017). We also found a positive effect for retrieval practice versus restudy in non-stress conditions, g = 0.37, CI [0.07, 0.67], 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 retrieval practice (H4) or control strategies (H1) are used, as seen by our equivalence tests showing that effects were statistically equivalent to zero within meaningful small-effect bounds. Thus, our

findings suggest that retrieval practice may offer a stable benefit compared to control strategies regardless of stress, rather than specifically protecting against stress. This was further supported by our meta-regression analyses which did not reveal a significant effect of stress scores on memory performance, suggesting the direct impact of stress levels on memory performance remains unclear. 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-and-fill procedure suggested no 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, and correction methods did not alter the adjusted effect sizes. This means the benefits of retrieval practice compared to restudy in both stress and non-stress conditions hold up 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-and-fill procedure imputed six "missing" small studies. Moreover, 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 as H1, with five missing studies imputed and significant PET-PEESE, suggesting positive corrected effects. However, due to the null findings for H1 and H4, the presence of publication bias does not necessarily indicate a hidden effect 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-and-fill methods are not robust when heterogeneity is high (H2 and H3) and when the power is low (Harrer et al., 2021), as is the case for all hypotheses.

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 across the board, suggesting that the positive effects of retrieval practice versus restudy are consistent and unaffected by various study characteristics and contexts. In H4, two-day delays showed stronger effects than one-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 also reversed in H2, which showed a significant difference for one-week delays compared to one-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 in H1 and word lists in H4. This suggests restudying 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 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). This, of course, does not mean that an effect does not exist, just that we did not detect one here.

One possible explanation is that the benefits of retrieval practice might be robust enough to negate the effects of stress altogether. This has already been shown in studies where retrieval practice successfully decreased test anxiety stressors (e.g., 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, 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 other responsibilities, could benefit from retrieval-based strategies to enhance memory resilience for daily life tasks like remembering a grocery list. The adaptability of retrieval practice makes it a practical and easy-to-use tool for promoting cognitive stability across various situations.

At the same time, high heterogeneity in the overall results for H3 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 (e.g., 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. Though the presence of a benefit even amid this variability underscores the robustness of retrieval practice, more research is needed to determine how specific design choices interact with cognitive and contextual factors to influence learning outcomes.

LIMITATIONS AND FUTURE DIRECTIONS

Restudy was the only control learning strategy utilized across all 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. All studies were conducted with student populations. This restricts the extent to which the findings can be generalized to other groups, such as professionals, older adults, or individuals from non-academic contexts. Caution is therefore warranted when interpreting these results beyond student samples, and future research should aim to replicate and extend these findings across other age groups and populations.

In addition, our included studies focused predominantly on the backward testing effect, however two studies reported effects also for the forward testing effect, which is the notion that retrieval practice enhances information for newly learned material (Pastötter et al. 2020). Pastötter and colleagues found that acute psychosocial stress did not affect the forward testing effect, suggesting that this effect is resilient to stress (effect sizes of g = 0.48 in Pastötter et al., 2020; g = 0.45 in Pastötter et al, 2023). The authors discuss this could be due to the different mechanism of action of the forward testing effect compared to the backward testing effect. So far, these are the only two studies examining the forward testing effect in stress conditions, and the consistent results obtained across both points to the forward testing effect being potentially even more beneficial in high-stress learning. This opens avenues for future research to further investigate the forward testing effect in these settings, focusing on its underlying mechanisms, practical applications and comparison against the backward testing effect.

Most studies exclusively utilized word lists for learning material, which were not consistently strong as seen in our moderator analyses. This raises 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 also found differing effects for test format on memory (i.e., cued recall vs. recognition tests) during stress (see Shields 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 research 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 shows that their findings may be less reliable and could contribute to variability in the overall results. Post-hoc power analysis revealed that the tests used in

each hypothesis were not sufficiently powered on their 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.

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 [0.17, 0.72], and a similar effect size for retrieval practice versus restudy in non-stress conditions, $g = 0.37$, CI [0.07, 0.67], meaning retrieval practice is more beneficial 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 is particularly vulnerable to the effects of stress when used alone. Mixed findings in moderator analyses across the Hs suggest factors like testing delay and material type may influence outcomes but not across all situations. Trends in publication bias analyses were generally mixed but suggested possible overestimation of effects. Additionally, the inadequately powered studies in this meta-analysis indicate 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 this is an emerging field requiring more sufficiently powered studies to clarify the conditions under which retrieval practice is or isn't protective.

ACKNOWLEDGEMENTS

This meta-analysis was conducted and reviewed through the *Peer Community In Registered Reports* (PCI RR) process. We are grateful to the PCI RR community, our recommender, Dr. Thomas Evans, and the reviewers for their feedback, guidance, and recommendation of this registered report, and for supporting open and rigorous research practices.

COMPETING INTERESTS

The authors declare no potential conflicts of interests with respect to the authorship of this article. The authors declare that Nicolas Rothen serves as Editor-in-Chief of *Swiss Psychology Open.*

AUTHOR CONTRIBUTIONS

Mariela Mihaylova conducted the meta-analysis as part of her PhD thesis, with Marie Zocca providing support. Both Matthias Kliegel and Nicolas Rothen supervised, guided, and provided edits for the meta-analysis at each step.

ROLE	MM	MZ	MK	NR
Conceptualization	x		x	x
Pre-testing	x	x		
Pre-registration	x			
Data curation	x			
Formal analysis	x	x		
Funding acquisition	*NA*	*NA*	*NA*	*NA*
Investigation	x	x		x
Pre-registration peer review/ verification			x	x
Literature search	x	x		
Datafile study/effect coding	x	x		
Reproducible code (e.g., RMarkdown)	x			
Contacting authors	x	x		
Data analysis peer review/ verification	x	x		
Methodology	x		x	x
Project administration				x
Resources			x	x
Software	x			
Supervision			x	x
Validation	x	x		
Visualization	x			
Writing-original draft	x	x		
Writing-review and editing	x		x	x

AUTHOR INFORMATION

Mariela Mihaylova is a healthcare and scientific communications consultant in the private sector. She obtained her PhD from the University of Geneva and UniDistance Suisse. Her research focuses on memory and learning.

Marie Zocca is a PhD candidate at the department of basic neurosciences of the University of Geneva. Her research focuses on the neurobiological underpinnings of social representations.

Matthias Kliegel is a professor of Psychology at the University of Geneva. His research focuses on prospective memory and aging.

Nicolas Rothen is a professor of Psychology and Rector of UniDistance Suisse. His research focuses on memory and learning.

AUTHOR AFFILIATIONS

Mariela Mihaylova orcid.org/0000-0003-3128-116X
Current work conducted while at the Faculty of Psychology, UniDistance Suisse, Switzerland; Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland

Marie Zocca
Faculty of Medicine, Department of Basic Neurosciences, University of Geneva, Switzerland

Matthias Kliegel orcid.org/0000-0002-2001-2522
Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland

Nicolas Rothen orcid.org/0000-0002-8874-8341
Faculty of Psychology, UniDistance Suisse, Switzerland

REFERENCES

Agarwal, P. K., et al. (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., Nunes, L. D., & **Blunt, J. R.** (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

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

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. https://doi.org/10.1037/amp0000191

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. https://doi.org/10.1177/2515245919838783

Begg, C. B., & **Berlin, J. A.** (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *151*(3), 419–445. https://doi.org/10.2307/2982993

Begg, C. B., & **Mazumdar, M.** (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088–1101. https://doi.org/10.2307/2533446

Borenstein, M., Hedges, L. V., Higgins, J. P., & **Rothstein, H. R.** (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Borenstein, M., & **Higgins, J. P. T.** (2013). Meta-Analysis and Subgroups. *Prevention Science*, *14*(2): 134–43. https://doi.org/10.1007/s11121-013-0377-7

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of cognition*, *2*(1). https://doi.org/10.5334/joc.72

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. https://doi.org/10.1177/2515245919847196

Cassady, J. C., & **Johnson, R. E.** (2002). Cognitive TA and Academic Performance. *Contemporary Educational Psychology*, *27*(2), 270–295. https://doi.org/10.1006/ceps.2001.1094

Cheung, M. (2014). Modeling Dependent Effect Sizes with Three-Level Meta-Analyses: A Structural Equation Modeling Approach. *Psychological Methods*, *19*(2), 211. https://doi.org/10.1037/a0032968

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

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. *Nature Neuroscience*, *3*(4), 313–314. https://doi.org/10.1038/73873

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

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

Duval, S., & **Tweedie, R.** (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

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

Fanelli, D., Wong, J., & **Moher, D.** (2021). What difference might retractions make? An estimate of the potential epistemic cost of retractions on meta-analyses: What difference might retractions make? An estimate of the epistemic impact of retractions on recent meta-analyses. *Accountability in Research*. https://doi.org/10.1080/08989621.2021.1947810

Feldman, G. (2019a). HKU Registered Report template: Supplementary.

Feldman, G. (2019b). Meta-Analysis Template Version: 2 (April 1, 2019).

Feltz, A., & **May, J.** (2017). The means/side-effect distinction in moral cognition: A meta-analysis. *Cognition*, *166*, 314–327. https://doi.org/10.1016/j.cognition.2017.05.027

Gagnon, S. A., Waskom, M. L., Brown, T. I., & **Wagner, A. D.** (2019). 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

Grames, E. M., Stillman, A. N., Tingley, M. W., & **Elphick, C.S.** (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution, 10*, 1645–1654. https://doi.org/10.1111/2041-210X.13268

Harrer, M., Cuijpers, P., Furukawa, T., & **Ebert, D. D.** (2019). Package 'dmetar'. *R Package Version 0.0.9000, 2019*.

Harrer, M., Cuijpers, P., Furukawa, T., & **Ebert, D.** (2021). Doing meta-analysis with R: A hands-on guide. Chapman and Hall/CRC. https://doi.org/10.1201/9781003107347

Hembree, R. (1988). Correlates, Causes, Effects, and Treatment of TA. *Review of Educational Research*, *58*(1), 47–77. https://doi.org/10.3102/00346543058001047

Higgins, J. P., & **Thompson, S. G.** (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P., Thompson, S. G., Deeks, J. J., & **Altman, D. G.** (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

***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. https://doi.org/10.1002/acp.3032

Hohn, R. E., Slaney, K. L., & **Tafreshi, D.** (2020). An Empirical Review of Research and Reporting Practices in Psychological Meta-Analyses. *Review of General Psychology*. https://doi.org/10.1177/1089268020918844

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

***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. https://doi.org/10.1037/a0030489

Iyengar, S., & **Greenhouse, J. B.** (1988). Selection models and the file drawer problem. *Statistical Science*, 109–117. https://doi.org/10.1214/ss/1177013012

Karpicke, J. D. (2017). Retrieval-Based Learning: A Decade of Progress. *Grantee Submission*. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., Lehman, M., & **Aue, W. R.** (2014). Retrieval-based learning: An episodic context account. In *Psychology of learning and motivation* (Vol. 61, pp. 237–284). Academic Press. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & **Roediger, III, H. L.** (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of experimental psychology: learning, memory, and cognition*, *33*(4), 704. https://doi.org/10.1037/0278-7393.33.4.704

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. *Neuropsychobiology*, *28*(1–2), 76–81. https://doi.org/10.1159/000119004

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

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*. https://doi.org/10.1016/j.nlm.2004.09.001

Lindsay, D. S. (2020). Seven steps toward transparency and replicability in psychological science. *Canadian Psychology/Psychologie canadienne*. Advance online publication. https://doi.org/10.1037/cap0000222

Lüdecke, D. (2019). Package 'esc'. *R Package Version 0.5.1, 2019*

Maassen, E., van Assen, M., Nuijten, M., Olsson-Collentine, A., & **Wicherts, J.** (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0233107

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

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & **PRISMA Group*, T.** (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Moreau, D., & **Gamble, B.** (2020a, May 4). *Meta-analysis templates and materials*. Retrieved from http://osf.io/q8stz

Moreau, D., & **Gamble, B.** (2020b, July 18). *Conducting a Meta-Analysis in the Age of Open Science: Tools, Tips, and Practical Recommendations*. https://doi.org/10.31234/osf.io/t5dwg

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

Nuijten, M. B., & **Polanin, J. R.** (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods*. 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. https://doi.org/10.1177/2515245920918872

Page, M. J. (2020, June 18). Updating the PRISMA reporting guideline for systematic reviews and meta-analyses. https://doi.org/10.17605/OSF.IO/P93GE

***Pastötter, B., von Dawans, B., Domes, G.,** & **Frings, C.** (2020). The forward testing effect is immune to acute psychosocial encoding/retrieval stress. *Experimental Psychology*. https://doi.org/10.1027/1618-3169/a000472

***Pastötter, B., von Dawans, B., Domes, G.,** & **Frings, C.** (2023). The forward testing effect is resistant to acute psychosocial retrieval stress. *Experimental Psychology*. https://doi.org/10.1027/1618-3169/a000571

Pick, J. L., Nakagawa, S., & **Noble, D. W. A.** (2018). Reproducible, flexible and high-throughput data extraction from primary literature: The metaDigitise R package. Biorxiv. https://doi.org/10.1101/247775

Piroozmanesh, A., & **Imanipour, M.** (2018). The effect of formative assessment on test anxiety of nursing students. *Journal of Medical Education Development*, *10*, 18–26. https://doi.org/10.29252/edcj.10.28.18

Polanin, J. R., Espelage, D. L., Grotpeter, J. K., Valido, A., Ingram, K. M., Torgal, C., ... & **Robinson, L. E.** (2020a). Locating unregistered and unreported data for use in a social science systematic review and meta-analysis. *Systematic Reviews, 9*, 1–9. https://doi.org/10.1186/s13643-020-01376-9

Polanin, J. R., Hennessy, E. A., & **Tsuji, S.** (2020b). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620906416

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. *Research Synthesis Methods, 10*(3), 330–342. https://doi.org/10.1002/jrsm.1354

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). https://doi.org/10.1177/25152459221147260

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Re, A. C. D. (2020). Package "compute.es". *R Package Version 0.2–5, 2020*.

Reber, T. P., & **Rothen, N.** (2018). Educational App-Development needs to be informed by the Cognitive Neurosciences of Learning & Memory. *npj Science of Learning, 3*(1), 22. https://doi.org/10.1038/s41539-018-0039-4

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. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rohatgi, A. (2020). WebPlotDigitizer user manual version 4.4. *URL* http://arohatgi.info/WebPlotDigitizer/app, *1–18*.

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

Schwabe, L., Haddad, L., & **Schachinger, H.** (2008). HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology*, *33*(6), 890–895. https://doi.org/10.1016/j.psyneuen.2008.03.001

***Sciarra, S.** (2018). *Do The Benefits Of Retrieval Practice Remain Under Stress?* (Doctoral dissertation).

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.** (2010). 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

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

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), 636–675. https://doi.org/10.1037/bul0000100

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. *Annual Review of Psychology, 70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

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. https://doi.org/10.1037/gpr0000140

***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. https://doi.org/10.1037/bne0000240

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

***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. https://doi.org/10.1016/j.bandc.2018.12.005

Smith, A. M., & **Thomas, A. K.** (2018). 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

Stanley, T. D., & **Doucouliagos, H.** (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

Sterne, J. A., & **Egger, M.** (2005). Regression methods to detect publication and other bias in meta-analysis. *Publication bias in meta-analysis: Prevention, assessment and adjustments*, 99–110. https://doi.org/10.1002/0470870168.ch6

Sterne, J. A., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., ... & **Higgins, J. P.** (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *bmj*, *366*. https://doi.org/10.1002/9781119536604.ch8

***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. https://doi.org/10.1002/acp.3363

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. https://doi.org/10.1073/pnas.1221764110

***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. https://doi.org/10.3389/fpsyg.2019.00969

van Lissa, C. J. (2017). MetaForest: Exploring heterogeneity in meta-analysis using random forests. Retrieved from https://psyarxiv.com/myg6s/. https://doi.org/10.31234/osf.io/myg6s

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Vogel, S., & **Schwabe, L.** (2016). Learning and memory under stress: implications for the classroom. *npj Science of Learning, 1*(1), 1–10. https://doi.org/10.1038/npjscilearn.2016.11

Walters, W. H. (2007). Google Scholar coverage of a multidisciplinary field. *Information Processing & Management*, *43*, 1121–1132. https://doi.org/10.1016/j.ipm.2006.08.006

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. https://doi.org/10.1007/s10648-023-09801-w

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

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? *Journal of Experimental Psychology: Applied*, *26*, 724–738. https://doi.org/10.1037/xap0000278

Yeung, S. K., & **Feldman, G.** (2021). Action-Inaction Asymmetries in Emotions and Counterfactual Thoughts: Meta-Analysis of the Action Effect [Registered Report Stage 1]. https://doi.org/10.17605/OSF.IO/ACM24

Yeung, S. K., Yay, T., & **Feldman, G.** (2021). Action and inaction in moral judgments and decisions: Meta-analysis of Omission-Bias omission-commission asymmetries. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/01461672211042315

Note: Studies with a "*" preceding them in the bibliography denote articles included in the meta-analysis.

TO CITE THIS ARTICLE:
Mihaylova, M., Zocca, M., Kliegel, M., & Rothen, N. (2025). Does Retrieval Practice Protect Memory Against Stress? A Meta-Analysis. *Swiss Psychology Open,* 5(1): 6, pp. 1–24. DOI: https://doi.org/10.5334/spo.101

Submitted: 08 November 2025 **Accepted:** 10 November 2025 **Published:** 27 November 2025

COPYRIGHT:
© 2025 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Swiss Psychology Open is a peer-reviewed open access journal published by Ubiquity Press.