Educational Psychology Review (2021) 33:959–987
https://doi.org/10.1007/s10648-020-09572-8

META-ANALYSIS

A Meta-Analytic Review of the Benefit of Spacing out Retrieval Practice Episodes on Retention

Alice Latimier[1] · Hugo Peyre[1,2,3] · Franck Ramus[1]

Accepted: 6 September 2020 / Published online: 7 October 2020

Abstract

Spaced retrieval practice consists of repetitions of the same retrieval event distributed through time. This learning strategy combines two "desirable difficulties": retrieval practice and spacing effects. We carried out meta-analyses on 29 studies investigating the benefit of spacing out retrieval practice episodes on final retention. The total dataset was divided into two subsets to investigate two main questions: (1) Does spaced retrieval practice induce better memory retention than massed retrieval practice? (subset 1); (2) Is the expanding spacing schedule superior to the uniform spacing schedule when learning with retrieval practice? (subset 2). Using meta-regression with robust variance estimation, 39 effect sizes were aggregated in subset 1 and 54 in subset 2. Results from subset 1 indicated a strong benefit of spaced retrieval practice in comparison with massed retrieval practice ($g = 0.74$). Results from subset 2 indicated no significant difference between expanding and uniform spacing schedules of retrieval practice ($g = 0.034$). Moderator analyses on this subset showed that the number of exposures of an item during retrieval practice explains inconsistencies between studies: the more learners are tested, the more beneficial the expanding schedule is compared with the uniform one. Overall, these results support the advantage of spacing out the retrieval practice episodes on the same content, but do not support the widely held belief that inter-retrieval intervals should be progressively increased until a retention test.

Keywords Retrieval practice · Spacing schedule · Learning · Memory

✉ Alice Latimier
alice.latimier@yahoo.fr

[1] Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Etudes Cognitives, ENS, EHESS, CNRS, PSL University, Paris, France

[2] Université Paris Diderot, Sorbonne Paris Cité, UMRS 1141, Paris, France

[3] Department of Child and Adolescent Psychiatry, Robert Debré Hospital, APHP, Paris, France

Introduction

Research on learning practices typically consists in studying the effects of a study phase on performance in a test phase. During the study phase, participants are exposed to learning contents, and review it under various modalities and on various time scales. Then, in a test phase occurring after a certain retention interval, they are tested for the retention of the initial learning contents. Such research has led to the demonstration of at least two major results: the benefits of retrieval practice and spacing learning (Bjork & Bjork, 2011; Brown et al., 2014).

The Benefits of Using Retrieval Practice

Numerous studies have shown that retrieval practice enhances long-term retention, compared with re-reading or re-exposure to the material (e.g., Roediger & Butler, 2011; Roediger & Karpicke, 2006). Retrieval practice refers to any activity that requires the learner to retrieve previously learnt information from memory. This may include free or cued recall, multiple choice questions, or application exercises.

Two recent meta-analyses have summarized the benefits of retrieval practice. They have shown a strong and positive mean effect of using retrieval practice during learning compared with re-studying. Rowland (2014) found a mean effect size of $g = 0.50$ [0.42, 0.58] from 159 effect sizes comparing retrieval practice with reading; Adesope et al. (2017) found a mean effect size of $g = 0.61$ [0.58, 0.65] in a comparison of retrieval practice with all other practices (restudying/re-reading, filler, no activity, or a combination). The retrieval practice effect is well established for both simple and complex materials (i.e., single word lists and prose passages; for a review see Karpicke & Aue, 2015), and in laboratory as well as in classroom settings (Bangert-Drowns et al., 1991; Karpicke & Grimaldi, 2012; McDaniel et al., 2011). Moreover, this learning strategy seems to provide knowledge transfer to untested but related information under certain circumstances (Chan et al., 2006; McDaniel et al., 2013a; Pan & Rickard, 2018).

The meta-analyses cited above have investigated the influence of several moderators on the magnitude of the retrieval practice effect. Rowland (2014) reported that it was stronger when the learning contents were more complex, when the type of retrieval practice was more effortful, and when feedback was given during practice. Adesope et al.'s main results (2017) suggested that the population of secondary students benefited more from retrieval practice than younger and older student populations, and that classroom experiments showed similar benefits as laboratory ones. They also reported that a mixture of different types of training tests (i.e., multiple choices + short answers or free recall) yielded the strongest retrieval practice effect. Finally, the two quantitative reviews agreed that one retrieval event is enough to elicit better retention than no testing at all, and that the benefit of using retrieval seems to be stronger for certain retention intervals before a final assessment (i.e., between 1 day and 1 month).

The Benefits of Using Distributed Learning

Distributed learning or spacing refers to the deliberate insertion of lags between learning episodes on the same content. Inserting time intervals between study episodes promotes better retention than massed practice (i.e., study episodes occurring in one single session without inter-study intervals) and optimally counteract forgetting by improving the consolidation of

newly learned information (Cepeda et al., 2009; McDaniel et al., 2013b). This benefit has been demonstrated in experiments where learners typically had to restudy or retrieve the same information repeatedly according to a massed or a spaced schedule; then participants' memory was assessed with one or more final tests (Bahrick, 1979; see for a recent review Wiseheart et al., 2019, pp. 550–583). The lag between two repetitions can be defined either in terms of intervening items or in terms of time between two study episodes for a given item ("item-based" versus "time-based" spacing). Only two meta-analyses of the spacing effect yielded estimates of effect sizes. The first one focused on motor learning (Lee & Genovese, 1989). The authors found that spacing the learning trials in time improved both the acquisition of new skills during the learning phase and their retention at the final assessment (an effect size of $d =$ 0.96 was computed; Hattie, 2008). The second one concluded that spaced practice conditions led to increased performance ($d = 0.46$ overall) relative to massed practice conditions in various learning contexts (e.g., discrete motor skills, Stroop task, video games, or music memorization; Donovan & Radosevich, 1999). These two meta-analyses encompassed a wide array of learning tasks, with a predominance of motor/performance tasks, and relatively few verbal learning tasks. Since then, a large literature on retrieval practice of verbal/educational content has emerged. Hattie (2008) conducted a synthesis of these meta-analyses and reported a large mean effect size of $d = 0.71$ for the spacing effect. An influential meta-analysis for verbal learning by Cepeda et al. (2006) focused more on spaced re-studying than on retrieval practice, and their main interest was the optimal interval between a first study event and a second one. Moreover, these authors did not estimate the mean effect size of the spacing effect itself. Recently, Wiseheart et al. (2019) gathered the meta-analytic data from Cepeda et al. (2006) and Moss (1996) and found an effect size of $d = 0.85$ for the benefit of spaced learning with verbal contents.

As hinted above, the spacing effect has been shown in different domains and population (Cecilio-Fernandes et al., 2018; Dail & Christina, 2004; Mumford et al., 1994). In verbal learning contexts, spacing has been shown to enhance the retention of both simple (e.g., word pairs) and more complex materials such as abstract science concepts (Gluckman et al., 2014; Vlach & Sandhofer, 2012). Spacing effects also occur across age groups, from children to healthy aging and individuals with memory impairments (Balota et al., 2006; Fritz et al., 2007; Kalenberg, 2017). Moreover, spaced learning also provides substantial improvements in long-term memory for learners in real educational settings (Carpenter et al., 2012; Karpicke et al., 2016; Larsen, 2018; Mettler et al., 2016; Seabrook et al., 2005; Sobel et al., 2011). Finally, spacing effects have been shown from very short inter-study intervals such as a few seconds to much longer intervals such as days or weeks (e.g., Dobson et al., 2016; Whitten & Bjork, 1977).

Donovan and Radosevich (1999) suggested that increasing the lag between learning episodes produced a greater spacing effect on both free recall and cued recall tasks. Later, Janiszewski et al. (2003) found that longer inter-study intervals between study episodes led to larger spacing effect sizes. Interestingly, Donovan and Radosevich (1999) and Cepeda et al. (2006) found that the spacing effect was not modulated by retention interval. Indeed, it seems that spaced conditions always lead to better retention than massed conditions, regardless of the retention interval between the learning and the final test phases (Benjamin & Tullis, 2010; Emilie Gerbier et al., 2015; Godbole et al., 2014; Kapler et al., 2015). However, recent work has shown that spaced learning might produce better retention on delayed assessments than massed practice does, while this latter practice might promote better retention on immediate assessments (Greving & Richter, 2018; Roediger III & Karpicke, 2011; Rohrer & Taylor, 2006). Overall, there is a relationship between the optimal spacing interval and the retention

interval. Specifically, the longer the retention interval, the longer the spacing interval associated with optimal performance on a final test (Cepeda et al., 2009).

Effects of the Type of Spacing Schedule

Relative spacing refers to how the repeated episodes are spaced relative to one another, i.e., how spacing is scheduled in time (Wiseheart et al., 2019, p 555). Two schedules of review have frequently been compared: the expanding spacing schedule versus the uniform spacing schedule (also called "equal" or "fixed" schedule). In the uniform spacing schedule, spacing intervals are kept constant throughout the study phase while in the expanding spacing schedule, spacing intervals increase after every re-exposure to an item. Similarly, one may also define a contracting schedule, whereby spacing intervals decrease with every repetition. Nevertheless, the contracting schedule has been investigated in relatively few studies and has been abandoned in more recent experiments because this schedule seems to be the least favorable to promote long-term retention, although it has shown benefits on short-term retention (Küpper-Tetzel et al., 2014; Mozer et al., 2009; Tsai, 1927).

Comparisons between expanding and uniform spacing schedules have been conducted in the context of repeated readings or presentations of the same material. Overall, results do not seem very consistent: the expanding schedule led to better final performance than the other two schedules in some cases (Gerbier & Koenig, 2012, experiment 1; Toppino et al., 2018), but not in others (Gerbier & Koenig, 2012, experiment 2; Gerbier et al., 2015). Results might depend on the retention interval. For example, using a categorization learning task with 3-year-old children, Vlach & Sandhofer (2012) found a superiority of the expanding schedule spaced presentation over the uniform one only at delayed but not at immediate final tests (i.e., expanding and uniform schedules were equivalent in term of final performances). Cepeda et al. (2006) meta-analyzed the comparison between expanding and uniform schedules (22 comparisons of retention performances and 8 effect size comparisons). They found that expanding intervals led to better performance than uniform ones. However, this result is difficult to interpret because large standard errors indicated a large between-study variability.

These two schedules have also been compared in the context of learning with retrieval practice, with diverse results. Several laboratory experiments did find a superiority of the expanding schedule (Cull et al., 1996; Kang et al., 2014; Maddox et al., 2011; Storm et al., 2010), but other studies found the opposite result (Cull, 2000; Karpicke & Roediger, 2007; Logan & Balota, 2008; Toppino et al., 2018 in the high-level initial training condition).

In addition, a non-negligible part of the literature reports no difference between the two schedules (Carpenter & DeLosh, 2005; Cull, 2000; Cull et al., 1996; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2010; Logan & Balota, 2008; Pyc & Rawson, 2007; Storm et al., 2010; Terenyi et al., 2018).

Thus, the remote date of the last meta-analysis and the many recent conflicting findings concur to suggest that a new meta-analysis of spacing schedules of retrieval practice is necessary.

The Present Study

Given their consistent benefits for long-term retention, retrieval practice and spaced learning are more and more cited as efficient learning strategies to apply in the classroom

(Brown et al., 2014; Dunlosky et al., 2013; Moreira et al., 2019; Rosenshine, 2010; Weinstein et al., 2018). We identified two gaps in the preceding literature review that we aimed to fill in. First, a non-negligible literature focused on the benefits of spacing out the learning sessions in the context of learning with repeated retrieval practice. Thus, the present meta-analytic review focused on the value of spaced retrieval practice relative to massed retrieval practice (subset 1). Second, and related to the same literature, we aimed at estimating the relative benefits of expanding versus uniform spacing schedules of repeated retrieval practice (subset 2). In order to be maximally relevant for educational research, we focus on semantic and verbal stimuli learning (including mathematics problems). Thus, studies on perceptual and motor learning were excluded. Depending on the degree of heterogeneity, we investigated possible moderator effects to measure whether various features of spaced retrieval practice have an effect on final performance, and to understand to what extent effect sizes are moderated by contextual and methodological features of the research.

Methods

Search Strategy and Inclusion Criteria

Electronic searches of scientific publication databases (ERIC, Web of Science, Scopus) were conducted using combinations of the following terms: spaced, distributed, retrieval, practice, testing ((spaced OR distributed OR retrieval) AND (practice OR testing)). Citation searches were also performed for existing review articles (Balota et al., 2007; Roediger III & Karpicke, 2011) to identify additional studies not captured by database searches as well as a regular update of the literature on spaced retrieval practice using Google Scholar alerts.

Thus, we tried to include as many references as possible by conducting the literature search until September 2017 to answer our two research questions (Kalenberg, 2017 was the last included reference in the final analyses).

Following these strategies, our searches generated a total of 3948 results. After removing duplicates, we applied a first screening based on titles and abstracts to select eligible studies. We eliminated off-topic references (i.e., those that did not investigate spaced or retrieval practice), those with patient populations only, literature reviews, and studies investigating perceptual and motor learning tasks only. We included laboratory as well as classroom experiments.

At the end of this initial stage, 42 articles were selected for the next screening stage (Fig. 1, *Screening*). Then, the second stage of screening was based on full-text reading, with the six following inclusion criteria for eligibility: (1) the population of interest was a neurotypical population with no age limit and no specific education level; (2) the experimental set-up included a training phase followed by an assessment phase of the memory retention; (3) the training phase included repeated retrieval practice attempts over time (at least two retrieval attempts for the same item); (4) the design included a control condition that consists in massed repeated retrieval practice. Learning is considered to be massed when the retrieval events for a given item are not separated by the retrieval events for other items AND/OR the design included an experimental manipulation of the type of spacing schedule where expanding and uniform were contrasted. Finally, (5) the participants had to be randomized into the experimental conditions and (6) all necessary information for effect size calculations must have been

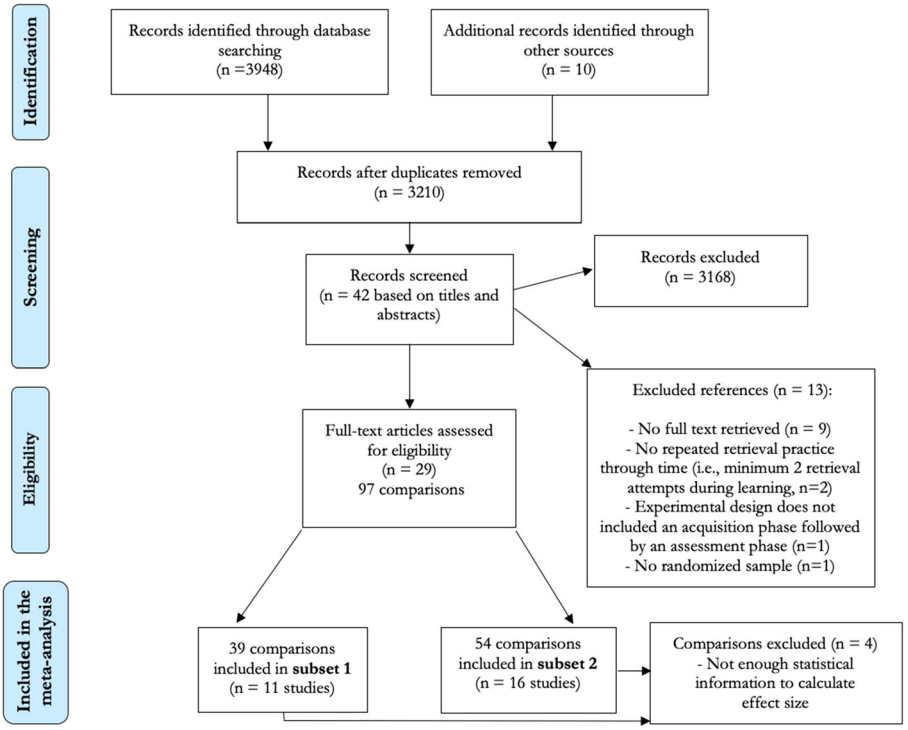

Fig. 1 PRISMA group flow diagram depicting study inclusion criteria (Moher et al., 2009). For each stage, we provide the number of included and excluded references, and the reason we excluded some of them

reported or derivable from other available data (i.e., sample sizes, means, SDs, and standard errors) either in as or in figures. When necessary and when available, graphs were digitized using the WebPlotDigitizer software[1] to recover missing relevant statistical information per condition (e.g., means, SDs and standard errors). Ultimately, 4 comparisons were excluded from final analyses because required data could not be obtained in any way (Fig. 1). We included unpublished references (i.e., dissertations) as long as they satisfied all inclusion criteria.

Thus, a total of 42 full-text articles were assessed for eligibility, 29 satisfied all criteria, including 97 comparisons in total (Fig. 1, *Eligibility*). During this stage, we gathered relevant information on each reference such as the outcome, the experimental set-up, and the main results, but we also coded all necessary information that were relevant for the moderator analyses for each comparison. The coding process was determined a priori and was completed by the first author for all studies included in the meta-analysis.

To answer our two main research questions, we separated the data into two subsets. Subset 1 included 39 comparisons between spaced and massed retrieval practice and subset 2 included 54 comparisons between an expanding and a uniform spacing schedule (Fig. 1, *Included in the meta-analysis*). Several comparisons could derive from the same study. This was the case for

[1] https://automeris.io/WebPlotDigitizer/.

the majority of the included studies in which multiple experiments were described. Moreover, one reference could be included in both subsets if the two main research questions were raised in the different experiments of this given reference.

Candidate Effect Size Moderators

When performing the full-text screening, all information relevant for the moderator analysis were extracted. These moderators were defined a priori apart from the moderator "Time of the first retrieval event" that was defined after reading the full text (subset 2). The categories for each of the categorical moderators were created a posteriori, based on observed distributions. We thus identified 11 different moderators.

Setting Setting type was coded as a categorical variable with two levels: laboratory versus classroom.

Education Level This moderator was coded as a categorical variable with two levels: less than 12 years (including preschool and elementary school) versus 12 years and more (including high school and undergraduates).

Type of Material (Stimuli) Due to the low number of comparisons available in any given subcategory, stimulus type was coded as a categorical variable with two levels: pairs (whatever the type—face–name pairs, translated word pairs…) versus others (including prose passages, word lists, classroom lectures, and maths problems).

Design Study design was coded as a categorical variable with two levels: between and within participant, referring to the spaced versus massed manipulation or expanding versus uniform schedule manipulation.

Test Type Used for the Training Phase The format of the retrieval practice used to learn was coded as binary categorical variable. We included cued recall, multiple choice tests, quizzes, and fill in the blanks in the first category (cued recall), and free recall in the second one (including short answers).

Final Test Type The same categories for the training test type were used to code this variable: cued recall (including fill in the blanks, multiple choices test, and quizzes) and free recall (including short answers).

Feedback The presence of feedback (corrective as well as elaborative) after the retrieval events (i.e., after each item or at the end of the training phase) was coded as a categorical variable with two levels: yes and no.

Retention Interval The duration between the end of the training phase (the last retrieval event) and the beginning of the final memory assessments was coded a continuous moderator in minutes and was then subjected to a logarithmic transformation.

Total Number of Exposures for a Given Item After a first exposure phase (initial study phase), several retrievals and restudy events were repeated over time for each item included in

the material. The total number of exposures for a given item was defined as the first presentation to the item and the number of retrieval events. It was first coded as a continuous variable, but we had to convert it into a categorical variable because the total number did not vary much from one study to another. For subset 1, we coded the variable as between two and four exposures versus more than four exposures. For subset 2, we coded the variable as four exposures versus more than four exposures.

Type of Spacing Schedule (Specific to Subset 1) In relation to the first research question, another moderator was coded: the type of spacing schedule compared with the massed schedule. It was coded as a categorical variable: expanding versus uniform.

Timing of the First Retrieval Event (Specific to Subset 2) Depending on the study, the first retrieval attempt occurred either immediately after the content exposure or after a given delay (e.g., number of intervening items). The timing of this retrieval attempt could be either the same whatever the spacing schedule or different (in most of the studies, the first retrieval attempt in the expanding schedule was immediate). We coded this moderator as a categorical variable: same versus different timing.

Analyses

Effect Size Calculations

In the present meta-analysis, each effect size indicates the standardized difference in performance in the final assessment between spaced and massed retrieval conditions for subset 1, and between expanding and uniform schedule conditions for subset 2. When effect sizes were not directly provided in the results section of the studies, we used available data to calculate each effect size as well as the standard error of the effect size following these formulas:

1 When only standard error *se* was available, SD *s* was calculated as

$$s = se^{*}\sqrt{n}$$

2 Cohen's *d* was computed as

$$d = \frac{M1-M2}{S}$$

where the pooled SD for a within-subject design was

$$S = \sqrt{\frac{s1^2 + s2^2}{2}}$$

and the pooled SD for a between-subject design was

$$S = \sqrt{\frac{(n1-1)(s1)^2 + (n2-1)(s2)^2}{n1 + n2-2}}$$

3. The standard error of the effect size for a within-subject design was computed with

$$d.se = \sqrt{\frac{\left(\frac{2(1-r)}{n}\right) + d^2}{2n}}$$

and the standard error of the effect size for a between-subject design was computed with

$$d.se = \sqrt{\frac{n1 + n2}{n1 * n2} + \frac{d^2}{2(n1 + n2)}}$$

M is the mean proportion correct for a given condition, *n* is the sample size for a condition, *s* is the SD for a condition, and *se* is the standard error for a condition. Moreover, *r* is the within-subject correlation between condition 1 scores (spaced retrieval or expanding schedule depending on the subset) and condition 2 scores (massed retrieval or uniform schedule depending on the subset). This correlation is rarely reported in most studies. Thus, a correlation of 0.5 was assumed for studies using a within-participant design as Rowland (2014) did in his own meta-analysis.

For small samples, Cohen's *d* might produce on overestimate of true effect size. Thus, we calculated Hedges' *g* for each of the included effect sizes in order to correct for this bias, following this formula (Hedges & Olkin, 1985):

$$g = d\left(1 - \frac{3}{4N - 9}\right.$$

where *N* is the total number of participants for within as well as between-subject designs.

Computation of Weighted Mean Effect Sizes

The R software (package *robumeta*; Fisher & Tipton, 2015) was used to conduct the meta-analysis. Each subset of comparisons was analyzed separately. We reported all analyses using effect sizes as measured by Hedges' *g*. The method we used to synthesize effect sizes was highly similar to the method used by Klingbeil et al. (2017) in their own meta-analysis using robust variance estimation (RVE). Effect size estimates were synthesized using RVE methods to address the problem that most studies contributed multiple and non-independent effect size estimates (Hedges et al., 2010). It is not generally reasonable to assume that effect size estimates based on a common sample are independent, which precludes the use of standard random effects models for meta-analysis. RVE method has several advantages to overcome this issue of non-independent effect sizes and gives a more accurate estimate of the standard errors of the effects of interest thus leading to smaller confidence intervals of the weighted mean effect sizes (Hedges et al., 2010). This approach also requires the specification of the correlation between within-study effects (Tipton & Pustejovsky, 2015). By default, the package *robumeta* set the correlation between effect sizes at $\rho = 0.80$; thus, we used this value. As the value of ρ might affect the value of the mean effect size and of the estimated between-study heterogeneity T^2, we conducted sensitivity analysis for each subset (Tanner-Smith & Tipton, 2014). This consists in varying the assumed within-study effect size correlation using values between $\rho = 0.0$ and $\rho = 1$.

For each subset of studies, analyses reported the weighted mean effect size and 95% confidence interval and the estimated between-study standard error (SE). In addition, we reported the estimated between-study heterogeneity T^2 that provides an estimate of the variance in the true effect sizes (Borenstein et al., 2009), and the magnitude of the heterogeneity I^2 (in %) among studies (as for the random effects model; Higgins et al., 2003). Given the relatively small number of included samples, small-sample adjustments for hypothesis tests and confidence intervals (CIs; Tipton & Pustejovsky, 2015) were used for our analyses. Due to the small number of comparisons yielding relatively few degrees of freedom, each moderator was analyzed separately. Finally, we reported publication bias analysis on the basis of using a funnel plot inspection and using Egger's regression test (Egger et al., 1997).

Results[2]

General Study Characteristics

Subset 1: Spaced versus Massed Repeated Retrieval Practice

We identified $n = 11$ studies involving $k = 39$ effect sizes for this comparison (Table 1). The great majority of the studies were in laboratory settings ($n = 9$), whereas a few were in classroom settings ($n = 4$). Studies mostly included a population of undergraduate students (more than 12 years of studies). About half of the studies involved learning of pairs ($n = 6$), while the other half used other types of materials (text passages, exercises, lists) ($n = 5$). Within-subject designs ($n = 7$) were used more frequently than between-subject designs ($n = 4$), and cued-recall tests ($n = 8$) were used more often than free recall tests ($n = 3$) during the training phase as well as for the final assessments (making the two moderators quite redundant). Half the comparisons measured final performance immediately after the end of the learning phase ($k = 23$), while the other half used longer retention intervals ($k = 15$, from 1 day to more than 1 month). Apart from Hopkins' study (2016), the studies included more than two exposures to a given item (4 exposures being the majority with 22 comparisons). Feedback was usually given during the training phase ($k = 20$), and two studies used the presence of feedback as an independent variable (Balota et al., 2006; Karpicke & Roediger, 2007). Finally, the spaced retrieval practice condition was most often implemented using a uniform schedule ($k = 23$), but also using an expanding or a contracting schedule.

Thus, the typical study comparing spaced versus massed retrieval practice is a laboratory study on adults, using a cued-recall task. Retention is assessed with the same test type, and the spaced retrieval condition is scheduled according to a uniform distribution with four retrieval attempts for a given item.

Subset 2: Expanding versus Uniform Spacing Schedule

We identified 16 studies involving 54 effect sizes for this comparison (Table 2). The great majority of the studies were in laboratory settings ($n = 13$), whereas a few were classroom settings ($n = 3$). Only one study included a child population, while the other studies included

[2] All data from included studies (subsets 1 and 2) and R script for analyses with *robumeta* package are available on OSF with the following link: https://osf.io/jbmq4/?view_only=a7678c23980c4914baefe2466c026632.

Table 1 Description and moderator information for studies included in subset 1 comparing spaced and massed retrieval practice

Subset 1 study	Effect size	*N*	Setting	Educational level	Design	Stimuli material	Training test format	Feedback	Final test format	Retention interval (min)	Number of exposures	Spacing schedule
Fishman et al. (1968)												
Experiment 1	0.44	29	L	5	WS	O	FR	Yes	FR	21,600	3	Un
Grote (1995)												
Experiment 1	0.56	36	L	11.5	WS	O	CR	Yes	CR	NR	3	NR
Caple (1996)												
Experiment 1	1.90	36	L	12+	BS	O	CR	Yes	CR	10,080	UC	NR
Carpenter and DeLosh (2005)												
Experiment 2	0.97	65	L	12+	WS	P	CR	No	CR	5	4	Exp
Experiment 2	1.27	65	L	12+	WS	P	CR	No	CR	5	4	Un
Experiment 3	1.22	69	L	12+	WS	P	CR	No	CR	5	4	Exp
Experiment 3	1.20	69	L	12+	WS	P	CR	No	CR	5	4	Un
Balota al. (2006)												
Experiment 1	0.50	29	L	12+	WS	P	CR	No	CR	5	6	Exp
Experiment 1	0.36	29	L	12+	WS	P	CR	No	CR	5	6	Un
Experiment 1	0.63	31	L	12+	WS	P	CR	No	CR	5	6	Exp
Experiment 1	0.63	31	L	12+	WS	P	CR	No	CR	5	6	Un
Experiment 2	1.04	37	L	12+	WS	P	CR	Yes	CR	5	6	Exp
Experiment 2	1.25	37	L	12+	WS	P	CR	Yes	CR	5	6	Un
Experiment 2	0.80	38	L	12+	WS	P	CR	Yes	CR	5	6	Exp
Experiment 2	0.72	38	L	12+	WS	P	CR	Yes	CR	5	6	Un
Karpicke and Roediger (2007)												
Experiment 1	0.86	24	L	12+	WS	P	FR	No	FR	10	4	Un
Experiment 1	0.46	24	L	12+	WS	P	FR	No	FR	10	4	Exp
Experiment 1	0.57	24	L	12+	WS	P	FR	No	FR	2880	4	Exp
Experiment 1	1.09	24	L	12+	WS	P	FR	No	FR	2880	4	Un
Experiment 2	2.13	24	L	12+	WS	P	FR	Yes	FR	10	4	Exp
Experiment 2	1.82	24	L	12+	WS	P	FR	Yes	FR	10	4	Un
Experiment 2	1.07	24	L	12+	WS	P	FR	Yes	FR	2880	4	Exp
Experiment 2	1.62	24	L	12+	WS	P	FR	Yes	FR	2880	4	Un
Fritz et al. (2007)												
Experiment 1	2.16	40	L	Preschool	BS	P	CR	Yes	CR	0	NR	Exp
Logan and Balota (2008)												

Table 1 (continued)

Subset 1 study	Effect size	*N*	Setting	Educational level	Design	Stimuli material	Training test format	Feedback	Final test format	Retention interval (min)	Number of exposures	Spacing schedule
Experiment 1	0.76	80	L	12+	WS	P	CR	NR	CR	0	4	Un
Experiment 1	1.78	24	L	12+	WS	P	CR	NR	CR	1440	4	Un
Experiment 1	1.09	66	L	12+	WS	P	CR	NR	CR	0	4	Un
Experiment 1	1.19	24	L	12+	WS	P	CR	NR	CR	1440	4	Un
Nakata (2015)												
Experiment 1a (short spacing)	1.06	64	L	12+	BS	P	CR	Yes	CR	0	4	Un
Experiment 1a	0.39	64	L	12+	BS	P	CR	Yes	CR	1440	4	Un
Experiment 1b (medium spacing)	1.38	64	L	12+	BS	P	CR	Yes	CR	0	4	Un
Experiment 1b	0.80	64	L	12+	BS	P	CR	Yes	CR	1440	4	Un
Experiment 1c (long spacing)	0.87	64	L	12+	BS	P	CR	Yes	CR	0	4	Un
Experiment 1c	0.84	64	L	12+	BS	P	CR	Yes	CR	1440	4	Un
Hopkins (2016)												
Experiment 1	0.30	40	C	12+	WS	O	CR	Yes	CR	57,148	2	Un
Experiment 1	2.10	86	C	12+	BS	O	CR	Yes	CR	57,148	2	Un
Dobson (2016)												
Experiment 1	− 0.34	60	C	12+	WS	O	FR	No	FR	0	6	Con
Experiment 1	0.94	60	C	12+	WS	O	FR	No	FR	1440	6	Con
Experiment 1	0.74	60	C	12+	WS	O	FR	No	FR	40,320	6	Con

Note. Effect sizes are indicated in Hedges' *g*. Retention interval durations are indicated in minutes. Educational level corresponds to the number of schooling years according to the USA level grade system

L laboratory, *C* classroom, *BS/WS* between or within subject, *P* pairs, *NR* not reported, *O* other materials, *CR* cued recall, *FR* free recall, *UC* until correct, *Un* uniform, *Exp* expanding, *Con* contracting

Table 2 Description and moderator information for studies included in subset 2 comparing expanding to uniform spacing schedule

Subset 2 study	Effect size	*N*	Setting	Educational level	Design	Stimuli material	Training test format	Feedback	Final test format	Retention interval (min)	Number of exposures	First retrieval attempt
Carpenter and DeLosh (2005)												
Experiment 2	−0.18	65	L	12+	WS	P	CR	No	CR	5	4	DT
Experiment 3	0.05	69	L	12+	WS	P	CR	No	CR	5	4	DT
Balota et al. (2006)												
Experiment 1	0.17	39	L	12+	WS	P	CR	No	CR	5	6	ST
Experiment 1	−0.03	31	L	12+	WS	P	CR	No	CR	5	6	ST
Experiment 2	−0.18	37	L	12+	WS	P	CR	No	CR	5	6	ST
Experiment 2	0.13	38	L	12+	WS	P	CR	No	CR	5	6	ST
Experiment 3	−0.14	10	L	12+	WS	P	CR	Yes	CR	5	4	DT
Experiment 3	−0.24	15	L	12+	WS	P	CR	Yes	CR	5	4	DT
Karpicke and Roediger (2007)												
Experiment 1	0.29	24	L	12+	WS	P	FR	No	FR	10	4	DT
Experiment 1	−0.47	24	L	12+	WS	P	FR	No	FR	2880	4	DT
Experiment 2	0.23	24	L	12+	WS	P	FR	Yes	FR	10	4	DT
Experiment 2	−0.39	24	L	12+	WS	P	FR	Yes	FR	2880	4	DT
Experiment 3	−0.15	28	L	12+	WS	P	CR	No	CR	10	6	ST
Experiment 3	0.07	28	L	12+	WS	P	CR	No	CR	10	6	ST
Experiment 3	−0.06	28	L	12+	WS	P	CR	No	CR	2880	6	ST
Experiment 3	−0.04	28	L	12+	WS	P	CR	No	CR	2880	6	ST
Pyc and Rawson (2007)												
Experiment 1	0.02	64	L	12+	BS	P	CR	Yes	CR	40	4	DT
Logan and Balota (2008)												
Experiment 1	0	80	L	12+	BS	P	CR	NR	CR	0	4	DT
Experiment 1	0.32	66	L	12+	WS	P	CR	NR	CR	0	4	DT
Experiment 1	−0.36	24	L	12+	WS	P	CR	NR	CR	0	4	DT
Experiment 1	−0.27	24	L	12+	WS	P	CR	NR	CR	0	4	DT
Storm et al. (2010)												
Experiment 1	−0.18	88	C	12+	BS	TP	FR	No	FR	10,080	5	DT
Experiment 1	−0.15	88	C	12+	BS	TP	FR	No	CR	10,080	5	DT
Experiment 2	0.99	30	C	12+	BS	TP	FR	No	FR	10,080	5	DT
Experiment 2	0.67	30	C	12+	BS	TP	FR	No	CR	10,080	5	DT
Experiment 3	0.64	16	C	12+	WS	TP	CR	No	CR	10,080	4	DT

Table 2 (continued)

Subset 2 study	Effect size	*N*	Setting	Educational level	Design	Stimuli material	Training test format	Feedback	Final test format	Retention interval (min)	Number of exposures	First retrieval attempt
Experiment 3	−0.21	18	C	12+	WS	TP	CR	No	CR	10,080	4	DT
Karpicke and Roediger (2010)												
Experiment 1	0.07	40	L	12+	BS	TP	FR	No	FR	10,080	4	ST
Experiment 1	−0.25	40	L	12+	BS	TP	FR	No	FR	10,080	4	ST
Experiment 2	0.08	32	L	12+	BS	TP	FR	Yes	FR	10,080	5	ST
Experiment 2	0.12	32	L	12+	BS	TP	FR	Yes	FR	10,080	5	ST
Experiment 2	−0.38	32	L	12+	BS	TP	FR	Yes	FR	10,080	5	ST
Experiment 2	0.08	32	L	12+	BS	TP	FR	Yes	FR	10,080	5	ST
Karpicke and Bauernschmidt (2011)												
Experiment 1	0.08	48	L	12+	WS	P	CR	No	CR	10,080	4	DT
Experiment 1	0.24	48	L	12+	WS	P	CR	No	CR	10,080	4	DT
Experiment 1	−0.13	48	L	12+	WS	P	CR	No	CR	10,080	4	DT
Maddox et al. (2011)												
Experiment 1	0.15	30	L	12+	WS	P	CR	No	CR	60	5	ST
Experiment 2	0.38	42	L	12+	WS	P	CR	No	CR	60	5	DT
Experiment 2	0.41	36	L	12+	WS	P	CR	No	CR	60	5	DT
Dobson (2012)												
Experiment 1	0.61	97	L	12+	BS	TP	CR	No	CR	41,760	5	ST
Experiment 1	0.67	93	L	12+	BS	TP	CR	No	CR	41,760	5	DT
Experiment 1	0.22	103	L	12+	BS	TP	CR	No	CR	41,760	5	DT
Experiment 1	0.47	99	L	12+	BS	TP	CR	No	CR	41,760	5	DT
Experiment 1	0.76	196	L	12+	BS	TP	CR	No	CR	41,760	5	NR
Dobson (2013)												
Experiment 1	−0.13	91	C	12+	BS	O	CR	No	CR	NR	10	ST
Kang et al. (2014)												
Experiment 1	0.20	37	L	NR	WS	P	CR	Yes	CR	80,640	4	ST
Küpper-Tetzel et al. (2014)												
Experiment 1	−0.11	34	L	12+	BS	P	CR	Yes	CR	15	UC	DT
Experiment 1	0.10	34	L	12+	BS	P	CR	Yes	CR	1440	UC	DT
Experiment 1	−0.27	34	L	12+	BS	P	CR	Yes	CR	10,080	UC	DT
Experiment 1	0.07	34	L	12+	BS	P	CR	Yes	CR	50,400	UC	DT
McGregor (2014)												

Table 2 (continued)

Subset 2 study	Effect size	*N*	Setting	Educational level	Design	Stimuli material	Training test format	Feedback	Final test format	Retention interval (min)	Number of exposures	First retrieval attempt
Experiment 1 Mettler et al. (2016)	−0.52	91	L	12+	BS	P	CR	No	FR	1020	NR	ST
Experiment 1	0.25	36	L	12+	WS	P	CR	Yes	CR	0	4	DT
Experiment 1 Kalenberg (2017)	0.05	36	L	12+	WS	P	CR	Yes	CR	10,080	4	DT
Experiment 1	−0.20	32	C	3 to 4	BS	P	CR	Yes	CR	10,080	4	DT

Note. Effect sizes are indicated in Hedges' *g*. Retention interval durations are indicated in minutes. Educational level corresponds to the number of schooling years according to the USA level grade system

L laboratory, *C* classroom, *BS/WS* between or within subject, *P* pairs, *TP* text passages, *O* other materials, *CR* cued recall, *FR* free recall, *UC* until correct, *NR* not reported, *ST* same time, *DT* different time

adults (more than 12 years of education). The most common design was a within-subject design ($k = 31$), using pairs as learning material ($k = 36$). A cued-recall task was most often used during the training phase as well as the final assessments ($k = 40$), and feedback was often not given ($k = 33$) after the retrieval practice event. Studies in subset 2 used longer retention intervals than studies in subset 1 (usually 1 week or more), and the total number of exposures for a given item was four or more. For more than half the comparisons ($k = 34$), the first retrieval attempt did not occur at the same time for the two spacing schedules: it usually occurred immediately after the initial exposure in the expanding schedule, whereas it was delayed in the uniform schedule. Thus, the typical study comparing two spacing schedules is a laboratory study with adult participants who learn pairs using cued recall, following a schedule of at least four repetitions for each item, and with long-term retention assessed using the same test type at relative long delays.

Effect Size Analyses

Subset 1: spaced retrieval practice versus massed retrieval practice.

Weighted Mean Effect Size - The dataset included 39 effect size estimates from 11 unique studies, with between one and eight effect sizes per study (min = 1, median = 3, max = 8). The overall weighted mean effect size across all 39 effect size estimates was $g = 1.01$ (95% CI [0.68, 1.34], $p < 0.0001$) with an estimated between-study SE of 0.15 (Table 3). Varying the assumed within-study effect size correlation (ρ) had no impact on g, ranging from 1.009 to 1.011 (Appendix 1). There was a minimal impact on the estimated between study-variance (T^2), which ranged from 0.204 when $\rho = 0$ to 0.215 when $\rho = 1$. Heterogeneity was moderate (Higgins' $I^2 = 49.09\%$).

Publication Bias Analysis - We estimated publication bias for this subset using a funnel plot and Egger's regression test. The significant Egger's regression test ($t = 4.41$, d.f. = 37, $p < 0.0001$) confirmed that the funnel plot was asymmetrical (Fig. 2a). This makes it likely that publication bias has occurred: some studies with negative or non-significant findings were probably not published and therefore were not included in this meta-analysis. Thus, the mean effect size may be overestimated. To estimate the mean effect size by taking in account the publication bias, we used the trim-and-fill method (Table 3 and Fig. 2b). The overall weighted mean effect size was reduced to $g = 0.74$ (95% CI [0.55, 0.91], $p < 0.0001$), with a moderate heterogeneity ($I^2 = 46.62\%$) too.

Table 3 Summary of the weighted mean effect sizes for subset 1 (including the trim-and-fill correction) and subset 2

Subset (k = number of effect sizes)	g	SE	d.f.	p	95% CI
Subset 1: spaced retrieval practice vs. massed retrieval practice ($k = 39$)	1.01	0.15	9.72	< 0.01***	[0.68, 1.34]
Subset 1: trim-and-fill correction ($k = 49$)	0.74	0.09	19.4	< 0.01***	[0.55, 0.91]
Subset 2: expanding vs. uniform retrieval practice schedule ($k = 54$)	0.034	0.06	13.7	0.62	[− 0.10, 0.17]

Note. Weighted mean effect size in terms of Hedges' g; SE: between-study standard error; d.f.: adjusted degrees of freedom; CI: confidence interval. Results are not reliable when d.f. < 4. Significance codes: < 0.01*** < 0.05** < 0.10*

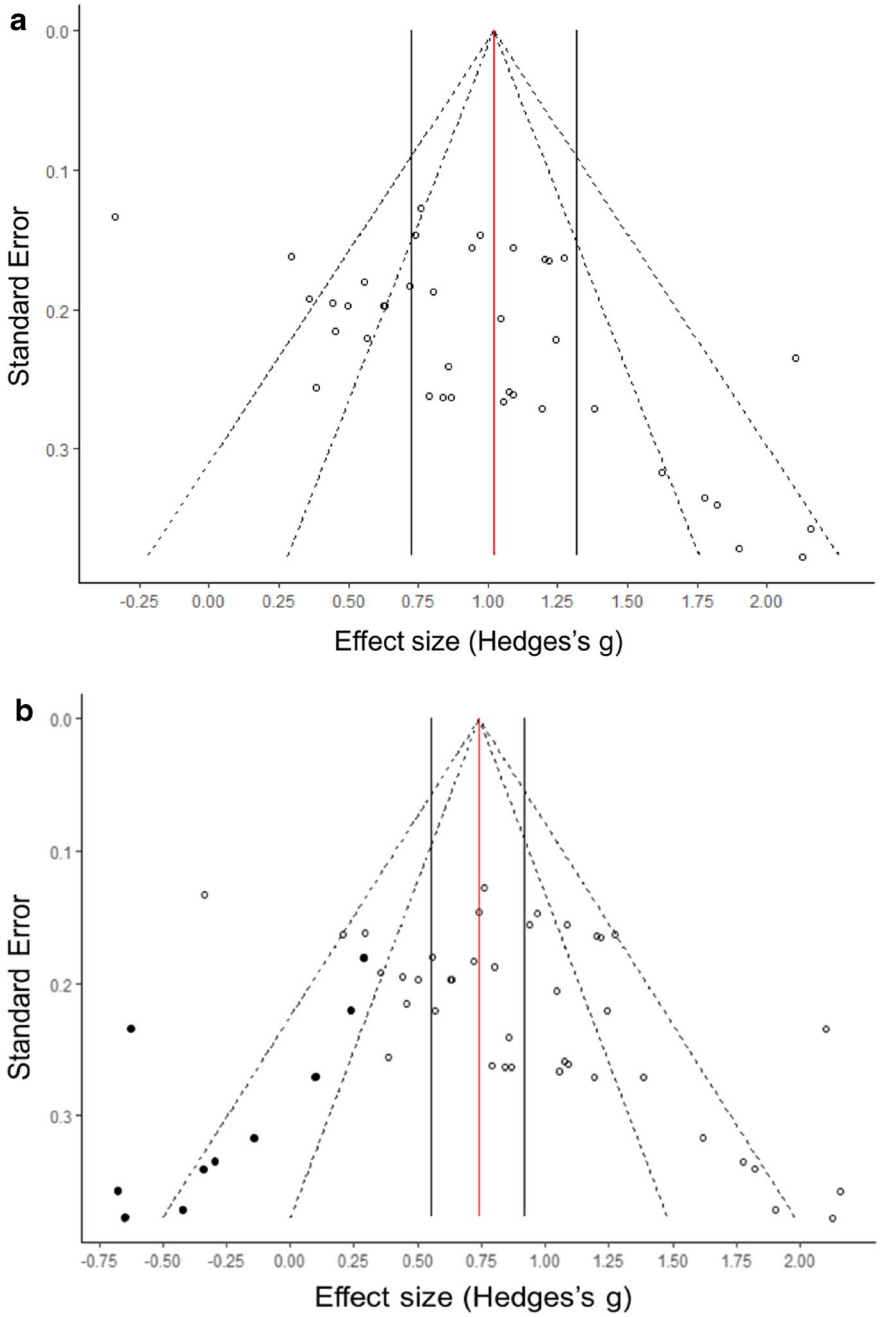

Fig. 2 Funnel plot for subset 1 (**a**) and trim-and-fill funnel plot (**b**) for the same subset. For publication bias correction, each point represents the effect size of one included comparison. For the trim-and-fill funnel plot, white dots are effect sizes from the included comparisons, while black dots are those added by the trim-and-fill procedure (10 new effect sizes). The *x*-axis represents Hedges' *g* for each comparison, and the *y*-axis is the corresponding standard error. Red solid line: mean effect size; black solid lines: CI for mean effect size; dashed lines: lower-limit and upper-limit values for the 95% CI and 99% CI regions

Subset 2: Expanding versus Uniform Retrieval Practice Schedule

Weighted Mean Effect Size The dataset included 54 effect size estimates from 16 unique studies, with between one and eight effect sizes per study (min = 1, median = 3, max = 8). The overall weighted mean effect size across all 54 effect size estimates was $g = 0.034$ (95% CI = [− 0.10, 0.17], $p = 0.59$) with an estimated between-study SE of 0.063 (Table 3). Varying the assumed correlation between within-study conditions had no impact at all on the mean effect size and no impact on the estimated between study variance (T^2) too (Appendix 1). Higgins test suggested no heterogeneity ($I^2 = 0\%$).

Publication Bias Analysis We estimated publication bias for this subset with a funnel plot representation (Fig. 3). Egger's regression test was not significant, suggesting that the funnel plot was symmetrical and therefore that there was no publication bias ($t = -0.44$, d.f. = 52, $p = 0.66$).

Potential Moderator Effects

We carried out moderator analyses for the two subsets (Table 4). For two of the moderators (settings and educational level), one category was largely predominant (laboratory settings, and more than 12 years of education), so we did not include it in the moderator analyses. In

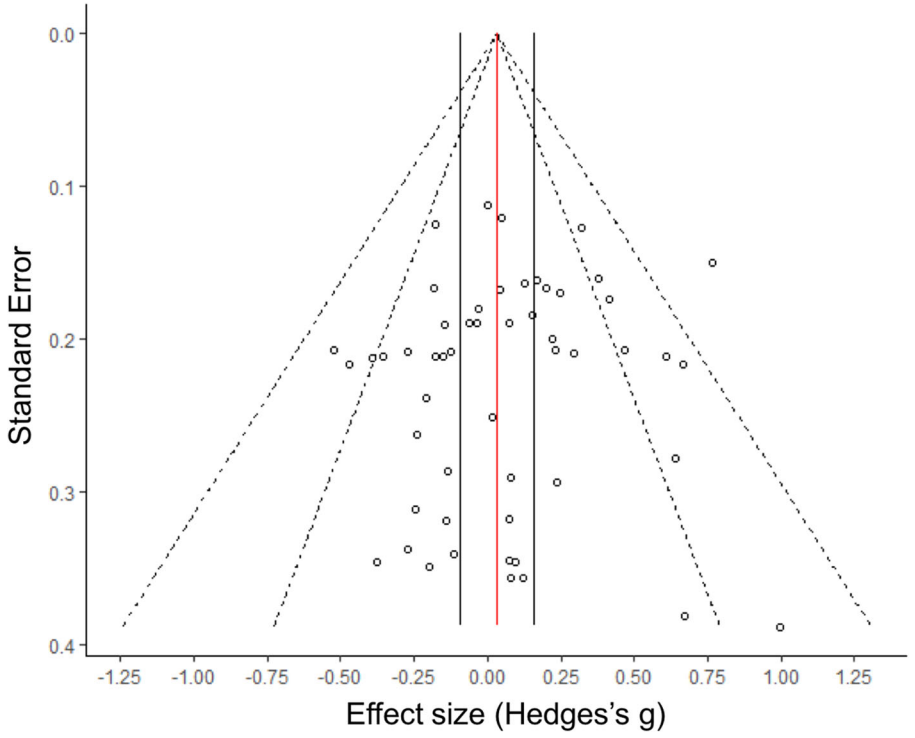

Fig. 3 Funnel plots for subset 2. Each point represents the effect size of one included comparison. The *x*-axis represents Hedges' *g* for each comparison, and the *y*-axis is the corresponding standard error. Red solid line: mean effect size; black solid lines: CI for mean effect size; dashed lines: lower-limit and upper-limit values for the 95% CI and 99% CI regions

Table 4 Moderator analyses for each subset of comparisons

Moderators	*k*	*β*	SE	95% CI		d.f.	*p*
				LL	UL		
Subset 1							
Retention interval (log minutes)	36	0.02	0.06	−0.12	0.16	5.33	0.78
Design (between vs. **within subjects**)	39	0.36	0.49	−1.54	2.27	2.26	0.53
Feedback (no vs. **yes**)	35	0.37	0.23	−0.20	0.94	5.47	0.16
Stimuli (pairs vs. **others**)	39	−0.32	0.23	−0.94	0.31	4.13	0.24
Number of exposures for a given item (2–4 vs. **more than 4**)	36	−0.40	0.16	−1.05	0.24	2.20	0.12
Training test type (cued-recall vs. **free recall**)	39	−0.15	0.34	−1.30	1.01	2.65	0.70
Spacing schedule (expanding vs. **uniform**)	39	−0.05	0.11	−0.38	0.28	3.19	0.66
Subset 2							
Retention interval (log minutes)	54	0.03	0.04	−0.07	0.13	7.43	0.53
Design (between vs. **within subjects**)	54	0.11	0.17	−0.29	0.50	8.06	0.54
Feedback (no vs. **yes**)	50	−0.13	0.10	−0.35	0.10	8.22	0.23
Stimuli (pairs vs. **others**)	54	0.24	0.17	−0.23	0.71	4.28	0.23
Number of exposures for a given item (4 exposures vs. **more than 4**)	49	0.22	0.11	−0.04	0.48	8.50	0.09 *
Training test type (cued recall vs. **free recall**)	53	−0.09	0.10	−0.43	0.24	2.93	0.44
Final assessment type (cued recall vs. **free recall**)	54	−0.17	0.10	−0.47	0.14	3.14	0.18
Placement of the first retrieval attempt (different vs. **same**)	54	0.25	0.23	−0.43	0.93	3.51	0.35

Notes. For discrete moderators, categories are indicated in brackets with the reference that is in normal type. *k*: number of comparisons included in the meta-regression; *β*: meta-regression coefficient; SE: standard error; CI: confidence interval; LL: lower limit for the interval of confidence; UL: upper limit for the interval of confidence; d.f.: adjusted degrees of freedom. Results are not reliable when d.f. < 4. Significance codes: < 0.01*** < 0.05 ** < 0.10*

addition, we computed univariate analyses only because degrees of freedom were insufficient for multivariate analyses.

For subset 1, no moderator came close to statistical significance at $\alpha = 0.1$. This analysis was limited by the degrees of freedom available for many moderators (at least 4 d.f.'s are necessary for a reliable analysis; Fisher & Tipton, 2015). For those with d.f.'s > 4, it is uncertain whether the lack of significance indicates a true lack of difference or insufficient power to detect an effect.

For subset 2, comparisons including more than four exposures for each item to learn ($n = 25$) were associated with increased effect sizes compared with those including four exposures ($n = 26$) ($p = 0.09$). This suggests that more than four exposures might be needed to differentiate expanding from uniform schedules. Indeed, when each item was presented more than four times during the training phase (initial exposure + retrieval practice episodes), $g = 0.2$ (95% CI [− 0.07, 0.47], $p = 0.12$) compared with $g = -0.04$ (95% CI [− 0.17, 0.094], $p = 0.54$) when it was presented four times or less. No other moderator was statistically significant at $\alpha = 0.1$. This analysis was also limited by the degrees of freedom available for many moderators.

Discussion

Our analysis of subset 1 (11 studies, 39 comparisons) using RVE indicated a large advantage of spaced retrieval practice over massed retrieval practice ($g = 1.01$, 95% CI

[0.68, 1.34]). There was evidence for publication bias, but even after correction with the trim-and-fill method, the effect remained substantial ($g = 0.74$, 95% CI [0.55, 0.91]). The width of the confidence interval does not overlap zero and the lower limit is far above zero; this informs us about how consistent is the benefit of the spaced retrieval practice (Valentine et al., 2010). This overall mean effect size is consistent with previous meta-analyses of spaced versus massed learning (Donovan & Radosevich, 1999; Hattie, 2008; Janiszewski et al., 2003). Furthermore, there was evidence for significant heterogeneity between studies, with effect sizes ranging from $g = 0.29$ to $g = 2.16$. However, moderator analyses failed to illuminate this heterogeneity.

Various if possible theories have been proposed to account for the benefits of spaced learning and could be considered to explain the robustness and consistency of the present result on subset 1 (Wiseheart et al., 2019). Several theories stated that a repetition of an item will remind the learners of the previous occurrence for this item, and that the repetition should be spaced out to increase the cognitive effort to retrieve the item (Küpper-Tetzel et al., 2014; Smolen et al., 2016; Toppino & Gerbier, 2014). Indeed, spacing makes retrieval more effortful, and as a consequence, this strengthens the trace in long-term memory. This is not the case with massed retrieval practice because the trace does not have to be reactivated since this is always present in working memory. Following this explanation, study-phase retrieval theory suggests that increasing the effort to retrieve an item at the second retrieval episode (i.e., increasing the difficulty in re-accessing the item) will improve the likelihood of remembering this item on a retention test (Braun & Rubin, 1998; Delaney et al., 2010). One parameter is important for the study-phase retrieval to take place: the interval between two retrieval episodes should be long enough to make retrieval effortful but not too long for the retrieval to be successful. The optimal interval might depend on the type of learning contents and timeline of the learning process, but also on the characteristics of the learner such as prior knowledge. Another interesting explanation on the benefit of spacing is the amplitude of the fluctuation of the learning context between two retrieval episodes (Glenberg, 1979). The elements of the learning context (e.g., mental states, environment, mental images) fluctuate over time and spacing helps to increase this variability. Thus, the greater the spacing between two retrieval episodes, the more diverse and numerous these contextual elements will be to increase the chances of retrieving the target item (Bjork & Allen, 1970; Gerbier, 2011). In relation with the Search of Associative Memory model, Raaijmakers (2003) proposed that the probability of correctly retrieving an item from long-term memory depends on the strength of association between the contextual cues and the information contained in the trace of the item, and this strength of association itself depends both on the inter-study interval and on the retention time. All three theories are consistent with the results of our meta-analysis. However, our moderator analyses did not allow us to distinguish these different explanatory hypotheses.

We also might have expected the number of exposures for a given item to significantly modulate the spacing effect. In their recent meta-analysis on the frequency effect in incidental vocabulary learning, Uchihara et al. (2019) found that massed learning conditions (i.e., sessions were completed within a single day) benefited more of the frequency effect than spaced learning conditions (i.e., treatment sessions lasted for more than 2 days). In other words, their result suggests that there is a smaller frequency effect in spaced learning conditions and that increasing the number of exposures for a given item affect less the learning outcome in spacing conditions. In the present study, we did

not find this effect. But the moderator analysis was not reliable, so we cannot conclude on this point.

A tentative compilation of subset 1 meta-analysis with previous estimates of both retrieval practice and spacing effects is shown in Fig. 4. Overall, this summary illustrates the well-established effects of retrieval practice in the context of a massed schedule and of spacing in the context of learning with reading only. This summary suggests that the spacing effect (g = 0.71) may be larger than the retrieval practice effect (g = 0.5 to 0.61 depending on the meta-analysis) and is consistent with the hypothesis that their effects are simply additive. However, the available data collected for the present review do not allow us to test directly the hypothesis of an interaction between retrieval practice and spaced learning. To do so, a compilation of studies including a comparison between massed reading and spaced retrieval practice would be necessary.

Our analysis of subset 2 (16 studies, 54 comparisons) indicated a non-significant advantage of the expanding over the uniform spacing schedule (g = 0.034, 95% CI [− 0.10, 0.17]). Effect sizes ranged from g = − 0.53 to g = 1.02, and 55% of effect sizes were positive (i.e., expanding schedule superiority), 43% were negative (i.e., equal-uniform schedule superiority), and one was equal to zero. The absence of publication bias and the narrow width of the confidence interval enhanced the reliability of this result. Moreover, the confidence interval includes zero which is consistent with the fact that we cannot conclude for a superiority of the expanding schedule over the uniform one. Thus, if the effect existed, it would small anyway. Overall, contrary to the apparent consensus in the literature, the expanding schedule did not seem consistently superior to the uniform schedule.

The moderator analysis suggested that the number of exposures to learning contents (initial exposures + training exposures) might have an effect on the difference between the two schedules (p = 0.09). Indeed, there was an advantage of the expanding over the uniform schedule when there were many exposures to each item (more than four),

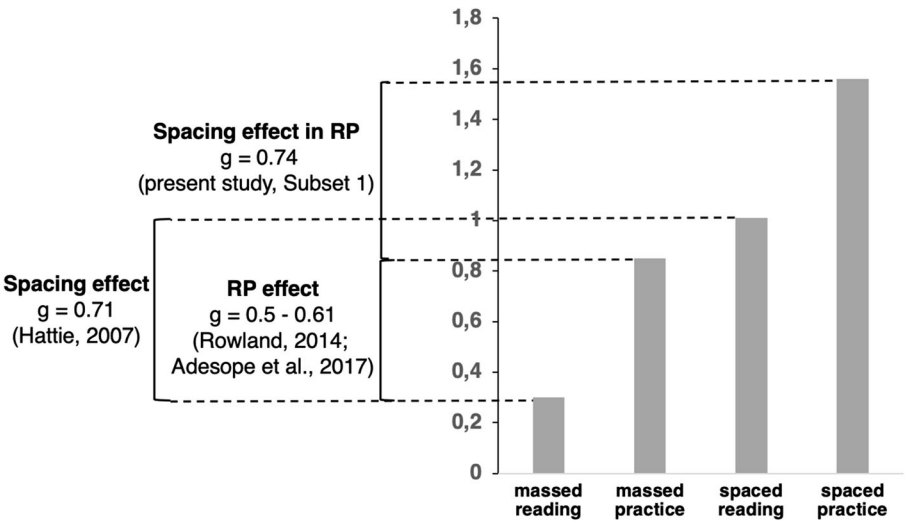

Fig. 4 Synthesis of present and previous meta-analyses with the comparison between four learning strategies. The y-axis is expressed in terms of mean effect size (Hedges' g), with an arbitrary value for the massed reading strategy (g = 0.3). "RP" is for retrieval practice

compared with when there were fewer, leading to a disadvantage of the uniform schedule at four exposures. Several models postulate that the optimal repetition schedule of a to-be-learned item depends on the memory strength of that item after initial encoding (Mozer et al., 2009; Pavlik & Anderson, 2005, 2008; Raaijmakers, 2003). If the memory strength is relatively high, the interval between the repetitions should be longer than when the memory strength is relatively low, in which case an immediate repetition is better. This model is relevant to explain the tendency for a superiority of the expanding schedule over the uniform one when participants have had more than four exposures during practice: the memory strength after the last occurrence was likely relatively high. However, the advantage of the expanding schedule at high repetition rates ($g = 0.2$, 95% CI [− 0.07, 0.47]) being non-significant, this conclusion remains tentative.

Other potential moderators of the spacing schedule effect have been proposed. Lindsey et al. (2009) compared the ACT-R model of Pavlik and Anderson (2005) to the MCM model of Mozer et al. (2009). Their study aimed at comparing the prediction made by the two models on the benefits of different spacing schedules (i.e., contracting, expanding, uniform). Each to-be-learned item was presented for three occurrences and was then tested after a retention interval varying between 10 s and 300 days. Interestingly, the two models made different predictions. The prediction made with the model of Mozer et al. (2009) suggested that the greater the retention interval, the more advantageous the expanding schedule for retention. This result was also demonstrated by empirical studies (Kang et al., 2014; Karpicke & Roediger, 2007; Storm et al., 2010). However, this hypothesis was not supported by our analysis of retention interval as a moderator ($\beta = 0.03$, 95% CI [− 0.07, 0.13]).

Toppino et al. (2018) suggested that the expanding schedule superiority might depend on how effortful the learning task is. Indeed, their results revealed an expanding-schedule superiority following spaced reading but not spaced retrieval practice. Since our subset 2 meta-analysis covered only spaced retrieval practice schedules, we were unable to evaluate this hypothesis. Thus, more studies comparing learning schedules will be necessary to conclude on the potential differences between them and on the conditions under which the expanding schedule might induce better retention than the uniform one. In the studies showing a superiority of the expanding schedule over other schedules, results suggested that the superiority of the expanding schedule seems to come from the rate of success from the first retrieval attempt rather than a higher retrieval difficulty across tests (Karpicke & Bauernschmidt, 2011; Roediger III & Karpicke, 2011). These results are directly linked to the study-phase retrieval account for the spacing effect. However, we were unable to test the assumption with the available moderators. It could also be argued that the best learning schedule is an adaptive one, taking into account each learner's performance on each item as well as learner's rates of forgetting across time and knowledge domains, thus making comparisons between generic uniform versus expanding schedules less relevant (Lindsey et al., 2014; Sense et al., 2016; Tabibian et al., 2019).

The present results of the two meta-analyses are complementary to those that were reported in previously published meta-analyses. Indeed, the intervals between the retrieval episodes were not taken into account, even as a moderator, in previous meta-analyses of the retrieval practice effect. Similarly, previous meta-analyses of the

spacing effect did not distinguish between repetitions of restudying and of retrieval practice. We did not find enough studies to perform a direct comparison between the two opposite learning strategies, namely, massed reading and spaced retrieval practice, to test a potential interaction between retrieval practice and spacing effects. Nevertheless, at this stage, our results are consistent with the idea that spacing and retrieval practice should have additive effects and that spaced reading shows a small advantage over massed retrieval practice (*g* difference between 0.1 and 0.2; see Fig. 4), but further research is needed to test this hypothesis.

The results from subset 2 highlight how crucial it is to conduct systematic, quantitative reviews taking into account all the studies comparing expanding versus uniform schedules. Indeed, in this case, our conclusion differs from those of qualitative literature reviews, which may be more at risk of neglecting studies reporting null results or results opposing the current consensus (Balota et al., 2007; Roediger III & Karpicke, 2011). As suggested by Karpicke and Roediger (2007), the placement of the first retrieval attempt might be more important than the specific schedule of subsequent retrieval attempts in maximizing long-term retention. Thus, equal-interval practice should be superior to expanding practice at longer retention intervals because the first retrieval attempt is more challenging or effortful (i.e., occurring after some delay rather than immediately after the initial presentation of the item).

Limitations

Obviously, our conclusions are limited by a number of factors. First, we did not make systematic efforts to uncover unpublished studies beyond dissertations registered in ERIC. There are both advantages and drawbacks associated with the inclusion of unpublished studies. The most obvious drawback is for the analysis to be biased by selective publication of positive effects. However, we evaluated that possibility. Indeed, we found a publication bias in favor of positive and significant effect sizes in subset 1. Nevertheless, we were able to calculate an effect size estimate adjusting for publication bias, using the trim-and-fill method. Second, there was significant heterogeneity between studies in subset 1, which our moderator analyses failed to explain. This could be explained by the insufficient number of effect sizes to test them all, and the ones that are tested show nothing. Tests of moderators using categorical models can have low statistical power. The consequence is that we cannot ensure that an effect is not significant due to a true lack of effect or a lack of power (Hempel et al., 2013). With low power, we should not conclude that there is no relationship between the moderator and the variability in the subset. We can only conclude that more studies would enhance the reliability (Harrer et al., 2019).

Concerning subset 2, our finding of no difference between spacing schedules is unlikely to result from a publication bias since this would probably have favored the expanding schedule. Furthermore, and this might explain the null results on moderator analysis for subset 1, the diversity of experimental settings (particular stimuli, test types, population) was limited, making it impossible to fully address the moderating effects of these factors. Ultimately, meta-analyses cannot make new results emerge that have not been sufficiently investigated in the experimental literature.

Practical Implications and Directions for Future Research

In 2007, the US Department of Education published a summary report with several recommendations for improving teaching to reinforce learning (Pashler et al., 2007). One of them was to *Space learning over time* ("We recommend that teachers arrange for students to be exposed to key course concepts on at least two occasions—separated by a period of several weeks to several months"). A second strong recommendation was to *Use quizzes to re-expose students to key content* ("Use quizzing with active retrieval of information at all phases of the learning process to exploit the ability of retrieval directly to facilitate long-lasting memory traces"). The level of evidence associated with these recommendations was indicated to be moderate and strong, respectively. The present meta-analyses confirm the evidence in support of both retrieval practice and spacing and suggests that they are best used combined with each other. Thus, strong recommendations to teachers and students in favor of spaced retrieval practice are warranted (Horvath et al., 2016; Kang, 2016; Weinstein et al., 2018), especially for mastering fact learning (Wiseheart et al., 2019, p. 571 for recommendations to implement spaced learning in classroom). In contrast, this report refrained from making recommendations regarding training schedules, and our results suggest that it did well. Our meta-analyses also highlight the need for more studies addressing the interaction between spacing and retrieval practice effects. For instance, a crossed design with retrieval practice (testing, reading) and spacing (massed, spaced) as factors would allow one to test whether spacing and retrieval practice are additive or not, and would provide more evidence on the potential interest of combining both practices. This would also allow one to quantify the effects whose estimation was impossible in the present meta-analysis (i.e., spaced reading vs. massed practice), thus providing a more direct comparison between the two practices (in case one has to choose). Finally, we call for new studies comparing different spacing schedules, both in restudying and in retrieval practice conditions, since the currently available evidence seems inconclusive.

Availability of Data and Material The data from the included studies are available on OSF with the following link: https://osf.io/jbmq4/?view_only=c05f2fcb93bc4d77b68fb11b1561d220.

Authors' Contribution A.L.: conceptualization, methodology, data curation, investigation, formal analysis, writing—original draft preparation. H.P.: methodology, formal analysis, writing—reviewing-editing. F.R.: conceptualization, methodology, supervision, writing—reviewing-editing.

Funding We acknowledge funding from Programme d'investissements d'avenir (Efran programme), Agence National de la Recherche (ANR-17-EURE-0017, ANR-10-IDEX-0001-02 PSL).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Code Availability R script for analyses with robumeta package is available on OSF with the following link: https://osf.io/jbmq4/?view_only=c05f2fcb93bc4d77b68fb11b1561d220.

Appendix 1

Sensitivity analysis for subset 1 and subset 2. This consists in varying the assumed within-study effect size correlation (ρ) and observing the impact on the mean effect size (Hedges' g) and on the estimated between-study variance (Tau^2)

Subset 1	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 1$
Mean effect size	1.009	1.009	1.010	1.010	1.010	1.011
Standard error	0.148	0.148	0.148	0.149	0.149	0.149
Tau^2	0.204	0.206	0.208	0.211	0.213	0.215

Subset 2	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 1$
Mean effect size	0.0343	0.0343	0.0343	0.0343	0.0343	0.0343
Standard error	0.0626	0.0626	0.0626	0.0626	0.0626	0.0626
Tau^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

References

References Marked with an Asterisk Indicate Studies Included in the Meta-Analysis

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

Bahrick, H. P. (1979). Maintenance of knowledge: questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*(3), 296–308.

* Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging, 21*(1), 19–31. https://doi.org/10.1037/0882-7974.21.1.19

Balota, D.A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In *The foundations of remembering: essays in honor of Henry L. Roediger, III* (p. 83-105). Psychology Press.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238. https://doi.org/10.3102/00346543061002213.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228–247. https://doi.org/10.1016/j.cogpsych.2010.05.004.

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

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior, 9*(5), 567–572. https://doi.org/10.1016/S0022-5371(70)80103-7.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386.

Braun, K., & Rubin, D. C. (1998). The spacing effect depends on an encoding deficit, retrieval, and time in working memory: evidence from once-presented words. *Memory (Hove, England), 6*(1), 37–65. https://doi.org/10.1080/741941599.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). Make it stick. Harvard University Press.

* Caple, C. (1996). *The effects of spaced practice and spaced review on recall and retention using computer assisted instruction*. Ann Arbor: University of Michigan Press.

* Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19*(5), 619–636. https://doi.org/10.1002/acp.1101.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: review of recent research and implications for instruction. *Educational Psychology Review, 24*(3), 369–378. https://doi.org/10.1007/s10648-012-9205-z.

Cecilio-Fernandes, D., Cnossen, F., Jaarsma, D. A. D. C., & Tio, R. A. (2018). Avoiding surgical skill decay: a systematic review on the spacing of training sessions. *Journal of Surgical Education, 75*(2), 471–480. https://doi.org/10.1016/j.jsurg.2017.08.002.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236–246. https://doi.org/10.1027/1618-3169.56.4.236.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology. General, 135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*(3), 215–235. https://doi.org/10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1.

Cull, W. L., Shaughnessy, J. J., & Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: toward confidence in applicability. *Journal of Experimental Psychology: Applied, 2*(4), 365.

Dail, T. K., & Christina, R. W. (2004). Distribution of practice and metacognition in learning and long-term retention of a discrete motor task. *Research Quarterly for Exercise and Sport, 75*(2), 148–155.

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. In Ross, B. H. (Éd.), *Psychology of learning and motivation: advances in research and theory, vol 53* (pp. 63–147). Elsevier Academic Press Inc.

* Dobson, J. L. (2012). Effect of uniform versus expanding retrieval practice on the recall of physiology information. *Advances in Physiology Education, 36*(1), 6–12. https://doi.org/10.1152/advan.00090.2011.

* Dobson, J. L. (2013). Retrieval practice is an efficient method of enhancing the retention of anatomy and physiology information. *Advances in Physiology Education, 37*(2), 184–191. https://doi.org/10.1152/advan.00174.2012.

* Dobson, J. L., Perez, J., & Linderholm, T. (2016). Distributed retrieval practice promotes superior recall of anatomy information. *Anatomical Sciences Education*, n/a-n/a., *10*(4), 339–347. https://doi.org/10.1002/ase.1668.

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. *Journal of Applied Psychology, 84*(5), 795–805. https://doi.org/10.1037/0021-9010.84.5.795.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest: A Journal of the American Psychological Society, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ [British Medical Journal], 315*(7109), 629–634.

Fisher, Z., & Tipton, E. (2015). Robumeta: an R-package for robust variance estimation in meta-analysis. *arXiv: 1503.02220 [stat]*. http://arxiv.org/abs/1503.02220. Accessed 26 July 2018.

* Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling drills. *Jounral of Educational Psychology, 59*(4), 290–296.

* Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: an effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology, 60*(7), 991–1004. https://doi.org/10.1080/17470210600823595.

Gerbier, E., & Koenig, O. (2012). Influence of multiple-day temporal distribution of repetitions on memory: a comparison of uniform, expanding, and contracting schedules. *Quarterly Journal of Experimental Psychology, 65*(3), 514–525. Scopus. https://doi.org/10.1080/17470218.2011.600806.

Gerbier, E.. (2011). Effet du type d'agencement temporel des répétitions d'une information Sur la récupération explicite. Lyon 2. http://www.theses.fr/2011LYO20029. Accessed 30 Sept 2016.

Gerbier, E., Toppino, T. C., & Koenig, O. (2015). Optimising retention through multiple study opportunities over days: the benefit of an expanding schedule of repetitions. *Memory, 23*(6), 943–954. https://doi.org/10.1080/09658211.2014.944916.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*(2), 95–112. https://doi.org/10.3758/BF03197590.

Gluckman, M., Vlach, H. A., & Sandhofer, C. M. (2014). Spacing simultaneously promotes multiple forms of learning in children's science curriculum. *Applied Cognitive Psychology, 28*(2), 266–273. https://doi.org/10.1002/acp.2997.

Godbole, N. R., Delaney, P. F., & Verkoeijen, P. P. J. L. (2014). The spacing effect in immediate and delayed free recall. *Memory, 22*(5), 462–469. https://doi.org/10.1080/09658211.2013.798416.

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: retrievability and question format matter. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02412.

* Grote, M. G. (1995). Distributed versus massed practice in high school physics. *School Science and Mathematics, 95*(2), 97–101.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). Doing meta-analysis in R: a hands-on guide: vol. https://doi.org/10.5281/zenodo.2551803. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/. Accessed 15 April 2020.

Hattie, J. (2008). Visible learning: a synthesis of over 800 meta-analyses relating to achievement. Routledge.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Academic Press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. https://doi.org/10.1002/jrsm.5.

Hempel, S., Miles, J. N., Booth, M. J., Wang, Z., Morton, S. C., & Shekelle, P. G. (2013). Risk of bias: a simulation study of power to detect study-level moderator effects in meta-analysis. *Systematic Reviews, 2*(1), 107. https://doi.org/10.1186/2046-4053-2-107.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ [British Medical Journal], 327*(7414), 557–560.

* 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. *Educational Psychology Review, 28*(4), 853–873. https://doi.org/10.1007/s10648-015-9349-8.

Horvath, J. C., Lodge, J. M., & Hattie, J. (2016). From the laboratory to the classroom: translating science of learning for teachers. Routledge.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: implications for research on advertising repetition and consumer memory. *Journal of Consumer Research, 30*(1), 138–149. https://doi.org/10.1086/374692.

* Kalenberg, K. (2017). Spaced and expanded practice: an investigation of methods to enhance retention. *Journal of Undergraduate Research at Minnesota State University, Mankato, 17*, 18.

Kang, S. H. K. (2016). Spaced repetition promotes efficient and effective learning. *Policy Insights From the Behavioral and Brain Sciences*, 2372732215624708, *3*(1), –19. https://doi.org/10.1177/2372732215624708.

* Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review, 21*(6), 1544–1550. https://doi.org/10.3758/s13423-014-0636-z.

Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: long-term benefits for factual and higher-level learning. *Learning and Instruction, 36*, 38–45. https://doi.org/10.1016/j.learninstruc.2014.11.001.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3.

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

Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: positive effects of retrieval practice in elementary school children. *Frontiers in Psychology, 7*, 350. https://doi.org/10.3380/fpsyg.2015.00350.

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: a perspective for enhancing meaningful learning. *Educational Psychology Review, 24*(3), 401–418. https://doi.org/10.1007/s10648-012-9202-2.

* Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology-Learning Memory and Cognition, 33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704.

* Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition, 38*(1), 116–124. https://doi.org/10.3758/MC.38.1.116.

Klingbeil, D. A., Renshaw, T. L., Willenbrink, J. B., Copek, R. A., Chan, K. T., Haddock, A., Yassine, J., & Clifton, J. (2017). Mindfulness-based interventions with youth: a comprehensive meta-analysis of group-design studies. *Journal of School Psychology, 63*, 77–103. https://doi.org/10.1016/j.jsp.2017.03.006.

* Küpper-Tetzel, C. E., Kapler, I. V., & Wiseheart, M. (2014). Contracting, equal, and expanding learning schedules: the optimal distribution of learning sessions depends on retention interval. *Memory and Cognition, 42*(5), 729–741. Scopus. https://doi.org/10.3758/s13421-014-0394-1.

Larsen, D. P. (2018). Planning education for long-term retention: the cognitive science and implementation of retrieval practice. *Seminars in Neurology, 38*(4), 449–456. https://doi.org/10.1055/s-0038-1666983.

Lee, T. D., & Genovese, E. D. (1989). Distribution of practice in motor skill acquisition: different effects for discrete and continuous tasks. *Research Quarterly for Exercise and Sport, 60*(1), 59–65.

Lindsey, R., Mozer, M., Cepeda, N. J., & Pashler, H. (2009). Optimizing memory retention with cognitive models. International conference on computing and Mission, 6.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science, 25*(3), 639–647. https://doi.org/10.1177/0956797613504302.

* Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: exploring different schedules of spacing and retention interval in younger and older adults. *Aging Neuropsychology and Cognition, 15*(3), 257–280. https://doi.org/10.1080/13825580701322171.

* Maddox, G. B., Balota, D. A., Coane, J. H., & Duchek, J. M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology and Aging, 26*(3), 661–670. https://doi.org/10.1037/a0022942.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: the effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399–414. https://doi.org/10.1037/a0021782.

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

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013b). Quizzing in middle-school science: successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*(3), 360–372. https://doi.org/10.1002/acp.2914.

* McGregor, K. K. (2014). What a difference a day makes: change in memory for newly learned word forms over 24 hours. *Journal of Speech, Language, and Hearing Research, 57*(5), 1842–1850. https://doi.org/10.1044/2014_JSLHR-L-13-0273.

* Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology. General, 145*(7), 897–917. https://doi.org/10.1037/xge0000170.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Moreira, B., Pinto, T., Starling, D., & Jaeger, A. (2019). Retrieval practice in classroom settings: a review of applied research. *Frontiers in Education*, 4. https://doi.org/10.3389/feduc.2019.00005.

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, 56*(9-B), 5204.

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R. V., & Vul, E. (2009). Predicting the optimal spacing of study: a multiscale context model of memory. *NIPS*.

Mumford, M. D., Costanza, D. P., Baughman, W. A., Threlfall, K. V., & Fleishman, E. A. (1994). Influence of abilities on performance during practice: effects of massed and distributed practice. *Journal of Educational Psychology, 86*(1), 134–144. https://doi.org/10.1037/0022-0663.86.1.134.

* Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning. does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition, 37*(4), 677–711. https://doi.org/10.1017/S0272263114000825.

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.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER* 2007–2004. National Center for Education Research. http://eric.ed.gov/?id=ED498555. Accessed 28 Aug 2016.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: an activation-based model of the spacing effect. *Cognitive Science, 29*(4), 559–586. https://doi.org/10.1207/s15516709cog0000_14.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology. Applied, 14*(2), 101–117. https://doi.org/10.1037/1076-898X.14.2.101.

* Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory and Cognition, 35*(8), 1917–1927 Scopus.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the SAM model. *Cognitive Science, 27*(3), 431–452. https://doi.org/10.1207/s15516709cog2703_5.

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.

Roediger III, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval: a resolution. In *Successful remembering and successful forgetting: a festschrift in honor of Robert A. Bjork* (pp. 23–47). Psychology Press.

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

Rosenshine, B. (2010). Principles of instruction. Educational practices series-21. UNESCO International Bureau of Education.

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.

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: from laboratory to classroom. *Applied Cognitive Psychology, 19*(1), 107–122. https://doi.org/10.1002/acp.1066.

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science, 8*(1), 305–321. https://doi.org/10.1111/tops.12183.

Smolen, P., Zhang, Y., & Byrne, J. H. (2016). The right time to learn: mechanisms and optimization of spaced learning. *Nature reviews. Neuroscience, 17*(2), 77–88. https://doi.org/10.1038/nrn.2015.18.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology, 25*(5), 763–767. https://doi.org/10.1002/acp.1747.

* Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: when and why expanding retrieval practice enhances long-term retention. *Memory & Cognition, 38*(2), 244–253. https://doi.org/10.3758/MC.38.2.244.

Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., & Gomez-Rodriguez, M. (2019). Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences, 116*(10), 3988–3993. https://doi.org/10.1073/pnas.1815156116.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5*(1), 13–30. https://doi.org/10.1002/jrsm.1091.

Terenyi, J., Anksorus, H., & Persky, A. M. (2018). Impact of spacing of practice on learning brand name and generic drugs. *American Journal of Pharmaceutical Education, 82*(1), 6179. https://doi.org/10.5688/ajpe6179.

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*(6), 604–634. https://doi.org/10.3102/1076998615606099.

Toppino, T. C., & Gerbier, E. (2014). About practice: repetition, spacing, and abstraction. In Ross, B. H. (Éd.), *Psychology of learning and motivation, vol 60* (pp. 113–189). Elsevier Academic Press Inc.

Toppino, T. C., Phelan, H.-A., & Gerbier, E. (2018). Level of initial training moderates the effects of distributing practice over multiple days with expanding, contracting, and uniform schedules: evidence for study-phase retrieval. *Memory & Cognition, 46*(6), 969–978. https://doi.org/10.3758/s13421-018-0815-7.

Tsai, L. (1927). The relation of retention to the distribution of relearning. *Journal of Experimental Psychology, 10*(1), 30–39.

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: a meta-analysis of correlational studies. *Language Learning, 69*(3), 559–599. https://doi.org/10.1111/lang.12343.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need?: A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*(2), 215–247. https://doi.org/10.3102/1076998609346961.

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: the spacing effect in children's acquisition and generalization of science concepts. *Child Development, 83*(4), 1137–1144. https://doi.org/10.1111/j.1467-8624.2012.01781.x.

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications, 3*(1), 2. https://doi.org/10.1186/s41235-017-0087-y.

Whitten, W. B., & Bjork, R. A. (1977). Learning from Tests : Effects of Spacing. *Journal of Verbal Learning and Verbal Behavior, 16*(4).

Wiseheart, M., Kim, A. S. N., Kapler, I. V., Foot-Seymour, V., & Küpper-Tetzel, C. E. (2019). Enhancing the quality of student learning using distributed practice. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (1re ed., pp. 550–584). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108235631.023.