# Does retrieval practice protect memory against stress? A meta-analysis [Stage 1 Registered Report]

^Mariela Mihaylova

Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland

mariela.mihaylova@etu.unige.ch

Matthias Kliegel

Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland

Matthias.kliegel@unige.ch

Nicolas Rothen

Faculty of Psychology, UniDistance Suisse

nicolas.rothen@fernuni.ch

^Corresponding author

Word: abstract – [XXX], manuscript - [XXXX]

**Corresponding author**

Mariela Mihaylova, Faculty of Psychology and Educational Sciences, University of Geneva
mariela.mihaylova@etu.unige.ch

**Author bios:**

Mariela Mihaylova is a PhD candidate at the University of Geneva. Her research focuses on memory and learning.

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

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

**Rights:**

**Declaration of Conflict of Interest:**

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

**Financial disclosure/funding:**

The author(s) received no financial support for the research and/or authorship of this article.

**Authorship declaration:**

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

## Contributor Roles Taxonomy

| Role | MM | MK | NR |
|---|---|---|---|
| Conceptualization | x | x | x |
| Pre-testing | x | | |
| Pre-registration | x | | |
| Data curation | x | | |
| Formal analysis | x | | |
| Funding acquisition | NA | NA | NA |
| Investigation | x | | x |
| Pre-registration peer review / verification | | x | x |
| Literature search | x | | |
| Datafile study/effect coding | x | | |
| Reproducible code (e.g., RMarkdown) | x | | |
| Contacting authors | 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 | | |
| Visualization | x | | |
| Writing-original draft | x | | |
| Writing-review and editing | x | x | x |

## Abstract

[Note: This is a Stage 1 Registered Report. All highlighted sections in Abstract will be replaced with actual results by Stage 2.]

Stressors such as test anxiety are known to decrease memory retrieval, whereas retrieval practice is the phenomenon that actively recalling information from memory enhances memory. Recent evidence suggests retrieval practice can protect memory against the negative effects of stress on memory (Agarwal et al., 2014; Smith & Thomas, 2016), however the findings are mixed (Yang et al., 2020). Determining the overall effects of using retrieval practice to counteract the negative effects of stress on memory could transform our understanding of memory resilience and help design new cognitive interventions to protect memory in stressful situations. This therefore 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* = [enter number of studies by Stage 2], total number of participants = [enter no. of participants by Stage 2]) of the impact of stress on learning with retrieval practice from [year/date of coverage], among [sample characteristics, if applicable, remove if irrelevant], using [databases and other information sources, Beller et al., 2013]. [Describe the eligibility criteria, Beller et al., 2013] We found [weak to no / mixed / substantial / strong] empirical evidence for the [Phenomenon name] hypothesis, [Hedge's *g* / Cohen's *d* / Other Effect Size Measure = X.XX, 95% CI [X.XX, X.XX]]., with [model(s), e.g., multivariate three-level model]. [Phenomenon name] is a meaningful effect for [measure(s)/dependent variable(s)]. Study heterogeneity was [low / low to medium / medium / medium to high / high], [*Q*(degrees of freedom) = XXX.XX, *p* = .XXX / < .001, $I^2$ = XX.XX%]. [Summarize results of publication biases tests]. We tested several moderators: [list of possible moderators tested]. We found that [list of moderator(s) with meaningful moderation, if there were any] moderated [Phenomenon name]. [Phenomenon name] was stronger [list of conditions in which the effects were stronger, if there was/were]. [Brief descriptions of strengths and limitations, Beller et al., 2013, and future research directions] We registered our meta-analysis here, with datafile, code and supplementary: https://osf.io/jwx4f/.

*Keywords:* retrieval practice, meta-analysis, registered report, memory, test anxiety, stress

# Does retrieval practice protect memory against stress? A meta-analysis

# [Stage 1 Registered Report]

## 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 (TA), also known as exam-related stress (Cassady & Johnson, 2002; Dickerson & Kemeny, 2004). A wealth of evidence suggests that retrieval stress, or stress occurring before memory recall, decreases memory subsequent learning (Gagnon et al., 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 (Roediger & Karpicke, 2006; Rowland, 2014). Recent evidence suggests that retrieval practice may have protective effects on memory against stress via memory strengthening mechanisms (Smith et al., 2016). These findings might suggest that memory may be made less sensitive against the detrimental effects of retrieval stress using an easy-to-use learning strategy. To date, the overall effects of retrieval stress on memory after learning with retrieval practice have not been examined in a meta-analytic approach. In this meta-analysis, we aim to explore the protective mechanisms of retrieval practice in the context of retrieval stress.

Stress can be experienced in many different forms. The one of interest for the current meta-analysis is psychological stress, which involves uncontrollable situations or events characterized by socio-evaluative threat, such as one’s performance being evaluated or judged negatively by others (Dickerson & Kemeny, 2004). Examples of these situations include: evaluative situations such as exams and test anxiety (Cassady & Johnson, 2002; Hembree, 1988); the Tier Social Stress Test (TSST), which consists of socially evaluative situations such as making a speech in front of others and being judged (Kirschbaum et al., 1993); or via instruction sets which mention that performance will be judged (Almazrouei et al., 2022). Stress can also be induced through procedures such as the Cold Pressor Test, where participants place their hands in cold water for a specific time, coupled with socio-evaluative elements (Schwabe et al., 2008; Schwabe & Schächinger, 2018). The abovementioned measures to induce stress are typically associated with increased cortisol levels or state anxiety responses which signal a stress response. And, importantly, a plethora of studies suggest that stress induced through these methods is associated with decreases in memory performance and memory retrieval (de Quervain et al., 2000; Kuhlmann, 2005; Kuhlmann 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 7 minutes and then either restudy (reread) the passage or take a short test where they wrote down as much as they could remember from the passage. After a retention interval of 5 minutes, 2 days or one week, participants were asked to recall as much as they could from the initial passages. Results revealed that for the longer retention intervals of 2 days and 1 week, participants in the testing, or retrieval practice, condition performed significantly better than the restudy condition (Roediger & Karpicke, 2006), highlighting the effectiveness of this strategy for long-term learning. Since then, the benefits of retrieval practice have been shown for a wide range of learning materials and retention intervals (Karpicke, 2017; Rowland, 2014;

Schwieren et al., 2017). Critically, retrieval practice is shown to be more effective than commonly used learning strategies such as restudying, highlighting, note-taking or elaborative techniques such as drawing concept maps (Moreira et al., 2019).

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

That is, until Smith and colleagues (2016) examined the protective effects of retrieval practice against retrieval stress for the first time. In their study, 120 participants were split into either a retrieval practice group, who learned material using the retrieval practice strategy, or a study practice group, who re-read the material. Twenty-four hours later, 30 participants from both the retrieval practice and study practice groups underwent TSST stress induction, and the other half underwent a non-stressful control task. At five minutes and 20 minutes into the stress induction, a memory test was administered to observe the immediate and delayed effects of the stress respectively. Results showed that participants who learned via retrieval practice and were exposed to stress outperformed those who restudied and were also exposed to stress (Smith et al., 2016). This study suggests that retrieval practice might protect memory from the otherwise detrimental effects of stress on memory, as well as carry over to other stressors such as during testing situations. Other evidence comes from Agarwal and colleagues (2014), who administered surveys to students in classes involved in a school-wide retrieval practice learning program. When asked if retrieval practice made students more or less nervous for tests and exams, 72% reported that retrieval practice made them feel less nervous for upcoming tests. When asked if they experienced more or less test anxiety for classes in which they underwent the retrieval practice intervention compared to classes where they did not use retrieval practice, only 19% indicated feeling more test anxiety, and over half of students (54%) reported that retrieval practice reduced their test anxiety. Taken together, these results suggest that retrieval practice can help protect memory against stressors.

However, other studies present contradictory findings regarding the protective role of retrieval practice on memory following stress exposure. For example, a recent study by Yang 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, suggesting that test anxiety does not significantly modulate retrieval practice effects. However, other evidence from Clark and colleagues (2018) suggests a positive relationship between test anxiety and using retrieval practice when external incentives are applied, suggesting memory can be protected by retrieval practice in the face of stressors like test anxiety.

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

To summarize, a wealth of evidence suggests that psychological stressors include situation-specific, socio-evaluative situations where individuals' performance is likely to be judged or evaluated such as test anxiety. Stressors experienced at retrieval decrease memory and learning. Retrieval practice has been consistently shown to boost memory and learning, however its protective effects in the face of retrieval stress are mixed. In this meta-analysis, we aim to answer the question of whether retrieval practice can 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 protect memory from 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., Moriera et al., 2019) under non-stressful conditions. We thus hypothesize that retrieval practice will yield memory benefits even in the absence of stress (H2). This hypothesis will be tested by comparing the effects of learning with retrieval practice versus a control strategy in non-stress conditions.

Third, using retrieval practice may make memory less sensitive against the detrimental impact of stress on memory (e.g., Smith et al., 2016). Thus, we expect that retrieval practice will outperform other control strategies in mitigating the memory impairments induced by stress (H3). This hypothesis will be tested by comparing the effects of learning with retrieval practice versus a control strategy in groups that underwent stress induction. And, in a second step, if H3 is confirmed, we will further explore this benefit by comparing the effects of retrieval practice on memory in a stress versus non-stress condition. Here, we expect relatively equal performance when using retrieval practice in a stress versus non-stress condition (H4), as the protective benefit of retrieval practice in the stress condition should make it equivalent with the benefit of the strategy already experienced in the non-stress condition (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 the second moderator according to the type of stressor: TSST or TA. Based on our current understanding of the literature, these are the two main types of stressor tasks. However, additional types of stressor tasks may be added at Stage 2 when we conduct the literature search.

> 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., Moriera 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 coding by categorizing each other strategy used (ie., restudy, highlighting, drawing diagrams.

> Other Strategies. We predict that 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 following learning to memory test (e.g., 1 day, 2 days, 1 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., 1 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 this learning strategy on different types of materials. Thus, we wanted to explore whether the type of learning material used impacts the overall effectiveness of retrieval practice. This moderator was coded as the different types of materials used (e.g., reading passages, word lists, questions, etc.).

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

## Methods

[Note: Written in past tense to demonstrate methods section after completion, but has yet to be conducted. Highlighted parts in yellow will be filled and updated after pre-registration and data collection.]

### Open Science Disclosures

We shared all procedures, materials, datasets, articles, and code on Open Science Framework (https://osf.io/jwx4f/). Systematic data collection has not commenced for this project. There are no other unreported/unlinked pre-registrations for this meta-analysis project. See Open Science Disclosures in Supplementary for details. The templates on OSF and the template for Stage 1 Registered Reports used in this meta-analysis have been adapted from the resources developed by Feldman (2019a, 2019b) and Yeung and colleagues (2021). We made all efforts recommended by the field to enhance reproducibility, openness, and transparency (Lindsay, 2020; Maassen et al., 2020; Moreau & Gamble, 2020b).

### Literature search

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

To find articles relevant on our topic, we used the following databases: Psycinfo, PubMed, JSTOR, SCOPUS, 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 et al., 2014; Schwieren et al., 2017) with the addition of the term "test anxiety." We selected experimental studies published in peer-reviewed journals in English between the years of 2006 – 2022. 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 featured unpublished studies and theses databases (e.g., Thesis Commons, ProQuest) using the same search terms as the database search. We reran the searches at least twice to ensure all literature was up to date. The date last searched was _____. The outcome was a total of YY prospective articles. Following deletion of duplicates, we had a total of XX 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 one additional round of search by skimming the reference sections of identified articles from our primary search. The date last searched was XX. The outcome was a total of YY articles.

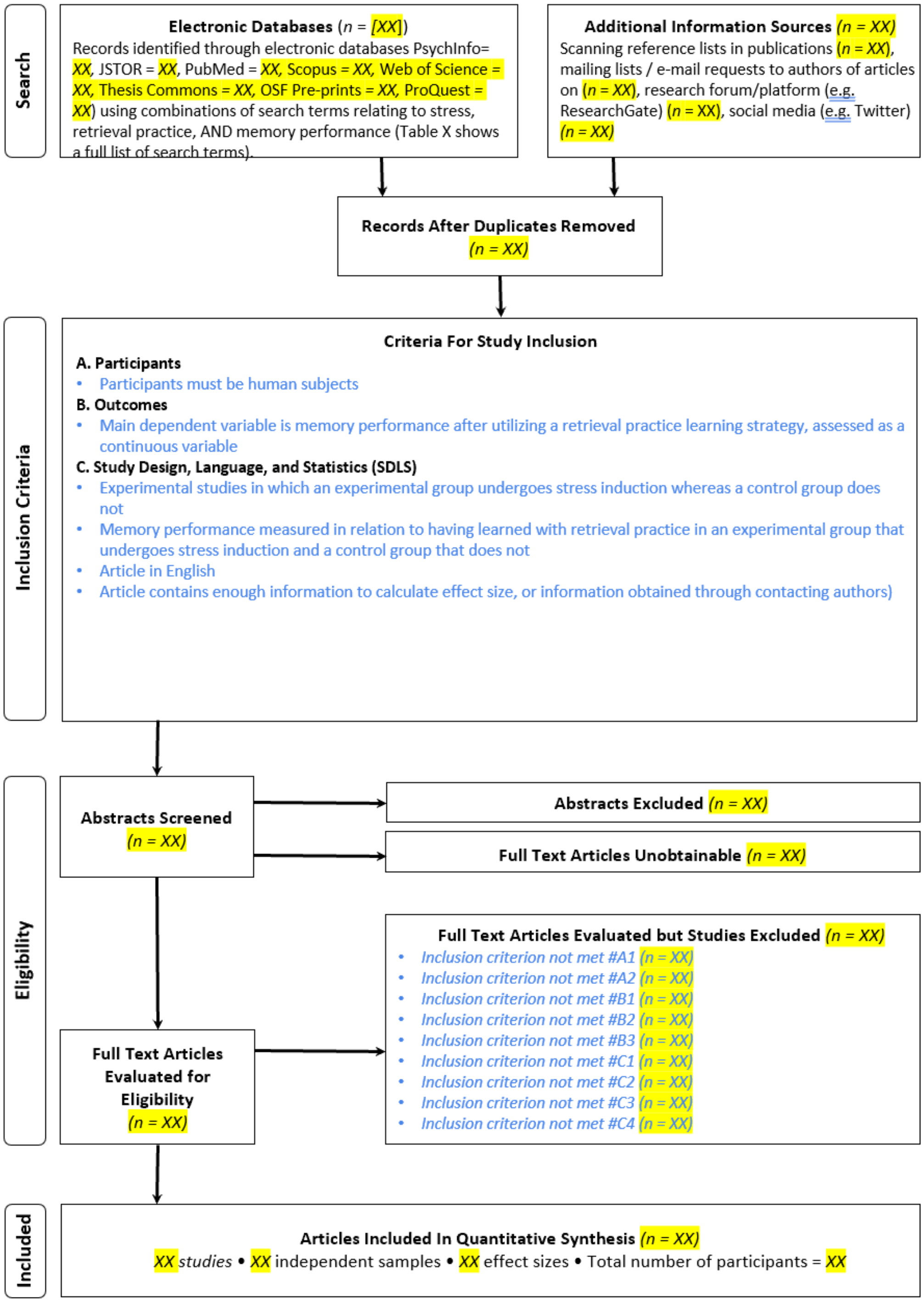
Furthermore, we identified authors in the field of the stress and memory literature along with authors of other identified articles and searched through their related publications. This ensured full coverage and maximized access to unpublished data and/or manuscripts that are also relevant (see Supplementary Materials - Template for Contacting Authors for Published and Unpublished Data for a mail-merge Word template to be sent to relevant authors identified). 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). We first included studies that require contacting the author for the dataset/further clarifications into the main coding sheet, but we documented them as to be excluded potentially, should the author not respond by a given date. We contacted authors of studies with missing necessary statistics for relevant datasets/information. If the original authors of studies provided the dataset, the researchers conducted needed analyses for coding. We documented this process and the relevant results in the "Contacting Authors" tab within the Full Coding Sheet (available on OSF: https://osf.io/jwx4f/). If included, we added the article record in the spreadsheet (tab name "Contacting Authors"). This process was performed on ___ (dates). This resulted in obtaining __ additional articles not found in our search.

In total, we contacted [number of authors, to be entered by Stage 2], [number of authors responded, to be entered by Stage 2] responded, and [number of authors that provided additional relevant papers, data or information, to be entered by Stage 2] provided [number of extra studies included through this search process, to be entered by Stage 2] additional data/papers that are eventually included in our meta-analysis (Appelbaum et al., 2018). Lastly, we issued a call for unpublished findings on online forums, research platforms, and social media (e.g., ResearchGate, listservs, social media) on [insert dates]. We set up a project on [forum(s) and/or platform(s)], and added all identified articles as references, where possible, to notify authors about this project, and to provide an open access list of available studies (link: [insert link]).

After the above search procedures, MM and ___ scanned all abstracts, tables, and method sections to identify the relevance of the sources (see Screening section). If the articles indicated relevance for our analysis, MM and __ read more of the articles to determine whether they met the inclusion criteria or whether articles had to be excluded based on our search criteria (see next paragraph). 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 XX articles, reducing our sample of studies to YY articles with a total of YYYY participants. We listed all the excluded articles in Table 2 of Supplementary and in the Full Coding Sheet.

**Figure 1**

*Systematic literature search flow diagram*



**Electronic Databases** (*n* = *[XX]*)
Records identified through electronic databases PsychInfo= *XX*, JSTOR = *XX*, PubMed = *XX*, Scopus = *XX*, Web of Science = *XX*, Thesis Commons = *XX*, OSF Pre-prints = *XX*, ProQuest = *XX*) using combinations of search terms relating to stress, retrieval practice, AND memory performance (Table X shows a full list of search terms).

**Additional Information Sources** *(n = XX)*
Scanning reference lists in publications *(n = XX)*, mailing lists / e-mail requests to authors of articles on *(n = XX)*, research forum/platform (e.g. ResearchGate) *(n = XX)*, social media (e.g. Twitter) *(n = XX)*

**Records After Duplicates Removed**
*(n = XX)*

**Criteria For Study Inclusion**

**A. Participants**
- Participants must be human subjects

**B. Outcomes**
- Main dependent variable is memory performance after utilizing a retrieval practice learning strategy, assessed as a continuous variable

**C. Study Design, Language, and Statistics (SDLS)**
- Experimental studies in which an experimental group undergoes stress induction whereas a control group does not
- Memory performance measured in relation to having learned with retrieval practice in an experimental group that undergoes stress induction and a control group that does not
- Article in English
- Article contains enough information to calculate effect size, or information obtained through contacting authors)

**Abstracts Screened**
*(n = XX)*

**Abstracts Excluded** *(n = XX)*

**Full Text Articles Unobtainable** *(n = XX)*

**Full Text Articles Evaluated but Studies Excluded** *(n = XX)*
- *Inclusion criterion not met #A1 (n = XX)*
- *Inclusion criterion not met #A2 (n = XX)*
- *Inclusion criterion not met #B1 (n = XX)*
- *Inclusion criterion not met #B2 (n = XX)*
- *Inclusion criterion not met #B3 (n = XX)*
- *Inclusion criterion not met #C1 (n = XX)*
- *Inclusion criterion not met #C2 (n = XX)*
- *Inclusion criterion not met #C3 (n = XX)*
- *Inclusion criterion not met #C4 (n = XX)*

**Full Text Articles Evaluated for Eligibility**
*(n = XX)*

**Articles Included In Quantitative Synthesis** *(n = XX)*
*XX studies* • *XX* independent samples • *XX* effect sizes • Total number of participants = *XX*

*Note*. *Highlighted Italic* - to be entered by Stage 2. The above template is adapted from Moher et al. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement (www.prisma-statement.org), as well as Moreau and Gamble (2020) 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.

## 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 protected memory following stress exposure, we established strict inclusion and exclusion criteria.

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

Second, the experimental studies we focused on had to include adequate statistical information for computing the effect size for the effects of retrieval practice on memory following stress induction. Namely, the article needs to report means, standard deviations, and sample sizes for both the experimental and control groups who learned with retrieval practice versus another strategy (if included) after undergoing the stress procedure. Alternatively, articles need to include the effect size that represents the magnitude of having learned with retrieval practice in a group that underwent a stressor versus a control or for the interaction effects between retrieval practice and a control strategy between stress and control groups. In cases of missing statistical data, we first attempted to contact the authors (Polanin et al., 2020a) and tried to extract required statistics from plots with WebPlotDigitizer (Rohatgi, 2020) or metaDigitise (Pick et al., 2018). If we were not able to obtain the required statistics, we excluded the articles even if the articles met all other search criteria. We excluded all correlational studies and other non-experimental studies.

Third, we excluded articles not written in English, unless we obtained all necessary data and information for coding in English, or we obtained such data and information from the authors. Fourth, we excluded retracted studies if the retraction is due to problems of data collection and data analysis (Fanelli et al., 2021).

### Screening

Studies collected through database searches and through contacting authors were assessed for their eligibility based on their titles, abstracts, and contents. Titles were first scanned to identify the relevance of the sources. Relevant titles were rated with a “1” and irrelevant titles with a “0” by two independent raters (MM and ___). If the titles indicated relevance for our analysis, the 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). 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 were carefully inspected to ensure they met all inclusion criteria and were eligible for inclusion.

The methods sections were independently checked by MM and ___ using separate spreadsheets. Raters met regularly to update on progress and discuss any disagreements at each stage. Disagreements were resolved through deliberations with a third senior member. Inter-rater reliability scores were assessed before and after rater deliberation. All decisions for inclusion and exclusion were documented clearly, transparently, and systematically in the excel spreadsheet Full Coding Sheet spreadsheet (tab name "Article List Inclusion and Exclusion Criteria," see OSF https://osf.io/jwx4f/). We saved all preliminary references of the studies in the total search into a list available on OSF. The open-access full-texts will be accessible on OSF at Stage 2.

### Coding and pre-testing

We developed a data coding sheet (tab name "Coding" in the Full Coding Sheet) and a Codebook (available on OSF) to keep a clear record of our decisions at different stages and enhance reproducibility (Arslan, 2019; Obels et al., 2020; Siddaway et al., 2019). Before we began with the coding process, we pilot-tested 2 randomly selected studies in two stages and refined it accordingly in every stage. MM and [number of coders] completed the coding process for the pretests to ensure a higher inter-rater reliability. We documented gaps and reported decisions in detail in the “Article and Decision” tab of the Full Coding Sheet (see OSF https://osf.io/jwx4f/). XX and XX coded all studies. 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 studies, MM and [names of responsible authors/coders] coded the studies independently using the Coding sheet. A Codebook with instructions (see OSF https://osf.io/jwx4f/) on how to code each column was provided to all coders during the coding process. All coders completed the coding individually and met regularly with the main contributor (MM) to discuss progress. 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 the statistical analyses with packages for meta-analysis such as metafor (Viechtbauer, 2010, v3.8-1). We used the analysis templates adapted from Yeung and colleagues (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* during analysis to facilitate comparison. Multiple effect sizes (i.e., different measures within the same study) were handled by computing a separate effect size for each different relevant scenario described in the article (Appelbaum et al., 2018). For missing data (e.g., effect size missing, but *M* and *SD* reported), we calculated using packages such as esc (Lüdecke, 2019, v0.5.1) or compute.es (Re, 2020, v0.2-5). Calculation or coding procedures, as well as all packages and functions used, were documented in the Full Coding Sheet ("Included Studies Effect Coding" tab).

Whenever standardized effect sizes were not available, we used either descriptive statistics or inferential statistics, such as Mean and Standard Deviation, Chi-Square Statistics, Count, *t*-statistics. We also verified statistical results from articles using statcheck (Nuijten & Polanin, 2020) to confirm internal consistency. If the original article did not directly report mean and standard deviation but simply provided graphs, we used WebPlotDigitizer (Rohatgi, 2020) or metaDigitise (Pick et al., 2018). 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 comparison strategies in stressed and non-stressed conditions, as well as of control strategies in stressed versus non-stressed conditions, we conducted the random-effects model for all primary hypotheses (H1, H2, H3, H4). For H1, we compared the effects of learning with control strategies in stress versus control, or non-stress conditions. For H2, we compared the effects of learning with retrieval practice versus a control strategy in a control or non-stress condition. For H3, we compared the effects of learning with retrieval practice versus other strategies in stress conditions. Lastly, for H4, we compared the effects of learning with retrieval practice in stress versus non-stress conditions. This breakdown allowed us to isolate and compare the effects of retrieval practice versus other strategies on memory performance. Because we are running models on all four scenarios, we decided not to conduct multivariate models, which are typically used to assess multiple correlated outcomes within the same study.

To determine the impact of the stressor on memory performance, we conducted additional analysis by applying meta-regression using participant's scores on the stress manipulation checks as moderators weighed on memory performance scores in the stress condition. This analysis was conducted as a sanity check to verify that the stress procedure was successful in the included studies. The stress scores were extracted from each study and reflected participant's self-reported stress score on a stress or anxiety measure taken before and after the stressor in both stress and control groups. This analysis will only be conducted if the stress measurements extracted from studies are sufficiently comparable at Stage 2.

We plotted forest plots presenting the effect size of each study. We presented the effect size with confidence intervals and sample size of each study. Statistical heterogeneity between studies was determined using the *Q* statistic and quantified with $I^2$ (Higgins & Thompson, 2002; Huedo-Medina et al., 2006). This global meta-analysis yielded a point estimate, confidence interval, and *p*-value, along with statistics for heterogeneity. We determined a threshold of $I^2$ of over 50% and a significant *Q* statistic as an indicator to perform subsequent moderator analysis (Harrer et al., 2021). If we obtain such results, we can assume that there are sources of variation other than sampling error in our sample, thus warranting further investigation. If there was indeed meaningful heterogeneity, we investigated and explored potential moderators.

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

Publication bias was assessed by first evaluating “small study effects” (Harrer et al., 2021). Small study effects refer to the phenomenon where studies with smaller sample sizes tend to show larger and more extreme effects compared to studies with larger sample sizes. Thus, small studies are more likely to get published while studies with non-significant results are more likely to be unpublished, creating skewed evidence. To assess small study effects, we first plotted the effect sizes and standard errors of each study, visually depicted in a funnel plot. Egger’s Test of the Intercept (Egger et al., 1997; Sterne & Egger, 2005) was then used to calculate whether asymmetry exists in the funnel plot. If Egger's Test is significant, this may be due to missing studies. To check this, we then applied the trim-and-fill procedure, which corrects for this asymmetry by filling in missing studies (Duval and Tweedy, 2000). We also conducted the Rank correlation test (Begg 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 strong correlations suggest publication bias.

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

The above publication bias methods are our preferred methods based on simulations of false positives, statistical power, and recommendations from the field (Carter, 2019; Harrer et al., 2021). We acknowledge that there are many different approaches to publication bias correction, on which we have limited information at Stage 1 prior to data extraction. We also acknowledge that heterogeneity and publication bias are closely intertwined, and that some measures of publication bias can be sensitive to underlying study heterogeneity (Harrer et al., 2021), which could affect the reliability and interpretation of our findings. One way in which we will disentangle the two in the current work involves conducting Egger's test to assess the presence of publication bias, while also evaluating heterogeneity using methods such as Cochran's $Q$ or $I^2$ statistic, as outlined above. Significant heterogeneity may indicate that studies are estimating different underlying effects, whereas significant results from Egger's test could suggest publication bias. Moreover, we will perform sensitivity analysis using the leave-one-out method (Harrer et al., 2021) where effect sizes are recalculated with one study removed each time to assess the robustness of findings and identify potential outliers. We will also consider the sample size and quality of included studies when interpreting results, recognizing that small sample sizes and low-quality studies are more vulnerable to biases and spurious results (Brysbaert, 2019), which may influence our understanding of potential publication bias. Additionally, we will also assess study level power to check whether publication bias is likely (Quintana, 2023).

### 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 current field is still emerging, we expected to include 10 studies. We expected the average sample size per study and condition to be 25 and we expected moderate-to-high heterogeneity. We conducted a priori-power calculation with dmetar 0.0.9000 package (Harrer et al., 2019, 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 will be performed once the meta-analysis has been conducted at Stage 2 by re-running our initial power analysis script above with the actual values obtained from our meta-analysis. As a complementary approach, we will also apply the metameta package (Quintana, 2023). The metameta package serves as a versatile tool for conducting post-hoc power analysis in meta-analysis, enabling researchers to determine the range of effect sizes reliably detectable within a body of studies. By utilizing data extracted from meta-analysis forest plots and tables, metameta calculates study-level statistical power and median statistical power based on published effect-size and variance data.

### Risk of Bias

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

## Results

We summarized our findings in Table 2, publication bias in Table 3, and moderator analysis in Table 4. We provided the list of studies/articles included in the meta-analysis in Table 5. We presented forest and funnel plots of the included studies in Figure 2 and Figure 3. In the following, we first present the main effect findings, followed by publication bias findings, and moderator analysis at the end. The results below and those in Supplementary are simulated with fake data for Stage 1. These will be replaced with the real results for Stage 2 following data extraction and analysis.

**Table 2**

*Summarized Results of the Meta-Analysis – will be updated for Stage 2*

| Hypotheses | Key findings / theories in the literature | Findings in the meta-analysis (Supported / |
|---|---|---|

| | |
|---|---|
| | Not Supported / Partially Supported) |
| *Main hypothesis* | |
| *Theoretical Moderator Hypotheses* | |

## Overall retrieval practice effects (simulated data)

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

We first examined the overall effect of having learned with a control learning strategy in a stress compared to non-stress condition. The mean effect was negative. We did not find support for the hypothesis, $k$ = 11, $g$ = -0.52, CI [-1.55, 0.52]. This suggests that across the selected studies, there seems to be a negative effect on memory performance when learning without specific learning strategies in a stressful situation compared to a non-stressful setting.

### *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 negative. We did not find support for the hypothesis, *k* = 11, *g* = -0.37, CI [-2.41, 1.65]. This suggests that across the selected studies, there seems to be a negative effect of learning with retrieval practice versus a control strategy on memory performance in participants who did not undergo stress.

### *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 performance in a stress condition. The mean effect was positive. We found support for the hypothesis, *k* = 11, *g* = 1.97, CI [1.03, 2.91]. This suggests that across the selected studies, there seems to be a strong benefit of learning with retrieval practice versus a control strategy in a stressful setting.

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

We then examined the overall effect having learned with retrieval practice in a stress versus non-stress condition. The mean effect was positive. We found support for the hypothesis, *k* = 11, *g* = 2.27 CI [0.66, 3.88]. This suggests that across the selected studies, there seems to be a strong benefit of learning with retrieval practice in a stress versus non-stress setting.

## Effect of Stressors (simulated data)

To examine the effect of the stress manipulation on memory performance, we submitted participant's stress scores to a meta-regression. The test of moderators produced a *QM* statistic of 3.2713 ($p$ = 0.0705), hinting at a significant influence of stress scores overall.

### Statistical Power

The obtained statistical power, based on effect size, average sample size, number of effect size, and heterogeneity is [TO BE ADDED FOR STAGE 2].

### Publication Bias (simulated data)

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 employed 6 different statistical approaches to examine a potential publication bias according to recommendations from the field (Harrer et al., 2021). These measures included including Egger's Test of the Intercept (Egger et al., 1997), the trim-fill procedure (Duval & Tweedy, 2000), the Rank correlation test (Begg & Mazumdar, 1994), PET-PEESE (Stanley & Doucouliagos, 2014), and a three-step parameter model (Iyengar & Greenhouse, 1988).

A summary of publication bias analyses is provided in Table 3 across all studies collapsed together. The bias findings were not conclusive, but they seem to be suggestive of a possible publication bias in favor of the effect, possibly leading to an overestimation of the effect. There were some discrepancies using the different methods of publication bias. The insignificant Egger's test suggests no funnel plot asymmetry, and an insignificant Kendall's tau suggests a low level of correlation between study ranks and their effect sizes. However, the trim-fill correction method identified 9 missing studies. Both the PET and PEESE methods suggest no significant hint of publication bias. The three-step selection model suggests the true effect size is still positive, indicating the pooled effect was not distorted by selective reporting.

**Table 3**

*Publication bias analyses (simulated data)*

| Publication bias analysis method | Results and adjusted models |
|---|---|
| *Small Study Effects* | |
| Trim and fill funnel plot asymmetry | 9 missing on the left side.<br>Adjusted model: *g* = 1.77, 95% CI [0.86, 2.67] |
| Egger's regression test | *z* = -0.60, *p* = 0.5472 |
| Rank correlation test<br>(Begg & Mazumdar, 1994) | Kendall's tau = 0.0994, *p* = 0.3490 |
| *Publication Bias Tests* | |
| Three-parameter selection model | Likelihood Ratio Test: 1.02, p = 0.3<br>Adjusted Model: *g* = 0.57, 95% CI [-0.48, 1.62] |

| | |
|---|---|
| PET | $b$ = 0.61 [-1.36, 2.59], $p$ = 0.991 |
| PEESE | $b$ = 0.70 [-0.35, 1.75], $p$ = 0.83 |

*Note*. 1) Values in parentheses indicate 95% confidence intervals [lower bound, upper bound]

### MetaForest moderator analyses (simulated data)

To address the problem of limited studies and lack of statistical power while without risk overfitting, we adopted MetaForest (van Lissa, 2017, v0.1.3). MetaForest uses "random forests," a machine learning technique, and bootstrapping to examine several possible moderators. We provide the detailed results in the Supplementary. Results suggested significant heterogeneity among the effect sizes and the main model indicator, R-squared (R-OOB) was 0.2006, meaning that the helps explain some of the variance observed. Stressor type and task type were the most important moderators. Delay and other strategies had a negligible impact.

### Moderator analyses (simulated data)

Statistical heterogeneity was determined using Cochran's $Q$ statistic and quantified with $I^2$ (Higgins & Thompson, 2002). The $Q$ statistic is significant ($Q$ = 262.3429, $p$ < 0.0001, $df$ = 10), suggesting the overall sample is highly heterogenous. This is further supported by an $I^2$ value of 97.46%, suggesting a 97% 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 two possible theoretical and methodological moderators according to a pre-registered criteria in the Full Coding Sheet: stressor type (TSST vs. TA stress), other learning strategies used (restudy vs. highlighting strategies), delay (i.e., 1 day, etc.), and type of learning task (i.e., reading materials, etc.). Results of moderator analysis are summarized in Table 4. Moderator analysis is shown for H3 only, as we are primarily interested in the effects of the moderators on learning with retrieval practice versus another strategy in the context of stress. Full moderator analyses (with simulated data) for H1, H2 and H4 are presented in Supplementary.

**Stressor Type**. Five studies for TSST stress had an effect size of *g* = 2.71, CI [1.02, 4.40], *p* = 0.0017. Six studies for TA stress had an effect size of *g* = 1.39, CI [0.47, 2.30], *p* = 0.0029. We used a fixed-effects contrast and MetaForest to test if there is a meaningful moderating effect. We found no support for a moderation effect of Moderator 1.

**Other Learning Strategies**. Four studies for restudying had an effect size of *g* = 2.11, CI [0.56, 3.66], *p* = 0.0077. Six studies for highlighting had an effect size of *g* = 1.48, CI [0.33, 2.63], *p* = 0.0118. One study for diagrams had an effect size of *g* = 4.41, CI [3.59, 5.22], *p* = <0.001. We used a fixed-effects two-level model and MetaForest to test if there is a meaningful moderating effect. We found support for moderation effect between highlighting and diagrams (*p* = 0.010).

**Delay.** Three studies for a delay of 1 day had an effect size of *g* = 2.79, CI [0.73, 4.84], *p* = 0.0079. Three studies for a delay of 2 days had an effect size of *g* = 2.43, CI [0.34, 4.51], *p* = 0.0223. Two studies for a delay of 1 week had an effect size of *g* = 0.57, CI [0.14, 1.01], *p* = 0.0097. Three studies for a delay of over 1 week had an effect size *g* = 1.71, CI [-0.16, 3.57], *p* =

0.0729. We used a fixed-effects two-level model and MetaForest to test if there is a meaningful moderating effect. We found no support for moderation effect of Moderator 3.

**Task Type.** Four studies for educational texts tasks had an effect size of *g* = 2.11, CI [0.56, 3.66], *p* = 0.0077. Three studies for word lists tasks had an effect size of *g* = 3.60, CI [2.15, 5.04], *p* = <0.001. Three studies for vocabulary tasks had an effect size of *g* = 0.73, CI [0.49, 0.98], *p* = <0.001. One study for math tasks had an effect size of *g* = 0.32, CI [-0.11, 0.75], *p* = 0.1451. We used a fixed-effects two-level model and MetaForest to test if there is a meaningful moderating effect. We found significant support for the moderation effect of between word lists and vocabulary learning material (*p* <0.001).

**Table 4**

*Summarized Results of Moderator Analysis (simulated data)*

Note. *k* = number of samples; *N* = total number of individuals in *k*; *g* = Hedge's g effect size, CI = lower and upper limits of 95% confidence interval, *tau2* = tau squared value, *I2* = I-squared value, * *p* < .05, ** *p* < .01, *** *p* <.001, (all two-tailed).

| Moderator | *k* | *Q* | *df* | *g* | 95% CI | *Tau2* | *I2* | Diff | *p* | *Categories* |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Stressor Type*** | | | | | | | | | | |
| TSST | 5 | 134 | 4 | 2.7130 | 1.0216, 4.4045 | 1.8918 | 96.95% | 1.829 | 0.176 | TSST vs. TA |
| TA | 6 | 118 | 5 | 1.3869 | 0.4748, 2.2990 | 1.1144 | 96.23% | | | |
| ***Other Learning Strategies*** | | | | | | | | | | |
| Restudying | 4 | 83 | 3 | 2.1102 | 0.5594, 3.6610 | 1.5513 | 96.43% | 0.406 | 0.524 | Restudy vs. highlighting |
| Highlighting | 6 | 84 | 5 | 1.4819 | 0.3288, 2.6351 | 1.4124 | 97.53% | 6.588 | 0.010* | Highlighting vs. diagrams |
| Diagrams | 1 | *NA* | *NA* | 4.4056 | 3.5888, 5.2224 | *NA* | *NA* | | | |
| ***Delay*** | | | | | | | | | | |
| 1 day | 3 | 63 | 2 | 2.7859 | 0.7290, 4.8429 | 1.7810 | 96.92% | 0.057 | 0.811 | 1 day vs. 2 days |
| 2 days | 3 | 50 | 2 | 2.4294 | 0.3454, 4.513 | 1.8118 | 97.06% | 2.994 | *0.084* | 2 days vs. 1 week |
| 1 week | 2 | 2 | 1 | 0.5737 | 0.1389, 1.0085 | 0.2556 | 65.21% | 1.344 | 0.246 | 1 week vs. <1week |
| 1 + week | 3 | 49 | 2 | 1.7069 | -0.1587, 3.572 | 1.6209 | 97.17% | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Type of Task*** | | | | | | | | | | |
| Passages | 4 | 83 | 3 | 2.1102 | 0.5594, 3.6610 | 1.5513 | 96.43% | 1.888 | 0.169 | Passages vs. Word lists |
| Word lists | 3 | 28 | 2 | 3.5967 | 2.1506, 5.0428 | 1.2123 | 90.98 | 14.641 | < .001*** | Word lists vs. vocabulary |
| Vocabulary | 3 | 2 | 2 | 0.7332 | 0.4877, 0.9786 | 0.1039 | 21.83% | 2.671 | 0.102 | Vocabulary vs. Math |
| Math | 1 | *NA* | *NA* | 0.3200 | -0.1104, 0.750 | *NA* | *NA* | | | |

**Table 5**

*Studies included in the meta-analysis (simulated data)*

| Number | Study | N | Country | Sample Population | Design | Publication status | DV type |
|---|---|---|---|---|---|---|---|
| 1 | Authors ABC (2013) | 120 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 2 | Authors ABC (2013) | 120 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 3 | Author ABC (2013) | 120 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 4 | Author ABC (2013) | 120 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 5 | Author B (2015) | 100 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 6 | Author B (2015) | 100 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 7 | Author B (2015) | 100 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 8 | Author B (2015) | 100 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 9 | Authors GHI (2012) | 160 | [To be inserted by Stage 2] | Student | Mixed-subject | Yes | DV1 |
| 10 | Authors GHI (2012) | 160 | [To be inserted by Stage 2] | Student | Mixed-subject | Yes | DV1 |
| 11 | Authors GHI (2012) | 160 | [To be inserted by Stage 2] | Student | Mixed-subject | Yes | DV1 |
| 12 | Authors GHI (2012) | 160 | [To be inserted by Stage 2] | Student | Mixed-subject | Yes | DV1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 13 | Author JK (2019) | 220 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 14 | Author JK (2019) | 220 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 15 | Author JK (2019) | 220 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 16 | Author JK (2019) | 220 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 17 | Authors LMN (2017) | 400 | [To be inserted by Stage 2] | Student | Mixed-subject | Yes | DV1 |
| 18 | Authors LMN (2017) | 400 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 19 | Authors LMN (2017) | 400 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 20 | Authors LMN (2017) | 400 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 21 | Authors OP (2020) | 228 | [To be inserted by Stage 2] | Student | Between-subject | No | DV1 |
| 22 | Authors OP (2020) | 228 | [To be inserted by Stage 2] | Student | Between-subject | No | DV1 |
| 23 | Authors OP (2020) | 228 | [To be inserted by Stage 2] | Student | Between-subject | No | DV1 |
| 24 | Authors OP (2020) | 228 | [To be inserted by Stage 2] | Student | Between-subject | No | DV1 |
| 25 | Author Q (2021) | 132 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 26 | Author Q (2021) | 132 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |

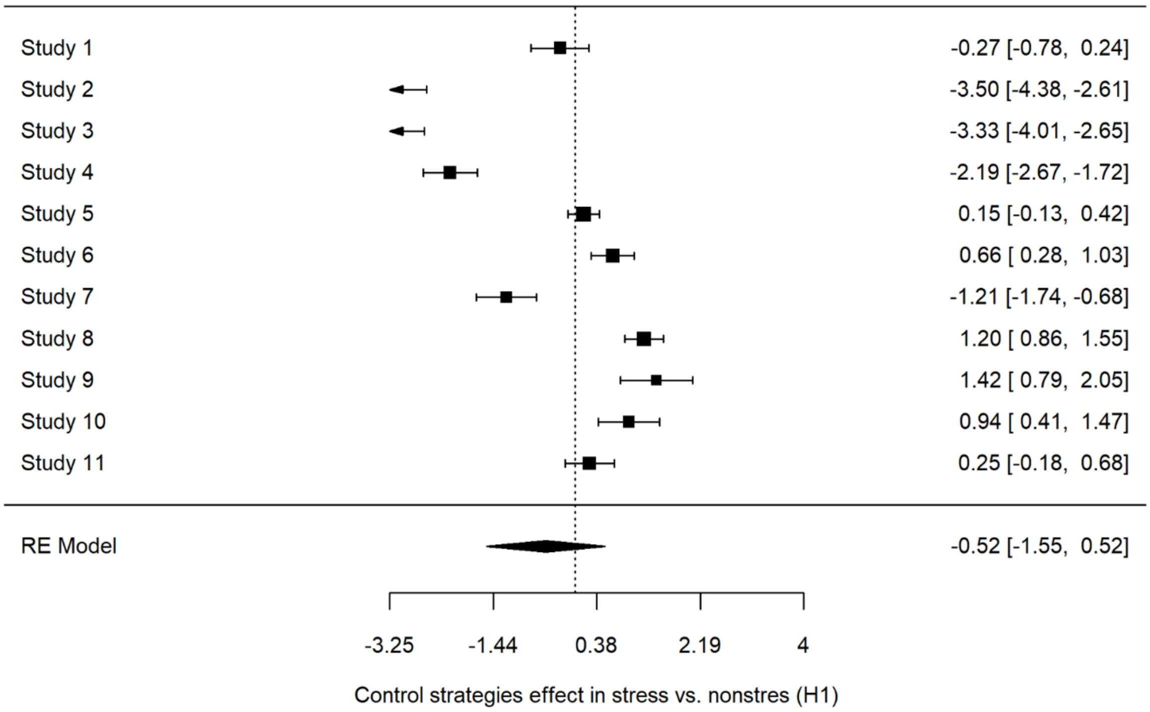
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 27 | Author Q (2021) | 132 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 28 | Author Q (2021) | 132 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 29 | Authors RST (2018) | 312 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 30 | Authors RST (2018) | 312 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 31 | Authors RST (2018) | 312 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 32 | Authors RST (2018) | 312 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 33 | Authors UV (2013) | 44 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 34 | Authors UV (2013) | 40 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 35 | Authors UV (2013) | 48 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 36 | Authors UV (2013) | 48 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 37 | Authors WXY (2017) | 60 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 38 | Authors WXY (2017) | 60 | [To be inserted by Stage 2] | Student | Between-subject | Yes | Dv1 |
| 39 | Authors WXY (2017) | 60 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |
| 40 | Authors WXY (2017) | 60 | [To be inserted by Stage 2] | Student | Between-subject | Yes | DV1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 41 | Author Z (2019) | 84 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 42 | Author Z (2019) | 84 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 43 | Author Z (2019) | 84 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |
| 44 | Author Z (2019) | 84 | [To be inserted by Stage 2] | Student | Mixed-subject | No | DV1 |

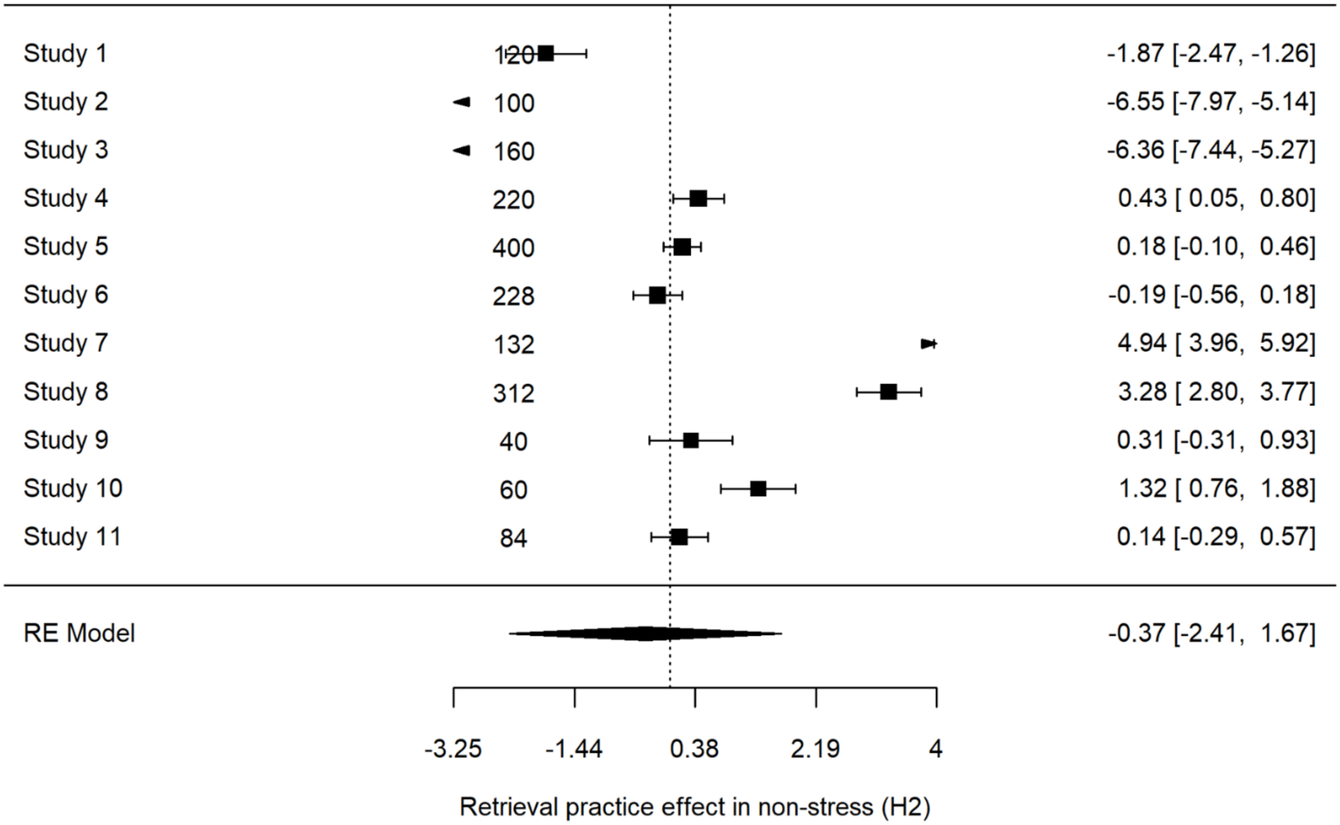
**Figure 2**

*Forest Plots of H1, H2, H3, H4 (simulated data)*

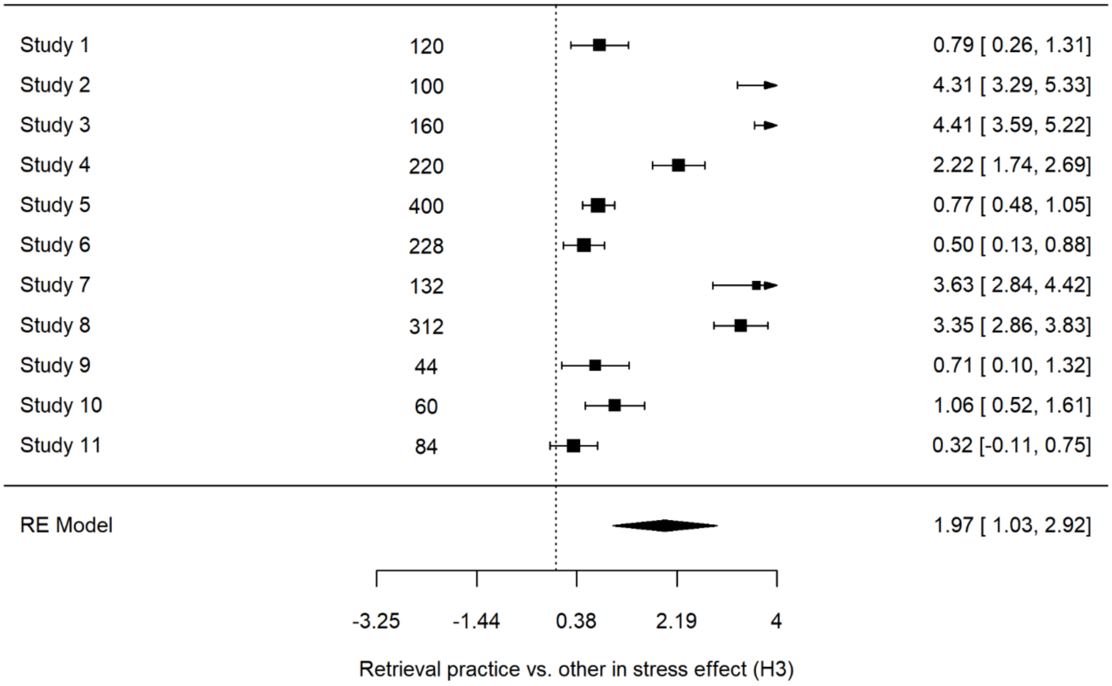
*2a. Forest plot of H1*



*2b. Forest plot of H2*



*2c. Forest plot of H3*



*2d. Forest plot of H4*

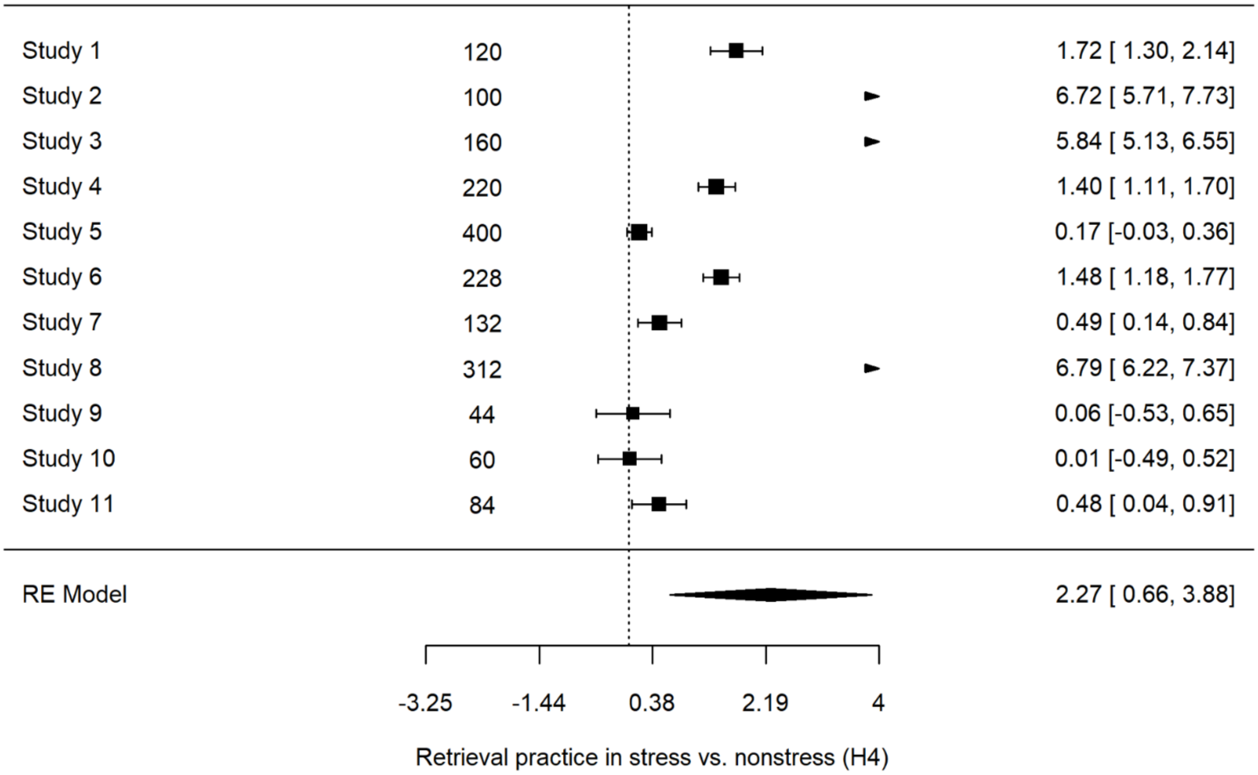
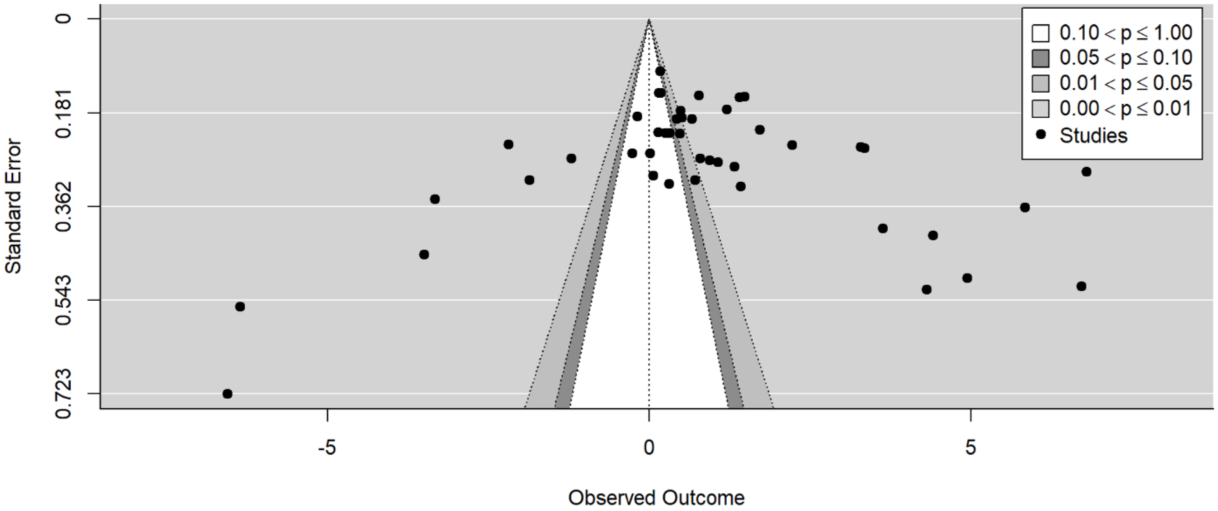

Study 1 120 1.72 [ 1.30, 2.14]
Study 2 100 6.72 [ 5.71, 7.73]
Study 3 160 5.84 [ 5.13, 6.55]
Study 4 220 1.40 [ 1.11, 1.70]
Study 5 400 0.17 [-0.03, 0.36]
Study 6 228 1.48 [ 1.18, 1.77]
Study 7 132 0.49 [ 0.14, 0.84]
Study 8 312 6.79 [ 6.22, 7.37]
Study 9 44 0.06 [-0.53, 0.65]
Study 10 60 0.01 [-0.49, 0.52]
Study 11 84 0.48 [ 0.04, 0.91]
RE Model 2.27 [ 0.66, 3.88]
-3.25
-1.44
0.38
2.19
4
Retrieval practice in stress vs. nonstress (H4)

**Figure 3**

*Funnel plot of all studies in meta-analysis (simulated data)*



## Discussion

[To be completed after data analysis, for Stage 2]

## References

Agarwal, Pooja K., et al. "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 (2014): 131-139. https://doi.org/10.1016/j.jarmac.2014.07.002

Allen, L., & O'Connell, A. (2014). CRediT - Contributor Roles Taxonomy. Retrieved May 13, 2020, from https://casrai.org/credit/

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

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.).

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

Beller, E. M., Glasziou, P. P., Altman, D. G., Hopewell, S., Bastian, H., Chalmers, I., ... & PRISMA for Abstracts Group. (2013). PRISMA for abstracts: reporting systematic reviews in journal and conference abstracts. *PLoS Med*, *10*(4), e1001419. https://doi.org/10.1371/journal.pmed.1001419

BMJ Open Science (2020). Registered Reports Guidelines. Retrieved July 21, 2020, from https://openscience.bmj.com/pages/registered-reports-guidelines/

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.

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).

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

Center for Open Science (n.d.). Registered Reports: Peer review before results are known to align scientific values and practices. Retrieved June 15, 2020, from https://www.cos.io/our-services/registered-reports

Cheung, M. W. L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review,* 1-10. https://doi.org/10.1007/s11065-019-09415-6

Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P., & Scott, A. M. (2020). A full systematic review was completed in 2 weeks using automation tools: a case study. *Journal of Clinical Epidemiology, 121*, 81-90. https://doi.org/10.1016/j.jclinepi.2020.01.008

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.*

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, EM, AN Stillman, MW Tingley, and CS Elphick (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*

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

Henmi, M., & Copas, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in medicine, 29*(29), 2969-2983. https://doi.org/10.1002/sim.4029

Higgins, J. P. (2011). Cochrane handbook for systematic reviews of interventions. Version 5.1. 0 [updated March 2011]. The Cochrane Collaboration. *www.cochrane-handbook.org.*

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557-560.

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., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons. https://doi.org/10.1002/9781119536604

Hohn, R. E., Slaney, K. L., & Tafreshi, D. (2019). Primary study quality in psychological meta-analyses: An empirical assessment of recent practice. *Frontiers in Psychology, 9*, 2667. https://doi.org/10.3389/fpsyg.2018.02667

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 I² index? *Psychological Methods*, *11*(2), 193. https://doi.org/10.1037/1082-989x.11.2.193

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*.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772-775. DOI: 10.1126/science.1199327

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

Kennedy, P. J., & Shapiro, M. L. (2004). Retrieving memories via internal context requires the hippocampus. *Journal of Neuroscience*, *24*(31), 6979-6985.

https://doi.org/10.1523/JNEUROSCI.1388-04.2004

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

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.

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

McKenzie, J. E., Brennan, S. E., Ryan, R. E., Thomson, H. J., Johnston, R. V., & Thomas, J. (2019). Defining the criteria for including studies and how they will be grouped for the synthesis. *Cochrane Handbook for Systematic Reviews of Interventions*, 33-65. https://doi.org/10.1002/9781119536604

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. J*ournal of Memory and Language, 112*, 104092.

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

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T., Mulrow, C. D., ... & Moher, D. (2020). Mapping of reporting guidance for systematic reviews and meta-analyses generated a comprehensive item bank for future reporting guidelines. *Journal of Clinical Epidemiology, 118*, 60-68. https://doi.org/10.1016/j.jclinepi.2019.11.010

Penkin, C., Haddaway, N., Kwong, J., Newman, P., & Ngwenya, M. (2019). Grey Literature Reporter [Chrome Plugin]. Retrieved from https://www.eshackathon.org/software/grey-lit-reporter.html

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

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 DS. A Guide for Calculating Study-Level Statistical Power for Meta-Analyses. Advances in Methods and Practices in Psychological Science. 2023;6(1). doi:10.1177/25152459221147260

Re, A. C. D. (2012). Package “MAc”. *R Package Version 1.1, 2012*

Re, A. C. D. (2020). Package “compute.es”. *R Package Version 0.2-5, 2020*

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

RStudio Team (2015). Integrated development for R. *RStudio, IncBoston, MA.*

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

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. DOI: 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., 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., & 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

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review, 8,* 203-220.

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

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p-values: reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science. 11*, 713–729. https://doi.org/ 10.1177/1745691616650874

Van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*(3), 293. https://doi.org/10.1037/met0000025

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

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., Sun, B., Potts, R., Yu, R., Luo, L., & Shanks, D. R. (2020). Do working memory capacity and TA 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