# A Meta-analytic Review of the Effectiveness of Spacing and Retrieval Practice for Mathematics Learning.

Ewan Murray1, Aidan J. Horner1,2, and Silke M. Göbel1,3

1Department of Psychology, University of York, UK

2York Biomedical Research Institute, University of York, UK

3CREATE & Department of Special Needs Education, University of Oslo

# Author Note

Correspondence concerning this article should be addressed to Ewan Murray, Department of Psychology, University of York, UK, Email: ewan.murray@york.ac.uk

# Abstract

Spaced retrieval practice harnesses two well-studied phenomena: the spacing effect, where spacing out practice over several sessions leads to a gain in retention compared to massed practice in one session; and the testing effect, where material that is tested is better retained than material that is restudied. This meta-analysis investigates if, and under what circumstances, spaced and retrieval practice can benefit mathematics learning. We found a robust small to medium effect of spaced versus massed practice overall (*g* = 0.28, 27 studies, 53 effect sizes). Those studies can be split into two subsets based on their experimental design, where material was either taught in isolation (10 studies, 27 effect sizes) or as part of a course (17 studies, 26 effect sizes). We found a larger, yet less robust, effect for the isolated learning (*g* = 0.43) than for course-embedded (*g* = 0.24). Our search also revealed 7 studies, 32 effect sizes, which manipulated testing versus restudy. The weighted mean effect of testing versus restudy was *g* = 0.18. However, the 95% confidence interval crossed zero, suggesting the testing effect is not robust. Overall, our results suggest that spaced practice can improve mathematics learning for material in isolation and within a course. However, the effect may be smaller than in other domains. Additionally, the current literature does not provide conclusive evidence for a consistent effect of retrieval practice for mathematics learning, possibly due to the smaller number of studies available.

*Keywords*: Spacing effect, Distributed practice, Testing effect, Retrieval practice, Mathematics

# A Meta-analytic Review of the Effectiveness of Spacing and Retrieval Practice for Mathematics Learning.

Many decisions go into the design of a program of learning. What to learn and how to learn it are thought about in painstaking detail to maximise the chances of success for the learners. A critical question to ask when designing a learning program is when should revision of prior material take place? Precisely when a session takes place during a day is often out of the hands of instructional designers and up to timetablers, but once time has been allotted there is often freedom to choose what specific questions to ask and when. When an instructor decides which questions to ask during the start of a lesson, during the main tasks or as homework, it is here that there is the opportunity to improve learning through the use of retrieval practice and spaced repetition. There has been increased interest in harnessing retrieval practice, actively retrieving information rather than restudying it, and spaced practice, spreading out practice over multiple sessions rather than a single session, to improve mathematics learning, both in research and in applied settings. Given this recent spike in interest, it is valuable to synthesize the current research and see if these techniques are as effective for mathematics learning as they are in other domains. The goal of this meta-analysis is to review and synthesise the current literature surrounding the spacing and testing effect in mathematics learning.

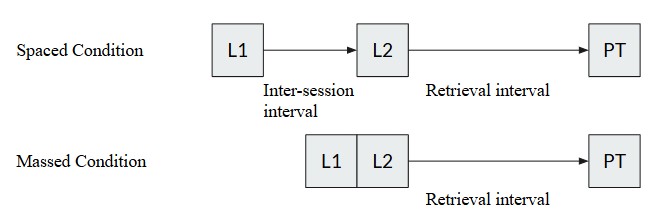
### Overview of Research on Spaced Practice

The spacing effect is a special case of the distributed practice effect, where the temporal spacing of practice or presentations of stimuli are varied and a change in retention is observed (Delaney et al., 2010). The spacing effect is found within session, by varying the time between presentations of a stimuli and across session, by varying the times between practice sessions (Küpper-Tetzel, 2014). All the spacing effect studies included in this meta-analysis are across-session studies, so we will define the spacing effect as the change in retention when an equivalent amount of practice is spread over multiple sessions versus a single massed session (see [Figure 1](#fig-spacing)). Broad ranging meta-analyses across all domains found medium to large effects of spacing versus massed practice (*g* = 0.46, Donovan & Radosevich, 1999), with some more specific ones such as the domain of L2 Language learning finding larger effects (*g* = 0.80, with a delayed post-test, Kim & Webb, 2022). The effect has been observed across many different domains, age groups and even species (Delaney et al., 2010).

Many theories have been proposed to describe the mechanisms underlying the spacing effect (Delaney et al., 2010; Küpper-Tetzel, 2014; Maddox, 2016; for reviews, see Toppino & Gerbier, 2014). The three most discussed theories are *study-phase retrieval*, *deficient processing*, *encoding variability,* or combinations of those three (Toppino and Gerbier, 2014). Study-phase retrieval suggests that successful recall of the initial presentation of the to be recalled item results in a boost to future recall (Hintzman, 2004; Thios & D’Agostino, 1976). Deficient processing suggests that repeated presentations of an item in the massed condition leads to lower overall processing, when compared to items spaced out over time, which in turn leads to a lower quality of learning overall (Hintzman, 1974). Encoding variability theories suggest that over time the context around an item drifts and is encoded when presented to the participant, when presentations of the item are spaced out then this increases the probability of the retrieval context being more similar to one of the spaced encoding contexts relative to two the massed encoding presentations (Glenberg, 1979). These theories and reviews informed our choice of moderators (see supplementary material), along with discussions of relevant moderators in recent studies and reviews (Emeny et al., 2021; Latimier et al., 2021). However, as our moderation analyses are unable to help distinguish between these theories we do not go into further detail.

Figure 1

Simple Spacing Effect Paradigm



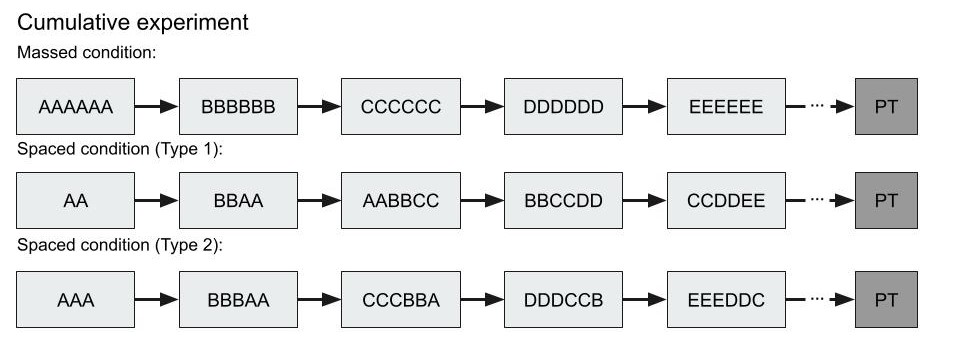
*Note*. A diagram outlining the overall procedure of a hypothetical spaced vs massed learning experiment. L = Learning session, PT = Posttest.

Within mathematics learning, spacing effect studies follow one of two experimental designs based on how the material is integrated into the larger educational environment: isolated learning versus material incorporated into a course. Some studies that use mathematics material require learners to practice isolated learning outside of the confines of a course. In these cases, the participants are taught a set number of mathematical procedures or are given practice solving problems in such a way that there is a clear inter-session interval and retrieval interval (Rohrer & Pashler, 2007; Rohrer & Taylor, 2006) and follow a similar experimental design to previous verbal learning studies (see [Figure 1](#fig-spacing)).

Another common structure for these studies is to embed the spacing schedule into a course already taking place. Suppose there are five learning objectives and six questions for each objective (see [Figure 2](#fig-cumulative)). There are various ways in which these practice questions can be presented to learners. One way is to present them in a massed manner where after learning about the topic the learners are presented with all available questions on that topic. This is how most mathematics textbooks are laid out (Rohrer et al., 2020), with few notable exceptions (Hake, 2012). On the other hand, they could be presented uniformly spaced across multiple sessions (see [Figure 2](#fig-cumulative), spaced condition type 1). Examples of this type include Lyle et al. (2020) and Hopkins et al. (2016). Alternatively, it is possible to start with more practice examples in the first instance and reduce the number of items in subsequent practice sessions; this type of spacing was used by (Holdan, 1986) (see [Figure 2](#fig-cumulative) spacing condition type 2). Others allow for variations in the exact spacing and amount of practice by implementing a formula for teachers to use (Hirsch et al., 1982). However, these course-embedded studies provide a less pure measure of spacing as interleaving, mixing practice of different topics and skills, is also involved.

Figure 2

Experimental Design for course-embedded spacing study



*Note*. A diagram outlining potential procedures for course-embedded spacing experiments.

It has been noted that increased *contextual interference* or *interleaving*, that is, shuffling the order of practice problems from tasks to increase the variability of practice, can lead to greater retention and transfer of skills versus blocked practice (Shea & Morgan, 1979). We do not include studies that purely investigated blocked versus interleaved practice. However, we did not exclude distributed practice studies which, due to overlapping practice schedules, interleaved practice material. In the included studies, the manipulation focused on the practice schedule of each item not carefully manipulating interleaving versus blocked practice throughout. We believed it was important to include these studies as this is likely how spacing would be implemented in a classroom. Previous mathematics learning interleaving experiments comparing interleaved with blocked practice have shown that alongside the benefit of spacing there is an additional benefit of discriminative contrast (Foster et al., 2019). A meta-analysis found a small to medium effect (*g* = 0.34) of interleaved versus blocked practice for mathematics learning materials (Brunmair & Richter, 2019). In an educational setting where the aim is to maximise learning with considerable time constraints, this is the way spacing would be implemented with old material intermixed with the new. Therefore, even though there is an effect of interleaving, experiments that follow this kind of structure were included, as all the studies claimed to be harnessing the spacing effect. As this is likely to be how spacing would be implemented in an actual educational setting, these studies bring increased ecological validity.

### Overview of Research on Retrieval Practice

The testing effect, or retrieval practice, is a boost in retention to material that has to be retrieved from memory compared to material that is restudied. Retrieval practice has been shown to be a powerful way to improve retention (Roediger & Butler, 2011), but there has been little research into its efficacy with mathematics. Multiple previous meta-analyses have investigated the testing effect in other domains. Rowland (2014) (*g* = 0.50, 95% CI [0.42, 0.58]), Adesope et al. (2017) (*g* = 0.61) and Yang et al. (2021) (*g* = 0.50, 95% CI [0.442, 0.557]) all found a medium to large boost to retention when retrieval practice was used. Yang et al. (2021) included 27 effect sizes from 14 studies that pertained to mathematics or statistics material, however, of those, only two studies compared testing to restudying and were included in our sample (Betsch et al., 2015; Dirkx et al., 2014). The other control conditions included in Yang et al. (2021) included more versus fewer questions, elaborative strategies, or no activity (just a distractor task after the initial study session), which were outside the scope of our pre-registration. There is also meta-analytic evidence that retrieval practice can enable transfer in a variety of contexts such as different test formats and problem types (Pan & Rickard, 2018), but perhaps not to untested information studied during the initial practice. However, there has been no meta-analysis that focused specifically on retrieval practice in mathematics versus restudy.

It is not immediately apparent what restudy would look like in mathematics, in comparison to verbal learning studies where a word pair could be presented again. Indeed, later in this meta-analysis we discuss the difficulty in understanding precisely how much students had to retrieve information, and when they were able to fall back on their notes. One way to manipulate retrieval versus restudy in mathematics is to compare completing practice where retrieval is necessary to worked example problem pairs (Fazio, 2018). Worked examples can be effective learning procedures as they allow learners to focus on the task at hand without the need to hold all parts of the problem in working memory at once (Sweller, 2006). For example, Fazio (2018) compared retrieval practice with worked example problem pairs. In the retrieval condition they were tested on the procedure, but in the worked example condition they were presented with a worked example they could follow along with, so they may not have had to retrieve the procedure. They found that worked examples increased performance on an immediate test, but the retrieval condition produced greater retention on a delayed test. Alternatively, Dirkx et al. (2014) compared reading a text on how to calculate a probability, versus alternating between reading and testing. They found that reading and testing led to both better fact recall and a better ability to apply the procedure.

Furthermore, spacing and testing are often combined into *spaced retrieval practice*. Latimier et al. (2021)’s previous meta-analysis suggested that the effect of spacing and testing is additive and is typically a large effect (*g* =1.01, 95% CI [0.68, 1.34]).

### The Present Meta-analysis

This meta-analysis focuses on studies that involve the learning of mathematics. We attempted to answer three key questions. First, does the spacing of mathematics practice lead to higher retention (versus massed)? Second, does retrieval practice of mathematics lead to higher retention (versus restudy)? Third, we planned to investigate whether the combination of spacing and retrieval lead to higher retention of mathematics knowledge (versus massed retrieval), however, this was not possible with the available studies. Additionally, we investigated whether any of the potential moderating variables (such as the length of the retrieval interval or type of material being studied) have any effect on the heterogeneity or mean effect sizes, however, as these exploratory analyses provided little informative information they are included only in the supplementary material.

Based on previous meta-analyses of spacing and retrieval practice with non-mathematical material we hypothesized that we would find a robust effect of testing and spacing for mathematics material and that the combination of spacing and retrieval would be additive.

## Method

### Selection Criteria

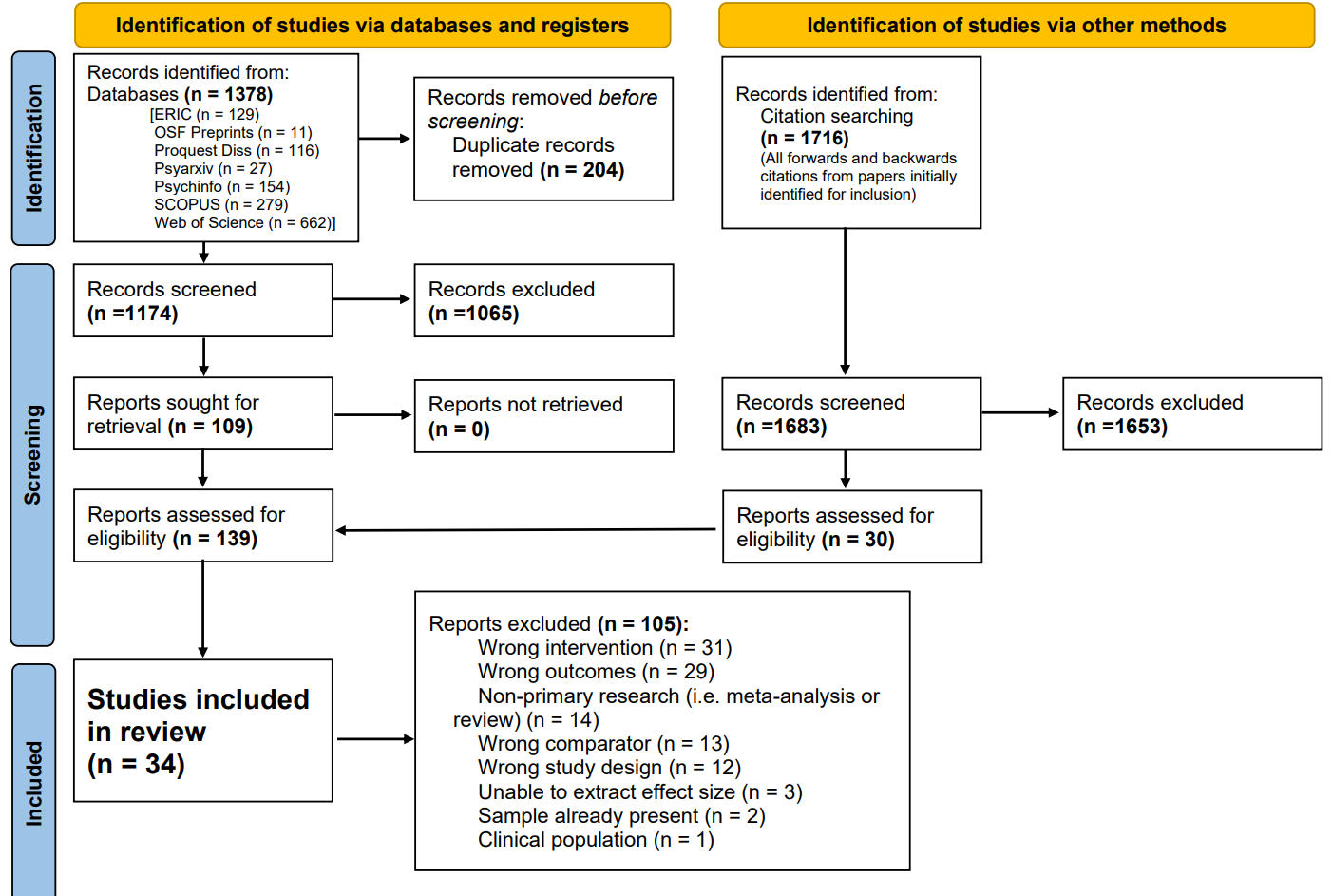
A librarian, with subject area expertise, was consulted to develop the Boolean search queries, as suggested in recent meta-analysis recommendations (Hansen et al., 2022; Steel et al., 2021). We first looked for any material related to distributed or retrieval practice (using a variety of terms) and then narrowed that down to only studies that mentioned mathematics. The exact query for each database has been made available to increase reproducibility in the supplementary material. At this stage, any article that mentioned the use of an intervention for mathematics education based on the spacing effect or retrieval effect was included. The database search was undertaken on the 11th of January 2023 and then updated after initial manuscript review on 25th of February 2025.

There are several papers that were excluded during the screening phase, that at first glance may have seemed applicable. We excluded three papers that used explicit timing designs to incorporate distributed practice because of their outcome measure: digits correct per minute (DCPM) (Bullard, 2020; Powell, 2022; Schutte et al., 2015). Digits correct per minute (a measure of fluency), did not seem comparable to a mean percentage score on a post-test (a measure of retention) which was the outcome specified in our pre-registration (<https://osf.io/qtfcu/>). An additional three papers were excluded where the mean percentage score, standard deviation or number of participants wasn’t provided and we were thus unable to extract them (Rea & Modigliani, 1985; Reed, 1924; Yazdani & Zebrowski, 2006), in each case we were not able to contact the authors.

### Data Collection

Figure 3

Prisma Diagram



*Note*. Prisma diagram

Alongside the database search a request for information was posted in the Mathematical Cognition and Learning Society mailing list but generated no replies. Additionally, authors included after full screening were asked via email whether they had any unreported data, to combat the file drawer problem (Rosenthal, 1979). Excluding those authors for whom no email was available, all but three authors responded to this question, however, no one reported any unpublished data.

After searching the databases, the title and abstract screening phase and full text reviews took place on *Covidence Systematic Review Software* (2023), this information is presented in the form of a PRISMA flowchart (see [Figure 3](#fig-PRISMA)). Two screeners then screened ten percent of titles and abstracts (n=88) to make sure that the inclusion and exclusion criteria (see [Table 1](#tbl-inc)) were being applied reliably and worked (88.6% inter-screener agreement rate). The inclusion and exclusion criteria were informed by the PICOS framework (see Appendix A). One screener then screened the remainder. Texts were then sent to full text review. Two screeners fully screened 10 of the papers (100% inter-screener agreement rate) before the remainder were screened by a single screener. The information required to estimate mean effect sizes, moderation and quality analyses was then extracted into an excel file containing the extraction table. During the extraction phase, the extraction table was evaluated by having two researchers extract the same five randomly selected papers, then meet up to discuss any discrepancies between the two sheets. This led to our terminology being clarified and we edited the extraction sheet to better fit the format of the studies being captured. A further three papers were extracted as a pair to check the updated extraction sheet was appropriate. Then a single screener extracted the remainder of the papers. The full data extraction sheet is available in the OSF repository (<https://osf.io/qtfcu/>). After the initial extraction nineteen authors were contacted regarding missing information. Nine replied (47%). Only one paper was unable to be included as critical information related to the calculation of effect sizes was unavailable and no reply was received. For six of the papers published before 2000, no current contact information for the researcher was found, but this did not prevent their inclusion.

Table 1

Inclusion and exclusion criteria

| Inclusion | Exclusion |
| --- | --- |
| The sample use either spaced repetition or retrieval practice (or both) | Non-experimental designs |
| Participants required to learn mathematics material | Clinical populations |
| Performance measured on a post-test | Effect size based solely on the mathematics material unable to be extracted |
| English abstract |  |

*Note*. These inclusion and exclusion criteria were visible to screeners during all phases of data collection.

### Computation of Effect Sizes

In all the studies included in this meta-analysis the recorded effect compares the difference in performance on a post-test between two groups. In the first analysis we have included the studies comparing spaced versus massed practice and in the second the studies which compared testing versus restudying.

The escalc function in the R module metafor (Viechtbauer, 2010) was used to calculate Hedges (1981)’s for each extracted effect. As most studies did not report the correlation for within-subject effects was used to compute the pooled standard error as used in previous spacing and retrieval meta-analyses (Latimier et al., 2021; Rowland, 2014). For between and within participant effects the was calculated as follows:

with the pooled standard deviation for between subject effects calculated as:

Where and are the number of participants in the first and second group, and are the observed means the first and second group and and are the observed standard deviations of the first and second group. The sample variance for between subjects was calculated using the following formula (Hedges, 1981):

and the standard error, .

For within participant effects the pooled standard deviation was using raw score standardization (4). The variance for within-subject effects was calculated as follows (Becker, 1988; Viechtbauer, 2010) (5).

Then for each subset of studies the overall weighted mean effect size was calculated using the metafor and clubSandwich modules in R. Using the guidelines provided by Pustejovsky and Tipton (2022) we chose to perform a Robust Variance estimation with correlated and hierarchical effects (CHE). We chose this method as many of the studies contribute multiple effect sizes but also often aim to find boundary conditions for the spacing or testing effect therefore within-study heterogeneity is expected. Using this method allows us to account for the correlation between effects from the same experiment and the hierarchical structure resulting from experiments reporting multiple effects. Heterogeneity is a measure of how different the effects are from one another, if there is a true effect with little variance then you would predict little heterogeneity.

Due to the small sample size the Q statistic would not have been an appropriate measure of heterogeneity (Gavaghan et al., 2000). Instead, we calculated using the dmetar package (Harrer et al., 2019), which uses the formula outlined in Cheung (2014) to calculate for each level of the hierarchical meta-analysis. In some cases, the initial analyses suggested there was very little heterogeneity, however visual inspection of the forest plots suggested this was unlikely. As can still be highly uncertain for between-study heterogeneity when sample size is small (Chung et al., 2013), confidence intervals were calculated to measure this uncertainty (Viechtbauer, 2017).

### Risk of Bias

To check for the risk of publication bias we performed Egger’s regression tests (Egger et al., 1997) on univariate versions of the meta-analyses. We also provide a corresponding funnel plot for each analysis. Additionally, we coded each source with quality indicators. The coding sheet was a modified version of Hjetland et al. (2020)’s coding sheet. We coded each extracted effect for sampling procedure used (0 - random, 1 - convenience), whether the test reliability was reported (0 - reported, 1 - not reported), if there were any floor or ceiling effects present (0 - not present, 1 - present), how missing data was dealt with (0 - better than list-wise deletion, 1 - list wise deletion), statistical power/sample size (0 - [N ≥ 150], 1 - [70 ≤ N <150], 2 - [N < 70] participants), whether attrition was reported (0 - reported, 1 - unreported). The scores were summed, and the overall quality score was used as a moderator to see if it accounted for a significant amount of heterogeneity.

Outliers were checked using the metafor R package (Viechtbauer, 2010) to calculate the Cook’s distance and difference in fit (DFFITS) values. The difference in fit value shows how many standard deviations the calculated mean effect size changes when a particular study is removed (Viechtbauer & Cheung, 2010).

### Potential Moderators

Due to small sample sizes and overly prevalent categories, we diverted from our preregistered method of coding multivariate moderators to, whenever possible, uni-variate form. This was because when the adjusted degrees of freedom (Satterthwaite, 1946) are below four, the moderator analyses tend to over-reject and are therefore unreliable (Tipton, 2015). For example, Latimier et al. (2021) also changed from multivariate to univariate due to small sample sizes. In the extraction sheet we have left the original multivariate categories, and the coding process is explained in full in the pre-registration. Despite this change none of the results were significant. As our moderating analyses were exploratory, we do not include them in this manuscript and instead they are available as a supplementary analysis.

## Results

This meta-analysis considers 34 studies with a total of 85 effect sizes. During the search and data collection processes it became clear that there were different ways that studies involving spacing could be broken down into subsets (isolated learning vs embedded into a course). We begin with an overview of all spacing studies before running further analyses on these two subsets. We then analyse retrieval studies that do not manipulate spacing.

### Spacing Effect in Mathematics Learning

Our first pre-registered analysis tested the effect of spaced versus massed practice in mathematics learning. The systematic review revealed 27 studies with 53 effect sizes. We found a weighted mean effect of *g* = 0.282 (se = 0.045) (95% confidence interval [0.188, 0.376]). However, upon further review it became clear that this analysis consisted of two different experimental designs. We first review studies that focus on learning a single, or small number of, mathematical skill(s) in isolation where all items have the same retrieval interval. We will refer to these studies as *isolated learning studies*. In contrast, we then review studies that involve spaced versus massed practice integrated into a course. We will refer to these studies as *course-embedded studies*. In course-embedded studies material covered at the beginning of the course has a longer retrieval interval than material covered at the end of the course and simpler material may also be incorporated into more complex material further along the course. Running a categorical meta-regression between isolated learning studies and course-embedded studies found that the effects were larger on average for isolated learning than course-embedded studies (see below), although they do not differ significantly , , . However, as we believe they are fundamentally measuring different effects, a pure measure of spaced practice (isolated learning) versus spaced and interleaved practice (learning embedded in a course), we ran all further analyses separately for the two subsets.

### Spacing Effect in Mathematics Learning - Isolated Learning

This subset consists of 10 studies with 27 effect sizes. Several studies focused on learning a simple combinatorial procedure (Ebersbach & Barzagar Nazari, 2020b; Emeny et al., 2021; Rohrer & Pashler, 2007; Rohrer & Taylor, 2006). Other areas of mathematics studied include arithmetic (Barzagar Nazari & Ebersbach, 2019; Chen et al., 2018), algebra (Chen et al., 2018; Gay, 1973) and statistics (Ebersbach & Barzagar Nazari, 2020a).

All the isolated learning studies were peer-reviewed articles. For the country moderator most of the studies were sampled from the USA, with others from Germany (Barzagar Nazari & Ebersbach, 2019; Betsch et al., 2015; Ebersbach & Barzagar Nazari, 2020a), the UK (Emeny et al., 2021) and China (Chen et al., 2018). Many of observed effects took place in a pre-university setting, for example, Emeny et al. (2021) in a secondary school and Chen et al. (2018) in a primary school. The mean age of participants ranged from 9.23 years (Chen et al., 2018) to 24.42 years (Ebersbach & Barzagar Nazari, 2020a).

The only study to use a within-subject design was Emeny et al. (2021). Retrieval intervals ranged from a single day (Chen et al., 2018) to six weeks (Barzagar Nazari & Ebersbach, 2019). The most common retrieval interval was one week. Inter-session intervals ranged from a single day (Chen et al., 2018) to two weeks (Gay, 1973). The most common inter-session interval was one day. Over half of the measured effects (61%) provided corrective feedback, the remainder provided no feedback. Under half of the studies had the first retrieval session immediately after the first-time participants learned the item, while the remainder had a delay. Each study had either two or three sessions in total. Mean performance at the first time point was on average 61% across all observations but ranged from 30% (Ebersbach & Barzagar Nazari, 2020a) to 95% (Rohrer & Pashler, 2007). The number of items (procedures, facts) participants had to learn, where reported, ranged from one to eight. The number of times the learning objective was exposed to participants ranged from three times to twelve (Emeny et al., 2021).

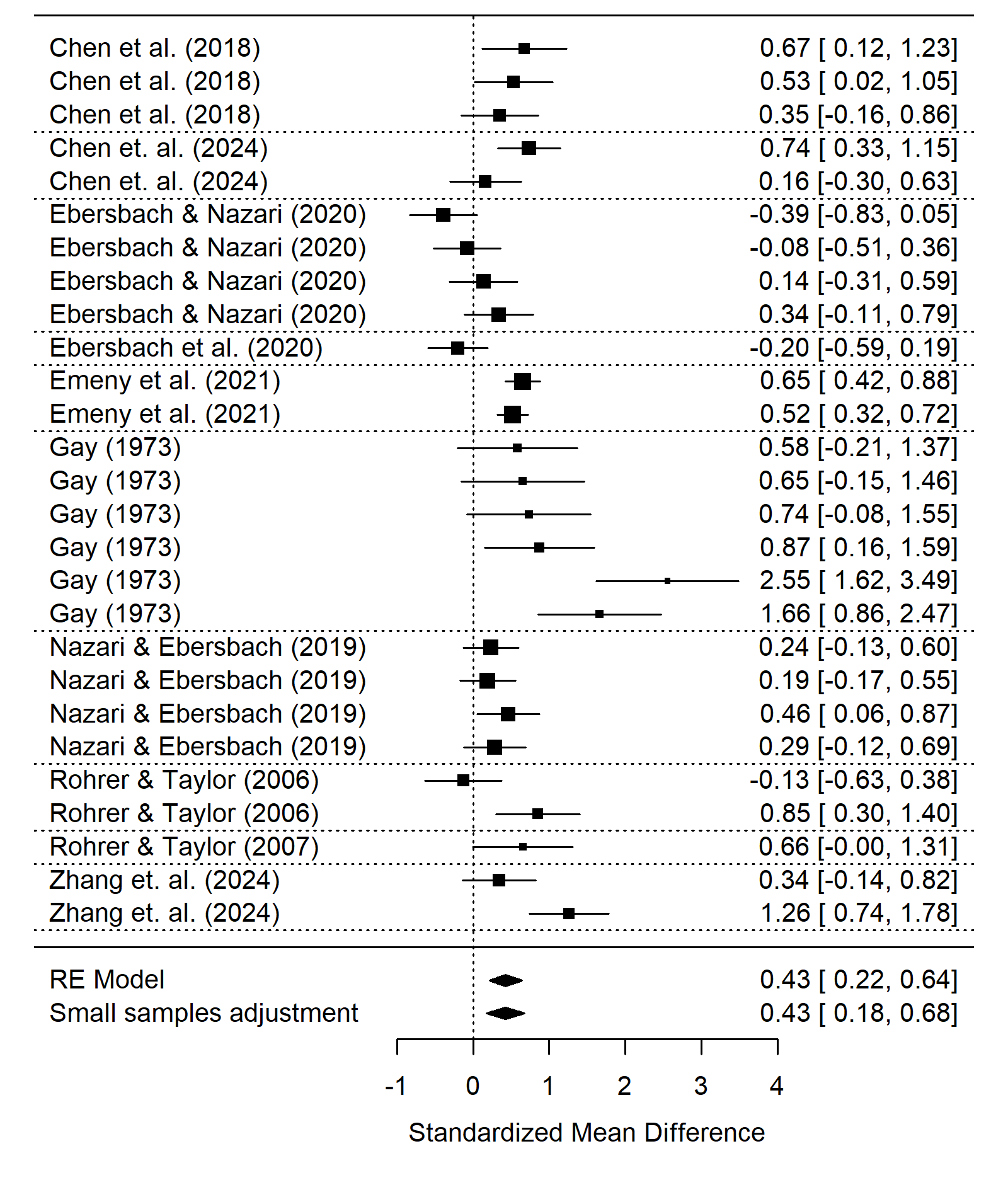
#### Overall Effect Size.

The overall weighted mean effect size of spaced versus massed practice for isolated learning is *g* = 0.427 (se = 0.107) (95% confidence interval [0.179, 0.675]). The hierarchical structure of the meta-analyses can allow us to see how much variance was associated with each level of the hierarchy. Firstly, = 24% of the variance was associated with the first level (sampling error), = 33% was associated with the second level (within study heterogeneity) (95% confidence interval [1.456, 82.675]) and = 43.287% associated with the third level (between-study heterogeneity) (95% confidence interval 0, 88.234]).

When the Cook’s distances for each effect were calculated none were greater than 0.5, suggesting that no single study was likely to be highly influential. In contrast, the difference in fit analysis shows that there are two dominant studies whose removal shifts the mean effect size by over half a standard deviation. Those studies are, firstly, Ebersbach and Barzagar Nazari (2020a) ( -0.708), which changes the mean effect size to *g* = 0.499 (se = 0.097) (95% confidence interval [0.268, 0.731]) and, secondly, Gay (1973) ( 0.638) whose exclusion changes the mean effect size to *g* = 0.368 (se = 0.096) (95% confidence interval [0.14, 0.596]). Critically, 95% confidence intervals remain greater than zero when either of these studies are removed, demonstrating the robustness of the effect.

Figure 4

Isolated Learning Studies - Forest Plot



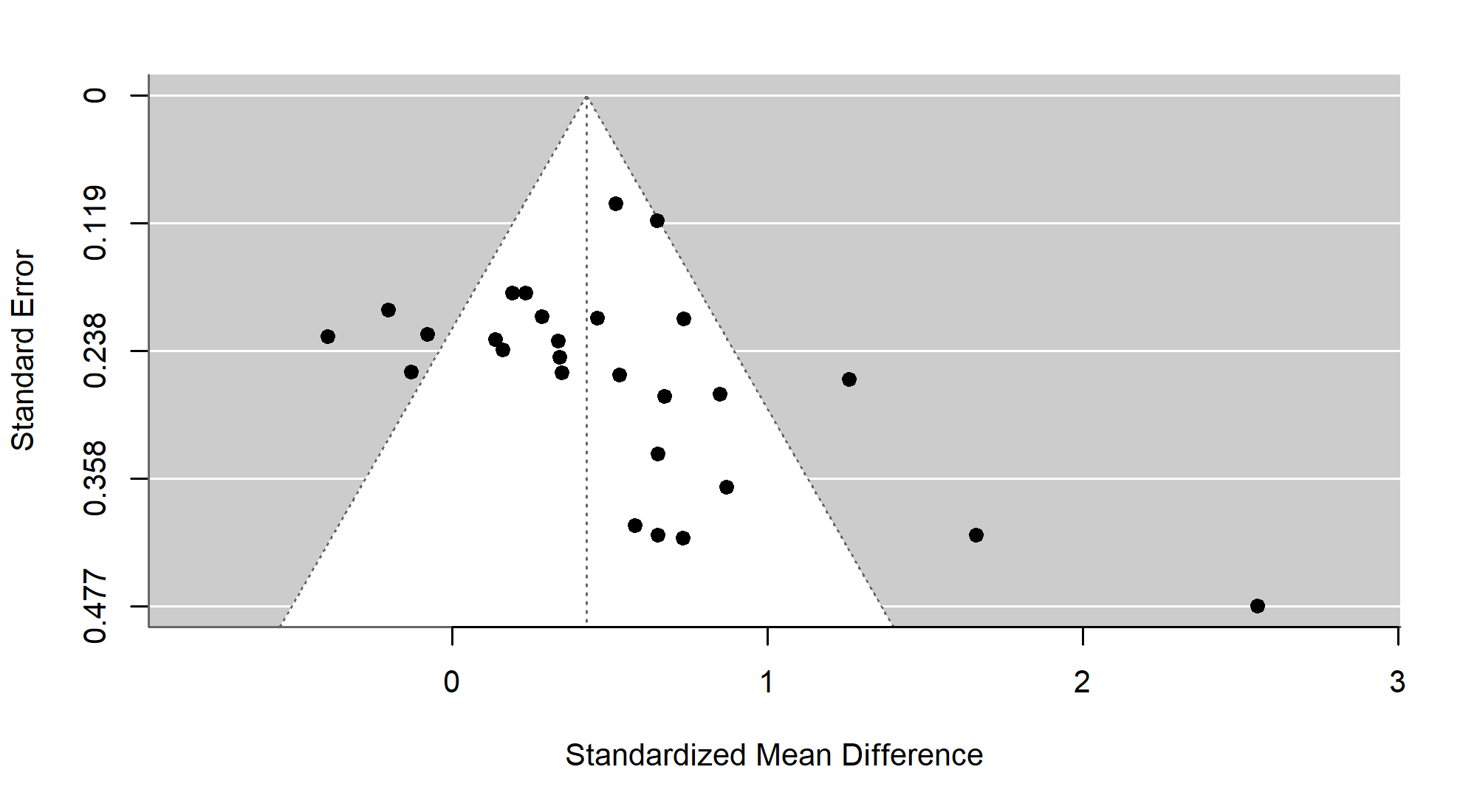
*Note*. A forest plot displaying the weighted mean effect sizes for each effect in the subset. The size of the square is proportional to the sample size, the error bars represent the 95% confidence interval. The diamond at the bottom represents the overall weighted mean effect size before and after the small sample adjustment.

#### Risk of Bias.

An insignificant Egger’s regression test for funnel plot asymmetry (*t* = 1.033, d.f. = 25, *p* = 0.312) does not suggest publication bias resulting from non-significant results remaining unpublished. In contrast, the funnel plot (see [Figure 5](#fig-ss1-funnel)) appears to show asymmetry. There appears to be a significant gap of lower powered observed spacing effects that were smaller or negative, which could be an indicator of publication bias. This is balanced by effects with a smaller standard error that were closer to zero or negative, which is why the asymmetry may not appear in the Egger’s regression test. Furthermore, when a trim and fill analysis was applied, it suggested there were no studies that needed to be added.

Figure 5

Isolated Learning Studies - Funnel Plot



*Note*. A funnel plot showing the effect size of spaced practice versus massed practice against the standard error for each effect in the sample. Large asymmetry would suggest publication bias.

Overall, these analyses suggest that there is a medium to large beneficial effect of spaced versus massed practice when a particular mathematical procedure/skill is taught in isolation under more controlled conditions.

### Spaced versus massed practice - Course-embedded Studies

This subset consists of 17 studies with 26 effect sizes. A variety of mathematical areas are used for the material including algebra (Camp, 1973; Goettl et al., 1996; Holdan, 1986; Lerma, 1990; Reed, 1924), arithmetic (Moss, 1996; Weaver, 1976), calculus (Beagley & Capaldi, 2016, 2020; Bego et al., 2017; Gorgievski, 2012; Lyle et al., 2020, 2022) and statistics (Crissinger, 2015).

All studies recruited participants in the USA. More than half (60%) used varying inter-session intervals based on a formula (e.g., Hirsch et al., 1982) to distribute questions on a particular topic or skill, while others had an inter-session interval interval for each learning objective that remained constant (Lyle et al., 2020). Three quarters of the studies provided some form of feedback while a quarter did not. Goettl et al. (1996) was the only study that was run in a lab setting, but was still structured as a course, while the remainder were embedded into a course in a classroom or as homework. The mean age of participants was only reported for 23% of observed effects, but for those effects the mean age was 17.4 years old.

Over half the effects were extracted from peer reviewed articles (54%), while the remainder were extracted from theses (i.e., Camp, 1973; Gorgievski, 2012) or a conference paper (Bego et al., 2017). Many of the studies took place in a university setting (65% of effects) while the remainder either took place in secondary schools or were drawn from an adult population not in University (Goettl et al., 1996). Studies that took place in the classroom or used untimed homework could have allowed pupils to look back at their work, making retrieval uncertain (53% of observed effects) while studies that used timed quizzes in class or at home reduced that chance and were coded as requiring retrieval. Most of the studies adopted an expanding inter-session interval design (68% of observations) while the remainder used a uniform design. Thirty percent of effects were within subjects while 70% were between subjects.

The number of items refers to the number of learning objectives that were intended to be learnt and tested at the final exam, this ranged from seven (Reed, 1983) to forty-eight (Hopkins et al., 2016) where reported. The number of exposures refers to how many times each learning objective is practiced and ranged from three (Hopkins et al., 2016) to nine times (Gorgievski, 2012). Mean performance at first time point ranged from 11% (Goettl et al., 1996) to 85% (Hirsch et al., 1982). Mean inter-session interval length ranged from one day (Goettl et al., 1996) to two weeks (Lyle et al., 2020). Time from final practice/retrieval session to exam ranged from two days (Camp, 1973) to five weeks (Hopkins et al., 2016; Lyle et al., 2020; Weaver, 1976) and five weeks was the most common delay.

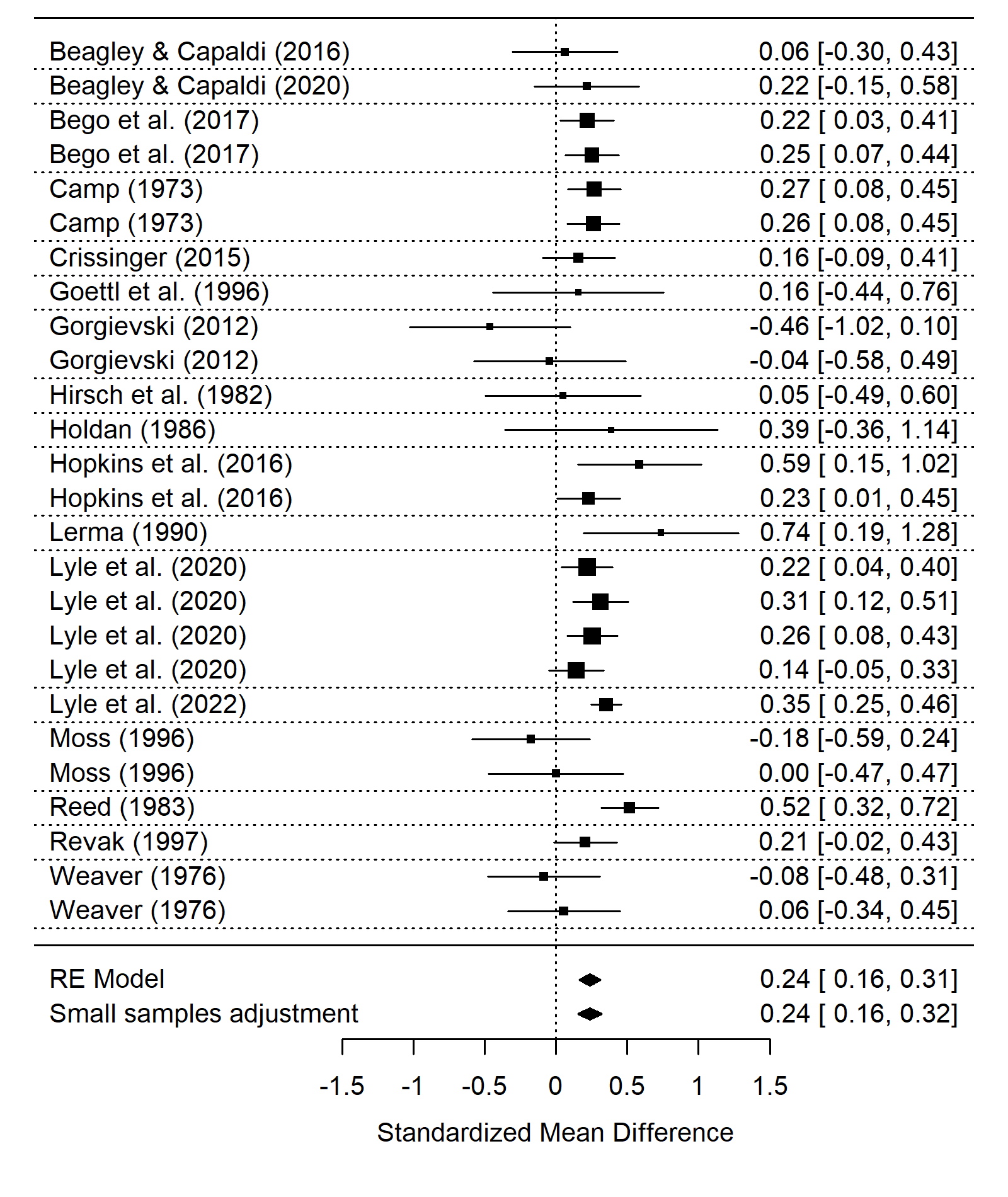
#### Overall Effect Size.

The overall weighted mean effect size of spaced versus massed practice for course-embedded material is *g* = 0.24 (se 0.038) (95% confidence interval [0.155, 0.324]). The hierarchical structure of the meta-analyses can allow us to see how much variance was associated with each level of the hierarchy. Firstly, = 57% of the variance was associated with the first level (sampling error), = 0% was associated with the second level (within study heterogeneity) (95% confidence interval [0, 50.555]). However, there was a large amount of uncertainty in the between-study heterogeneity due to the small sample size (Chung et al., 2021). Finally, = 43.065% associated with the third level (between-study heterogeneity) (95% confidence interval [0, 78.745]).

Again, the Cook’s distances for each effect were calculated and none were greater than 0.5 which suggested that no single study was likely to be highly influential. The difference in fit analysis shows that study 15 (Lyle et al., 2022) shifts the mean effect size just under half a standard deviation ( 0.491), the removal of this study changes the mean effect size to *g* = 0.24 (se = 0.038) (95% confidence interval [0.155, 0.324]). The studies that produced the next largest difference in fit were Hirsch et al. (1982) ( 0.638) and Gorgievski (2012) ( 0.147). The removal of Hirsch et al. (1982) changes the mean effect size to *g* = 0.256 (se = 0.036) (95% confidence interval [0.176, 0.337]).

Figure 6

Course-embedded Studies - Forest plot



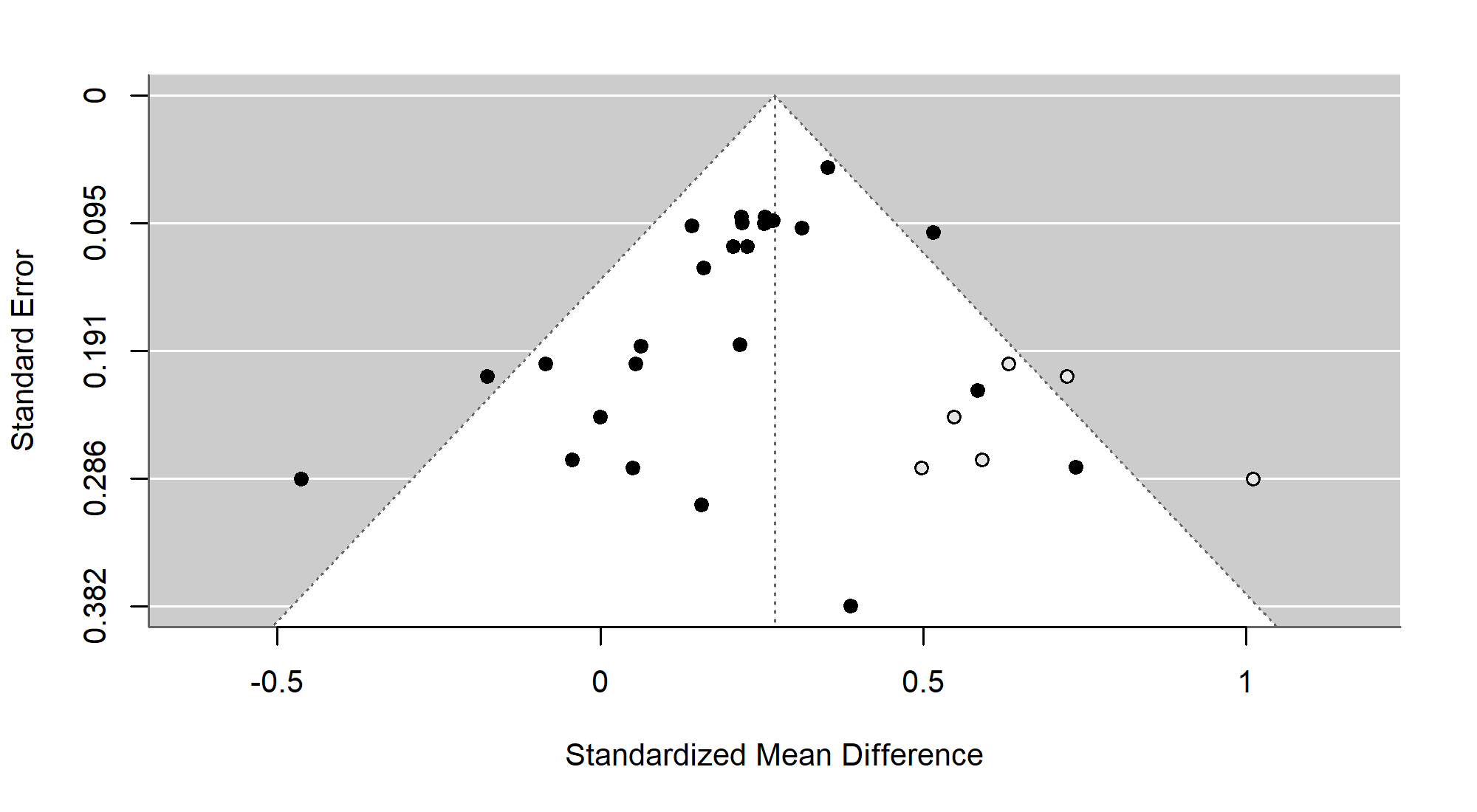
*Note*. A forest plot displaying the weighted mean effect sizes for each effect in the subset. The size of the square is proportional to the sample size, the error bars represent the 95% confidence interval. The diamond at the bottom represents the overall weighted mean effect size before and after the small sample adjustment.

#### Risk of Bias.

Egger’s regression test for funnel plot asymmetry is significant (*t* = -2.4, d.f. = 24, *p* = 0.025) suggesting publication bias resulting from insignificant results remaining unpublished. The negative t-value would suggest under-reporting of results that exceed the mean effect size. The funnel plot (see [Figure 7](#fig-ss2-funnel)) does not show obvious asymmetry. When the trim and fill method (Duval & Tweedie, 2000) is applied to the uni-variate version of the model it adds in 6 studies and provides a corrected effect of 0.27[0.215, 0.326].

Figure 7

Course-embedded Studies - Funnel Plot



*Note*. A funnel plot showing the effect size of spaced practice versus massed (for course-embedded studies) against the standard error for each effect in the sample. Large asymmetry would suggest publication bias. The black circles represent the studies included in the main analysis, while grey studies are studies suspected to be missing by the trim and fill method.

In summary, there is a small to medium positive effect of spaced versus massed practice when the material is incorporated into a course structure, it appears to be a numerically smaller effect than for isolated learning.

### Retrieval versus Restudying in Mathematics Learning

This subset consists of 7 studies with 32 effect sizes. Fazio (2018) focused on learning and applying a particular procedure (fraction multiplication), while the others used a mixture of learning procedures alongside concepts or using the knowledge in problem solving tasks. The most common topic was statistics (Eustace et al., 2020; Lyle & Crawford, 2011; Yeo & Fazio, 2019), with the other studies focusing on probability (Dirkx et al., 2014), geometry (Betsch et al., 2015) or arithmetic (Fazio, 2018).

Half of the studies were undertaken in the USA, with the other three sampling from populations in the Netherlands (Dirkx et al., 2014), Germany (Betsch et al., 2015) and Ireland (Eustace et al., 2020). All studies took place in a university setting except one in a Secondary setting (Dirkx et al., 2014) and one in a primary setting (Betsch et al., 2015). Where reported, the mean age of participants ranged from approximately nine (Betsch et al., 2015) to twenty-two years old (Lyle & Crawford, 2011). All the included studies were peer-reviewed articles except for Fazio (2018) which was published as a book chapter.

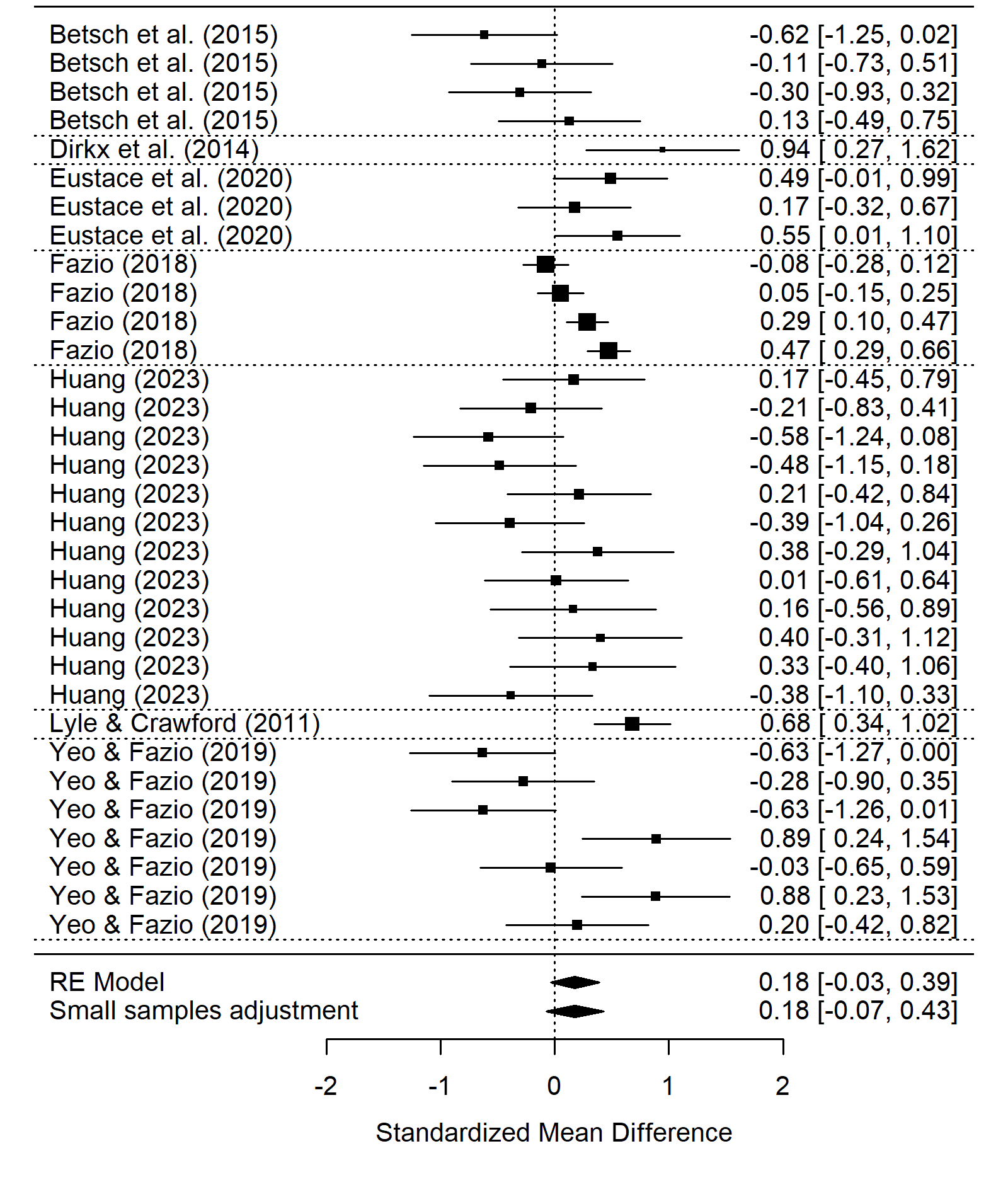
Where reported, participants were provided with three or four exposures to each item to be learnt, for example, in Fazio (2018) (Experiment 1) participants had either four study opportunities (SSSS) or one study and three test opportunities (STTT). Each study or testing session ranged from four (Yeo & Fazio, 2019) to eight minutes (Dirkx et al., 2014). For the retrieval interval, 42% of the observations had a retrieval interval of less than one day, while the remainder were longer. Eustace et al. (2020), Betsch et al. (2015) and Lyle and Crawford (2011) provided feedback and Yeo and Fazio (2019) gave feedback in one condition in their second experiment, while the other studies did not. Just under half of the studies (46%) received their retrieval opportunity immediately, while the remainder had it after a delay. Two studies were lab experiments (Fazio, 2018; Yeo & Fazio, 2019) while the remainder were performed in a classroom setting. Due to the nature of the testing effect most studies did not include an initial test, as even being tested before learning material can affect subsequent learning, this is known as the *pretesting effect* (Pan & Carpenter, 2023), the exceptions were Fazio (2018) and Dirkx et al. (2014). Fazio (2018) was also the only study that adopted a within-subject design.

#### Overall Effect size.

Calculating the impact of testing versus restudying for mathematics learning, the overall weighted mean effect size is *g* = 0.184 (se = 0.095) (95% confidence interval [-0.069, 0.436]). Note that the 95% confidence intervals crossed the zero line. The hierarchical structure of the meta-analyses can allow us to see how much variance was associated with each level of the hierarchy. Firstly, = 30% of the variance was associated with the first level (sampling error), = 50% was associated with the second level (within study heterogeneity) (95% confidence interval [26.417, 83.001]) and = 20.105% associated with the third level (between-study heterogeneity) (95% confidence interval [0, 89.458]).

Figure 8

Retrieval versus Restudying - Forest plot



*Note*. A forest plot displaying the weighted mean effect sizes for each effect in the subset. The size of the square is proportional to the sample size, the error bars represent the 95% confidence interval. The diamond at the bottom represents the overall weighted mean effect size before and after the small sample adjustment.

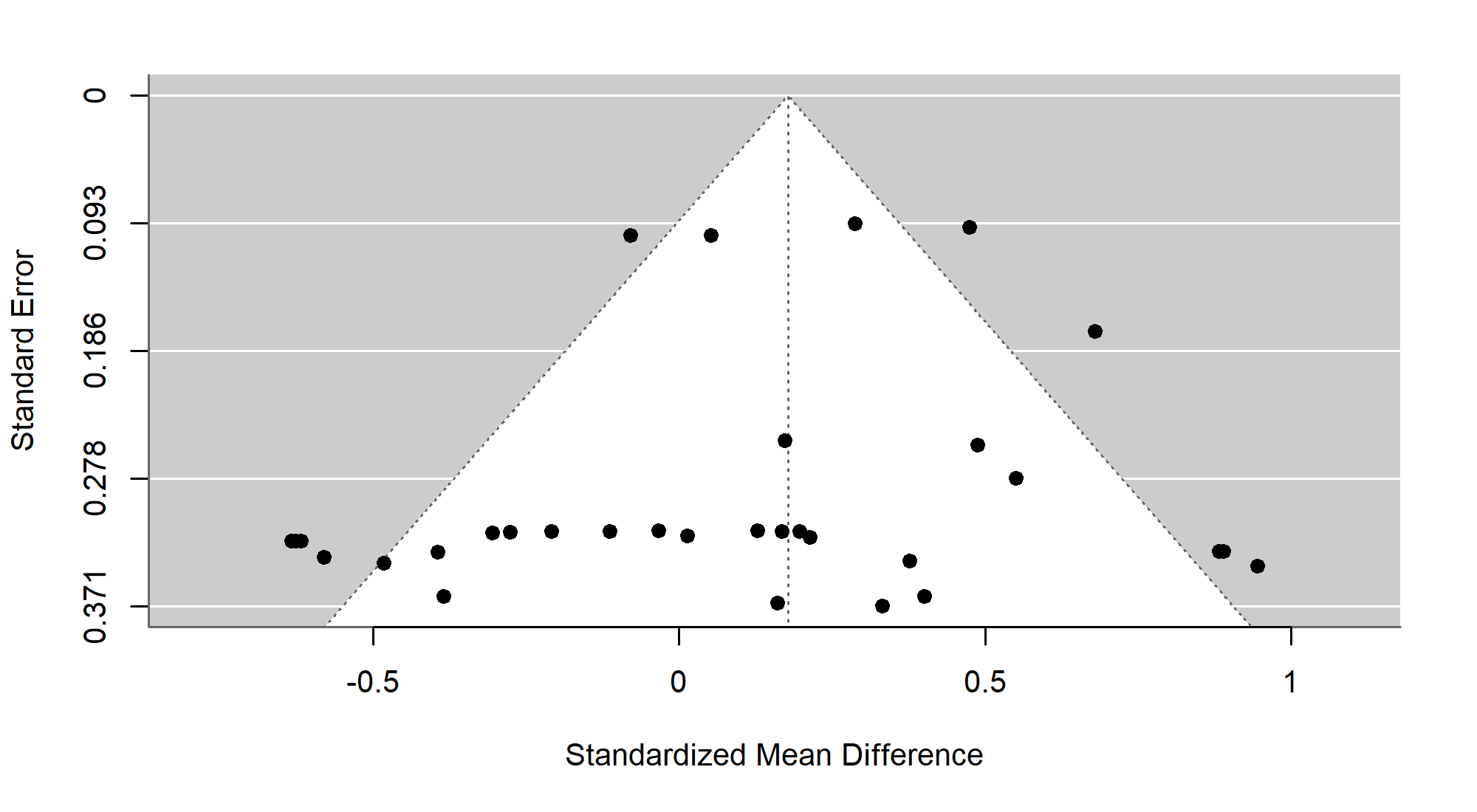
When the Cook’s distances for each effect were calculated none were greater than 0.5 which suggested that no single study was likely to be highly influential. However, the difference in fit analysis shows that the removal of one study (Betsch et al., 2015) shifts the mean effect size by almost half a standard deviation ( -0.483), which changes the mean effect size to *g* = 0.239 (se = 0.096) (95% confidence interval [-0.038, 0.516]). These results suggest the lack of robustness is not due to any one studies inclusion as the 95% confidence interval always crosses zero regardless of which study is removed.

#### Risk of Bias.

An insignificant Egger’s regression test for funnel plot asymmetry (*t* = -1.279, d.f. = 30, *p* = 0.211) does not suggest publication bias resulting from insignificant results remaining unpublished. Additionally, there is not clear asymmetry in funnel plot (see [Figure 9](#fig-ss3-funnel)).

Figure 9

Retrieval versus Restudying - Funnel Plot



*Note*. A funnel plot showing the effect size of retrieval practice versus restudy against the standard error for each effect in the sample. Large asymmetry would suggest publication bias.

Our analyses were hampered by the small sample of studies that explicitly compare testing to restudy in mathematics and the overall effect was smaller than previously found and was not robust.

## Discussion

This meta-analysis had three purposes. First, to synthesize the current evidence regarding the efficacy of spaced versus massed practice with regards to mathematics learning. We found a robust small to medium effect of spacing, versus massed practice, overall, and in both subsets. This spacing effect was larger for material taught in isolation and smaller for material embedded into a course. Second, to investigate the efficacy of retrieval practice versus restudy within the domain of mathematics learning. We found a small to medium effect of retrieval versus restudy, however, this was not robust as the 95% confidence interval crossed zero. Third, we intended to examine the effect of spaced retrieval practice versus massed retrieval practice, however, this was not possible due to a lack of studies.

Our first aim was to answer the question: does the spacing of mathematics practice lead to higher retention than massed practice? Our preregistered analysis looked at the weighted mean effect size of all studies that included a spaced versus massed manipulation. We found a small to medium effect of spacing (*g* = 0.26) and the 95% confidence interval suggests this is robust. However, this effect size is smaller than typically found effects of spacing in other domains such as L2 Language learning (*g* = 0.80, with a delayed post-test, Kim & Webb, 2022), however it is similar to more general reviews of spaced versus massed practice (*g* = 0.46, Donovan & Radosevich, 1999). This suggests that spacing is effective for mathematics learning, but some aspect of the material or the way it is taught may lead to reduced efficacy in comparison to other domains.

Furthermore, two subsets of studies with distinct paradigms contributed to this overall effect. Isolated learning studies followed a traditional spacing paradigm where massed practice is compared to spaced practice (over multiple sessions) before being tested after a specified delay. In contrast, course-embedded studies were embedded in a course structure, each individual item following a spacing schedule, however a crucial difference is that material at the start of the course may have a much longer retrieval interval than material at the end of the course. Additionally, the material is interleaved with the other questions. Interleaving provides a bonus effect of increasing learners’ ability to discriminate between question types, a skill unnecessary in isolated learning studies, but very relevant in course-embedded studies. Furthermore, due to the cumulative nature of mathematics material, the end of the course typically builds on prior material, so the prior material may receive more practice than reported. For example, in a course on calculus, basic rules for differentiation may be tested explicitly at the beginning, but then also used to answer more complex questions at the end, resulting in additional practice.

For isolated learning studies, our analysis consisted of nine studies with thirty-five effect sizes. For isolated learning, there was a significant small to medium effect of spacing (*g* = 0.427), which was much larger than the overall spacing effect, however, the lower limit of the 95% confidence interval is closer to zero than the overall spacing effect. There were two studies whose removal would cause a significant change in the overall mean effect, however neither caused the confidence interval to cross zero. The increased experimental control available in isolated learning studies provides a purer measure of the spacing effect and is more comparable to past spacing effect meta-analyses such as the one conducted by Donovan and Radosevich (1999).

There are hints of publication bias in the isolated learning studies. The funnel plot is asymmetrical, there are no negative, or close to zero, effects for studies with larger standard errors than 0.3, while there are seven studies that found a significant effect with a standard error larger than 0.3, which hints at potential publication bias. However, while the funnel plot asymmetry hints at a possible under-reporting of smaller effects, neither Egger’s regression test nor the trim and fill method suggested significant statistical evidence of publication bias. This suggests an overall lack of strong evidence for publication bias for isolated learning studies.

Seventeen course-embedded studies with twenty-six effect sizes were analysed. We found a robust significant effect of spacing for course-embedded studies (*g* = 0.24). The difference in fit analysis showed that the removal of no single study would cause the 95% lower limit to drop below zero, providing further confidence in this study effect. As these studies implemented spacing into a course, they also benefited from interleaving which we would have expected to have produced an additional beneficial effect and increased performance on the post-test.

One potential explanation for the smaller mean effect (though not significantly smaller) in course-embedded studies could be linked to the study-phase retrieval hypothesis. The study-phase retrieval hypothesis suggests a key mechanism underlying the spacing effect is the retrieval of the initial study-phase. In a course setting in classrooms or as homework, students likely have prior material they can access during practice. If students have access to the materials, they may not have to retrieve the original memory, instead relying on external guides. Two isolated learning studies make clear statements as to which materials participants had access to, for example, “While working on the practice sheets, each student had access to a summary sheet containing examples, including solutions, from the introductory lesson.” (Barzagar Nazari & Ebersbach, 2019, pg. 291) and “Throughout each practice session, students could see their written work for practice problems that had appeared previously in the session.” (Emeny et al., 2021 pg. 1085). However, even in these cases it is difficult to understand to what extent participants made use of these materials. It has been previously noted that distributing practice may benefit medium performers best (Barzagar Nazari & Ebersbach, 2019). For these students, they may feel equipped to attempt the problem before checking the available materials, using them for feedback rather than undercutting the retrieval process. It would be valuable for future experiments to both control for and directly manipulate what materials students have access to, and when, during spaced practice of mathematics.

Within a course structure both the nature of mathematics learning and the lack of experimental control may have led to a smaller effect. The cumulative nature of mathematics learning means that, even in the massed conditions, simpler material may be practiced after a delay when more complex information is introduced, introducing an element of spacing into the massed condition. Further, sometimes spacing has been observed to increase accuracy when learners provide judgments of learning (Emeny et al., 2021; Logan et al., 2012), however, it is often clear whether you have successfully completed a mathematics problem so this beneficial effect may be minimised in the spaced condition. On the other hand, there are multiple reasons to predict the spacing effect could have been stronger within subset two. First, as these studies were embedded into a course structure, the learners may have been more intrinsically motivated as their grade counted for something, unlike most of the experiments in subset one. Second, in the spaced condition, the material was often also interleaved with material that had been previously learned, which could have provided a bonus ability for students to discriminate between question types and could have improved performance (Foster et al., 2019). Overall, there is no clear reason why spacing may be less effective when embedded in a course.

The analysis of studies comparing retrieval practice versus restudy sought to help us answer the question: does retrieval practice of mathematics lead to better learning (than just restudy)? It consisted of six studies with twenty effect sizes. Our analysis suggested the testing effect in mathematics may provide some advantage (*g* = 0.184), however, the 95% confidence interval crosses zero, which suggests that the effect is not robust.

We had planned to investigate whether the combination of spacing and retrieval lead to higher retention of mathematics knowledge (versus massed retrieval)? However, there was insufficient information in most studies to decide whether the spacing study required retrieval or not. Ultimately, we deviated from our pre-registered retrieval moderator (yes or no) to retrieval (certain or uncertain). If it was a timed online test or under test conditions, we said it was likely to require retrieval, otherwise we could not be sure. The moderator analyses for whether retrieval was certain or uncertain were not significant.

Every effort was made to follow current best practices in meta-analysis (Steel et al., 2021) including pre-registering, checking for publication bias, quality indicators and using difference of fits analyses. We calculated the overall effect using robust variance estimation with correlated hierarchical effects. This allowed us to account for multiple comparisons within the same study (Pustejovsky & Tipton, 2022). We made systematic efforts to uncover unpublished studies, however, no authors we contacted reported any unpublished studies.

Given the current available literature, we were not able to state whether there is or is not a robust effect of testing for mathematics material. Similarly, the use of practice questions in mathematics and the lack of clarity as to whether learners were able to view past work meant that we were not able to accurately code whether studies required retrieval or not. We did, however, find significant evidence for a beneficial effect of spacing for mathematics material. This effect was smaller in studies where the material was embedded in a course, in comparison to stand alone experiments. Overall, the current state of the literature allowed us to answer one of our three questions, spacing is effective for mathematics material. However, we were unable to answer whether retrieval practice is effective for mathematics learning and how the testing and spacing effects interact. Future studies should aim to control, or directly vary, whether participants have to retrieve the information during practice or have it available to them. We believed that while not differing statistically, the isolated and course-embedded subsets implement the spacing effect in two theoretically distinct ways. By carefully manipulating the order of practice problems, into blocked, interleaved and remote interleaved sets, Foster et al. (2019) were able to calculate the relative contributions to learning from spacing and interleaving. However, in their experiment the practice took place in a single session with no significant temporal delays between sets, so isn’t directly applicable to the difference between isolated and course-embedded material. Future experiments involving course-embedded spacing should take care to structure problem questions to allow the relative contributions of spacing and interleaving to be extracted. More testing effect studies that use testing effect best practices, rather than pushing the effect to find boundary conditions, are required to boost confidence in the utility of the testing effect to benefit mathematics material.

## Conclusion

We found robust evidence for a positive small to medium effect of spaced rather than massed practice for mathematics learning. This effect was numerically larger for studies where participants learnt isolated material in contrast to when the material was embedded in a course, but importantly, the effect was robust in both subsets. We found a small to medium mean effect of testing versus restudy, however this was not robust. Overall, there is sufficient evidence to promote the use of spacing for mathematics learning as a means to improve mathematics learning. Given the robust nature of retrieval practice in other domains, there is insufficient evidence to suggest that retrieval practice is not a useful pedagogical tool for mathematics learning, however, more evidence is required to ensure the robustness of the testing effect.

## References

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

Barzagar Nazari, K., & Ebersbach, M. (2019). Distributed practice in mathematics: Recommendable especially for students on a medium performance level? *Trends Neurosci Educ*, *17*, 100122. <https://doi.org/10.1016/j.tine.2019.100122>

Beagley, J. E., & Capaldi, M. (2016). The effect of cumulative tests on the final exam. *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *26*(9), 878–888. <https://doi.org/10.1080/10511970.2016.1194343>

Beagley, J. E., & Capaldi, M. (2020). Using cumulative homework in calculus classes. *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *30*(3), 335–348. <https://doi.org/10.1080/10511970.2019.1588814>

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 257–278. <https://doi.org/10.1111/j.2044-8317.1988.tb00901.x>

Bego, C. R., Lyle, K. B., Ralston, P. A., & Hieb, J. L. (2017). *2017 IEEE frontiers in education conference (FIE)*. *2017-October*, 1–5. <https://doi.org/10.1109/FIE.2017.8190463>

Betsch, T., Quittenbaum, N., & Lüders, M. (2015). On the robustness of the quizzing effect under real teaching conditions. *Z. Entwicklungspsychol. Padagog. Psychol.*, *29*(2), 109–114. <https://doi.org/10.1024/1010-0652/a000149>

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. <https://doi.org/10.1037/bul0000209>

Bullard, J. C. (2020). Evaluating explicit timing modifications to maximize student learning rates: A comparison of distributed practice and goal setting with performance feedback. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *81*(4-A), No–Specified. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc17&NEWS=N&AN=2020-28119-064>

Camp, J. S. (1973). The effects of distributed practice upon learning and retention in introductory algebra. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *7328190*(5-A), 2455–2456. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1976-03010-001>

Chen, O., Castro-Alonso, J. C., Paas, F., Sweller, J., & John;. (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educ. Psychol. Rev.*, *30*(2), 483–501. <https://doi.org/10.1007/s10648-017-9426-2>

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. <https://doi.org/10.1037/a0032968>

Chung, Y., Rabe-Hesketh, S., & Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, *32*(23), 4071–4089. <https://doi.org/10.1002/sim.5821>

*Covidence systematic review software*. (2023). Veritas Health Innovation.

Crissinger, B. R. (2015). The effect of distributed practice in undergraduate statistics homework sets: A randomized trial. *J. Stat. Educ.*, *23*(3). <https://doi.org/10.1080/10691898.2015.11889743>

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). *Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature* (Vol. 53, pp. 63–147). <https://doi.org/10.1016/S0079-7421(10)53003-2>

Dirkx, K. J. H., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *J. Educ. Res.*, *107*(5), 357–364. <https://doi.org/10.1080/00220671.2013.823370>

Donovan, J. J., & Radosevich, D. J. (1999). *A meta-analytic review of the distribution of practice effect: Now you see it, now you don’t*.

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>

Ebersbach, M., & Barzagar Nazari, K. (2020a). Implementing distributed practice in statistics courses: Benefits for retention and transfer. *J. Appl. Res. Mem. Cogn.*, *9*(4), 532–541. <https://doi.org/10.1016/j.jarmac.2020.08.014>

Ebersbach, M., & Barzagar Nazari, K. (2020b). No robust effect of distributed practice on the short- and long-term retention of mathematical procedures. *Frontiers in Psychology*, *11*, 811. <https://doi.org/10.3389/fpsyg.2020.00811>

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

Emeny, W. G., Hartwig, M. K., & Rohrer, D. (2021). Spaced mathematics practice improves test scores and reduces overconfidence. *Appl. Cogn. Psychol.*, *35*(4), 1082–1089. <https://doi.org/10.1002/acp.3814>

Eustace, J., Bradford, M., & Pathak, P. (2020). Evaluating flipped and traditional pedagogy with retrieval practice in mathematics for computing. *Journal of Computers in Mathematics and Science Teaching*, *39*(2), 149–167. <https://www.learntechlib.org/primary/p/210971/>

Fazio, L. K. (2018). *The effects of retrieval practice on fraction arithmetic knowledge* (pp. 169–182). Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9781315200446-11>

Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., & Dunlosky, J. (2019). Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice. *Mem. Cognit.*, *47*(6), 1088–1101. <https://doi.org/10.3758/s13421-019-00918-4>

Gavaghan, D. J., Moore, R. A., & McQuay, H. J. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain*, *85*(3), 415–424. <https://doi.org/10.1016/S0304-3959(99)00302-4>

Gay, L. R. (1973). Temporal position of reviews and its effect on the retention of mathematical rules. *J. Educ. Psychol.*, *64*(2), 171–182. <https://doi.org/10.1037/h0034595>

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Mem. Cognit.*, *7*(2), 95–112. <https://doi.org/10.3758/bf03197590>

Goettl, B. P., Yadrick, R. M., Connolly-Gomez, C., Regian, J. W., & Shebilske, W. L. (1996). Alternating task modules in isochronal distributed training of complex tasks. *Hum. Factors*, *38*(2), 330–346. <https://doi.org/10.1518/001872096779048048>

Gorgievski, N. (2012). The impact of the spacing effect and overlearning on student performance in calculus. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *3464319*(9-A), 3178. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc9&NEWS=N&AN=2012-99050-453>

Hake, S. (2012). *Saxon math. Course 1*. Houghton Mifflin Harcourt Publishing Company.

Hansen, C., Steinmetz, H., & Block, J. (2022). How to conduct a meta-analysis in eight steps: A practical guide. *Management Review Quarterly*, *72*(1). <https://doi.org/10.1007/s11301-021-00247-4>

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *Dmetar: Companion R package for the guide ’doing meta-analysis in R’* [Manual].

Hedges, L. V. (1981). Distribution Theory for Glass’s Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–128. <https://doi.org/10.3102/10769986006002107>

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. *Theories in Cognitive Psychology: The Loyola Symposium.*, *386*. <https://psycnet.apa.org/fulltext/1975-00291-001.pdf>

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Mem. Cognit.*, *32*(2), 336–350. <https://doi.org/10.3758/bf03196863>

Hirsch, C. R., Kapoor, S. F., & Laing, R. A. (1982). Alternative models for mathematics assignments. *Internat. J. Math. Ed. Sci. Tech.*, *13*(3), 243–252. <https://doi.org/10.1080/0020739820130301>

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervåg, M. (2020). Preschool pathways to reading comprehension: A systematic meta-analytic review. *Educational Research Review*, *30*, 100323. <https://doi.org/10.1016/j.edurev.2020.100323>

Holdan, E. G. (1986). A comparison of the effects of traditional, exploratory, distributed, and a combination of distributed and exploratory practice on initial learning, transfer, and retention of verbal problem types in first-year algebra. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *46*(9-A), 2542. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1986-55259-001>

Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. S. (2016). Spaced retrieval practice increases college students’ short- and long-term retention of mathematics knowledge. *Educ. Psychol. Rev.*, *28*(4), 853–873. <https://doi.org/10.1007/s10648-015-9349-8>

Kim, S. K., & Webb, S. (2022). The Effects of Spaced Practice on Second Language Learning: A Meta-Analysis. *Language Learning*, *72*(1), 269–319. <https://doi.org/10.1111/lang.12479>

Küpper-Tetzel, C. E. (2014). Understanding the distributed practice effect. *Zeitschrift Für Psychologie*, *222*(2), 71–81. <https://doi.org/10.1027/2151-2604/a000168>

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educ. Psychol. Rev.*, *33*(3), 959–987. <https://doi.org/10.1007/s10648-020-09572-8>

Lerma, A. (1990). *The effects of a cumulative review homework assignment schedule upon achievement in intermediate algebra and attitude toward mathematics* [PhD thesis]. <https://www.proquest.com/dissertations-theses/effects-cumulative-review-homework-assignment/docview/303918610/se-2>

Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: the role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning*, *7*(3), 175–195. <https://doi.org/10.1007/s11409-012-9090-3>

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2020). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. *Educ. Psychol. Rev.*, *32*(1), 277–295. <https://doi.org/10.1007/s10648-019-09489-x>

Lyle, K. B., Bego, C. R., Ralston, P. A. S., & Immekus, J. C. (2022). Spaced retrieval practice imposes desirable difficulty in calculus learning. *Educ. Psychol. Rev.*, *34*(3), 1799–1812. <https://doi.org/10.1007/s10648-022-09677-2>

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teach. Psychol.*, *38*(2), 94–97. <https://doi.org/10.1177/0098628311401587>

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *J. Cogn. Psychol.*, *28*(6), 684–706. <https://doi.org/10.1080/20445911.2016.1181637>

Moss, V. D. (1996). The efficacy of massed versus distributed practice as a function of desired learning outcomes and grade level of the student. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *9603493*(9-B), 5204. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc3&NEWS=N&AN=1996-95005-375>

Pan, S. C., & Carpenter, S. K. (2023). Prequestioning and pretesting effects: A review of empirical research, theoretical perspectives, and implications for educational practice. *Educ. Psychol. Rev.*, *35*(4), 97. <https://doi.org/10.1007/s10648-023-09814-5>

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

Powell, S. L. (2022). A comparative analysis of math facts fluency gains made through massed and distributed practice with varied inter-session intervals. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *28150257*(3-B), No–Specified. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc18&NEWS=N&AN=2021-94593-063>

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prev. Sci.*, *23*(3), 425–438. <https://doi.org/10.1007/s11121-021-01246-3>

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research & Applications*, *4*(1), 11–18. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1986-07610-001>

Reed, B. W. (1983). *Incremental, continuous-review versus conventional teaching of algebra*. search.proquest.com. <https://search.proquest.com/openview/b35a0125a60bc0d0ad5733df1e9dae4b/1?pq-origsite=gscholar&cbl=18750&diss=y>

Reed, H. B. (1924). Distributed practice in addition. *J. Educ. Psychol.*, *15*(4), 248–249. <https://doi.org/10.1037/h0070683>

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>

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020). The scarcity of interleaved practice in mathematics textbooks. *Educ. Psychol. Rev.*, *32*(3), 873–883. <https://doi.org/10.1007/s10648-020-09516-2>

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Curr. Dir. Psychol. Sci.*, *16*(4), 183–186. <https://doi.org/10.1111/j.1467-8721.2007.00500.x>

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Appl. Cogn. Psychol.*, *20*(9), 1209–1224. <https://doi.org/10.1002/acp.1266>

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>

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

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110–114. <https://doi.org/10.2307/3002019>

Schutte, G. M., Duhon, G. J., Solomon, B. G., Poncy, B. C., Moore, K., & Story, B. (2015). A comparative analysis of massed vs. Distributed practice on basic math fact fluency growth rates. *J. Sch. Psychol.*, *53*(2), 149–159. <https://doi.org/10.1016/j.jsp.2014.12.003>

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(2), 179–187. <https://doi.org/10.1037/0278-7393.5.2.179>

Steel, P., Beugelsdijk, S., & Aguinis, H. (2021). The anatomy of an award-winning meta-analysis: Recommendations for authors, reviewers, and readers of meta-analytic reviews. *Journal of International Business Studies*, *52*(1), 23–44. <https://doi.org/10.1057/s41267-020-00385-z>

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction*, *16*(2), 165–169. <https://doi.org/10.1016/j.learninstruc.2006.02.005>

Thios, S. J., & D’Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning & Verbal Behavior*, *15*(5), 529–536. <https://doi.org/10.1016/0022-5371(76)90047-5>

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychol. Methods*, *20*(3), 375–393. <https://doi.org/10.1037/met0000011>

Toppino, T. C., & Gerbier, E. (2014). *About practice: Repetition, spacing, and abstraction* (Vol. 60, pp. 113–189). Elsevier. <https://www.sciencedirect.com/science/article/pii/B9780128000908000044>

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>

Viechtbauer, W. (2017). *[R-meta] confidence interval for i^2 in multilevel model*. <https://stat.ethz.ch/pipermail/r-sig-meta-analysis/2017-August/000150.html>

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. <https://doi.org/10.1002/jrsm.11>

Weaver, J. R. (1976). The relative effects of massed versus distributed practice upon the learning and retention of eighth grade mathematics. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *7624394*(5-A), 2698. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1978-10974-001>

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

Yazdani, M. A., & Zebrowski, E. (2006). *Spaced reinforcement: An effective approach to enhance the achievement in plane geometry*. <http://ww.w.msme.us/2006-1-8.pdf>

Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *J. Educ. Psychol.*, *111*(1), 73–90. <https://doi.org/10.1037/edu0000268>

## Appendix A

## PICOS table

Table 2

PICOS Table

| PICOS.Term | Inclusion | Exclusion |
| --- | --- | --- |
| Population | Any age range or education level. | Clinical populations (classroom studies with particular participants with additional needs will be included, but not exclusively clinical samples) |
|  | Lab or classroom based. |  |
| Intervention | An intervention that bases its design on the use of the spacing effect, retrieval effect or a combination of the two and uses mathematics material. | Studies which use multiple subjects (for example Science questions and mathematics questions) where the results of the mathematics section is not reported in sufficient detail to calculate an effect size. |
| Comparators | Massed versus spaced practice | No comparator or control condition |
|  | Greater versus lesser spacing |  |
|  | Retrieval versus restudying |  |
| Outcome | Performance on a post-test |  |
| Study design | Includes a post-test after the condition or control intervention | Non-experimental designs |
|  |  | No comparator or control condition |
|  | Is the data already reported in another paper included in the review? |  |

*Note*. Desctiption of inclusion and exclusion criteria following the Population, Intervention, Comparison, Outcomes and Study (PICOS) framework